Anonymisation Standard for Publishing Health and Social Care Data

Supporting Guidance: Drawing the line between identifying and non-identifying data

Author – *Malcolm Oswald*

Document Control

Change Control History

Version	Change Summary	Date	Change author
0.1	First incomplete draft	06/12/11	M Oswald
0.2	Second incomplete draft for working group meeting of 12 Dec 2011	07/12/11	M Oswald
0.3	Third incomplete draft	29/12/11	M Oswald
0.4	Fourth incomplete draft	6/01/12	M Oswald
0.5	Fifth incomplete draft for working group meeting of 23 Jan 2012	16/01/12	M Oswald
0.6	Sixth draft, and first full draft for wider review	3/02/12	M Oswald
0.7	Seventh draft, taking into account comments from v06 reviewers	13/02/12	M Oswald
0.8	Eighth draft, restructured, with reduced scope to exclude releases to a controlled environment	02/03/12	M Oswald
0.9	Ninth draft, following comments from Information Standards Board appraisers	03/04/12	M Oswald
0.92	Tenth draft following comments from NIGB and others	29/05/12	M Oswald
0.93	Eleventh draft following comments from ISB appraisers	19/06/12	M Oswald
0.94	Twelfth draft to reflect comments from Technology Office	26/06/12	M Oswald
0.95	Minor changes made as a result of ISB reviewer comments	05/07/12	M Oswald
0.96	Changes made following draft submission to ISB Board	25/09/12	M Oswald
0.97	Changes made to take account of phase 1 testing. – version used for phase 2 testing	30/10/12	M Oswald
0.98	Changes made to take account of phase 2 testing results and ISB appraiser comments	20/12/12	M Oswald
0.99	Changes made to take account of more ISB comments	18/01/13	M Oswald
1.0	Publication copy	21/02/13	Information Standards Management Service

Reviewers

Version	Reviewer	Role
0.1, 0.3, 0.7	Phil Walker	Sponsor
0.1, 0.3, 0.7	Clare Sanderson	Chair of working group
0.2, 0.5, 0.7	De-identification working group (several reviewers)	Informatics Managers from Health and Social Care
0.4, 0.7	Iain Bourne	Information Commissioner's Office
V.06, v.08	Karen Thomson	NIGB IG Lead, and ISB Policy Lead, and member of De-identification Working Group
v.06	Ralph Sullivan	NHS IC Primary Care Lead, and member of De-identification Working Group
v.06	Mary Hawking, Ian Herbert, Rob Navarro	BCS PHCSG members

v.06	Michael Wilks	NIGB member
v.07	Peter Singleton	
v.06	Tony Calland, Paul Cundy, Sophie Brennan, Rachel Merritt et al	BMA
v.06	Jean Roberts	UKCHIP
v.08, v.092, 0.97	Ian Shepherd, Sue Dennis, Sarah Bradley	Information Standards Board appraisers
v.06, v.09, v0.92	National Information Governance Board	NIGB
V.092	Mark Penny, Danny Solomon, David Doran-Hughes	Technology Office, Department of Health
v.096	Information Centre reviewers and testers	Information Centre

Approvals and Reviews:

Name	Organisation & Role	Version	Review/ Approval
Information Standards Managers	Information Standards Management Service	v.08, v0.92	R
Stakeholder Groups	Health and Social Care	0.2, 0.5, 0.7	R
Appraisers	Information Standards Board	v.08, v0.92	R
Phil Walker	Department of Health – Sponsor – Head of IG Policy	0.1, 0.3, 0.7	R
Clare Sanderson	The Information Centre for Health and Social Care – Lead Developer – Executive Director of IG	0.1, 0.3, 0.7	R
Phil Walker	Department of Health – Sponsor – Head of IG Policy	0.98	A
Clare Sanderson	The Information Centre for Health and Social Care – Lead Developer – Executive Director of IG	0.98	A
Technology Office	NHS Connecting for Health	v.092	A
NIGB	National Information Governance Board	V0.6 V0.9, v0.92	R
Information Standards Board	NHS Connecting for Health	V0.99	A
Information Standards Board	NHS Connecting for Health	V1.0	A

CONTENTS

Preface

The law and the need for a anonymisation standard for health and social care

The law pulls in two opposite directions. Human Rights and Data Protection legislation, along with our domestic common law duty to respect confidentiality, require us to protect information that could identify an individual. The Freedom of Information Act requires public authorities to release information about their activities, and this message is reinforced by the government's transparency agenda (although that policy cannot override a public authority's legal duty to protect personal and confidential data).

Detailed care records about patients and service users are confidential and must be protected. However, because they also act as important records of what a public authority has done, and are a rich source of information for improving health and social care in the future, law and policy require public authorities to derive non-identifying information from care records for many purposes, and where it is both lawful and appropriate, to make it publicly available. Therefore, it is crucial to be able to recognise the point at which information can no longer identify individuals.

Although the law makes a clear distinction between identifying and non-identifying data, where that line should be drawn may be far from clear in practice. Some data, like the full medical records held by Anytown General Practice, are clearly identifying, whereas a figure for the total number of patients with diabetes in the whole of Anytown clearly counts as non-identifying information. However, what about a list of diabetic patients from the practice, showing just their gender, height and age? If Anyown General Practice received a Freedom of Information Act request for just that information, would the law require them to release it or protect it? The answer depends on several factors: on the actual content of the information listed, on the availability of other information now and in the future that could be used to reveal the identity of diabetic patients on the list, and on the likelihood that someone will get hold of that other information and use it to learn something about one of the patients represented on the list. Some of these factors cannot be measured, only assessed, and public authorities may be uneasy about making these judgements.

This guidance, and the associated process standard ("Anonymisation standard for publishing health and social care data") are needed in order to address these difficult issues. Their purpose is to provide organisations with an agreed and standardised approach, grounded in the law, for distinguishing between identifying and non-identifying information, and to specify a set of standard tools for ensuring, as far as it is reasonably practicable to do so, that any information published (for example, as part of the transparency agenda) cannot identify individuals.

The anonymisation standard documents

The document "Anonymisation standard for publishing health and social care data specification" specifies the steps required to select an appropriate anonymisation plan and to assess re-identification risk. Its scope is the publication of non-identifying information.

The standard is based on an interpretation of the law and policy set out in this guidance document: "Drawing the line between identifying and non-identifying data". Guidance on anonymiseing data in response to Freedom of Information Act requests, and on releasing information that is only non-identifying within controlled environments (i.e. not released into the public domain) are planned for the future.

1 Introduction

1.1 Purpose

1. The purpose of the *Anonymisation Standard for Publishing Health and Social Care Data* is to set out a standard set of steps for health and social care organisations to follow to enable them to make a justifiable distinction between identifying and non-identifying data, and use non-identifying data when publishing information. Turning identifying data into non-identifying data protects personal privacy, and enables published information to be used for public benefit,

2. The main purpose of this supporting document is to explain the law underpinning the standard, and in particular what the law says about where the line should be drawn between identifying and non-identifying data. It also contains a set of scenarios illustrating how anonymisation, and the standard in particular, may be put in practice.

1.2 Scope

3. The scope of the standard is set out in the standard specification. The scope of this document includes the information necessary to fulfil the purposes set out above. Although the standard covers only publishing, the legal guidance in this document also considers other circumstances in which non-identifying data may be disclosed. Two such circumstances are:

   - Responding to Freedom of Information Act requests;
   - the release of non-identifying data into controlled environments (e.g. where access to data is restricted); anonymisation guidance for such disclosures is under development.

1.3 Background

4. Health and social care professionals create and maintain care records to deliver safe and effective care to their patients / service users. However, the information in care records is also used for a variety of purposes other than direct care. In such circumstances, unless individuals have consented explicitly or there is another lawful justification (see section 2.1), the law requires that disclosures of information should not identify individuals.

5. Therefore, when releasing information, health and social care organisations and their staff must understand what the law says about where to draw the line between identifying and non-identifying information, and be able to apply it. The aim of the standard and this supporting guidance is to assist organisations to do that.

6. Surveys of the general public have consistently highlighted concerns about the privacy of their medical records[1]. However, these surveys show that people also understand the importance of the information within their records to valuable work for the public good like research and epidemiology. They want such work to continue but want to be told how their records are being used, and to be asked

[1] See, for example, "Public and Professional attitudes to privacy of healthcare data - A Survey of the Literature", , available at: http://www.gmc-uk.org/GMC_Privacy_Attitudes_Final_Report_with_Addendum.pdf_27007284.pdf

before information that identifies them is accessed[2]. Thus, the public has a clear interest in where the line between identifying and non-identifying information is drawn.

7. Distinguishing between identifying and non-identifying information becomes increasingly important as the government moves towards publishing as much relevant information as possible about public services. Two linked policy agendas (on transparency and enabling research) are described briefly below. Whilst these policy initiatives matter, it is important to keep in mind that all government policies and actions, and responses by health and social care organisations, are constrained by what the law determines may be done with personal data or confidential information (described in section 2).

1.3.1 The government's transparency agenda

8. In both health and social care, and throughout public services, the government is pressing for greater data transparency. It has published 14 Public Data Principles[3]:

> "Public Data" is the objective, factual, non-personal data on which public services run and are assessed, and on which policy decisions are based, or which is collected or generated in the course of public service delivery."

9. The draft principles include the following:

   a) "Public data policy and practice will be clearly driven by the public and businesses who want and use the data, including what data is released when and in what form…";

   b) "Public data will be published in reusable, machine-readable form…"

   c) "Public data will be released under the same open licence which enables free re-use, including commercial re-use…"

   d) "Release data quickly, and then re-publish it in linked data form – Linked data standards allow the most powerful and easiest re-use of data. However most existing internal public sector data is not in linked data form. Rather than delay any release of the data, our recommendation is to release it 'as is' as soon as possible, and then work to convert it to a better format"

   e) "Public data will be freely available to use in any lawful way…"

   f) Public bodies should actively encourage the re-use of their public data – in addition to publishing the data itself, public bodies should provide information and support to enable it to be re-used easily and effectively.

10. This puts pressure on all health and social care organisations to publish more, and more often, useful data that might be of interest to the public or business in an electronic form, ready for processing.

11. Furthermore, the Health and Social Care Act 2012 obliges the Health and Social Care Information Centre to distinguish between confidential information that identifies a person and information that it collects that does not identify a person. It

---

[2] See, for example, page 69 of "Private Lives", available at: http://www.demos.co.uk/files/Private_Lives_-_web.pdf. Note also that some people want to be able to opt out of such uses of their data.

[3] See: http://data.gov.uk/blog/public-data-statement-of-principles

must not publish the former, and has a duty to publish the latter (other than in specific circumstances)[4].

1.3.2 Enabling research and working with the pharmaceutical industry

12. Linked to the drive for transparency, on 5th December 2011, the Prime Minister announced that the government sought to enable better use of patient data in research, work more closely with the pharmaceutical industry, and stimulate drug discovery and economic growth[5]. He stressed that data made available for researchers would be anonymous i.e. non-personal data.

13. These plans are consistent with existing practice, and with the law. Companies may already lawfully make use of non-personal data for research[6] and some do. What is new is that the government is determined to make such data widely available. Aggregate data are of limited value to researchers; the government seeks to make individual-level data much more widely available, on the basis that such data can be derived from source data using computers into non-personal data and then released.

1.4 Glossary

Term	Definition
Aggregate data	Data derived from records about more than one person, and expressed in summary form, such as statistical tables.
Anonymisation	Any processing that minimises the likelihood that a data set will identify individuals. A wide variety of anonymisation techniques can be used; some examples of such processing are explained in the Standard Specification. Also commonly referred to as "de-identification".
Caldicott Guardian	A senior person responsible for protecting the confidentiality of patient and service user information and enabling appropriate information sharing. Caldicott Guardians were mandated for NHS organisations by Health Service Circular HSC1999/012 and later for social care by Local Authority Circular LAC 2002/2. General practices are required by regulations to have a confidentiality lead[7]. Note that a Caldicott Guardian is an individual, whereas

[4] See in particular clauses 256 and 260 of the Act, available at: http://www.legislation.gov.uk/ukpga/2012/7/contents/enacted/data.htm

[5] See: http://www.bbc.co.uk/news/uk-16028836

[6] as evidenced in *R v Department of Health, ex parte Source Informatics* [2000] 1 All ER 793

[7] The definition provided in the glossary that forms part of the Information Governance Toolkit, available at: https://www.igt.connectingforhealth.nhs.uk/Resources/Glossary.pdf

	a data controller is a "legal person" (invariably an organisation such as a general medical practice or foundation trust).
Confidential information	Information to which a common law duty of confidence applies.
Cell	An entry in a table of aggregate data.
Data	Data means information which – (a) is being processed by means of equipment operating automatically in response to instructions given for that purpose, (b) is recorded with the intention that it should be processed by means of such equipment, (c) is recorded as part of a relevant filing system or with the intention that it should form part of a relevant filing system, (d) does not fall within paragraph (a), (b) or (c) but forms part of an accessible record as defined by section 68, or (e) is recorded information held by a public authority and does not fall within any of paragraphs (a) to (d)[8].
Data controller	A person[9] who (either alone or jointly or in common with other persons) determines the purposes for which and the manner in which any personal data are, or are to be, processed[10].
Data processor	Any person (other than an employee of the data controller) who processes the data on behalf of the data controller[11].
Direct identifier	Name, address, widely-used unique person or record identifier (notably National Insurance Number, NHS Number, Hospital Number), telephone number, email address, and any other data item that on its own could

[8] The definition in the Data Protection Act 1998; see: http://www.legislation.gov.uk/ukpga/1998/29/part/I

[9] Note that this is a "legal person", and in the context of health and social care, this will be a legal entity such as a local authority, NHS trust, or general practice rather than an individual working for such a body.

[10] The definition in the Data Protection Act 1998; see: http://www.legislation.gov.uk/ukpga/1998/29/part/I

[11] The definition in the Data Protection Act 1998; see: http://www.legislation.gov.uk/ukpga/1998/29/part/I

	uniquely identify the individual.
Disclose	To provide information to specific recipient(s).
Duty of confidence	A duty of confidence arises when one person discloses information to another (e.g. patient to clinician) in circumstances where it is reasonable to expect that the information will be held in confidence. It – a. is a legal obligation that is derived from case law; b. is a requirement established within professional codes of conduct; and c. must be included within NHS employment contracts as a specific requirement linked to disciplinary procedures[12].
Identifying data	The same meaning as personal data, but extended to apply to dead, as well as living, people.
Indirect identifier	A data item (including postal code, gender, date of birth, event date or a derivative of one of these items) that when used in combination with other items could reveal the identity of a person. Also referred to as "quasi-identifier".
Individual-level data	Data that have not been aggregated[13] and that relate to an individual person, and/or to events about that person. The data may or may not reveal the identity of a person, and thus may or may not be identifying data. An example is a request for an investigation that accompanies a blood test, with NHS Number, date of birth, and details of the sample and tests required[14].
Information	See definition of "data". Within this document, the two terms are used synonymously.
k-anonymity	A criterion to ensure that there are at least k records in a data set that have the same quasi-identifier values. For example, if the quasi-identifiers are age and

[12] Definition taken from page 7 of NHS Code of Practice on Confidentiality, available at: http://www.dh.gov.uk/en/Publicationsandstatistics/Publications/PublicationsPolicyAndGuidance/DH_4069253

[13] Note though that a record about an individual may contain aggregate counts (such as a data item of "Number of hospital admissions for patient in 2011".

[14] In this particular example, the data would be identifying because of the presence of the NHS Number and date of birth.

	gender, then it will ensure that there are at least k records with 45-year old females[15]. Note that it is necessary to remove direct identifiers in order to satisfy k-anonymity.
Non-identifying data	Data that are not "identifying data" (see definition above). Non-identifying data are always also non-personal data.
Non-personal data	Data that are not "personal data". Non-personal data may still be identifying in relation to the deceased (see definition of "identifying data" and "personal data").
Personal data	Data which relate to a living individual who can be identified – (a) from those data, or (b) from those data and other information which is in the possession of, or is likely to come into the possession of, the data controller, and includes any expression of opinion about the individual and any indication of the intentions of the data controller or any other person in respect of the individual[16].
Pseudonymisation	A technique that replaces identifiers with a pseudonym[17] that uniquely identifies a person. In practice, pseudonymisation is typically combined with other anonymisation techniques.

[15] The definition provided at page 47 of 'Best Practice' Guidelines for Managing the Disclosure of De-Identified Health Information', available at: www.ehealthinformation.ca/documents/de-idguidelines.pdf

[16] The definition in the Data Protection Act 1998; see: http://www.legislation.gov.uk/ukpga/1998/29/part/I

[17] A pseudonym is a fictitious name or code.

Publish	To disseminate to the public[18]. Note that "disseminate" is sometimes given a meaning similar to that of "disclose" above, although its dictionary meaning used here is quite different: "to spread abroad" and "to disperse throughout"[19].
Public authority	A body defined under, and subject to, the Freedom of Information Act 2000[20]. It includes government departments, local authorities, the NHS, state schools and police forces. A non-public authority, such as a company, carrying out health and social care activities on behalf of, and under contract to, a public authority may be required to assist the public authority in satisfying requests under the Freedom of Information Act.
Quasi-identifier	See entry for "indirect identifier" above.
Redact	To censor or obscure (part of a text) for legal or security purposes[21]. In the context of responding to Freedom of Information Act requests, redaction should be explicit and permanent, making clear that information has been withheld.
Re-identification	The process of discovering the identity of individuals from a data set by using additional relevant information.
Statistical disclosure control	Techniques for obscuring small numbers (e.g. less than "5") that appear in aggregate tables so as to prevent re-identification.

[18] A definition provided by Merriam-Webster Dictionary at: http://www.merriam-webster.com/dictionary/publish. To "Disseminate" here means unrestricted distribution or availability, and "the public" includes any person or group of people.

[19] A definition provided by Merriam-Webster Dictionary at: http://www.merriam-webster.com/dictionary/publish

[20] For a full definition, see Schedule 1 of the Freedom of Information Act at: http://www.legislation.gov.uk/ukpga/2000/36/schedule/1?view=plain

[21] The definition provided by Oxford Dictionaries at: http://oxforddictionaries.com/definition/redact

2 The law underpinning the proposed standard

14. This section sets out and analyses relevant law, and reaches a series of conclusions of particular relevance to anonymisation. These can be found in section 2.7.

2.1 The legal context for anonymisation

15. In the eyes of the law, data either identify a person or they do not (even though in practice the distinction may be difficult to make - see section 2.5). The distinction matters because very different rules apply in each case. When data are identifying, there are significant constraints on how they can be used. For example, as little as possible identifying information must be used, and even that can only be justified in certain circumstances. In contrast, when data are not identifying, then in general there are no constraints on use.

16. So when does the law allow use of identifying data, and when is anonymisation necessary? For a fully comprehensive answer to this question, consult national guidance on law and policy[22]. However, in general, health and social care organisations may use identifying data in one of the following circumstances:

    a. With the explicit consent[23] of the individual concerned;

    b. For the purposes of direct patient care and/or local clinical audit of the individual's care (where the policy is that consent can be inferred);

    c. Where there is statutory authority for a particular use; or

    d. In exceptional circumstances, where it is justified in the public interest[24].

17. Where no lawful justification exists, non-identifying data must be used, which is where the need for anonymisation arises. Typically, this is in situations where data are to be used for a purpose other than direct care, such as epidemiology, health and social care research, and service management. In such circumstances, data may be required about large groups or populations, and it may be impractical to gain the explicit consent of all the individuals concerned. Section 251 of the NHS Act 2006 can justify use of identifying data for such purposes in certain circumstances and where specifically approved[25], but normally data used for purposes other than direct care should be de-identified.

---

[22] See the NHS Code of Practice on confidentiality at: http://www.dh.gov.uk/en/Managingyourorganisation/Informationpolicy/Patientconfidentialityandcaldicottguardians/DH_4100550, and for a relatively recent overview of relevant law, see: the Department of Health's "NHS Information Governance Guidance on Legal and Professional Obligations", available at: http://www.dh.gov.uk/en/Publicationsandstatistics/Publications/PublicationsPolicyAndGuidance/DH_079616

[23] For a brief discussion of consent, see: http://www.ico.gov.uk/for_organisations/privacy_and_electronic_communications/consent.aspx

[24] Note that, for secondary uses, a public interest justification may be difficult to sustain given the existence of section 251 of the NHS Act as a statutory means of justifying the use of confidential information for secondary uses. For further guidance on the public interest, see the supplementary guidance to the NHS Code of Practice on Confidentiality, available at: http://www.dh.gov.uk/en/Publicationsandstatistics/Publications/PublicationsPolicyAndGuidance/DH_4069253

[25] See: http://www.nigb.nhs.uk/s251/abouts251

18. Therefore, health and social care organisations need to understand where to draw the line between identifying and non-identifying data, and how to anonymise data.

2.2 Personal data and confidential information

19. Whether health and social care data or information[26] identify a person matters: keeping information confidential is fundamental to maintaining public trust in health and social care services. Whether or not data identify a person is also a significant fact in legislation and common law - this distinction determines how data or information must be treated.

20. The Data Protection Act 1998 defines "personal data"[27] as:

> "data which relate to a living individual who can be identified –
>
> (a) from those data, or
>
> (b) from those data and other information which is in the possession of, or is likely to come into the possession of, the data controller,
>
> and includes any expression of opinion about the individual and any indication of the intentions of the data controller or any other person in respect of the individual."[28]

21. As "personal data" are data that can identify a person, it follows that data that cannot identify a person are not personal data (which hereafter will be referred to as "non-personal data"). The European Directive on Data Protection (on which the Data Protection Act is based) recognises that judgement is often required to decide whether data are personal[29]:

> "Sometimes it is not immediately obvious whether an individual can be identified or not, for example, when someone holds information where the names and other identifiers have been removed. In these cases, Recital 26 of the Directive states that, whether or not the individual is nevertheless identifiable will depend on "all the means likely reasonably to be used either by the controller or by any other person to identify the said person"."

22. "Personal data" is also relevant to the Freedom of Information Act 2000. When a person requests information from a public body, the information must be provided unless refusal is for a valid reason. The Act specifies the valid reasons. One valid reason is to avoid breaching the Data Protection Act. Therefore, personal data (other than in specific circumstances) should be withheld when satisfying Freedom

---

[26] No special meaning is given here to the terms "data" and "information"; dictionary definitions apply.

[27] This, and other key definitions from the Act, may found on the Information Commissioner's Office website at: http://www.ico.gov.uk/for_organisations/data_protection/the_guide/key_definitions.aspx

[28] Note that this is very similar to the concept of "personal information" in the Statistics and Registration Service Act 2007 information: "information identifies a particular person if the identity of that person—(a) is specified in the information, (b) can be deduced from the information, or (c) can be deduced from the information taken together with any other published information". See: http://www.legislation.gov.uk/ukpga/2007/18/section/39

[29] "Determining what is personal data", available at: http://www.ico.gov.uk/upload/documents/determining_what_is_personal_data/whatispersonaldata2.htm

of Information Act requests, and a redacted version of the requested information supplied[30].

23. The concept of "confidential information" comes from case law rather than legislation, and is information provided in circumstances "importing a duty of confidence"[31], such as the relationship between a patient and a doctor[32],[33],[34]. This common law duty of confidence is now reinforced by, and interpreted in the light of, Article 8 of the Human Rights 1998 (which protects privacy). Where confidential information is about a person, then the information must reveal something about that person for it to be confidential.

24. The test for identifiability is not specified explicitly in confidentiality case law or the Human Rights Act, although there is no reason to expect it to be any stronger or weaker than the identifiability test specified for personal data in the Data Protection Act. If confidential information about a person is altered in such a way that it no longer identifies that person, then normally it will not be possible to learn something about that person from the information, and so it ceases to be confidential and may be used without restriction[35].

25. However there is an important exception to that. It is possible to breach confidence by learning something new about a group of people, and inferring information about one individual. For example, suppose there are three records without direct identifiers that relate to three people, and it is known that one of the three relates to Fred. The records may by chance all reveal a diagnosis of cancer: one breast cancer, one lymphoma, and one lung cancer. It would then be known that Fred has a diagnosis of cancer – confidential information – even though it would not be known which type of cancer. This issue does not apply solely to individual-level data; an aggregate report with a breakdown of information about people with cancer could contain a cell of "3" and reveal the same information about Fred.

---

[30] For further explanation, see the advice from the Information Commissioner's Office at: http://www.ico.gov.uk/upload/documents/library/freedom_of_information/practical_application/redactingandextractinginformation.pdf

[31] A-G v Guardian Newspapers Ltd [1988] 3 All ER 545 at 624; see: http://www.bailii.org/uk/cases/UKHL/1988/6.html

[32] Stephens v Avery [1988] 2 All ER 477 at 482, described at: http://www.lawgazette.co.uk/news/tort-breach-confidence-stephens-v-avery-and-others

[33] Note that it is a matter of debate as to whether all information in a person's health record is confidential. For example, it can be argued that certain information, such as a person's date of birth, is not provided in confidence.

[34] A potential difference between "personal data" and confidential information is that the latter does not necessarily exclude dead people. This is discussed further in section 3.

[35] This is the interpretation of the law by the Department of Health in the 2003 NHS Confidentiality Code of Practice, and in guidance from the British Medical Association and General Medical Council, following the landmark case of R v Department of Health, ex parte Source Informatics (2000). For a description of the Source Informatics case, see: http://business.highbeam.com/437582/article-1G1-201654159/r-v-department-health-ex-p-source-informatics-ltd

2.2.1 What *Department of Health v Information Commissioner* tells us about drawing the line between personal and non-personal data

26. Although only one court case in the body of case law, *Department of Health v Information Commissioner*[36] provides important insight into where and how to draw the line between personal and non-personal data, particularly when public authorities are disclosing and publishing data. The case concerned the publication of abortion statistics and a Freedom of Information Act request from a pro-life organisation. The Department had applied disclosure control techniques to the aggregate abortion statistics published for England – suppressing small numbers in published statistical tables as recommended by the Office for National Statistics. The Freedom of Information Act request sought the full aggregate figures, with no small number suppression. The Information Commissioner found that the full statistics were not personal data, and so they should be released. The Department of Health appealed the Information Commissioner's decision to the Information Tribunal. The tribunal agreed with the Information Commissioner's judgement. The Department of Health challenged the tribunal's ruling in the High Court.

27. Although the judgment makes clear that some of the tribunal's reasoning was flawed, the judge (Cranston J.) endorsed the tribunal's decision that the full abortion statistics were not personal data and so should be released[37].

28. Specifics matter, and judgements as to whether data are personal or non-personal should be made on a case-by-case basis. However, important points of wider relevance that emerge from this case include:

- **Even where statistics relate to matters of a very sensitive nature (as with abortion), these are not grounds for withholding statistics that pose only a remote risk of revealing a person's identity[38];**
- The existence of the identifiable abortion information from which the statistics were derived and which were also held by the Department does not affect the status of the full abortion statistics as non-personal data[39], and were the data pseudonymised, the fact that the Department (the data controller) holds the key does not necessarily mean that the pseudonymised data as released are personal data[40];

---

[36] [2011] EWHC 1430 (Admin). The full judgment is available at: http://www.bailii.org/ew/cases/EWHC/Admin/2011/1430.html

[37] The judge drew on the House of Lords ruling in *Common Services Agency v Scottish Information Commissioner:* [2008] 1 W.L.R. 1550

[38] The judgment states: "In my view, this ground of appeal goes nowhere. As I have described, the Tribunal accepted the devastating consequences of identification. While it placed great weight on them, it concluded that these consequences were all dependent upon a patient being identified. The Tribunal was satisfied that this was extremely remote, that being the key conclusion at the end of the many paragraphs discussing the issue. Essentially, this was a matter of judgment for the Tribunal. It then acted entirely properly in reaching an overall assessment of the likelihood of identification arising from the publication of requested information."

[39] The judgment of Cranston J. included: "data would not be personal data if the other information was incapable of adding anything, and the data itself could not lead to identification, or if the data had been put into a form from which individuals to whom they related could not be identified at all, even with the assistance of the "other information" from which they were derived"

[40] Cranston J. noted the minority judgment of Baroness Hale in House of Lords case of Common Services Agency v Scottish Information Commissioner [2008] 1 W.L.R. 1550 "in short, Baroness Hale recognised difficulties with the statutory definition of personal data, but concluded that if the data could be anonymised in such a way that third parties could not identify the individuals to whom it

- **When deciding what to release, part of the role of data controllers disclosing and publishing data is risk assessment: whether the recipients are likely to seek and find additional information to enable them to identify individual(s) from within the data disclosed or published by the data controller[41];**
- **if there is little or no possibility for the recipient of the data to identify a person from the data they receive, then those data are not personal data[42].**

2.3 Applying the law: drawing the line between personal and non-personal data

2.3.1 Drawing the line between personal and non-personal data

29. Consider now the implications of the law to data, and specifically what data are personal data - identifying living individuals - and what data are non-personal data. The records of people that have died are considered in section 2.6. Assume for now that it is possible to assess accurately the intrinsic identifiability of data.

30. In the figure 1 below, imagine A is "821,013" - the total number of people living in England who were born in 1953, and C is "John Smith, born 18/6/1953, 20 Wakefield Road, Leeds LS1 1FD"[43]. It is clear that the former are non-personal data, and the latter are personal data. A does not identify any individual, and C identifies a particular John Smith.

31. Let us suppose that, whatever the context, everything to the left of point 1 is always non-personal data, (i.e. not identifying), and that everything to the right of point 2 is always personal data (identifying). Remember that point 1 is based on the test for personal data – a very low, but not zero, risk of re-identification, Let us now consider B. B is "John Smith, Leeds, born 18/6/1953, has Huntington's Disease". B falls in

---

related, it did not matter that the agency had the key which linked it back to the individual patients". Note however, that the European Commission's Data Protection Article 29 Working Party advises that if the pseudonymisation process is intentionally reversible, so that, for example, a person may be traced and contacted were a particular health problem revealed through the pseudonymised data, then they would count as personal data; see pages 18-20 of Opinion 4/2007 at: http://ec.europa.eu/justice/policies/privacy/docs/wpdocs/2007/wp136_en.pdf

[41] Cranston J.'s judgment took into account the fact that "there was no example within the past of identification from published statistical information, nor was there any evidence of information in the public domain that could be used in conjunction with these statistics so as to identify individual patients and doctors". The judge also concluded: ""To begin, the issue before the Tribunal was one of assessment: the likelihood that a living individual could be identified from the statistics. That was in my judgment only partly a question of statistical expertise, as regards matters such as the sensitivity of the data. Partly, also, it was a matter of assessing a range of every day factors, such as the likelihood that particular groups, such as campaigners, and the press, will seek out information of identity and the types of other information, already in the public domain, which could inform the search. These are factors which the Tribunal was in as good a position to evaluate as the statistical experts, a point which one of the Department of Health's experts conceded."

[42] Cranston J relied on Lord Hope's analysis in the House of Lords case of *Common Services Agency v Scottish Information Commissioner* [2008] 1 W.L.R. 1550 that "The question is whether the data controller, or anybody else who was in possession of the barnardised data, would be able to identify the living individual or individuals to whom the data in that form related. If it were impossible for the recipient of the barnardised data to identify those individuals, the information would not constitute 'personal data' in his hands". Note that the test in law is not as strong as "impossible" – a very small possibility of revealing identity would not make the data "personal data". However, the law does not provide an explicit probability level for the test.

[43] John Smith is of course a fictitious person.

between the two lines – whether B is identifiable depends on whether additional information is likely to be available and used to discover John's identity. To restate this in the terms used in the Data Protection Act, it depends on what other relevant information that could identify an individual is "in the possession, or likely to come into the possession, of the data controller".

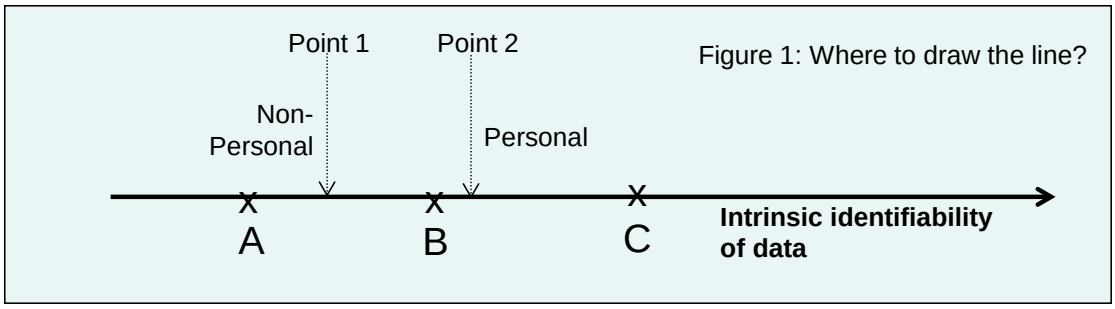

Figure 1: Where to draw the line?

32. If by chance, my brother is John Smith, born on 18/6/1953, living in Leeds, and with Huntington's Disease, then I would know that this information relates to him. So B would be personal data to me, but not personal data to you, because you do not have the extra information that I have, and no reason for, or likelihood of, discovering it. B may also be personal data for a journalist because she is able, and motivated, to gain possession of other sources of information and discover that there is only one person called John Smith living in Leeds, born on 18/6/1953, with Huntington's. Therefore, whether data are personal is determined not just by the nature of the data set itself, but also by the people who can access it and whether they are likely to get hold of additional information to discover personal identity. In other words, **whether data are personal or non-personal depends not only on the intrinsic identifiability of the data, but on the context in which they are used.**

33. Therefore, in figure 2 below, drawing on the definition of personal data, we can add another dimension to the diagram: a vertical axis of "risk of access to additional relevant information". This additional second axis is not just about what theoretical access a person might have to additional relevant information that could reveal identity, but also whether that person is likely to *make access* and thus come into possession of the additional information (because, for example, they are motivated to do it). We know that everything to the left of point 1 is non-personal data, and everything to the right of point 2 is personal data, so we can focus on the "grey area" between these two points. In a context where the risk of additional relevant information being used is at its highest point (e.g. where publishing details of the bonuses of senior banking executives), relatively little data may be published, and the maximum intrinsic identifiability of the data published is at point 1. This is shown as D on the figure 2 below.

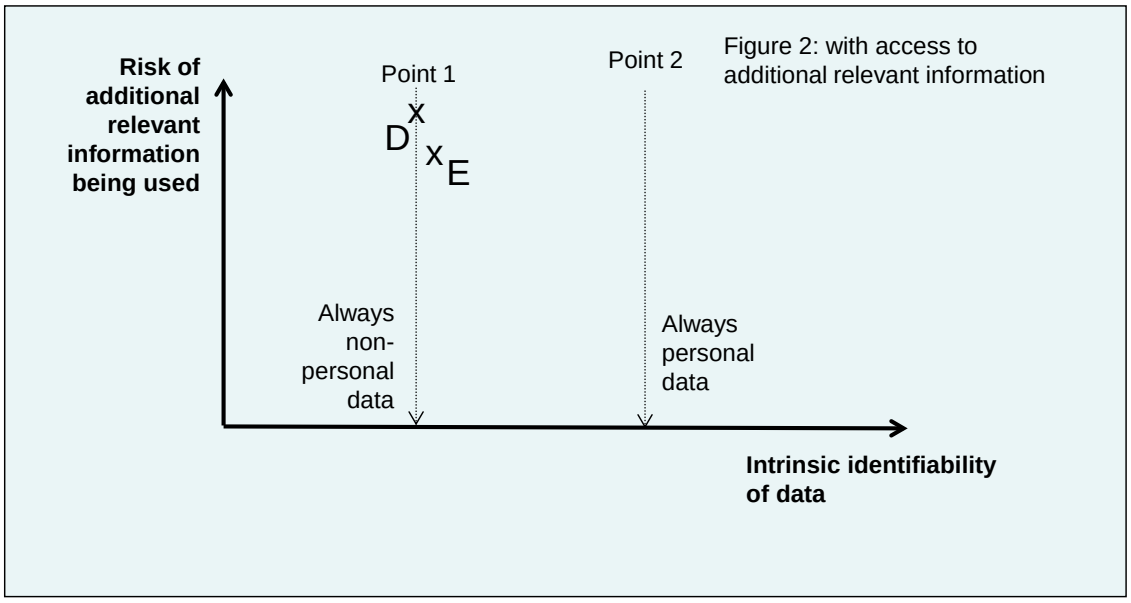

Figure 2: with access to additional relevant information

2.3.2 Drawing the line between personal and non-personal data when publishing information

34. Consider now the publication of data that will attract as little interest as possible - perhaps a breakdown of care home residents with in-growing toenails by age and local authority of residence. Although also published into the public domain, the risk that people will use additional relevant information to reveal identity is lower than when publishing figures for bankers' bonuses. Therefore, the maximum intrinsic identifiability of the data that can be published as non-personal data is a little higher than for D – shown as point E in figure 2 above. Nevertheless, even though this is the lowest risk possible risk for publishing, the risk of additional information being used is still very significant because information is being released into the public domain. Because published data are freely available to anyone, even the lowest risk publication carries significant risk. Point E marks a threshold of risk when publishing data in the "grey area" between point 1 and point 2– no release of non-identifying data into the public domain could be lower down the y-axis.

35. Therefore, **when publishing information to an unrestricted audience (e.g. on a public website), a data controller has to assume that there may be people who are motivated to seek additional relevant information to use to try to discover the identity of individuals within the information to be published. The risk of this happening will vary according to the nature of the information to be published[44] (e.g. the more value to be gained from discovering identity, the greater the threat).** Nevertheless, whenever information is published, the risks of re-identification are always relatively high.

2.3.3 Drawing the line between personal and non-personal data when processing Freedom of Information Act requests

36. The Act specifies the reasons permitting or requiring information to be withheld[45] when responding to a Freedom of Information Act request. One requirement is to

[44] Cranston J. assessed this risk in his judgment on abortion statistics – see section 2.1.

[45] See the Information Commissioner's website at: http://www.ico.gov.uk/for_organisations/freedom_of_information/guide/refusing_a_request.aspx

prevent the release of personal data that would contravene the Data Protection Act 1998. However, as the Information Commissioner's website explains[46]: "When refusing a request for information, you cannot withhold an entire document because some of the information contained within it is exempt. You must provide a redacted version of the document along with a refusal notice stating why some of the information cannot be released".

37. If a request is made for a data set that contains personal data (such as all the patient records held by a general practice), then the data are exempt, and the request should be refused. However, if a request is made under the Act for a data set containing both personal and non-personal data, then the data controller responding has to anonymise data to the extent that no personal data are revealed. Equally, the data controller must withhold no more than is necessary.

38. **The risk posed when releasing a data set in response to a Freedom of Information Act request is the same as it would be for the publication of the same data set because "disclosure under FOIA is effectively an unlimited disclosure to the public as a whole, without conditions"**[47]. Public authorities must work on the assumption that a recipient may publish the data received from the Freedom of Information Act request. Therefore, if the information on patients with in-growing toenails broken down by local authority was provided in response to a Freedom of Information Act request, it would be just as though that data were published, at point E in figure 2 above. The risk of additional information being used to reveal identity has to be assessed at the same level it would were it published, and so the maximum intrinsic identifiability (on the x-axis in figure 2) is the same.

39. Furthermore, the public authority must release to the requestor as much of what is requested as possible, so if they are asked for data close to the boundary between personal and non-personal data, the intrinsic identifiability of the data released must be close to the maximum allowable for non-personal data. **The law pulls in two opposite directions. If too much data are released, personal data are disclosed breaching the Data Protection Act and potentially confidentiality. If too little are released, the public authority may fall foul of the Freedom of Information Act. A public authority must assess where the boundary between personal and non-personal data lies in a particular case, and either publish or protect accordingly.**

2.3.4 Drawing the line between personal and non-personal data when disclosing information

40. A data controller may choose to disclose non-personal data: providing the data to a specific recipient for the use of that recipient. That recipient might be a separate organisation and data controller, a different part of the same organisation and thus the same data controller (e.g. a finance department), or a separate organisation acting as a data processor under a written contract with the data controller. Where the data are personal data, the disclosure must be justifiable under the Data Protection Act and other law. However, where the data are not personal data, those justifications are not relevant or required. In such circumstances, it is also important to understand where to draw the line between personal and non-personal data.

---

[46] http://www.ico.gov.uk/for_organisations/freedom_of_information/guide.aspx

[47] As concluded in the tribunal of *Guardian & Brooke v The Information Commissioner & the BBC*. For more information, see the Information Commissioner's website at: http://www.ico.gov.uk/foikb/PolicyLines/FOIPolicyDisclosuretopublic.htm

41. As when publishing non-personal data, the maximum intrinsic identifiability of the data disclosed will fall somewhere between point 1 and point 2 in the diagrams above. However, **the level of risk with disclosure is more controllable, and thus lower than, the previous cases of publishing and Freedom of Information Act requests (all other things being equal).** The risk of additional relevant information being used to reveal identity will be relatively low if the data recipient agrees to protect the data, and not release or publish the data in the form in which it is provided.

42. The risk that individuals' identity are revealed from the data set can be reduced through, for example, the data recipient agreeing in writing to:

- Deploy secure storage of, and restricted access to, the data set (such as through role-based access controls);
- Never publish or disclose the data in full, and only publish or disclose non-personal data derived from the data set; and
- Ensure that all individuals with access to the data agree not to publish the data, or seek additional information to reveal identity, and ensure that sanctions (e.g. dismissal) are used to deter miscreants; and
- Monitor access made to the data, and report and take action in response to any data misuse.

43. As part of judging risk, the data controller should make an assessment of the extent to which the agreed controls placed on the data will be enforced by the recipient.

44. When risk is reduced in this way, the maximum intrinsic identifiability of the data that can be disclosed as non-personal data increases, and as a general rule, the more intrinsically identifying the data, the greater its potential utility. However, even where data are disclosed into the most highly-controlled, secure environment, a significant residual risk of additional relevant information being used to reveal identity remains – indicated by point F in figure 3 below. The maximum intrinsic identifiability is greater than for publishing, but still some distance from point 2 because significant risk inevitably remains (because, for example, some people will have access and despite controls may seek to identify individuals using additional relevant information).

45. If the controls in place in the recipient's domain exist, but are less stringent, then the risk increases, and the maximum intrinsic identifiability of the data is not as high – shown in figure 3 as point G. It is clear from points plotted in the diagrams that, in theory at least, we are starting to identify a trend: the boundary between personal and non-personal data. That boundary reflects a trade-off between the two axes. **As the risk that additional relevant information being used goes up, the maximum intrinsic identifiability of the data to be released must come down. The greater the risk, the less identifying the information that can be released.**

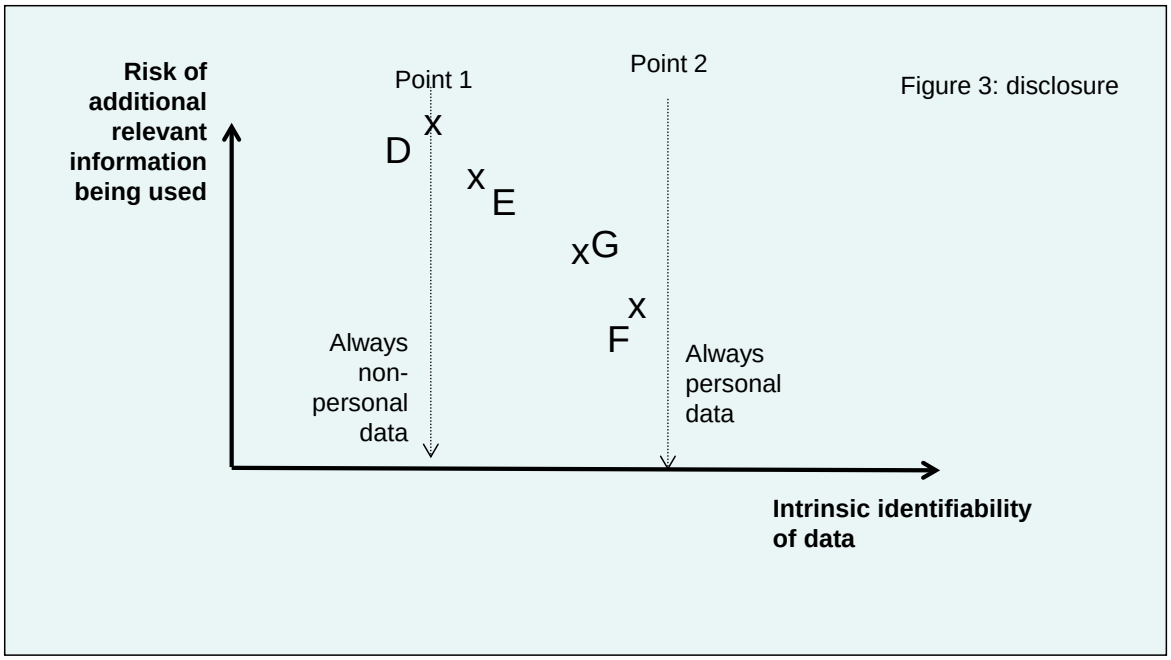

Figure 3: disclosure

46. This analysis does not try to explain the more complicated question about the precise nature of that trade-off relationship; it does not explain whether a line drawn through points D, E, F and G is straight, or curved in some way[48]. The important conclusion here is that there is a trade-off.

Providing query access to non-personal data

47. Rather than sending a non-personal data set to a recipient, a data controller may wish to provide a third party with query access to a database of personal data that is controlled by the data controller, but only allow that third party to retrieve non-personal data. Ensuring that the only data that are retrieved are non-personal data is challenging to achieve, requiring sophisticated controls on query software[49]. However, in principle the case is similar to the previous case of disclosing data, where controls can be imposed on the third party user to reduce the risk of additional information being used to reveal identity.

2.4 Individual-level data or aggregate data?

48. The implication of the Freedom of Information Act is that, if requested, public bodies controlling records that contain some personal data must be prepared to disclose as much detail as possible from these records without revealing the identity of individuals. Similarly, public bodies are being encouraged by the government to publish as much data as possible as part of the transparency agenda (see section 1). How should public authorities respond to these information requirements?

[48] Note that there is a dependency between the x-axis (intrinsic identifiability) and the y-axis (risk of additional relevant information being used) that makes more complicated the drawing of a boundary line between personal and non-personal data. However, that inter-dependency is not crucial to this analysis.

[49] For further information on some of the challenges involved, see Security Engineering by Ross Anderson, and particularly chapter 9, available at: http://www.cl.cam.ac.uk/~rja14/book.html

Should they only ever release aggregate data, or would it be lawful to publish individual-level data as non-personal data?

2.4.1 Releasing person-level data as non-personal data

49. Consider the case of a Freedom of Information Act request for full details of all complaints made to a hospital trust. The hospital holds a small simple computerised database of complaints. If the computerised complaint records contained all the details of the complaints, the patient(s) involved, and their stays in hospital, then these records could be considered as personal data about patients, and the request could be refused outright.

50. Suppose now that the database holds only summary details of each complaint, and that the full complaint records are held in paper files. The database contains a link to the paper file, summary details of the complaint like key dates and classifiers of the complaint, and a free text field. In some of the records, the free-text field contains identifying information. These records could be considered to contain a mixture of personal data and non-personal data. In principle, it would be relatively simple for the hospital to satisfy the Freedom of Information Act request by extracting all the fields except the link to the paper file and the free-text field. The extract from the database would not constitute creating new information[50], as long as the extracted data did not identify individuals.

51. However, such record-level database extracts may indeed identify individuals, and data controllers must be cautious when providing such extracts because:

- Individual-level data derived from full records about individuals are likely to pose a more than negligible risk of revealing the identity of individuals (and thus represent personal data) unless the data set released:
    a. Contains relatively few data items (for example, perhaps only three or four items) because risk of identifying an individual increases as the number of data items increases, AND
    b. Contains only a relatively small number of individual-level records (for example, fewer than 100) because the risk that some individual is identifiable increases with the number of records;
- "outliers" – rare values – can arise and need to be controlled in any data item (in the case of the complaints example, the staff member involved, the date the complaint was made, or the hospital ward the patient was on, might all be unique).

52. The above conclusions apply equally to the publication of non-personal data as well as Freedom of Information Act requests. **When data are released into the public domain, the risks of a person's identity being revealed are such that relatively little individual-level data can be released[51].** The standard specifies steps for establishing what individual-level data may be released when publishing.

---

[50] See pages 4-5 of "Do I have to create information to answer a request?" from the Information Commissioner's Office available at: http://www.ico.gov.uk/for_organisations/freedom_of_information/information_request/~/media/documents/library/Freedom_of_Information/Detailed_specialist_guides/INFORMATION_FROM_ORIGINAL_SOURCES.ashx

[51] One study found that of 1000 medical records, three data items were safe, while with four data items one individual record could be found, and with 10 data items most records could be isolated. See section 9.3.3.5 in chapter 9 of "Security Engineering" by Ross Anderson, available at: http://www.cl.cam.ac.uk/~rja14/book.html

2.4.2 Releasing aggregate data as non-personal data

53. The clear implication of *Department of Health v Information Commissioner*[52] (see section 2.2.1) is that small numbers in cells need only be altered to prevent the identification of individuals from tables of data in certain circumstances. **Whether publishing, responding to a Freedom of Information Act Request, or disclosing non-personal data, an assessment has to be made of the likelihood of additional information being used to reveal identity. When publishing, it has to be assumed that such information is likely to be used if it is available.** However, as in the *Department of Health v Information Commissioner* case, there are instances when modifying cells so as not to risk revealing the identity of data subjects is irrelevant because of the scale of the population to which the data relate. For example, where data relate to the whole of England, then there will be no risk, or negligible risk, of a person's identity being revealed. When data are broken down, for example by geographical area, the risk may be such that disclosure control is required. This issue is addressed directly in the standard[53].

2.5 The difficulties of accurate risk assessment

54. The analysis above illustrates the importance of risk assessment when determining the data to be released when publishing, disclosing data to specific recipient(s), or responding to a Freedom of Information Act request. Assessing the likelihood that personal identity could be revealed is an essential step in meeting the requirements of the law. The analysis above assumes that accurate risk assessment is possible, and that the intrinsic identifiability of data can be measured. **In practice, assessing the risk that additional relevant information will be used by others to reveal identity is difficult because of lack of reliable information about the variables influencing risk.** For example, a data controller is very unlikely to know:

- All of the information that is already publicly and privately available that might be used in conjunction with the data to be released to reveal identity;
- All of the information that will become publicly and privately available in future that might be used in conjunction with the data to be released to reveal identity;
- The potential value of the data to be released for those who might use it to reveal identity (if that were possible);
- The variety of motivations of, and all of the techniques[54] used by, people that might wish to use the data released to reveal personal identity.

---

[52] [2011] EWHC 1430 (Admin). The full judgment is available at: http://www.bailii.org/ew/cases/EWHC/Admin/2011/1430.html

[53] A rule of thumb is provided from the Chief Statistician, on page 8 of GSS / GSR Disclosure Control Policy for Tables Produced from Administrative Data Sources available at: http://www.ons.gov.uk/ons/guide-method/best-practice/disclosure-control-policy-for-tables/index.html . It suggests a threshold population size to which a published cell relates of around 100,000, allowing for maximum risk. However, the guidance was published before the case of *Department of Health v Information Commissioner*, and is now subject to review. The anonymisation standard sets lower thresholds, dependent on risk.

[54] For useful guidance on potential methods, see chapter 9 of "Security Engineering" by Ross Anderson, available at: http://www.cl.cam.ac.uk/~rja14/book.html

55. **Furthermore, where a data set is relatively large (thousands or millions of records), the data controller is very unlikely to know enough about the content of data held to accurately assess the intrinsic identifiability of the data.**

56. **Nevertheless, the law requires a public authority to make reasonable efforts to gather relevant information so as to make a fair assessment of risk based on the information available, and decide where to draw the line between personal and non-personal data.** *Anonymisation Standard for Publishing Health and Social Care Data* provides a standard approach for reaching a justifiable assessment of risk.

2.6 Applying the law to records of the deceased

57. The discussion so far has ignored the records of dead people. "Personal data", as defined in the Data Protection Act 1998, applies to living people, and excludes data about the deceased. This implies that health and social care records of dead people count as "non-personal data". In law, the confidentiality of health and social care records "arguably"[55] continues after a person's death. Furthermore, records of the deceased may contain information about other people e.g. living relatives. Department of Health policy[56], General Medical Council guidance to doctors[57], and common practice within health and social care, is to treat records of the deceased in the same way as those of living individuals, with certain exceptions allowing access set out in the Access to Health Records Act 1990.

58. Therefore, for the purposes of the standard[58], the conclusion is drawn that the law does protect the confidentiality of the health and social care records of the deceased. On that basis, it would be unlawful to publish these records, and thus they would be exempt from Freedom of Information Act requests. In other words, **for the purposes of anonymisation, the records of the deceased can be treated in just the same way as the records of living people.**

59. For this reason, the anonymisation processes that follow in the standard do not refer to "personal data" and "non-personal data", but instead to "identifying data" and "non-identifying data". **"Identifying data" is exactly the same as the concept of "personal data" defined in the Data Protection Act, except that "identifying data" also includes data about the deceased.**

---

[55] See "Judge allows disclosure of workers' medical records after death", BMJ2008;337doi: 10.1136/bmj.a1794(Published 23 September 2008), and also Lewis v Secretary of State for Health (Defendant) & Michael Redfern QC (Interested Party), [2008] EWHC 2196 (QB), reported at: www.1cor.com/1315/?form_1155.replyids=1171. See also *Bluck v ICO & Epsom and St Helier University Hospital NHS Trust* [EA/2006/0090], available at: http://www.informationtribunal.gov.uk/DBFiles/Decision/i25/mrspbluckvinformationcommissioner17sept07.pdf, where, for the circumstances of the case, it was found that confidentiality was protected after death.

[2008] EWHC 2196 (QB)

[56] See section 1.2 of "NHS Information Governance - Guidance on Legal and Professional Obligations" published in October 2007, and available at: http://www.dh.gov.uk/en/Publicationsandstatistics/Publications/PublicationsPolicyAndGuidance/DH_079616

[57] See paragraph 30 of the General Medical Council has produced guidance for doctors entitled: Confidentiality: Protecting and Providing Information. It is available at: http://www.gmc-uk.org/guidance/current/library/confidentiality.asp

[58] A court might decide differently in certain specific circumstances – for example, records of people dead for more than 50 years may be deemed as not confidential.

2.7 Identifying but not confidential data

60. It is unlawful to publish data that is identifying if it reveals something confidential about a person. In some circumstances, a publication could reveal the identity of data subjects but no confidential information about them, and be lawful. For example, publishing a list of NHS Numbers with their district postcode and/or age band (with no other information) would identify people but not release anything confidential about them, because their age and address is already in the public domain (e.g. through the electoral register). However, the data would still be personal data and so processing would have to be necessary (e.g. to serve a public function). But as long as there was such a public interest reason to release the data, publication would not be unlawful.

61. Similarly, suppose a publication contained information about the prevalence of a particular diagnosis. If a person can self-identify, and recognise themselves and their diagnosis from the data, no confidence is being broken.

62. A publication that listed consultant code, or the identifier of a treating ward nurse, along with data about hospital episodes that identified no patients, would not be releasing confidential information even if the staff identifiers and names were available in the public domain. It would reveal only the kind of activity carried out by the nurse and consultant as part of their work[59]. Whether that data can be released depends on whether it breaches the Data Protection Principles; useful guidance on this has been published by the Information Commissioner[60]. The guidance makes clear that releasing sensitive personal data about a person (such as the ethnic category or disability status of the consultant or nurse) is almost certainly unfair and thus breaches Data Protection Principles. Information about the role, work activities and salaries of employees, and especially senior staff, is often publishable[61] as long as it is justifiable in the public interest.

2.8 Conclusions about the law and applying the law

63. The main conclusions drawn about the law are highlighted in bold throughout section 2. They are reproduced below.

    a. Even where statistics relate to matters of a very sensitive nature (as with abortion), these are not grounds for withholding statistics that pose only a remote risk of revealing a person's identity.

    b. When deciding what to release, part of the role of data controllers disclosing and publishing data is risk assessment: whether the recipients are likely to seek and find additional information to enable them to identify individual(s) from within the data released.

    c. If there is little or no possibility for the recipient of the data to identify a person from the data they receive, then those data are not personal data.

---

[59] Although, in exceptional circumstances (such as for staff who might be under threat if known to carry out abortions), this could be withheld in the public interest.

[60] See (in particular pages 6-17 of) the guidance on Freedom of Information and personal data, available at : http://www.ico.gov.uk/Global/faqs/~/media/documents/library/Freedom_of_Information/Detailed_specialist_guides/PERSONAL_INFORMATION.ashx

[61] See page 5 of the Information Commissioner's Office Freedom of Information Act guidance at: http://www.ico.gov.uk/upload/documents/library/freedom_of_information/detailed_specialist_guides/awareness_guidance_1_-_personal_information.pdf

d. Whether data are personal or non-personal depends not only on the intrinsic identifiability of the data, but on the context in which it is used.

e. When publishing information to an unrestricted audience (e.g. on a public website), a data controller has to assume that there may be people who are motivated to seek additional relevant information to use to try to discover the identity of individuals within the information to be published. The risk of this happening will vary according to the nature of the information to be published (e.g. the more value to be gained from discovering identity, the greater the threat).

f. The risk posed when releasing a data set in response to a Freedom of Information Act request is the same as it would be for the publication of the same data set because "disclosure under FOIA is effectively an unlimited disclosure to the public as a whole, without conditions".

g. The law pulls in two opposite directions. If too much data are released, personal data are disclosed breaching the Data Protection Act and potentially confidentiality. If too little are released, the public authority may fall foul of the Freedom of Information Act. A public authority must assess where the boundary lies in a particular case, and either publish or protect accordingly.

h. When disclosing data into a controlled domain, the level of risk with disclosure is more controllable, and thus lower than, the previous cases of publishing and Freedom of Information Act requests (all other things being equal).

i. When data are released into the public domain, the risks of a person's identity being revealed are such that relatively little individual-level data can be released.

j. A trade-off exists: as the risk that additional relevant information being used goes up, the maximum intrinsic identifiability of the data to be released must come down. The greater the risk, the less identifying the information that can be released.

k. Whether publishing, responding to a Freedom of Information Act Request, or disclosing non-personal data, an assessment has to be made of the likelihood of additional information being used to reveal identity. When publishing and responding to a Freedom of Information Act Request, it has to be assumed that such information is likely to be used if it is available.

l. In practice, assessing the risk that additional relevant information will be used by others to reveal identity is difficult because of lack of reliable information about the variables influencing risk. Furthermore, where a data set is relatively large (thousands or millions of records), the data controller is very unlikely to know enough about the content of the data held to accurately assess the intrinsic identifiability of the data. Nevertheless, the law requires a public authority to make reasonable efforts to gather relevant information so as to make a fair assessment of risk based on the information available, and decide where to draw the line between personal and non-personal data.

m. For the purposes of anonymisation, the records of the deceased can be treated in just the same way as the records of living people (as "identifying data). "Identifying data" is exactly the same as the concept of "personal data" defined in the Data Protection Act, except that "identifying data" also includes data about dead people.

n. In some circumstances, a publication could reveal the identity of data subjects but no confidential information about them, and be lawful.

64. The *Anonymisation Standard for Publishing Health and Social Care Data* has been developed directly from these conclusions about the law.

3 Anonymisation scenarios

3.1 Introduction

65. The purpose of this section is to illustrate through three scenarios how the anonymisation standard can be applied by health and social care organisations when publishing non-identifying data. In each sub-section below, a problem scenario is posed, and is followed by a response to the scenario.

66. To enable the responses to the scenarios about publishing to be matched back to the standard anonymisation processes in the *Anonymisation Standard for Publishing Health and Social Care Data*, relevant processes and process steps are referenced [**in square brackets**]. The first scenario illustrates how a simplified process may be applied to a simpler publishing situation.

3.2 Anytown General Practice responds to a Freedom of Information Act request

Problem scenario

67. Dr. Finlay, the senior partner in Anytown General Practice, has received a letter from Mr. Amir Khan, a patient registered with the practice, who is asking, for the clinic sessions scheduled for the coming week, a list of all of the booked and all of the unbooked slots. For each booked slot, he asks for the name, date of birth, and ethnic origin of the booked-in patient. He says that he is frustrated that he always has to wait several days for an appointment, despite the urgency of his case. Mr. Khan says he wishes to understand whether the practice has too few sessions, and is chronically overbooked, or whether he is unsuccessful because he is the subject of discrimination because he is elderly and of Asian descent.

Response to scenario

68. Dr. Finlay, as information governance lead, reviews Mr. Khan's request. She had never considered whether general practices had to respond to Freedom of Information Act requests, and confirms that they do[62]. It is clear that the request cannot be met in full because some of the data that Mr. Khan is asking for would identify individual patients. However, even though the practice does not currently store specifically a list of booked slots of the kind referred to in the request, the relevant information is held in a database and could be extracted without great difficulty or great cost, and so she cannot reject the request as a whole on the grounds that the information is not held[63]. Nor could she reject the request on the grounds that booking slots are entirely identifying data and so exempt. However,

[62] See paragraph 44 in Schedule 1 of the Act at: http://www.legislation.gov.uk/ukpga/2000/36/schedule/1?view=plain and the guidance from the Information Commissioner at: http://www.ico.gov.uk/for_organisations/freedom_of_information/application.aspx

[63] See pages 4-5 of "Do I have to create information to answer a request?" available at: http://www.ico.gov.uk/for_organisations/freedom_of_information/information_request/~/media/documents/library/Freedom_of_Information/Detailed_specialist_guides/INFORMATION_FROM_ORIGINAL_SOURCES.ashx

some of the data Mr. Khan seeks are identifying data and need to be withheld from the response. [**Identify nature of information to publish and data source(s)**].

69. Dr. Finlay identifies that both name and date of birth could reveal the identity of individual patients. She decides to withhold name, and to substitute date of birth with five year age range. She decides to withhold booking time from the list of slots because Mr Khan, or anyone else with whom he shares the list, could discover the identity of someone booked into a particular time, and discover some new information about that person from the list (i.e. self-declared ethnicity). Ethnic origin on its own will not identify a patient in Dr. Finlay's practice, so it can be included, as can the other information he seeks about booking slots. This revised data set provides Mr. Khan with relevant information to address his expressed concerns. She specifies the revised data set to be extracted from the practice database, and asks the practice manager to run the query. [**Assess risk and specify data anonymisation**].

70. Dr. Finlay runs her eyes over the list of slots generated by the query. [**Derive data from data sources**]. She confirms that it does not contain identifying data.[**Review/test data provided are non-identifying**] She asks the practice manager to respond to Mr. Khan with the list, along with a brief explanation of why certain information was withheld.[**Publish**]

3.3 Anytown Local Authority plans to publish information on care home MRSA

Problem scenario

71. Transparency of local services is an important objective of Anytown Local Authority. After a series of reports in the local newspaper of people living in care homes bringing Methicillin-Resistant Staphylococcus Aureus (MRSA) into the local hospital, the chief executive instructs Jim, the Chief Information Officer, to research and then publish as much detailed information as possible on the incidence of MRSA amongst care home residents who are funded, or part-funded, by the local authority.

Response to scenario

72. After a considerable amount of time, Jim receives the information he needs from all of the care homes in the authority: the number of local authority-funded people in the care home and how many of those have MRSA (when last seen by their doctor). He considers what information he can publish, and decides that as long as it was non-identifying, he would like to publish a breakdown of MRSA by care home [Identify nature of information to publish and data source(s)].

73. He thinks about the interest that members of the public and journalists might have in identifying individual care home residents with MRSA. Despite the recent publicity, he does not consider the identification of individuals as being newsworthy or of special interest, and so he assesses the interest in the data as low and scores the threat level as "normal" [**Assess threat level associated with data and its release**].

74. Jim recognises that the breakdown of MRSA by care home could be matched against other data already published about care homes and their residents, and were this done, it could potentially lead to the identification of individual residents with MRSA. However, he knows this is always true whenever publishing data about residents of the authority, and there were no special factors to make him think that the information available made the re-identification risk especially high. Considering this along with the threat level, he scores the overall risk of extra information being used to try to reveal identity as "normal".[**Assess risk of extra information being used to try to reveal identity.**]

75. Given this, Jim has a choice of three standard anonymisation plans:

- Plan 1: Where cells to be published relate to population > 1,000 people, derive aggregate data without statistical disclosure control
- Plan 2: Where cells to be published relate to population≤1,000 people, derive aggregate data with statistical disclosure control
- Plan 3: Derive individual-level data to "weak" k-anonymity.

76. Having read a little about K anonymity, he knows that because he is publishing about a relatively small number of individuals, and there are less than five local authority-funded residents with MRSA in most of the care homes, he realises that k-anonymity will not allow him to publish individual-level data, which include the care home of residents with MRSA. Each care home has only a small number of residents, so he realises that if he were to publish MRSA data broken down by care home, statistical disclosure control under plan 2 would mean that he has to obscure all of the cells under five, which would be virtually every cell in the published tables. Therefore, he decides to publish aggregate data for the local authority as a whole, showing the number of local authority-funded residents, along with the number of those who have MRSA, adopting plan 2 because of the small population basis [**Select anonymisation plan**]. He knows that the numbers are large enough that statistical disclosure control will not affect anything, and so the final publication will still be useful.

77. He runs through his thinking with the authority's Caldicott Guardian [**Refine anonymisation plan and specify anonymisation, Consult Caldicott Guardian / SIRO**], and then pulls together the figures [**Derive non-identifying data from data sources**]. He confirms they are as he expected [**Review/test data provided are non-identifying**] and agrees with the chief executive to issue them through a press release to the local papers [**Publish**].

3.4 The National Centre for Hospital Health publishes individual-level data extracted from Hospital Admissions Records

Problem scenario

78. The chief executive of the National Centre for Hospital Health is keen to meet the challenge of the government's data transparency agenda, and asks Maria, the organisation's information governance lead, to anonymise and publish Hospital Admissions Records, not in aggregate form, but as individual-level data that could be used for research, public health and other medical purposes. He asks her to report back later that week. The Centre is data controller for the Hospital Admission Records. Maria is uneasy about the request, as the Centre has never published individual-level clinical data before. Indeed, she does not remember any comparable publication by a national body. She agrees to investigate the matter, liaise with the Caldicott Guardian, and come back to the chief executive as soon as possible.

Response to scenario

79. Maria knows the kind of information that the chief executive is looking for, and confirms that all the necessary data are accessible, including the reference files [**Identify nature of information to publish and data source(s)**]. She begins by considering the threats. Hospital Admission Records exist for almost every person in England, and contain sensitive and revealing information that, were a person's identity to be revealed, could be of great interest and value to journalists and some other members of the public. Given this, she scores the threat level as "high" [**Assess threat level associated with data and its release**].

80. Maria is all too aware that individual-level data extracted from the Hospital Admissions Records can be matched with numerous possible data sources to reveal identity. Some of the conditions are rare and not evenly distributed across the population, and so especially vulnerable to re-identification. Because of this, and when taking account of the threat level, she has little doubt that the risk of extra information being used to try to reveal identity is "high" [**Assess risk of extra information being used to try to reveal identity**].

81. In order to meet the chief executive's objective, she can only pick anonymisation plan 6: "Derive individual–level data to strong k-anonymity". She knows enough about k-anonymity to know that this will only allow diagnosis plus a few quasi-identifiers like age range and district postcode to be published. She decides to go directly to the Centre's Caldicott Guardian to check that he agrees with her analysis. He broadly accepts her assessment [**Consult Caldicott Guardian / SIRO**], but is concerned that the risks of publishing are still too high. He knows, for example, that populations are not evenly distributed, and in some areas, just a handful of people may live a district postcode. He advocates refining the standard anonymisation plan, to publish individual-level data with especially strong k-anonymity (for example, if k=10, and with all published attributes including diagnosis were controlled by k-anonymity) [**Refine anonymisation plan and specify anonymisation**]. Maria feels caught between the Caldicott Guardian and the Chief Executive. She tells the Caldicott Guardian that she will run the k-anonymity software both under the standard plan [**Derive non-identifying data from data source(s)**], and under the especially stringent requirements sought by the Caldicott Guardian, to discover what data sets would be publishable under each plan [**Review/test data provided are non-identifying**].

82. Under the Caldicott Guardian's anonymisation plan, they find that what is publishable is very limited indeed, whereas Maria feels that by deploying the standard anonymisation plan 6, they could publish data of value. She and the Caldicott Guardian realise that before publication, more sophisticated risk assessment and testing would be required, requiring further time and resources. Before going any further, they decide to go together to talk to the chief executive to try to resolve what anonymisation plan they should deploy, and what data they should aim to publish (if anything).