Chapter 4

Law, governance and security

Chapter 4 – Law, governance and security

Chapter overview

This chapter discusses the effectiveness of legal, technical and administrative measures to protect privacy in the face of advances in information technology and data science.

Privacy is protected by a number of overlapping legal measures, principally: formal privacy rights, which guarantee freedom from interference; rules of data protection, which control the 'processing' of various kinds of 'personal data'; and duties of confidentiality, which bind people in certain professional relationships.

A number of technical measures are used to prevent the identification of individual subjects, including aggregation, anonymisation or pseudonymisation of data. While de-identification measures may help to protect privacy they may not make re-identification impossible. The risk of re-identification is both difficult to quantify and may become greater over time. De-identification should therefore be combined with further controls on the access to and uses of data.

A standard control is to limit access to data in accordance with consent. Broad consent allows subjects to set certain parameters for the use of the data but it may be hard to foresee the relevance of certain implications and the scope of consent given when data are collected may become unclear in changing circumstances, especially over long time periods. Continuing involvement of subjects through 'dynamic' forms of consent can address this but is potentially demanding.

Neither anonymisation nor compliance with consent offer sufficient protection from potentially harmful consequences of data use. Additional controls on the use of data – on who is permitted access, for what purposes, and how they must conduct themselves – are required. These have both administrative and technical aspects.

Data initiatives are increasingly caught in a double bind by the obligation to generate, use and extend access to data while, at the same time, being obliged to protect privacy as a moral obligation and a requirement of human rights law.

Introduction

4.1 The legal framework applicable to data use in biomedical research and health care recognises, broadly, two sorts of measures that may be applied to protect the interests of citizens against potentially injurious misuse of data. First, it recognises operations that alter the data in order to de-identify them so that their use no longer poses a direct risk to data subjects through them being identified. Second, it recognises controls on access to data so that the data are only made available to authorised users, in circumstances in which they are expected not to be misused or to otherwise result in harm to data subjects. These measures are often used in combination. The law permits and prohibits data processing according to the kinds of additional measures taken. In this chapter we will consider the kinds of measures in use and their principal shortcomings in the face of advances in information technologies and data science, and the changing data environment. We will also consider the way in which conventional measures have been modified to address these difficulties.

Legal framework for use of biological and health data

> **Proposition 14**
>
> How data are managed, used and re-used is as morally relevant as how they are classified or how they were obtained.

4.2 Many of the data used in biomedical research and health care come from people. At the point at which they are collected from a person they are *personal data*, data that are related to that individual.[177] The processing of data that relate to living individuals is governed by rules set out in data protection law. Some personal data may also be *private*, data to which the individual does not wish others to have access. Access to and disclosure of private data is governed by privacy norms that refer to relationships between those individuals (or groups) and others. Some disclosures of data may potentially cause harm to individuals. To disclose data without proper respect for the individuals concerned may infringe their rights. The overlapping legal frameworks governing the use of data are engaged variously by whether or not data are personal data (data protection), by the transgression of established norms (confidentiality) or the infringement of privacy rights. We discuss these frameworks in outline below; information governance measures for specific health and research systems are described in chapters 6 and 7 respectively.

A legal right to privacy

4.3 Given the moral significance of privacy and the consequences of violating privacy norms, some of these norms and the means of protecting them have been set down in law. The legal traditions both in the US and Europe, though very different in many respects, have nevertheless evolved protections for privacy by way of concerns for individual liberty and human dignity.[178]

4.4 The legal right to privacy arose initially as a defence of property in the early modern age.[179] The utilitarian philosophers, Jeremy Bentham and John Stuart Mill, subsequently developed the defence of a sphere of self regulation and private action against interference by others, and especially by public authorities (see chapter 3). The development of distinct informational privacy rights came about in the context of developments in mass communication technologies in the late 19th Century (such as mass-circulation newspapers, photography and telegraphy). Reflecting on the injury to individuals that resulted from uncontrolled dissemination of information about them, the US jurists Warren and Brandeis influentially sought to frame a new right to privacy as a 'right to be let alone'.[180] This right was distinct from the right to property (since

[177] 'Personal data' is a technical (and contested) concept in data protection. See paragraph 4.8.

[178] See Whitman JQ (2004) The two Western cultures of privacy: dignity versus liberty (Yale Law School Faculty Scholarship Series, paper 649, available at: http://digitalcommons.law.yale.edu/cgi/viewcontent.cgi?article=1647&context=fss_papers.

[179] Locke's *Second treatise of* government (1690) suggests that the primary purpose of government is to protect property rather than to pursue common ends.

[180] Warren SD and Brandeis LD (1890) The right to privacy *Harvard Law Review* **4(5)**: 193-220, available at: http://www.jstor.org/stable/1321160?seq=1#page_scan_tab_contents. They note that a right such as they propose "has already found expression in the law of France" (the *Loi Relative à la Presse*, 11 Mai 1868). Warren and Brandeis consider the desirability of criminal protection but their proposal is for a civil tort, pending further legislation. The law on privacy has been developed by the US courts since it was first formulated.

their concerns went beyond the theft of intellectual property), and was not based on any implied contract or trust (since privacy rights are not exercisable against a specific individual but are 'rights against the world'[181]). Instead it was based on "the more general right to the immunity of the person – the right to one's personality", although it was recognised that this right must be limited by public interest.[182]

4.5 In the 20th Century a European right to respect for private life was provided by the European Convention for the Protection of Human Rights and Fundamental Freedoms (ECHR) and a number of other high-level legal and regulatory instruments.[183] The citizen's 'right to respect for his private and family life, his home and his correspondence', guaranteed by Article 8 of the ECHR, is not absolute but is qualified to permit interference with the right when it is necessary for the protection of the rights and freedoms of others and a number of other specific public interest purposes. Where privacy rights are engaged, determining whether they are violated requires balancing the claims of the victim against the justification offered for the infringement and assessing whether the infringement is necessary and proportionate to the achievement of those aims according to supposed norms. The idea that informational privacy is connected to the right to one's personality has been developed in more recent jurisprudence from the European Court of Human Rights that takes up aspects of the 'right to privacy and family life', guaranteed by the ECHR.[184]

4.6 A related idea, which acknowledges the indelibility, indefinite reproducibility and ease of recall of digital information, is the so-called 'right to be forgotten', that is, to have personal information – including public information – expunged from records. The recent decision of the European Court of Justice in the *González* case established a right to have information provided by Internet search companies removed if it infringed individual privacy.[185] This is significant in that it implicitly acknowledges the impact on privacy of features of information technologies that are not relevantly new in kind but extraordinarily greater in power (the significance does not lie in the difference between the informational value, the persistence or truth value of electronic records compared to, for example, paper records, but their power to impinge on personality).[186]

[181] Warren and Brandies note that "since the latest advances in photographic art have rendered it possible to take pictures surreptitiously, the doctrines of contract and of trust are inadequate to support the required protection, and the law of tort must be resorted to." (ibid., at page 211).

[182] Ibid., at page 207. There must be a "line at which the dignity and convenience of the individual must yield to the demands of the public welfare or of private justice"; more general guidance is suggested to stem from jurisprudence relating to libel and slander, as well as intellectual property. (ibid., at page 214).

[183] ECHR Article 8 (available at: http://www.echr.coe.int/Documents/Convention_ENG.pdf). The Convention is transposed into UK law by the Human Rights Act 1998 (available at: http://www.legislation.gov.uk/ukpga/1998/42). Similar rights are included in related instruments (such as other Council of Europe Conventions and the Charter of Fundamental Rights of the EU (available at: http://www.europarl.europa.eu/charter/pdf/text_en.pdf) and UN Declarations.

[184] It is even more strongly embodied in the concept of *Persönlichkeitsrecht* developed in the German courts; see Consultation response by Atina Krajewska and Ruth Chadwick, available at: www.nuffieldbioethics.org/project/biological-health-data/evidence-gathering/.

[185] *Google Spain SL and Google Inc. v. Agencia Española de Protección de Datos (AEPD) and Mario Costeja González* (Case C-131/12). The court found that search engines must consider requests for delisting of results that 'appear to be inadequate, irrelevant or no longer relevant, or excessive in relation to those purposes and in the light of the time that has elapsed' (para.93), subject to exceptions relating to public figures and to balancing the data subject's fundamental rights with the rights of others to information. Google subsequently established a procedure through which people might apply to have their names removed from Google's index and received a large number of applications.

[186] See paragraph 3.8.

Data protection law

4.7 Data protection law does not concern privacy as such, but rather the 'processing' of 'personal data'. Personal data are, broadly, data that relate to a living individual who can be identified from those data or from a combination of those and other available data. [187] An early impetus for data protection legislation was the fear that governments would increasingly develop centralised computer 'data banks' containing information about their citizens.[188] But, as the U.S. Privacy Protection Study Commission concluded in 1977: "The real danger is the gradual erosion of individual liberties through the automation, integration, and interconnection of many small, separate record-keeping systems, each of which alone may seem innocuous, even benevolent, and wholly justifiable."[189]

4.8 In the UK and in Europe data protection has been successively framed by a set of relatively stable principles.[190] The central tenet of data protection law is that personal data should be processed fairly and lawfully. The requirement for fairness places stress on the fact that the person processing the data (the 'data controller') has made reasonable efforts to ensure that those to whom the data relate ('data subjects') are aware of who is processing the data and for what purposes.[191] The legislation furthermore treats certain categories of data, including health data as being, for the most part, *sensitive* personal data.[192] Consequently, more exacting requirements apply to the processing of these data. The requirement for fairness links the acceptability of different types of processing to the understanding and expectations of the people to whom the data relate. A number of legal grounds for processing data are given (broadly, where the processing is necessary for a number of prescribed purposes or where the processing is carried out with the consent of the data subject or in their own vital interests). Furthermore, the laws of most countries acknowledge that there are circumstances in which the objections of data subjects may justly be disregarded or overridden, for reasons ranging from the protection of minors to the notification of serious infectious diseases. In Europe, exceptions must be made by means of law that is sufficiently clear for its consequences to be foreseeable.

---

[187] The concept of 'personal data' is frequently contested, particularly in relation to the extent of other information that may reasonably be expected to be available to be combined with the information in question in order to identify the subject, i.e. the context of processing (or possible processing) is important. Not all personal data are collected from a subject: some may be generated from non-personal data. Furthermore, some personal data may relate to more than one individual – data about members of the same family, for example.

[188] See, for example, the Younger Committee (1972) *Report of the committee on privacy,* Cmnd. 5012 (London: HMSO).

[189] U.S. Privacy Protection Study Commission (1977) *Personal privacy in an information society*, available at: https://www.ncjrs.gov/pdffiles1/Digitization/49602NCJRS.pdf, at page 533.

[190] The UK's Younger Report (op.cit.) had 10 principles; the OECD (1980) Guidelines on the protection of privacy and transborder flows of personal data honed eight (available at: http://www.oecd.org/internet/ieconomy/oecdguidelinesontheprotectionofprivacyandtransborderflowsofpersonaldata.htm), which found their way into successive domestic Data Protection Acts (1984, (http://www.legislation.gov.uk/ukpga/1984/35/pdfs/ukpga_19840035_en.pdf) and 1998 (http://www.legislation.gov.uk/ukpga/1998/29; Schedule 1 Part 1) and the EU data protection Directive 95/46/EC (http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=CELEX:31995L0046:en:HTML).

[191] See Data Protection Act 1998, Schedule1, Part 2.

[192] See Data Protection Act 1998, s.2. Although the legislation specifies 'sensitive personal data' as personal data that fall into a number of pre-defined categories there is some support for a construction that makes the context of processing relevant to whether data is 'sensitive' as well as whether it is 'personal': see *Common Services Agency v. Scottish Information Commissioner* [2008] UKHL 47 *per* Lord Hope, at 40 (available at: http://www.publications.parliament.uk/pa/ld200708/ldjudgmt/jd080709/comm-1.htm). Whether or not it is the case in law, from a commonsense perspective, not only whether data is 'identifying' but also whether it is about 'health', can depend on the context in which it is placed, as we argued above.

4.9 The multinational nature of contemporary commercial organisations, health systems and biomedical science mean that data will often be expected to travel across jurisdictional boundaries and their associated protective measures.[193] At the time of writing a draft General Data Protection Regulation (GDPR) is making progress through the European lawmaking procedure.[194] The intention of the GDPR is both to update EU law to account for advances in information technology and to harmonise its implementation across the Union. Unlike the existing Directive, which must be transposed into national law, allowing account to be taken of national legal traditions, a Regulation becomes directly applicable law in each Member State. The final form of the Regulation is currently unclear, although its principal mechanisms are likely to be similar to those of the existing Directive. It may, nevertheless, have significant impact on the extension of access to health data, depending on the provisions adopted with regard to consent.[195]

Common law

4.10 Confidentiality is an important way of codifying expectations about how data will be handled. These expectations may be created by professional relationships (such as that between a doctor and a patient) or through explicit undertakings or contracts. Where they are not made explicit in this way, the legitimacy of expectations about the use of data relies not only on a subjective element (the individual's own expectations) but also a social element (whether society is prepared to recognise that expectation as reasonable).[196] Case law establishing a tort (civil offence) of the misuse of 'private information' has been developing in England and Wales, for which the threshold test is whether the person publishing information knows or ought to know that there is a reasonable expectation that the information in question will be kept confidential.[197] What is reasonable will depend on the context and the moral interests at stake.

4.11 The common law in England and Wales, and developing case law in Scotland, provides a duty of medical confidentiality (essentially a ground to sue for any breach of confidentiality). This is further codified in professional guidance, as well as through contractual agreements, that allow the conditional disclosure of information for specific purposes and provide assurance that it will not be disclosed further than necessary for that purpose or used for other purposes (especially those that might cause detriment to the patient).

4.12 The duty of confidentiality cannot be absolute: a strong justification, particularly one that involves the protection of others, can license a breach of confidence. There is a body of case law that addresses the balance of competing interests for lawful breach of

[193] The EU Data Protection Proposals restrict transfers of personal data outside the European Economic Area (EEA), except with explicit consent, though only addresses data held in the EU. The Regulation will be extra-territorial when data is held on EU citizens overseas (http://ec.europa.eu/justice/data-protection/data-collection/data-transfer/index_en.htm). This mirrors the reach of the US Patriot Act, which can require any US company to provide personal data held by them (whether as controller or processor) to the National Security Agency.

[194] See: http://ec.europa.eu/justice/data-protection/review/index_en.htm.

[195] In the version adopted by the European Parliament, for example, specific consent would be required for the use of data (including pseudonymised data) in research. Research organisations have claimed that this will have a serious negative impact on the conduct of research that is currently lawful (see footnote 124).

[196] The concept of 'legitimate expectation' is one that has arisen in administrative law in England and Wales; the test of reasonable expectation of privacy was applied in *Campbell v. Mirror Group Newspapers Ltd* [2004] UKHL 22 (available at: http://www.publications.parliament.uk/pa/ld200304/ldjudgmt/jd040506/campbe-1.htm); a similar 'concept of reasonable expectation of privacy' has been developed by the US Supreme Court (see: *Smith v. Maryland* [1979] 442 U.S. 735, 740, available at: https://supreme.justia.com/cases/federal/us/442/735/case.html).

[197] See *Campbell v. MGN Limited* [2004] UKHL 22 *per* Hale L.J at 134. This was confirmed recently in *Vidal Hall and Ors v Google Inc* [2014] EWHC 13 (QB) in which 'personal' and 'private' information are considered separate 'types' of information.

confidentiality.[198] Furthermore, there are recognised situations, provided for in legal instruments along with additional controls, in which confidentiality may (and in some cases, must) be set aside. (For example, in England and Wales, the Police and Criminal Evidence Act 1984 provides that a judge may order that the police may have access to medical records for the purpose of a criminal investigation).[199] Perhaps most relevantly, section 251 of NHS Act 2006 (formerly section 60 of Health & Social Care Act 2001) creates a power, exercised under Regulations, to set aside the common law duty of confidentiality in certain circumstances, permitting the processing of confidential patient information for medical purposes, without the consent of the data subject, in specified circumstances and subject to various controls.[200]

Security of data

Operations designed to prevent the identification of data subjects

Aggregation

4.13 A great deal of useful research, particularly in the area of public health, can be carried out using aggregated data. Indeed this has been the major underpinning of much epidemiological and aetiological research that has led the better understanding of health and disease. This use is analogous to the way that data produced by the UK Office for National Statistics (ONS) supports development of government policy and secondary academic research. Of course, the data have to be collected in the first place, so will be personal data prior to their aggregation but, once aggregated, the privacy interests of research participants are usually thought to be protected.

4.14 Nevertheless, it is sometimes possible to pick data relating to individuals out of aggregate data. As a simple example, professorial salaries at most universities are confidential but aggregate data may be available. If a department has only one female professor, and publishes the average salary for all professors, then it cannot publish the average salary for all male professors, since that would allow the female professor's salary to be worked out with ease. A statistic that leaks individual data is

[198] See *W v. Edgell* [1990] 1 All ER 835 (in which a psychiatrist sent a confidential expert opinion on the fitness of a criminal to be moved from a secure hospital to the medical director of the hospital and to the Home Office in the public interest); *X v. Y* [1988] 2 All ER 648 (in which a Health Authority successfully sought to prevent publication by a newspaper of the names of doctors receiving treatment for AIDS).

[199] http://www.legislation.gov.uk/ukpga/1984/60, s.9. Other statues mandate the submission of otherwise confidential information to a public authority, such as the Public Health (Control of Disease) Act 1984 (http://www.legislation.gov.uk/ukpga/1984/22, s.11) and Public Health (Infectious Diseases) Regulations 1988 Reg.6 (http://www.legislation.gov.uk/uksi/1988/1546/made) (provides that a doctor must notify the relevant local authority officer if they suspect that a patient has a 'notifiable disease'), the Human Fertilisation and Embryology Act 1990 (http://www.legislation.gov.uk/ukpga/1990/37, s.31) (creates a statutory register of gamete donors and fertility treatments), the Abortion Regulations 1991 Reg.4 (http://www.legislation.gov.uk/uksi/1991/499/contents/made) (mandatory notification of abortion procedures), the Births and Deaths Registration Act 1953 (http://www.legislation.gov.uk/ukpga/Eliz2/1-2/20) (mandatory procedures for informing a relevant authority about births and deaths), and the Children Act 2004 (http://www.legislation.gov.uk/ukpga/2004/31, s.12) (duties on authorities to co-operate in order to safeguard or promote the welfare of children).

[200] See: http://www.legislation.gov.uk/ukpga/2006/41/section/251. The relevant regulations are the Health Service (Control of Patient Information) Regulations 2002 (http://www.legislation.gov.uk/uksi/2002/1438/made). This power is only applicable in England & Wales – Scotland's legislators were not moved to provide a similar mechanism. Following the passage of the Care Act 2014 (http://www.legislation.gov.uk/ukpga/2014/23/contents/enacted), the procedure depends formally on advice from the Confidentiality Advisory Group, an independent committee hosted by the Health Research Authority that advises the HRA/Secretary of State for Health on the merits of data access applications (see: http://www.hra.nhs.uk/resources/confidentiality-advisory-group/). In all cases the Data Protection Act will apply, especially the 7th Principle. See also: Health Research Authority (2012) Principles of advice: exploring the concepts of 'public interest' and 'reasonably practicable', available at: http://www.hra.nhs.uk/documents/2014/12/v-2_principles_of_advice_april_2013.pdf.

called a 'tracker'. Trackers have been studied for over 30 years and are increasingly relevant to medicine. In 2008 and 2009 the *Public Library of Science (PLoS) Genetics* published a series of papers demonstrating how an individual subject could be identified in aggregate genomic data.[201] While this does not in itself imply that the individuals identified could be traced from that data alone or identified in another context, the individual-level data extracted could potentially be linked with other datasets leading to positive re-identification (see below), thereby making it potentially personal data.[202] For this reason we have to consider anonymisation and pseudonymisation more carefully.

Anonymisation

4.15 Anonymisation – literally the removal of the name – used to be done by simply blanking out a person's name and address from a paper record. This gives some privacy from casual inspection but dates of birth, postcodes and other distinctive elements of linked information can be used to re-identify individuals with relative ease. For practical identification, phenotype data, photographs (which are common in some medical databases) and even behavioural data (an individual's 'mobility pattern' for example) can effectively identify individuals.[203] In a birth cohort study where the week of a child's birth is already known, a little further information, such as sex, birth weight, mode of delivery, etc. is probably sufficient to pick out an individual in the dataset. To achieve 'anonymity' increasing amounts of associated data must be stripped away to give confidence that re-identification is no longer possible.

4.16 Research is often carried out on anonymised data, such as Genome-Wide Association Studies (GWAS), in which researchers have genome data from two populations, one with a trait of interest and the other without, which they compare in order to identify variations correlated with the trait. The holding of genomic datasets often raises concerns because they constitute "a biometric that can be used to track and identify individuals and their relatives."[204] While the identification of blood relatives may not be feasible with other biomarkers, such as the proteome and microbiome, they may be equally informative in other ways and both offer, in effect, as precise a personal

[201] The first of these was Homer N, Szelinger S, Redman M, *et al*. (2008) Resolving individuals contributing trace amounts of DNA to highly complex mixtures using high-density SNP genotyping microarrays *PLoS Genetics* **4(8)**: e1000167, available at: http://www.plosgenetics.org/article/info%3Adoi%2F10.1371%2Fjournal.pgen.1000167#pgen-1000167-g003. Similar articles followed and these prompted the NIH to amend their anonymisation policy at the time (http://www.genomicslawreport.com/index.php/2009/10/28/back-to-the-future-nih-to-revisit-its-genomic-data-sharing-policies/; http://www.sciencemag.org/content/322/5898/44.1.long#ref-2). Schadt, Woo and Hao (2012) support the assumption that some people can be identified in most individual level biomedical and health record data sets, see: Schadt EE, Woo S and Hao K (2012) Bayesian method to predict individual SNP genotypes from gene expression data *Nature Genetics* **44(5)**: 603-8.

[202] See Ehrlich Y and Narayanan A (2014) Routes for breaching and protecting genetic privacy *Nature Reviews Genetics* **15(6)**: 409-21, available at: http://www.nature.com/nrg/journal/v15/n6/full/nrg3723.html.

[203] de Montjoye Y-A, Hidalgo CA, Verleysen M, and Blondel VD (2013) Unique in the crowd: the privacy bounds of human mobility *Scientific Reports* **3**: 1376, available at: http://www.nature.com/srep/2013/130325/srep01376/full/srep01376.html?utm_source=feedburner&utm_medium=feed&utm_campaign=Feed%3A+mediaredef+%28jason+hirschhorn%27s+Media+ReDEFined%29. Whether or not data can be 'intrinsically identifying' is a conceptual problem that turns on the propositional/recognitional meaning of 'identifying'. It is possible to argue that no data set is identifying (without a specific context) or, alternatively, that every datum is identifying in some context.

[204] Consultation response by GeneWatch UK, available at: www.nuffieldbioethics.org/project/biological-health-data/evidence-gathering/. See Heeney C, Hawkins N, De Vries J, Boddington P and Kaye (2010) Assessing the privacy risks of data sharing in genomics *Public Health Genomics* **14(1)**: 17-25, available at: http://www.karger.com/Article/Abstract/294150: "once genomic data is publicly released, it is virtually impossible to retrieve it or to make it private again, or even to know who has the information or to what use it is being put." (at page 22).

'fingerprint' as the genome itself.[205] The most important difference is not only that we can sequence the genome efficiently but that we also have large and growing databases (such as those held by most national police forces, and those of firms like 23andMe) to link genomic data to identifiable people. For anonymisation to fail unique data is insufficient; it must also be capable of being linked to a living person (see Box 4.3 below).

Box 4.1: Re-identification: some examples

Case A: During the height of the Bovine spongiform encephalopathy (BSE)/'Mad Cow disease' scare, a doctor interviewed on television mentioned that he had seen a teenage vegetarian girl who had contracted new variant Creutzfeldt-Jakob disease (nvCJD). The media succeeded in identifying the girl within a few days, and the doctor was subsequently brought before the GMC Disciplinary Committee for breach of confidence. He was cautioned by the Committee, despite having attempted (unsuccessfully) to hide her identity by speaking in generalities.

Case B: An 'anonymous' sperm donor in the USA was identified and traced by a 15-year-old who had been born through the use of his donation. By using a genetic ancestry test and a commercial database a surname was suggested for men who shared his Y chromosome characteristics. That could then be used to narrow the search around information that his mother had been given at the time of treatment which, finally, led to the identification of the donor.[206]

Case C: Group Insurance Commission (GIC), a purchaser of health insurance for employees, released records of state employees to researchers, having removed names, addresses, social security numbers, and other identifying information, in order to protect the privacy of these employees. As a demonstration, Latanya Sweeney purchased voter rolls, which included name, zip code, address, sex, and birth date of voters in Cambridge MA (USA) and, by combining the voter roll information with GIC's data, was able to identify data relating to the Massachusetts governor who had assured residents of their privacy. (From GIC's databases, only six people in Cambridge were born on the same day as the governor, half of them were men, and the governor was the only one who lived in the zip code provided by the voter rolls.) The information in the GIC database on the Massachusetts governor included medical diagnoses and prescriptions.[207]

Case D: Researchers found the individuals to whom 50 'anonymous' DNA sequences belonged that were posted online for the purposes of scientific research by querying other public databases. The researchers were not only able to assign names to the DNA sequences but could also begin to identify family traits.[208]

[205] Hawkins AK and O'Doherty KC (2011) "Who owns your poop?": insights regarding the intersection of human microbiome research and the ELSI aspects of biobanking and related studies *BMC Medical Genomics* **4**: 72, available at: http://www.biomedcentral.com/1755-8794/4/72/.

[206] See: http://www.bionews.org.uk/page_12558.asp.

[207] Sweeney L (2002) k-anonymity: a model for protecting privacy International Journal on Uncertainty, Fuzziness and Knowledge-based Systems 10(5): 557-70, available at: https://epic.org/privacy/reidentification/Sweeney_Article.pdf.

[208] Gymrek M, McGuire AL, Golan D, Halperin E, and Erlich Y (2013) Identifying personal genomes by surname inference *Science* **339(6117)**: 321-4, available at: http://data2discovery.org/dev/wp-content/uploads/2013/05/Gymrek-et-al.-2013-Genome-Hacking-Science-2013-Gymrek-321-4.pdf.

Pseudonymisation

4.17 In some cases, it may be desirable for the process of de-identification to be reversible, for example, to feed back information to an individual within a cohort who is discovered to be at particular risk, or to validate an analytical procedure, or to enable further data about individuals to be added over time.[209] This is possible where, rather than being removed, identifiers are replaced with a unique code. A simple approach that was recommended by the first Caldicott report was the use of the NHS number in place of a patient's name.[210] In the context of communications within health services this was an improvement on using names, as the data were at least not obviously identifiable, and so reduce the risk of accidental disclosure, although in this case a very large number of people have access to the key meaning that re-identification would be easy for insiders. This could be improved by removing overtly identifying information on a medical record, such as name, address, postcode, hospital number and NHS number, and replacing it with an encrypted NHS number, encrypted using a key held by a Caldicott guardian.[211] Pseudonymisation mechanisms are often much cruder than this, however.

4.18 Using more complex pseudonymisation mechanisms at the scale of a health service is not straightforward. Pseudonymisation at source can work well in some instances, such as adverse drug reaction reporting, where records from different sites do not need to be linked. If it is done centrally, the question arises of whether the local health care providers (and the patients) trust the centre to do it properly. It is possible to use a 'trusted third party'; for example, the Icelandic health service had a system whereby care records were sent from GPs and hospitals to the data protection authorities, who removed patient identifiers, replaced them with an encrypted version of the patient's social security number and sent the record on to the secondary uses database. However, even that system was vulnerable to data insertion attacks; by adding a new record to a patient's file and then looking at the secondary database, an insider could still identify patients there.

4.19 It is usual with either anonymisation or pseudonymisation to redact or obfuscate other fields as well as removing direct identifiers, e.g. limiting to postal area and to age (or age group) rather than full postcode and date of birth. Furthermore, data are routinely encrypted to protect communications between web browsers and web servers, and encryption offers an additional layer of security for data held in cloud storage to support collaborative research.[212] Data values may also be randomly perturbed to maintain

---

[209] Deryck Beyleveld, for example, argues that, given a wide concept of privacy (and other rights that can apply), true anonymisation can violate privacy rather than protecting it. For example, the right to privacy arguably includes a right to know the personal implications for oneself of research but this is rendered impossible by anonymisation; individuals arguably have a privacy right (under suitable conditions) for medical research to be conducted. Beyleveld D (2011) Privacy, confidentiality and data protection, in *The SAGE handbook of healthcare ethics*, Chadwick R, Ten Have H, and Meslin EM (Editors) (London: SAGE), pp95-105.

[210] Department of Health, The Caldicott Committee (1997) *Report on the review of patient-identifiable information* (recommendation 8), available at: http://webarchive.nationalarchives.gov.uk/20130107105354/http://www.dh.gov.uk/en/Publicationsandstatistics/Publications/PublicationsPolicyAndGuidance/DH_4068403. It is worth noting that this approach is not as strong when used with the Scottish CHI number which includes embedded information about gender and date of birth.

[211] See paragraph 2.43.

[212] These approaches allow statistical analysis to be carried out on encrypted data without direct access by researchers, who only see the results. Such approaches have been used, for example, to study data from individuals affected by stigmatising conditions such as antimicrobial resistant organisms and HIV. See El Emam K, Arbuckle L, Essex A, *et al*. (2014) Secure surveillance of antimicrobial resistant organism colonization or infection in Ontario long term care homes *PLoS ONE* **9(4)**: e93285, available at: http://www.plosone.org/article/info%3Adoi%2F10.1371%2Fjournal.pone.0093285.

statistical validity but reduce the risk of re-identification.[213] Such techniques were used in the system that was the subject of the *Source Informatics* case, which started to establish UK case law on anonymisation.[214]

4.20 There are many other techniques that can be used in statistical disclosure control. For example, one may answer each query based on only a sample of the population data, so that slightly different queries are answered on the basis of quite different sets of data, and tracker attacks therefore become difficult.[215] However, no technique is without vulnerabilities.

Weaknesses of de-identification

> **Proposition 15**
>
> De-identification of individual-level data is, on its own, an unsafe strategy for ensuring the privacy of individuals to whom the data relate. This can only be expected to become more unsafe with the continued accumulation of data (see Proposition 1) that makes potentially identifying linkages possible and with the increasing power and availability of analytical tools (see Proposition 2) that can realise this potential.

4.21 If enough identifying data are removed from a dataset then one may be able to assert that the data are sufficiently anonymous to pose little risk of re-identification. For example, a file consisting of just the gender of every person in the UK (60 million records of just one field of one character) would clearly be anonymous, provided the fields are only coded 'M' or 'F'.[216] The problem is that such a redacted file is virtually useless.

4.22 There are a few applications where anonymised data can be and are safely used. The classic case is the system that was the subject of the *Source Informatics* case. This is used to analyse doctors' prescribing habits to generate information that is then sold to drug companies so they can assess the effectiveness of their sales representatives. Neither doctors nor patients are identified and repeat prescriptions are not linked. In effect the system records how many prescriptions each doctor wrote for each drug in each time period, with the data being perturbed (deliberately altered) to prevent inference attacks.[217]

4.23 While unlinked episode data can be used for some purposes, most researchers want to link up successive episodes of care so they can analyse health outcomes. Hence the use of pseudonymisation to ensure that data from different datasets can be properly

[213] The procedure of randomly altering data values prior to publication to prevent identification is known as Barnardisation after the mathematician, George Alfred Barnard. The question of what constitutes 'anonymised' data, and the status of 'barnardised' data under the provisions of the Data Protection Act 1998 was considered by the House of Lords in *Common Service Agency v Scottish Information Commissioner* [2008] UKHL 47, available at: http://www.publications.parliament.uk/pa/ld200708/ldjudgmt/jd080709/comm-1.htm.

[214] *R v Department of Health, ex parte Source Informatics* [2001] QB 424. See paragraph 4.22.

[215] Greater use of sampling methods was one of the approaches advocated in response to our consultation (Professor Sheila M Bird OBE FRSE, see: www.nuffieldbioethics.org/project/biological-health-data/evidence-gathering/).

[216] A few cases coded for transgender conditions might permit some of the records to be associated with particular individuals but with no actual disclosure as one would need to know all the relevant information anyway to achieve the identification.

[217] For technical details, see: Matyáš V (1998) Protecting the identity of doctors in drug prescription analysis *Health Informatics Journal* **4(3/4)**: 205-9.

correlated, to ensure that individuals are not counted twice, and to allow the validation of analyses performed on them. However, linking is only possible with access to the original algorithm or key-file, so someone must hold what are, in effect, personal data. There are three possibilities:

- the data can be linked by the source
- the data can be linked by the recipient
- the data can be linked by a third party

4.24 How appropriate each of these approaches may be will depend to a great extent on the circumstances of the particular initiative, although the use of third parties is becoming increasingly accepted as good practice for data linking. (We will consider specific examples in chapters 6 and 7 below.) However, even if the technical mechanisms for removing identifiers or replacing them with pseudonyms are sound, the increasing richness of the digital data environment, combined with the availability of analytical tools, presents a significant challenge by increasing the risk of re-identification.

Box 4.2: Technical anonymity

The *anonymity set* is the set of all individuals with whom a data subject might be confused; thus if instead of being named, someone is merely described as "a Member of Parliament" the anonymity set consists of the set of Members of Parliament.

The *privacy set* is the set of people to whom a data subject requires that a given sensitive datum not be disclosed. For most data subjects and most sensitive data, the privacy set may consist of friends, family, colleagues and enemies – perhaps a hundred individuals (though for celebrities and in some particular contexts the privacy set may be essentially everyone). For any recorded datum, the privacy set may change substantially over time, as an individual's circumstances change.

There is a failure of anonymisation if the anonymity set is reduced to one from the viewpoint of anyone in the privacy set. This will happen if the dataset is available to someone in the privacy set (although privacy will remain so long as no one in the privacy set has the means or inclination to perform the re-identification or access to the necessary data).

4.25 Individuals can be identified within an anonymity set by processes of deduction or inference. Where the anonymity set is small, additional brute force inquiries may also work. The basic risk of deductive re-identification arises because a person disclosing information often does not know what other information is available to the person to whom the information is disclosed.[218] This is complicated by the fact that they may originally disclose it in the context of a specific relationship (e.g. to a hospital administrator) but be unaware of other relationships in which the information recipient may stand with respect to the subject (neighbour, family member, etc.), at that time or

[218] The UK Government has been criticised for failing adequately to transpose Recital 26 of the Data Protection Directive, which stipulates that "to determine whether a person is identifiable, account should be taken of all the means likely reasonably to be used either by the controller or by any other person to identify the said person" (see: http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=CELEX:31995L0046:en:HTML). As presently drafted, the EU General Data Protection Regulation contains a similar recital (recital 23: "The principles of protection should apply to any information concerning an identified or identifiable person. To determine whether a person is identifiable, account should be taken of all the means likely reasonably to be used either by the controller or by any other person to identify the individual.", http://ec.europa.eu/justice/data-protection/document/review2012/com_2012_11_en.pdf). See also the Information Commissioner's Code of Practice on anonymisation: https://ico.org.uk/media/for-organisations/documents/1061/anonymisation-code.pdf.

in the future. However, this is not merely a risk arising from the 'linking' of study data with 'environmental' data available to people handling the data but arises in the controlled context of linkage studies themselves. As one of our consultation respondents argued:

> "Linking can be done very nearly as well with data pseudonymised at source as with identifiable data and so no longer requires identifying data, and would not per se usually require patient consent. Very few secondary uses require identifying data for any other reason. However linked data is richer than the data from any single source, and may well be so potentially identifiable that it has to be treated as identifying data, as the DPA (Data Protection Act) 1998 states it should."[219]

4.26 The significance of the data context is recognised in the way that anonymisation is understood in legal instruments. For example, the Article 29 Working Party (the European advisory body on data protection established under Article 29 of the European Data Protection Directive) describe an effective anonymisation solution as one that "prevents all parties from singling out an individual in a dataset, from linking two records within a dataset (or between two separate datasets) and from inferring any information in such a dataset."[220] As they note, this implies that simply removing directly identifying elements is generally not enough. Consequently, additional measures, depending on the context, will usually be required to prevent individual identification or record linking. These measures must take the data context into account, so standardised anonymisation protocols will usually be insufficient and anonymisation must therefore be sensitive to risk of re-identification.[221] Furthermore, future-proofing is bound to be difficult where the data are to be retained for long periods. So as well as anonymisation, some further maintenance and control of the context will also be required.

4.27 It seems difficult to conclude that the privacy of a data subject can be guaranteed by any predetermined set of de-identification measures. The privacy of the data subject depends upon what tools and other information are available to those who have access to the data, and whether the potential viewer is a benign researcher or disinterested administrator, or a malicious and motivated attacker. If it is therefore not tenable to consider data simply as either identifying or not based on the nature of the data alone, we have to think in terms of a continuum that includes:

- data that are identifying in most contexts (proper names, and addresses, photographic portraits, etc.);

[219] Consultation response by Ian Herbert, available at: www.nuffieldbioethics.org/project/biological-health-data/evidence-gathering/. The HSCIC is currently looking into the utility and security of data pseudonymised at source. See: https://www.whatdotheyknow.com/request/review_of_pseudonymisation_at_so#incoming-496410; https://www.gov.uk/government/publications/data-pseudonymisation-review. See also: Data linkage and Data Quality sub-group (https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/385954/Data_Linkage__Data_Quality_Sub_Group_Terms_of_Reference_V1.1.pdf; Report (http://www.hscic.gov.uk/media/14828/HSCIC-Data-Pseudonymisation-Review--Interim-Report/pdf/HSCIC_Data_Pseudonymisation_Review_-_Interim_Report.pdf).

[220] Article 29 Data Protection Working Party (2014) *Opinion 05/2014 on anonymisation techniques*, available at: http://ec.europa.eu/justice/data-protection/article-29/documentation/opinion-recommendation/files/2014/wp216_en.pdf.

[221] The response of the NIH to the article by Homer et al. to generally restrict access to GWAS data was criticised as to harsh (http://www.nature.com/news/2008/080904/full/news.2008.1083.html). See also Erlich Y and Narayanan A (2014) Routes for breaching and protecting genetic privacy *Nature Reviews Genetics* **15(6)**: 409-21, available at: http://www.nature.com/nrg/journal/v15/n6/full/nrg3723.html.

- data that are contingently identifying in conjunction with readily/publicly available data;
- data that are contingently identifying in conjunction with not-readily-available data but data that may be available, either already or in the foreseeable future, to someone seeking to re-identify individuals from the data.

4.28 One might think of these as data that are identifying to: (1) anyone, (2) a nosey neighbour, and (3) a motivated attacker, perhaps with the kind of resources available to national security services.[222] However, the critical thing is the extent to which advances in data science and information technology may narrow the gap between (2) and (3) and, indeed, shift all the boundaries. This is particularly relevant for data that may remain sensitive for a long time into the future. The distinctions between these three segments of the spectrum concern contingent technical thresholds and thresholds of confidence (*e.g.* the absence of an adequate combination of skill and will to misuse data) rather than robust categorical distinctions. This has implications for governance: the unsettled and indefinite limitations of privacy through anonymity mean that there will be a need for continuous monitoring and reflective control of disclosures.

4.29 There seems to be a broad consensus, which is supported by respondents to our consultation, that irreversible anonymisation of meaningful data is practically unattainable given the availability of data tools and environmental data for contextualisation.[223] This is now increasingly accepted in policy circles, too.[224] Re-identification now has to be considered not only as a theoretical possibility but also as a practical one. However, this risk is very difficult to quantify because a number of factors will usually be uncertain, such as the nature and availability of contextual information, the range of people in the 'privacy set' who have an interest in re-identifying an individual, their motives, intentions, resources and technical capabilities, and how all these things may change over time. Nevertheless, the days when both policymakers and researchers could avoid privacy issues by simply presuming that anonymisation was an effective privacy mechanism are drawing to a close. Henceforth, claims that privacy can be assured through anonymisation when data are accessible to a large or indefinite number of people should be treated with suspicion. People who provide data in the context of health care and research will need to be made aware of this.

Controlling data access and use

4.30 Within health care and biomedical research, the conventional approach to any extension of data access (for example, when health information is communicated outside the immediate context of confidentiality created by the provision of treatment) was encapsulated in the injunction 'consent or anonymise'. Where the purpose could be achieved with sufficiently de-identified data, this was often preferred over seeking consent as this is assumed to be most convenient and to minimise the risk to the data

[222] The ICO Code of Practice on anonymisation introduces the concept of a 'motivated intruder' test (https://ico.org.uk/media/for-organisations/documents/1061/anonymisation-code.pdf, pp22-4), although this is limited to an intruder who is a member of the public (not enjoying any specific legal powers), does not possess any specialist knowledge such as computer hacking skills, or have access to specialist equipment or resort to criminality in order to gain access to data that is kept securely. A difficulty is identifying the scope of motivated intruders for the lifetime of a data resource, which may be as long, or longer, than the lifetime of the data subject.

[223] A 2010 paper by Paul Ohm (see:Ohm P (2009) Broken promises of privacy: responding to the surprising failure of anonymization *UCLA Law Review* **57**: 1701-77, available at: http://www.patents.gov.il/NR/rdonlyres/E1685C34-19FF-47F0-B460-9D3DC9D89103/26389/UCLAOhmFailureofAnonymity5763.pdf) gave rise to a debate in legal and policy circles on the appropriate response to computer science research on re-identification techniques.

[224] See, for example, evidence to the House of Commons Health Select Committee on 1 July 214 (available at: http://www.parliament.uk/business/committees/committees-a-z/commons-select/health-committee/inquiries/parliament-2010/cdd-2014/).

subject.[225] As we have seen, there is a genuine difficulty establishing what level of de-identification will produce reliably 'anonymous' data.

Consent

4.31 Given the limited utility of effectively anonymised datasets and the technical difficulties of linking pseudonymised data, data initiatives that re-use health and biomedical data have often found it necessary to seek the consent of data subjects to provide them with a legitimate ground for their activities.[226] The practice of obtaining consent for the use of medical information was historically poor, although it is improving. For many years, the NHS accepted that simple compliance could be used as a sufficient signal of consent: when the phlebotomist asked the patient to roll up their sleeve, doing so could be taken as consent to the drawing of blood. But the drawing of blood is not an end in itself, and consent to a number of data processing and data generating (e.g. pathology) procedures are also implicit in the sleeve-rolling request. 'Implicit consent' was used to recognise data processing that was necessary in order to provide health care and treatment to an individual, in terms of direct care and the running of health care services.[227] There may be cases in which the underlying norms are so well established, and the implications so broadly accepted, as to make implicit consent a legitimate default.

4.32 A problem arises when incompatible norms are in play. Patients may assume that data about them will be used to support their own direct care, while health care professionals and medical researchers may operate under assumptions that secondary use of patient data is routine and unproblematic. Following the first Caldicott report matters started to improve and a typical GP practice now has a notice in the waiting room informing patients that their data may be used for research unless they opt out.[228] In such cases, only a few people may make use of the opt-out, which is unlikely to frustrate the objectives of the data use. Underlying this is a keen awareness on the part of policymakers that most people will accept the offered defaults. (If consent to secondary uses of health records is opt-in, few people will bother and medical research will be compromised; if consent is opt-out, again few people will bother, and medical research can proceed freely.)[229] For this reason, choosing the default option is a morally significant decision. There is also the practical problem of finding a way to

[225] The 'consent or anonymise' rule was established by the Patient Information Advisory Group (later the Ethics and Confidentiality Committee of the National Information Governance Board and now the Confidentiality and Advisory Group of the Health Research Authority).

[226] There are other grounds for the lawful processing of personal data in most relevant data protection legislation (e.g. Directive 95/46/EC) but, arguably for reasons of compatibility with the ECHR, consent is generally sought.

[227] It is sometimes suggested that sleeve-rolling can be seen as tantamount to seeking to enter into a contract. A problem with relying on implicit consent is that the EU Data Protection Directive did not consider the seeking of healthcare services from a state provider to be seeking or entering into a contract (unlike private healthcare) nor do Member State laws recognise these consequential processes within their own laws. In other words, the underlying norms to which a procedure needs to refer for its legitimacy are formally absent.

[228] For the first Caldicott report, see The Caldicott Committee (1997) *Report on the review of patient-identifiable information*, available at: http://webarchive.nationalarchives.gov.uk/20130107105354/http://www.dh.gov.uk/en/Publicationsandstatistics/Publications/PublicationsPolicyAndGuidance/DH_4068403. Patients who have attempted to opt out have, however, faced very significant hurdles. Following pressure from privacy campaigners, the Health and Social Care Information Centre is now intending to produce more effective opt-out mechanisms.

[229] This can be seen in the context of government interest in 'libertarian paternalism', which means allowing citizens to choose, but setting the defaults so that those who are not motivated to choose otherwise will end up with what is considered 'good for them'. See: Thaler RH and Sunstein CR (2009) *Nudge: improving decisions about health, wealth, and happiness* (New Haven: Yale University Press), and the Behavioural Insights Team, or 'Nudge Unit': https://www.gov.uk/government/organisations/behavioural-insights-team.

support what are effectively registers of those who do not wish their data to be used or to be contacted for research studies, and to ensure that the opt-outs are respected across all systems.

4.33 So-called 'broad consent' is a solution that invites people to agree to parameters for the use of data without specifying the fine detail. It is sought, for example, from volunteers who sign up for a biobank study and give blood samples as well as consent for their records to be used for all kinds of activities falling within a general description of 'medical research'. Broad consent is not necessarily the opposite of 'specific' consent since it may be both broad and specific – covering a wide range of activities for specified purpose such as research into the causes of complex diseases just as much as it may be narrow (for very limited uses) but vague. Nevertheless, broad consent typically operates at a higher level of abstraction in contrast to more narrow consent that has a clearly defined method and aim in sight. Crucially, it also contains the possibility of consenting to unforeseen and possibly as-yet-unimagined uses as long as their morally relevant features are encompassed within the description of what has been 'consented to'.[230] This is why, even if the scope of the consent to use of data can be circumscribed (e.g. by a general criterion such as 'for medical purposes') there can be serious ethical issues if, for example, the data are used selectively for private gain rather than public good.

4.34 An alternative way of overcoming the scope problems of 'one-off' consents (narrow or broad), and that does not require data users to constantly seek new or refreshed consent, is to engage the active participation of the data subjects. So-called 'dynamic consent' allows control of data access by individuals, enabled by mechanisms such as consent portals.[231] These mechanisms also provide a way of informing participants about opportunities for, and outcomes of, the use of data, and can be configured to allow them to set a number of preferences and choose the level of their engagement. Continuing participation can have the advantage of allowing participants to shape the possibilities of research through their decisions about what uses of data to permit by effectively 'voting' for those uses by consenting to them. However, some commentators have raised concerns that dynamic consent may not be suitable or serviceable if the data are used for many purposes, such as reuse of health data for service planning.[232]

4.35 Dynamic approaches to consent may have the advantage of being consistent with the more stringent data protection requirements currently being proposed in the GDPR, in

---

[230] See Manson NC and O'Neill O (2007) *Rethinking informed consent in bioethics* (Cambridge: Cambridge University Press).

[231] See Kaye J, Whitley EA, Lund D, *et al*. (2014) Dynamic consent: a patient interface for twenty-first century research networks *European Journal of Human Genetics* (advance online publication), available at: http://www.nature.com/ejhg/journal/vaop/ncurrent/full/ejhg201471a.html and Bernal P (2010) Collaborative consent: harnessing the strengths of the Internet for consent in the online *environment International Review of Law, Computers and Technology* **24(3)**: 287–97, available at: https://ueaeprints.uea.ac.uk/28370/1/Collaborative_Consent.pdf. Consumer facing companies have begun to emerge whose business models range from purchased online heath record services, to free services where the company exists to exploit the data they control on behalf of their customers. Companies like Miinome and Allfiled provide platforms to secure compensation or other benefits for data subjects in return for allowing use of personal data by third parties: "Technology Company Allfiled believes there is money to be made in providing a platform that gives each individual consumer control of his or her own data." (http://www.forbes.com/sites/trevorclawson/2014/03/14/data-disruption-putting-control-of-information-in-the-hands-of-consumers/).

[232] See objections to opt-in for care.data where it is argued that losing a small percentage of records would lead to selection bias. Tim Kelsey, National Director for Patients and Information, NHS England, giving evidence to the House of Commons Health Committee on 1st July 2014: "The evidence is really clear that the people who need health services most receive them least, and they are also the least likely to opt in if we were to offer that service. If we want an inclusive national health service, we have to be able to plan in the interests of the entire community, particularly for those who would be least likely to opt in to the care.data scheme.", available at: http://data.parliament.uk/writtenevidence/committeeevidence.svc/evidencedocument/health-committee/handling-of-nhs-patient-data/oral/11192.pdf, at page 49.

particular the requirement for explicit consent to the use of sensitive personal data.[233] They are being explored by international 'big data' initiatives such as those of the International Medical Informatics Association (IMIA) and the Global Alliance for Genomics and Health.[234] It may be argued that, in a context of greater personalisation, 'responsibilisation' and 'consumerisation', individuals will be able – and may be required – to exercise greater choice over how they interact with health systems, public services and other actors.[235] The National Programme for IT (NPfIT, now defunct) experiment with HealthSpace was discouraging, but other applications such as PatientsLikeMe and HealthUnlocked report high levels of engagement from motivated patients, not to mention the popularity of health-related 'apps'.[236] Mechanisms of this sort have been promoted as enablers of new forms of participant-driven research or 'citizen science'.[237] (We discuss participant-led research initiatives further in chapter 7, below.)

4.36 Concerns about the scope of consent – although they are not the only concerns – might be obviated if people donate data or tissue samples on a completely unlimited basis. A model is offered by the use of 'portable legal consent'.[238] This provides a kind of open source licence for the use of data. However, only a highly altruistic minority of people are likely to be prepared to give completely unlimited permission of this sort. By analogy, in the world of open-source software, some code is licensed without limits (for example, under the FreeBSD license) while much more code is published subject to the condition that people who adapt it must also share their adaptations, which encourages its use in collaborative or cooperative contexts. It is quite possible that many people would agree to their data being used only in not-for-profit research but have a different opinion if research is conducted by a private company (see chapter 5 below).

Difficulties for consent in data initiatives

> **Proposition 16**
>
> Where a person providing data about themselves cannot foresee or comprehend the possible consequences when data are to be available for linkage or re-use, consent at the time of data collection cannot, on its own, be relied upon to protect their interests.

[233] See GDPR (draft), http://ec.europa.eu/justice/data-protection/document/review2012/com_2012_11_en.pdf.

[234] For the IMIA, see: http://www.imia-medinfo.org/new2/node/10. For the Global Alliance, see: http://www.ebi.ac.uk/sites/ebi.ac.uk/files/shared/images/News/Global_Alliance_White_Paper_3_June_2013.pdf. The 'Global Alliance' proposal, which embodies a form of dynamic consent, seems, on its face, to be compliant with the Albrecht amendments to the GDPR, for example. See also services like Mydex and ID3 which can be seen as part of a general movement to put individuals in control of their own data (and, possibly, anticipate opportunities to capitalise it, too).

[235] On personalisation, 'responsibilisation' and 'consumerisation', see paragraph 2.6 and, more generally, Nuffield Council on Bioethics (2010) *Medical profiling and online medicine: the ethics of 'personalised healthcare' in a consumer age,* available at: http://www.nuffieldbioethics.org/personalised-healthcare-0.

[236] For an evaluation of HealthSpace in the context of the NPfIT, see: Greenhalgh T, Stramer K, Bratan T, *et al*. (2010) *The devil's in the detail*, available at http://www.ucl.ac.uk/news/scriefullreport.pdf.

[237] See Vayena E and Tasioulas J (2013) Adapting standards: ethical oversight of participant-led health research, *PLoS Medicine* **10(3)**: e1001402, available at: http://www.plosmedicine.org/article/info%3Adoi%2F10.1371%2Fjournal.pmed.1001402.

[238] See: http://sagecongress.org/WP/wp-content/uploads/2012/04/PortableLegalConsentOverview.pdf; http://del-fi.org/consent. For a discussion see Vayena E, Mastroianni AC, and Kahn JP (2013) Caught in the web: informed consent for online health research *Science Translational Medicine* **5(173)**:173fs6, available at: http://stm.sciencemag.org/content/5/173/173fs6.full.

4.37 The orthodox view of consent is that it is valid if, and only if, it is voluntarily and freely given. This means not just that it is un-coerced but that it is deliberate.[239] The process of informing people giving consent (i.e. providing the information necessary to ensure that the decision to consent is genuinely informed) can be expensive, and people often misunderstand to what they are consenting.[240] (These difficulties apply equally to 'privacy notices' that have developed as an important way of ensuring that data processing is 'fair and lawful'.)[241] Where the further information that may be generated by additional uses of data is unpredictable or indefinite, 'fully informed' consent is difficult to solicit meaningfully. This is of particular concern with research that involves searching databases for potentially significant correlations rather than to confirm or falsify a specific hypothesis since it may turn up findings which have unanticipated implications.[242] A meaningful 'consent' process in these circumstances (as used occasionally, for example, with biobanks) involves the data subject making a decision that they are willing, in effect, to give undefined researchers unconditional and irrevocable permission to use the data they provide in perpetuity, in ways to be determined by others.

4.38 There is a further complication in the case of data that, while freely obtained from one individual, may also relate significantly to the privacy interests of other individuals, including those not yet born. (Genomic data offer a good example but by no means the only one.) Thus, a DNA sequence may reveal probabilistic information or, in rare cases, disease traits or other characteristics in biological relatives, not merely about the person from whom it was obtained. In such cases, there is a divergence between the scope of autonomy (who gives or withholds permission for data access) and the scope of privacy (those whose privacy interests are affected by this permission) that current data protection mechanisms find difficult to manage.[243] This presents a very difficult challenge for existing security and privacy techniques that tend to rely on the exclusive relation of data to an individual subject in order to enforce protection.

4.39 Against this background, there are still debates in health research about obtaining 'consent for consent' (consent to be approached to take part in research). This arises because only those who have legitimate access to data in the first place may be able to identify candidate subjects for research or be permitted to seek a subject's consent to be approached for possible enrolment in research, or for their data to be disclosed to researchers. For example, a medical researcher may wish to enrol patients in an aggregated dataset in a clinical trial but cannot approach the patient directly. Researchers must then rely on, and possibly pay, those with legitimate access to the data (GPs, for example, in the case of primary care records) to contact the patient to

[239] Various constructions ('informed', 'fully informed', 'freely given', 'express', etc.) appear in different instruments. The DP Directive states: "'the data subject's consent' shall mean any freely given specific and informed indication of his wishes by which the data subject signifies his agreement to personal data relating to him being processed." (Art.2(h), available at: http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=CELEX:31995L0046:en:HTML).

[240] See, for example, Pentz RD, White M, Harvey RD, *et al*. (2012) Therapeutic misconception, misestimation, and optimism in participants enrolled in phase 1 trials. *Cancer* **118(18)**: 4571-8, available at: http://onlinelibrary.wiley.com/doi/10.1002/cncr.27397/pdf. Seeking consent can even be counterproductive to research if the culture of consent-seeking generates an unwarranted suspicion about reasons why the person seeking consent is apparently keen to shift the burden of liability, although this may misunderstand the function of consent.

[241] This was a part of the concerns expressed by GP bodies (representing data controllers) in relation to the transfer of patient data to the HSCIC as part of the care.data programme, i.e. that reasonable efforts had not been made to inform patients of the purposes for which their data would be used. See: http://www.pulsetoday.co.uk/nhs-england-bows-to-confidentiality-concerns-and-launches-2m-national-publicity-campaign-on-caredata/20004748.article#.VKe57iusWSo.

[242] Wolf, SM, Lawrenz FP, Nelson CA, *et al*. (2008) Managing incidental findings in human subjects research: analysis and recommendations *Journal of Law, Medicine and Ethics* **36(2)**: 219-48, available at: http://onlinelibrary.wiley.com/doi/10.1111/j.1748-720X.2008.00266.x/abstract.

[243] See Gertz R (2004) Is it 'me' or 'we'? Genetic relations and the meaning of 'personal data' under the Data Protection Directive *European Journal of Health Law* **11(3)**: 231-44.

ask if they are willing to participate in research. This increases the cost and difficulty of research and GPs may be too busy to act as intermediaries.[244] Furthermore the processing of the data for this purpose (to identify candidates for research) must itself have a legitimate ground and take place in accordance with applicable data protection law.

4.40 While there are both practical and conceptual difficulties with obtaining consent, there are, equally, difficulties with withdrawing it. If consent cannot abolish the underlying rights, legal controls or norms that it waives, it should, therefore, be capable of withdrawal or of modification by further conditions. (This does not mean, of course, that once data has been used it can be 'un-used'. For example, if the health data of a consenting research subject contribute to results published in a research paper, withdrawal of consent does not mean that they should be able to get an order for Google to delist any scientific papers based on their data. But it should mean that they should be able to prevent the researchers making further use of that data.)[245] It can be both difficult and costly to extract a subject's data from a dataset, especially if the data have been aggregated and distributed.[246] Moreover, withdrawal may undermine the purpose for which the data are being used if that purpose depends on having an appropriate sample. Further practical difficulties with exercising a right of withdrawal, might arise if the data subject is unaware of the ways in which data relating to them have been propagated.[247]

The limited role of consent

> **Proposition 17**
>
> It is a continuing moral duty of data custodians and users to promote and protect the legitimate rights and interests of those who have provided data about themselves irrespective of the terms of any consent given.

4.41 The existence of consent for the use of data does not, in itself, reduce the risk of harm to data subjects. While, it apparently foregrounds the authority of the data subject, it does so by redistributing the burden of responsibility for outcomes from sole reliance on the data user's probity under moral and legal responsibilities to a rule-governed model in which the data subject may set some of the rules (or, at least, make a limited choice among those on offer). If used cynically, however, it may be simply an attempt to shift the moral responsibility for using data fairly from the data user to the data subject. Where there is a pre-existing expectation of privacy, seeking consent may be a requirement to show respect for persons; however, consent alone is not sufficient (or

[244] Recognising this as a difficulty, in 2013 the Health Research Authority invited researchers to submit models of good practice for identifying research participants. See http://www.hra.nhs.uk/news/2013/12/11/call-good-practice-models-identifying-potential-participants-research-studies/).

[245] We should also note that there may be cases in which someone refuses consent to information disclosure that is nevertheless legitimate (i.e. where they have unreasonable expectations of privacy because no underlying privacy right or norm exists that would prevent the disclosure).

[246] "Once an individual's data has been used in aggregate in analysis, it is effectively impossible to remove that information inherently from aggregate analysis." Anonymous Consultation response, available at: www.nuffieldbioethics.org/project/biological-health-data/evidence-gathering/. See also note 407 at paragraph 7.4 with regard to UK Biobank.

[247] And, in particular, if they are the data subject of data provided by someone else (as may arguably be the case with some genetic data).

necessary) to protect privacy. For this, some additional governance mechanism is required, which may ensure that consent is complied with and provide redress where it is not but should, in any case, provide overarching protections.

4.42 The limitations of anonymisation or consent mechanisms for secondary uses of data, particularly where data are to be disclosed to third parties, linked with other datasets and/or used for indefinite further purposes, has necessitated the search for more satisfactory legal bases for data processing and for suitable measures and processes to give effect to them.[248] Researchers and other secondary users may therefore increasingly seek to use statutory exemptions such as those provided by section 261 of the HSCA 2012, particularly if the secondary uses might be objected to on compatibility grounds, if consent might be refused, and anonymisation is no longer trusted.[249] In this light, the ethical appropriateness of such an approach requires all the more urgent consideration.

Governance and security

4.43 Because of the risk of misuse and consequential privacy infringements, de-identification and consent measures may be supplemented by further governance arrangements. These usually take the form of some additional control to limit data access to authorised users. These arrangements usually have related managerial (e.g. data access committees) and technical (e.g. safe havens) aspects. We consider some specific examples of governance practices in chapters 6 and 7.

Authorisation of data access

4.44 A number of bodies provide governance of information access at different levels and in different ways. Within the terms of applicable law, data access or disclosure (supplying extracts from databases) may be subject to approval by functional elements within the information governance infrastructure of institutions, with or without independent advice or oversight. Research Ethics Committees (RECs) and Data Access Committees may provide scrutiny of specific applications or for specific data collections, although RECs do not necessarily monitor compliance with the terms of any agreement following their opinion or decision. Institutional oversight committees (such as the UK Biobank's Ethics and Governance Council – see chapter 7) do provide continuing scrutiny but have limited powers. The Data Access Advisory Committee of the Health and Social Care Information Centre and the Health Research Authority's Confidentiality Advisory Group (formerly the Ethics and Confidentiality Committee (ECC) of the National Information Governance Board) provide advice in relation to specific cases but do not hold formal authority to approve access. While ethics committees often have one or more participant representative, additional advice may also be sought from separate panels of representatives of participant and patient communities, particularly with long-term research programmes such as biobanks.

---

[248] As US computer scientist Arvind Narayanan points out: "Data privacy is a hard problem. Data custodians face a choice between roughly three alternatives: sticking with the old habit of de-identification and hoping for the best; turning to emerging technologies like differential privacy that involve some trade-offs in utility and convenience; and using legal agreements to limit the flow and use of sensitive data. These solutions aren't fully satisfactory, either individually or in combination, nor is any one approach the best in all circumstances." (https://freedom-to-tinker.com/blog/randomwalker/no-silver-bullet-de-identification-still-doesnt-work/). But see also Sethi N and Laurie G (2013) Delivering proportionate governance in the era of eHealth: making linkage and privacy work together *Medical Law International* **13(2-3)**: 168-204, available at: http://mli.sagepub.com/content/13/2-3/168.

[249] Section 261 of the Health and Social Care Act 2012 (available at: http://www.legislation.gov.uk/ukpga/2012/7) provides, inter alia, for the HSCIC to disseminate (but not publish) identifying information if it considers doing so to be in the public interest.

4.45 Authorising bodies have regard to higher level strategic advice and professional guidance that is provided by a variety of bodies such as the Expert Advisory Group on Data Access (EAGDA), research funders (such as the MRC), various professional organisations (e.g. the BMA), regulatory bodies (e.g. the GMC) and Royal Colleges. General guidance, adjudication of complaints and enforcement is provided in the UK by the Information Commissioner's Office (ICO) and by the case law established by the First-tier Tribunal (Information Rights) of the General Regulatory Chamber and the courts.

4.46 Institutional bodies, although they often fulfil a quasi-judicial function are, nevertheless, open to the criticism that they are not always independent. It is a possible criticism of bodies like research ethics committees and data access committees that they diffuse responsibility for upholding the rights of patients and research subjects, making it very much less likely that researchers who abuse data will be sued or prosecuted. In particular, the fact that a project has received ethics approval makes it highly unlikely that researchers would be found to have criminal intent (*mens rea* – an essential element for successful prosecutions in criminal offences), and standardising practices frustrates civil actions that might be judged 'by the standards of the industry'.[250] This criticism draws attention to the possible consequences of relying on institutions and orthodoxies rather than critically examining underlying moral norms of access and disclosure; we will return to this and to the potential need for broader forms of accountability in the next chapter.

Limiting data access

4.47 There are additional technical mechanisms that provide greater security, such as that data linkage may be performed within a regulated safe haven. The original safe haven was the hospital library where records were kept and where researchers went to interrogate them. The records remained in the library, and the researcher emerged with only some notes of the aggregated results. They were extended to health authorities, which had facilities for the safe storage of paper records that could be reviewed for administrative purposes.

4.48 The use of safe havens was advocated by the second Caldicott review, whereby a secure centre provides a pseudonymisation and linkage service.[251] A variant is the trusted third party (TTP) which was envisaged to provide economies of scale and have no incentive to interfere with the data. The now defunct National Programme for IT (NPfIT) had specified a 'Pseudonymisation Service', until the contractors realised how difficult it would be to anonymise data effectively and that this would be operationally more complex than had been envisaged. The formalisation of a system of accredited safe havens for health data in England is, at the time of writing, in train under proposed regulations.[252] Systems that enable third party linking and access to linked data via safe havens or the equivalent have been developed in Scotland (SHIP) and Wales

[251] The Caldicott Committee (2013) *Information: to share or not to share? The information governance review*, available at: https://www.gov.uk/government/publications/the-information-governance-review.

[252] Department of Health (2014) Protecting health and care information (consultation) (HMSO), available at: https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/323967/Consultation_document.pdf.

(SAIL), and for specific initiatives (e.g. GeL). (We discuss some examples in chapter 6.)[253]

Limitations to data use

4.49 Formal agreements (Data Sharing Agreements, Data Re-use Agreements and Material Transfer Agreements, depending on the nature of the procedure) may be used to set out the terms on which data access or disclosure is to take place. These may restrict the use of the data to an approved class of users, for approved purposes and forbid further disclosure or attempts to re-identify individuals in the dataset. Penalties for breach of an agreement (that do not otherwise constitute criminal offences) remain comparatively lenient, however. They may include, in theory, refusal to provide further data access in future. In health systems such as the NHS, data-sharing agreements or service contracts allow the NHS to commission services from third-party suppliers or provide information externally.

4.50 Agreements may not be effective or well managed, however: failures in the management of data sharing agreements were identified in the 2014 *Review of data releases made by the NHS Information Centre* ('Partridge review').[254] The HSCIC has talked about a 'one strike and out' principle but this does not appear to have been adopted fully and an external enforcement mechanism is lacking. Furthermore, there are no practical mechanisms available for other stakeholders, such as patients, to take enforcement action independently of the data controllers.

4.51 The enforcement of data sharing agreements and contracts relies on the possibility of detection, and on effective sanctions. These depend both for credibility and efficacy on the existence of systems of audit, inspection, regulation and enforcement that can detect and remedy the mischief.[255] Where fundamental rights are at stake, and they cannot be protected by private action, there is a reasonable argument that at least the most egregious breaches should be brought within the scope of the criminal law. This is why, in chapter 2, we made a series of recommendations in relation to the identification of possible harms, mapping of information flows, reporting of breaches and the creation of an offence of deliberate misuse of data. For this latter, we found much support among those we consulted in the preparation of this report.[256]

Conclusion

4.52 From the point of view of knowledge discovery (whether in health care or biomedical research), for which the widest access to the richest data is implicitly desirable, those designing data initiatives find themselves in something like a double bind, a demand that they do two mutually contradictory things at the same time:[257]

---

[253] For SHIP, see chapter 6, below. For a description of the SAIL process, see: http://www.saildatabank.com/faq.aspx#.

[254] PwC LLP (2014) Data release review (HSCIC), available at: http://www.hscic.gov.uk/datareview. The review found that the NHS Information Centre (the forerunner of the present Health and Social Care Information Centre) was unable to locate agreements relating to releases of individual-level data so it was not possible to determine to whom the data had been released, and there was no evidence that a company contracted to the Information Centre to manage releases had obtained appropriate clearance.

[255] Penalty schemes may be applied at the level of re-identification (before any discriminatory treatment has been visited on individuals), or of misuse of data (preparatory to discriminatory treatment), although as we have observed (see chapter 2) these are not easily detectable.

[256] See: www.nuffieldbioethics.org/project/biological-health-data/evidence-gathering/.

[257] On the 'double bind' see Bateson G (1972) *Steps to an ecology of mind: collected essays in anthropology, psychiatry, evolution, and epistemology* (San Francisco: Chandler). This is presented as a double bind rather than a simple tension

- researchers and administrators are encouraged to generate, use and extend access to data (because doing so is expected to advance research and make public services more efficient); however,
- there is a similarly strong imperative, and a requirement of human rights law, to protect privacy (and the more access to data is extended, the greater are the risks of abuse).

4.53 In this chapter we have discussed the difficulties that may arise for data initiatives in effectively anonymising individual-level data (and even aggregate data) and in determining whether consent is effective and valid for a proposed use of data. We have also discussed the need for governance to have broader accountability. We draw three main conclusions from our discussion. First, 'anonymisation' is unlikely on its own to be sufficient to protect privacy as it is simply too hard to prevent re-identification. Second, the consent of data subjects is not always morally necessary (the use of personal data may not affect privacy interests) and is never sufficient to secure their moral interests (consent to use of data does not make harm arising from that use impossible, nor does it offer any direct say in what options are available). Third, while governance provides an essential, enabling condition for data initiatives, the form it should take and the way in which it should be deployed cannot be determined without reference to the norms and interests at stake in a particular data initiative, and without wider forms of accountability. In the next chapter we move from these largely negative conclusions to a more positive account of how a set of morally reasonable expectations may be defined and met in the context of data initiatives.

because there is an imperative to do both things simultaneously (i.e. share more information *and* make information more secure), not merely to find a balance between them (e.g. share less information so that what is shared is more secure).