Chapter 2

Data opportunities and threats

Chapter 2 – Data opportunities and threats

Chapter overview

Given the UK's strong research base in the biomedical sciences and the unique resource and infrastructure of the UK's National Health Services, the use of health data has become a strategic focus.

There is a clear public interest in the responsible use of data to improve wellbeing through improved health advice, treatment and care, as well as through increasing economic prosperity more generally. These objectives are being pursued in three main ways:

- increasing efficiency and transforming service delivery
- generating improvements in medical treatment
- generating economic growth from the life sciences

Policy orientations to achieve these outcomes include:

- increasing IT intensity and introducing new infrastructure in health systems
- establishing partnerships between the public and private sectors
- centralising data resources
- promoting 'open data' and 'data sharing'
- investing in 'big data'

However, there are a number of risks and fears, including:

- misuse of data leading to harms to individuals and institutions
- discriminatory treatment of individuals and groups
- fear of state surveillance of citizens

The negative impacts of data misuse are potentially much wider than are those recognised by legal and regulatory systems. Furthermore, the nature of privacy harms and of the judicial and regulatory systems means that they are likely to be under-reported by the victims.

A number of recommendations are made relating to understanding data use, research into data misuse, preventing fraudulent access to data, reporting abuses of data and penalties for deliberate misuse of data.

Introduction

2.1 In the previous chapter we discussed developments in data production and data analysis. These developments have a range of possible consequences, many of which are significant but very few of which are inevitable. The scientific, technological and clinical factors that shape them comprise an interacting and evolving system along with policy, economic and social conditions. This chapter examines the economic and policy drivers of further developments in the use of data in biomedical research and health care. Considerable enthusiasm has been generated by the potential for using information to produce transformative efficiencies in services, generate new knowledge and promote innovation. This has led to substantial public investment and an enabling policy environment in the UK and elsewhere. 'Data sharing', 'big data', 'open data' and 'data revolution' have become familiar buzzwords in public and policy discourse. Here we describe the main dimensions of opportunity opened in biomedical research and health care by data science and technology, consider the policy orientations of the UK

Government and others to realise them, and identify some of the costs and risks that these might entail.

Opportunities for linking and re-use of data

> **Proposition 6**
>
> The continuing accumulation of data (see Proposition 1) and the increasing power and availability of analytical tools (see Proposition 2) mean that new opportunities arise, and will continue to arise, to extract value from data. There is a public interest in the responsible use of data to support the development of knowledge and innovation through scientific research and to improve the well-being of all through improved health advice, treatment and care.

The 'value proposition'

2.2 The global financial crisis of 2007-8 and the subsequent economic downturn, focussed the attention of governments on the extraction of value from existing assets, the search for greater efficiency and the promotion of economic growth building on areas of existing strength. In the UK this focus has fallen on, among other things, the exploitation of public sector data (PSD), IT innovation, and the strong research base in the life sciences.[61]

2.3 The promotion of national economic growth on the back of public sector data was a theme of the *Shakespeare Review* (2013), which envisaged that Britain could 'be the winner' of 'phase 2' of the digital revolution. America, it said, had won the first phase which was about connectivity and access to information and efficiency gains; the second phase would be about extracting value from the data. A 2011 report from the McKinsey Global Institute provides some idea of context. It suggests that, in 10 years, given the right strategic innovations, data use in the US health sector could generate $300 billion of value per year (two thirds in efficiency savings) and that "medical clinical information providers, which aggregate data and perform the analyses necessary to improve health care efficiency, could compete in a market worth more than $10 billion by 2020."[62] Shakespeare argued that the vast advantage enjoyed by the USA in terms of its domestic market size, the west-coast entrepreneurial culture, and the existence of firms like Google, Apple, Microsoft, Amazon and eBay, could be offset by making public-sector data available to innovative firms in the UK.

> "We should remain firm in the principle that publicly-funded data belongs to the public; recognise that we cannot always predict where the greatest value lies but know there are huge opportunities across the whole spectrum of PSI [public sector information]; appreciate that value is in discovery (understanding what works), better

[61] For a discussion of the link between research investment in the biosciences and national economic growth see Nuffield Council on Bioethics (2012) Emerging Biotechnologies: technology, choice and the public good, available at: http://www.nuffieldbioethics.org/emerging-biotechnologies, especially chapter 7 ('Research and Innovation Policy').

[62] McKinsey Global Institute (2011) *Big data: The next frontier for innovation, competition, and productivity*, available at http://www.mckinsey.com/insights/business_technology/big_data_the_next_frontier_for_innovation, at page 6. MGI studied the US healthcare sector – along with 4 others – and concluded that there are opportunities to generate $300 billion/ year through big data (as distinct from simple automation).

management (tracking effectiveness of public administration), and commercialisation (making data practically useful to citizens and clients); create faster and more predictable routes to access; and be bold in making it happen."[63]

2.4 While the large administrative datasets such as those held by Companies House, the Land Registry, the Met Office and the Ordnance Survey offer an abundance of 'ripe, low hanging fruit', public sector health data have, for a long time, been seen as a prized asset with exploitable potential, albeit (in the UK) one that has been hampered by a lag in introduction of IT systems compared to other industries.[64] The value proposition of data initiatives in biomedical research and health care has essentially three dimensions: generating significant service efficiencies through the better use of business intelligence, generating improvements in the practice of medicine, and generating value through science and innovation. These three dimensions are interrelated: all data initiatives in biomedical research and health care can be located within the volume that they describe.

Efficiency and transformation of service delivery

2.5 Pressure to make wider use of individual health information has come from the evolving professional and institutional organisation of health systems (which includes more complex treatment pathways on one hand, and attempts at IT-driven administrative simplification and cost control on the other) as well as the long-recognised opportunities for research. In the UK, resource constraints facing the NHS, in the context of more general austerity policies and the burdens of an ageing population, have led to the need to find significant efficiency savings which present serious challenges to the NHS.[65] IT innovation and more effective use of data are placed at the heart of the response to these challenges, both in terms of more efficient processes and the use of evidence to improve clinical decisions.[66] The NHS England 'care.data' programme, for example, has been described as necessary in order to secure the future of an affordable NHS in England. (We discuss this argument in chapter 6.)

2.6 The aims of policy initiatives in this area are, however, more ambitious than simply the more efficient and widespread use of electronic records and systems. Information technology and data science is envisaged as disruptive technology that will

---

[63] Stephan Shakespeare (2013) *Shakespeare Review: an independent review of public sector information*, available at: https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/198752/13-744-shakespeare-review-of-public-sector-information.pdf, at page 6. The Shakespeare Review followed the Government Growth Review 2011 (https://www.gov.uk/government/news/autumn-statement-growth) which outlined plans to establish the Open Data Institute (http://theodi.org/) with some Government funding, and the White Paper (Cm 8353) *Open data: unleashing the potential* (2012), available at: https://www.gov.uk/government/publications/open-data-white-paper-unleashing-the-potential.

[64] In his 2002 Report, *Securing our future health: taking a long-term view* (http://si.easp.es/derechosciudadania/wp-content/uploads/2009/10/4.Informe-Wanless.pdf), Derek Wanless highlighted the poor state of ICT use in the UK health service and that significant investment was required to improve informational infrastructure. His follow-up report for the King's Fund, *Our future health secured? A review of NHS funding and performance* (2007) (http://www.kingsfund.org.uk/sites/files/kf/field/field_publication_file/our-future-health-secured-review-nhs-funding-performance-full-version-sir-derek-wanless-john-appleby-tony-harrison-darshan-patel-11-september-2007.pdf), concluded that, although there had been some positive developments, further improvement in ICT systems was still needed. See also chapter 6.

[65] The 'Nicholson Challenge' was first set out by Sir David Nicholson, (then) chief Executive of the NHS, in the NHS Chief Executive's Annual Report for 2008-09 (NHS, May 2009) and refers, in effect, to the need for the NHS to achieve efficiency savings of £15-20 billion by 2014/15. The QIPP (Quality, Innovation, Productivity and Prevention) policy agenda set out a programme of actions designed to meet this challenge (see https://www.evidence.nhs.uk/qipp).

[66] For example, a 2014 report from a big data solutions company claims that better use of data analytics could free between £16.5 billion and £66 billion worth of NHS capacity. Bosanquet N and Evans E (2014) *Sustaining universal healthcare in the UK: making better use of information* (http://volterra.co.uk/wp-content/uploads/2014/09/Final-EMC-Volterra-Healthcare-report-web-version.pdf). The biggest single projected saving (£5billion) relates to time saved searching for missing records.

revolutionise the way in which health care is delivered.[67] They are expected to enable a shift towards prediction, prevention, personalisation and 'responsibilisation' in health care, to be facilitated by a range of e-Health initiatives.[68] According to the European Commission, 'e-Health' is a portmanteau policy area that includes:

> "information and data sharing between patients and health service providers, hospitals, health professionals and health information networks; electronic health records; telemedicine services; portable patient-monitoring devices, operating room scheduling software, robotized surgery and blue-sky research on the virtual physiological human."[69]

2.7 The e-Health vision is built on the foundation of electronic care records, which are fed with data from a variety of sources, including patients, who are expected to access them easily and routinely. According to the UK Department of Health these will "progressively become the source for core information used to improve our care, improve services and to inform research."[70] The European Union's 2012 eHealth Task Force Report, *Redesigning health in Europe for 2020*, sets out five 'levers for change': patients taking control of their data, liberating data for business intelligence and research, integrating systems to add value and drive out error, 'revolutionising' health by making it responsive to patient needs and ensuring that no one is excluded.[71] However, alongside these positive ambitions there are also warnings that if governments do not act to secure the public interest they may cede control of the overall direction of innovation to giant commercial Internet companies.[72] (This is an aspect that we attend to through the approach we develop in the remainder of this report.)

Generating improvements in medical treatment

2.8 Ethical imperatives relating to data in health care and biological research have traditionally pulled in opposite directions. In health care, the primary reason for patients to share personal information with their doctors was to optimise the care they received. The privileged relationship of confidentiality between the patient and doctor has meant, at least since the time of Hippocrates, that only the strongest reasons could justify broader disclosure.[73] This principle has been generally accepted down the ages in Western medicine, notwithstanding the growth in the number and variety of those

---

[67] See, for example, the 3 key imminent shifts in medical practice identified by Simon Stevens (NHS England CEO): "a coming revolution in biomedicine, in data for quality and proactive care, and in the role that patients play in controlling their own health and care" (speech to the annual conference of the NHS Confederation, 4 June 2014). See http://www.england.nhs.uk/2014/06/04/simon-stevens-speech-confed/.

[68] On 'responsibilisation' see the 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.

[69] See http://ec.europa.eu/health/ehealth/policy/index_en.htm.

[70] Department of Health (2012) *The power of information: putting all of us in control of the health and care information we need*, available at: https://www.gov.uk/government/publications/giving-people-control-of-the-health-and-care-information-they-need, at page 5. See also: National Information Board, Department of Health (England) (2014) *Personalised health and care 2020: using data and technology to transform outcomes for patients and citizens. A framework for action*, available at: https://www.gov.uk/government/publications/personalised-health-and-care-2020.

[71] See http://www.president.ee/images/stories/pdf/ehtf-report2012.pdf.

[72] The EU eHealth Task Force Report (2012) Redesigning health in Europe for 2020 presents e-Health opportunities in the face of the threat that giant internet corporations might replace governments as the rule setters.

[73] The Hippocratic Oath, as usually understood, contains the following precept: "What I may see or hear in the course of the treatment or even outside the treatment in regard to the life of men, which on no account one must noise abroad, I will keep to myself holding such things shameful to be spoken about." For a further discussion of confidentiality, see chapters 3 and 4.

involved in the provision of health care (including, for example, various clinical specialisms, administrators, medical secretaries and auditors).

2.9 For a long time, decisions about the treatment of patients relied on the training and experience of individual doctors, learning informally from the experience of others, case reports in specialist publications, and advice from Royal Colleges or health care delivery organisations. This would be brought to bear on how the patient presented in the clinic, the patient's phenotype, and their recorded or recounted medical history. Two sets of developments, involving 'wider' and 'deeper' data, have transformed the practice of medicine in the last half century.

2.10 The first development came about as a result of combining data to compare the effectiveness of different interventions on significant numbers of patients in relevantly similar circumstances. In the second half of the 20th Century, the randomised, controlled trial (RCT) became the 'gold standard' approach to determining the effectiveness of medical interventions. When they are well constructed, RCTs allow the effect of a therapeutic intervention to be isolated from circumstantial factors that might affect the outcome.

2.11 The publication of data from trials made possible the further step of meta-analysis or systematic review, which can increase statistical power and confidence in the findings.[74] This enabled the clinical judgement of individual doctors to be supported by an 'evidence base' of carefully collected and interpreted data.[75] However, considerable skill is required in interpreting and applying evidence to clinical situations: evidence from trials concerns the *efficacy* of a treatment in optimised conditions but not its *effectiveness* for a particular patient in real world circumstances. The availability of data from clinical trials does not annul the value of observational studies and 'real world' data. Furthermore, the effectiveness of treatment will depend not only on the interaction between the intervention and the disease, but also on the patient, whose values and preferences are not only important to the 'success' of the treatment but also contribute to what 'success' means.[76]

2.12 A second development in the field of biomedicine has focussed on overcoming the limitations of evidence-based medicine (EBM) by considering more data in order to understand variations in patient response. This is achieved by stratifying the ideal patient population based on additional dimensions of information. The focus of stratified or personalised medicine was initially on integrating genomic data, and this remains an important pillar, although increasingly as part of more complex 'knowledge networks'.[77]

---

[74] Most meta-analyses deal with efficacy (a positive difference attributed to the intervention in a carefully controlled trial situation) and few with serious, uncommon or rare adverse events, which the underlying RCTs are seldom sufficiently powered to detect.

[75] For systematic review, see: http://www.cochrane.org/. Nevertheless, meta-analyses of RCTs, which are designed to increase the reliability of inferences drawn from the study data, may offer only relatively weak support for those inferences compared to an adequately powered (and randomised) trial. See: Turner RM, Bird SM, and Higgins JPT (2013) The impact of study size on meta-analyses: examination of underpowered studies in Cochrane reviews *PLoS ONE* **8(3)**: e59202, available at: http://www.plosone.org/article/infoper cent3Adoiper cent2F10.1371per cent2Fjournal.pone.0059202.

[76] This is recognised in evidence-based practice (EBP), which integrates three components: the best relevant research evidence, the professional expertise of the clinician and the values and preferences of the patient. It begins by framing the clinical question to be addressed from the care needs and preferences of the patient.

[77] Compare Department of Health (2003) *Our inheritance, our future: realising the potential of genetics in the NHS*, NHS Genetics White Paper (Cm5791-II), available at: http://webarchive.nationalarchives.gov.uk/20130107105354/http:/www.dh.gov.uk/prod_consum_dh/groups/dh_digitalassets/@dh/@en/documents/digitalasset/dh_4019239.pdf with (US) Committee on a framework for developing a new taxonomy of disease; National Research Council (2011) *Toward precision medicine: building a knowledge network for biomedical research and a new taxonomy of disease* (Washington, DC: National Academies Press), available at: https://www.ucsf.edu/sites/default/files/legacy_files/documents/new-taxonomy.pdf.

The contemporary vision is for medicine based on large data resources as well as better standardisation of data and linkage between phenotype and genotype data with additional lifestyle, environmental and social data.[78] Putting these together can provide a detailed picture of the nature of the disease affecting a patient, how and why it is manifesting itself in that individual as it is, and that patient's likely response to any treatments. The same data can help to identify risk factors for disease or those individual characteristics, practices or treatments associated with the best health outcomes. This is a different approach from orthodox EBM, which seeks to avoid questions about phenotypic and environmental variability using statistical control techniques. While EBM simplifies, the big data approach is to embrace complexity.

Generating economic growth through the life sciences

2.13 The network of databases within the National Health Services in the UK provides a source of abundant longitudinal phenotypic and pathology data that could give rise to significant new insights. A key axis of research policy in the UK and, indeed, of research policy's contribution to national industrial policy, has been the combination of NHS infrastructure and genome science. This has been a consistent theme in both health and science policy since the Human Genome Project, building on the possibilities of population genetics and personalised medicine.[79]

Box 2.1: Data intensive bioscience: the Human Genome Project

The Human Genome Project offers an example of ambitious bioscience as 'big science': large-scale projects, usually involving international consortia and with multiple research sites often distributed internationally, usually funded (or part funded) directly or indirectly on a vast scale by national governments in the public interest.

Extracting value from knowledge of the human genome has, however, turned out to be more difficult than most (certainly most policy makers) expected. It has both required and accelerated the development of computational biology and the demand for biological data – deeper genotyping and phenotyping, and the inclusion of clinical data. The establishment of data sharing resources with strong links between health services, academic institutions and industrial partners is seen as a key element of research programmes of this sort.[80] This is echoed in almost every area of the biosciences.

2.14 The potential of research capability in the NHS was emphasised in 2003 by a report for the Department of Trade and Industry, *Bioscience 2015 – Improving National Health, Increasing National Wealth* and in the Genetics White Paper *Our inheritance, our future: realising the potential of genetics in the NHS*, which stressed the value of an "Integrated Care Records Service (ICRS) – the standard patient record, one per

[78] Kohane IS (2014) Deeper, longer phenotyping to accelerate the discovery of the genetic architectures of diseases *Genome Biology* **15**:115, available at: http://genomebiology.com/2014/15/5/115.

[79] See: Department of Trade and Industry (1999) *Genome Valley: the economic potential and strategic importance of biotechnology in the UK report*, available at: http://webarchive.nationalarchives.gov.uk/+/http://www.dti.gov.uk/genomevalley/report.htm; see also: Fears R and Poste G (1999) Building population genetics resources using the U.K. NHS *Science* **284(5412)**: 267-268.

[80] Academy of Medical Sciences (2013) *Realising the potential of stratified medicine*; available at: http://www.acmedsci.ac.uk/more/news/realising-the-potential-of-stratified-medicine/.

patient, which will hold all health and social care data".[81] The emphasis on research in health policy documents continued in the 2012 *The Power of Information* report and the subsequent White Paper *Equity and Excellence: Liberating the NHS* and, to an extent, through the NHS Constitution.[82] Indeed, the choice to articulate an NHS constitution at all and the 'social contract' mode in which it is articulated (through pledges, rights and responsibilities) signals a more reciprocal view about the contribution of patients both to their own care and to the wider public interest in national health and wealth. Rather than simply paying taxes and receiving health care when they need it, patients now implicitly become morally enjoined contributors to a public data resource. The exploitation of the NHS as both a data source and research infrastructure was at the centre of the 2010 *Strategy for UK Life Sciences* which argued for an amendment to the NHS constitution to introduce a "default assumption (with ability to opt out): for data collected as part of NHS care to be used for approved research, with appropriate protection for patient confidentiality"; and "that patients are content to be approached about research studies for which they may be eligible."[83] The initiative represented by the *Strategy for UK Life Sciences* was further consolidated in 2014 by the formation of a refreshed and expanded Office for Life Sciences (OLS) jointly by the Department for Business, Innovation and Skills (BIS) and the Department of Health (DH), with the intention of making the UK attractive as a place to invest in life science research and facilitating cooperation between basic research and the NHS.[84] (It is taking concrete shape in the current '100,000 Genomes' project to be delivered by Genomics England Ltd. We discuss this case in more detail in chapter 6.)

2.15 As of 1 April 2013, the Secretary of State for Health has a statutory duty to promote research (and the use of research evidence) in exercising functions in relation to the health service.[85] A similar duty applies at all levels of the NHS. This formal requirement consolidates the orientation of the NHS not simply towards becoming a 'learning' health system (through which data are fed back into commissioning and service development) but a combined care and research system.[86]

---

[81] Bioscience Innovation and Growth Team (BIGT) (2003) *Bioscience 2015 – improving national health, increasing national wealth*, available at: http://www.bioindustry.org/document-library/bioscience-2015/1bia-1103-bioscience-2015.pdf; Department of Health (2003) Genetics White Paper (Cm 5791 – II) *Our inheritance, our future: realising the potential of genetics in the NHS*, available at: http://webarchive.nationalarchives.gov.uk/20130107105354/http://www.dh.gov.uk/prod_consum_dh/groups/dh_digitalassets/@dh/@en/documents/digitalasset/dh_4019239.pdf, at page 53.

[82] *Liberating the NHS*, see https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/213823/dh_117794.pdf; *The power of information*, see: https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/213689/dh_134205.pdf; Department of Health (2013) The NHS Constitution for England, available at: https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/170656/NHS_Constitution.pdf.

[83] Department for Business, Innovation and Skills (2011) *Strategy for UK life sciences*, available at: https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/32457/11-1429-strategy-for-uk-life-sciences.pdf, at page 32. This was underlined in a speech by the British Prime Minister in December 2011, in which he argued for changes to the NHS Constitution to make every NHS patient a "research patient" with their medical details "opened up" to private healthcare firms (see: https://www.gov.uk/government/speeches/pm-speech-on-life-sciences-and-opening-up-the-nhs). See also: NHS England (2011) *Innovation, health and wealth: accelerating adoption and diffusion in the NHS*: "It is a key goal of the NHS for every willing patient to be a research patient, enabling them to access novel treatments earlier. The greater the number of patients involved in research, the wider the public benefit.", available at: http://www.england.nhs.uk/wp-content/uploads/2014/02/adopt-diff.pdf, at page 17.

[84] See: https://www.gov.uk/government/news/a-bigger-better-office-for-life-sciences.

[85] These positive duties were provided in the Health and Social Care Act 2012, s.6. See: http://www.legislation.gov.uk/ukpga/2012/7/contents/enacted.

[86] The National Institute of Health Research (NIHR) provides a key means of giving effect to this obligation. See: http://www.nihr.ac.uk/documents/about-NIHR/Briefing-Documents/1.1-The-National-Institute-for-Health-Research.pdf.

Policy orientations

2.16 There are potentially different ways of realising the opportunities presented by data science and technology in biomedical research and health care. Here we abstract some of the main orientations that characterise contemporary policy in the UK and some other countries.

IT intensity

2.17 Computerisation has transformed areas of life such as banking and administration, as well as research, through advances in computational speed, network communications and digital storage. The tools and practices of bioinformatics have radically transformed the pace of discovery in biomedical research and biology more generally, and the nature of the research enterprise and the professional skills involved.[87] Similar gains in health care have, however, proved more elusive.

2.18 The promise of efficiency savings and collateral benefits from the implementation of information technology has proved enduringly appealing to policy makers faced with essentially intractable problems of conflicts over resources. This appeal has been burnished by the impressive projections of IT companies and consultants.[88] Furthermore it has endured despite evidence of public sector organisations, in both the UK and overseas, having long running difficulties with IT systems, with many projects being late, over budget or failing to deliver the promised functions or savings. (We discuss the implementation of information technology in health care in more detail in chapter 6 below).

2.19 The disappointments of previous experience may be manifestations of a 'productivity paradox', which suggests that simply implementing more efficient technologies (replacing paper files with electronic ones, for example) will not yield significant benefits without a more substantial reconfiguration of the way in which they are used.[89] It is therefore to be expected that attempts to digitise health care will take longer, cost more and save less than those with pressing political deadlines might wish. The strategy proposals from the National Information Board (the body responsible for commissioning informatics services for health and social care in England), set out in *Personalised Health and Care 2020*, however, continue the IT-intensive approach with increasing expectations placed on e-Health initiatives to deliver benefits.[90]

---

87 Schatz, MC (2012) Computational thinking in the era of big biology *Genome Biology* **15**: 177, available at: http://genomebiology.com/2012/13/11/177; Thessen AE and Patterson DJ (2011) Data issues in the life sciences *ZooKeys* **150**: 15-51, available at: http://zookeys.pensoft.net/articles.php?id=3041&display_type=list&element_type=12.

88 See note 66 above.

89 The productivity paradox is pithily summed up in a quip by to the economist, Robert Solow: "You can see the computer age everywhere but in the productivity statistics." (New York Times Book Review (12 July 1987) *We'd better watch out*). See also: David, PA (1990) The dynamo and the computer: an historical perspective on the modern productivity paradox *American Economic Review* **80(2)**: 355-61, available at: http://eml.berkeley.edu/~bhhall/e124/David90_dynamo.pdf; Brynjolfsson E (1993) The productivity paradox of information technology *Communications of the Association for Computing Machinery* **36(12)**: 66-77; Jones SS, Heaton PS, Rudin RS, and Schneider EC (2012) Unraveling the IT productivity paradox – Lessons for Health Care *New England Journal of Medicine* **366:** 2243-5, available at: http://www.nejm.org/doi/full/10.1056/NEJMp1204980.

90 National Information Board, Department of Health (2014) *Personalised health and care 2020: a framework for action*, available at: https://www.gov.uk/government/publications/personalised-health-and-care-2020. For a brief critical response see: Greenhalgh T and Keen J (2014) "Personalising" NHS information technology in England (editorial) *British Medical Journal* **349**: g7341.

Public-private partnerships

2.20 The relationship between industry and universities in the biosciences has been close through most of the 20th Century, with the pharmaceutical industry making use of academic science as the basis for the development of successful medicines. The relationship has been cemented by the crossing of individuals between the academic and commercial sectors and the institutional collaborations that characterised the early phase of the biotechnology industry in the 1990s. While a tension arose between the public and private sectors when human gene sequencing led to an initial rush to secure intellectual property rights through potentially valuable patents, this was largely resolved by adaptations in patent law.[91]

2.21 When the complexity of gene function for complex diseases became evident, the need to link gene sequence information to clinical data to identify the relationship between genetic variation and disease risk (e.g. through genome-wide association studies) led to recognition of the value of research based around large-scale biobanks.[92] Although some commercial biobanks that allowed the linking of phenotypic and pathology (e.g. genetic biomarker) data appeared, the long timescales and uncertainties, which are the norm in biotechnology, underscored the importance of contributions from the public and charitable sectors (e.g. the Wellcome Trust, Cancer Research UK), who could invest for the 'long haul'.[93]

2.22 Medical research charities provide an important function in the funding ecosystem, and both the size and orientation of their influence is significant. It is inevitably easier to raise money for some conditions (such as cancers and heart disease) than others, which has caused difficulties for rare disease research. The biggest charities, like the Wellcome Trust, which dispenses more money than the Medical Research Council in the UK, can have a significant effect on the direction of research. The Wellcome Trust has consistently, and perhaps critically, advanced genome research and helped to build the UK's infrastructure and expertise in this area, as well as promoting data sharing and open data.[94] As they are not politically accountable for their strategic decisions, charities are not as vulnerable to external political or economic pressures as governments and firms, although they may be susceptible to different forms of public and sectional interest. UK Biobank, which depends substantially on Wellcome Trust funding, was established explicitly to support research that is in the 'public interest' (see chapter 7).[95]

---

[91] Hopkins M, Mahdi S, Patel P, and Thomas SM (2007) DNA patenting: the end of an era? *Nature Biotechnology* **25**: 185-87, available at: http://www.nature.com/nbt/journal/v25/n2/pdf/nbt0207-185.pdf. See also: Cook-Deegan R and Chandrasekharan S (2014) Patents and genome-wide DNA sequence analysis: is it safe to go into the human genome? *Journal of Law, Medicine and Ethics* **42(s1)**: 42-50, available at: https://aslme.org/media/downloadable/files/links/0/4/04.SUPP_Cook-Deegan.pdf.

[92] Biobanks were established in quick succession in many parts of the world, though particularly in the U.S. and Europe around the time of the completion of the Human Genome Project. See: Vaught J, Kelly A, and Hewitt R (2009) A review of international biobanks and networks *Biopreservation and Biobanking* **7(3)**: 143-50; Hewitt, RE (2011) Biobanking: the foundation of personalized medicine *Current Opinion in Oncology* **23**: 112-9; Wichmann, H-E, Kuhn KA, Waldenberger M, et al. (2011) Comprehensive catalogue of European biobanks *Nature Biotechnology* **29**: 795-7. Scott CT, Caulfield T, Borgelt, E and Illes, J (2012) Personal medicine – the new banking crisis *Nature Biotechnology* **30(2)**: 141-7.

[93] See the small number of gene-based diagnostics and drugs derived from genomic targets currently available. See, for example, Hopkins MM, Martin P, Nightingale P, and Kraft A (2008) Living with dinosaurs: genomics, and the industrial dynamics of the pharmaceutical industry, conference paper, available at: http://www2.druid.dk/conferences/viewpaper.php?id=3847&cf=29.

[94] For example, through the Sanger Institute in Hinxton/Cambridgeshire, which led both the practical work and the political orientation of the UK contribution to the Human Genome Project.

[95] See UK Biobank Coordinating Centre (2011) Access Procedures: application and review procedures for access to the UK Biobank resource, available at: http://www.ukbiobank.ac.uk/wp-content/uploads/2011/11/Access_Procedures_Nov_2011.pdf.

2.23 Although the public sector generates most of the data and has a near monopoly on collecting certain sorts of data, data analysis and innovation will probably continue to be pushed out to the private sector owing to the lack of public sector IT capability and political decisions about the shape and balance of the innovation ecosystem, including the fostering of diverse research approaches.[96] The public and charitable sectors have therefore progressively taken on at least three distinct functions over the course of the last three decades, in support of anticipated delivery of biomedical products by the private sector.

- funding of 'underpinning research' and skilled workforce (academic institutions)
- funding of major data resources (e.g. UK Biobank)
- funding of infrastructure/ capacity (e.g. National Programme for IT)

Centralisation of data

2.24 Whereas standardisation is desirable and technical interoperability essential for linking separately collected and maintained datasets, the drive towards exploitation of public data in the UK has, additionally, involved the consolidation and centralisation of some data resources in so-called 'safe havens' for health and public sector data, such as the Health and Social Care Information Centre (HSCIC).[97]

2.25 Although centralisation is convenient for the extraction of value through the application of data analysis, consolidated databases create large targets for unauthorised technical access, unauthorised access by insiders, or abuse of authorised access at the behest of powerful lobbyists.[98] Centralisation of data is not the only way of achieving the objectives of research. For many years, for example, GP systems had mechanisms for researchers to send queries to practices and receive aggregated answers. The centralised approach to health data taken by the HSCIC in England is conspicuously different from that adopted in Scotland, for example. (These approaches are discussed in more detail in chapter 6.)

Open data

2.26 For several decades it has been recognised that clinical trials and other research studies that do not show a clear difference between medicines (or interventions) are less likely to be published in the medical literature. As a result, systematic reviews, using only published outcomes of medical studies, can reach misleading conclusions. [99] Pharmaceutical companies, in particular, have attracted scrutiny and suspicion as clinical trials results are not always made public in a timely fashion and some –

---

[96] For a discussion of recent UK research policy, see Nuffield Council on Bioethics (2012) Emerging Biotechnologies: technology, choice and the public good, available at: http://www.nuffieldbioethics.org/emerging-biotechnologies, especially chapter 7 ('Research and Innovation Policy').

[97] Department of Health (2014) *Protecting health and care information: a consultation on proposals to introduce new regulations*, available at: https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/323967/Consultation_document.pdf. There are further arrangements for distributed accredited safe havens as recommended by the Caldicott review. See: 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, at s.6.5.

[98] See discussion of 'data threats' at paragraph 2.32.

[99] See, for example, Easterbrook PJ, Gopalan R, Berlin JA, and Matthews DR (1991) Publication bias in clinical research *The Lancet* **337(8746)**: 867-72.This publication bias or 'file-drawer problem' has been observed in many scientific disciplines. See: Scargle JD (2000) Publication bias: the "file-drawer" problem in scientific inference *Journal of Scientific Exploration* **14(1)**: 91-106, available at: http://scientificexploration.org/journal/jse_14_1_scargle.pdf.

especially negative results unfavourable to the company – not published at all.[100] This has led to pressure for all trials to be registered and results published, and has also contributed to a more general argument in science for data to be 'open', that is, made available publicly for independent validation of findings and for secondary research.[101] Indeed, if data derived from national resources such as the NHS or administrative databases are to be used for research there is a strong moral argument that their use should not be restricted to only those who can pay for them (e.g. to publication in academic journals with access restricted by paywalls) or to those with particular kinds of interest.

2.27 Open data has been defined as data that anyone is free to access, use, modify, and share, so long as it is correctly attributed and further use is not constrained.[102] Open data is therefore unlikely to contain individual-level data or data that might be subject to data protection measures.[103] The open data movement nevertheless suggests a strengthening of the ethical 'imperative' for data sharing, albeit without weakening the imperative to protect individuals' privacy.[104] It is also partly a response to concerns about research inefficiency (e.g. unwitting duplication or failing to exploit synergies) and the need for raw data in order to test the reproducibility of research findings, and about turning around poor practice and misconduct (e.g. withholding unfavourable research results).

2.28 Although the open data movement is self-consciously modelled on the 'open source' software movement, its organisation and driving forces are different.[105] While the open source and free software movements developed 'from the bottom up' enjoy broad support and are involved in the maintenance of much of the planet's digital infrastructure, and while many scientists strongly support open access publication, the open data movement has less substantial voluntary support.

2.29 Open data policy is currently being supported enthusiastically by some governments, notably in the UK and the USA, through initiatives to put 'public data' into the public domain.[106] They argue that this is the best way of extracting economic value from the data (in contrast to the model now adopted by commercial data brokers such as Google and Microsoft).[107] In some cases this policy is enshrined in legislation: the Health and Social Care Act 2012 (HSCA) placed an obligation on the HSCIC to publish information it holds that is not subject to privacy restrictions.

---

[100] See, for example, Goldacre B (2012) *Bad pharma: how drug companies mislead doctors and harm patients* (London: Fourth Estate).

[101] Already many clinical trials have been registered on open access sites (especially if publically funded) and in September 2013 it became a requirement to register when getting Research Ethics Committee approval in the UK (see http://www.hra.nhs.uk/news/2013/09/10/trial-registration-to-be-condition-of-the-favourable-rec-opinion-from-30-september/). In April 2014, the European Parliament adopted a new Clinical Trials Regulation, which requires all trials in Europe to be registered before they begin, and trial results to be published within a year of their end. See: http://ec.europa.eu/health/files/eudralex/vol-1/reg_2014_536/reg_2014_536_en.pdf. There is now a global movement for registration of all clinical trials (http://www.alltrials.net/).

[102] See http://opendefinition.org/od/.

[103] See, however, the discussion of the Personal Genome Project in chapter 7.

[104] See Royal Society (2012) *Science as an open enterprise*, available at: https://royalsociety.org/policy/projects/science-public-enterprise/Report/; see the Bethesda statement on open access publishing (2003), available at: http://legacy.earlham.edu/~peters/fos/bethesda.htm. Some funders, such as the Wellcome Trust, make it a condition of grants that the findings of research are published in open access journals. See: http://www.wellcome.ac.uk/About-us/Policy/Policy-and-position-statements/WTD002766.htm.

[105] The charge of "open-washing" is sometimes levelled at those who conflate data sharing and open data in order to imply that ethically ambiguous sharing initiatives have the laudable 'transparency' of open data.

[106] See Business, Innovation and Skills Committee (2013) *3rd special report – open access: responses to the committee's fifth report of session 2013-14*, available at: http://www.publications.parliament.uk/pa/cm201314/cmselect/cmbis/833/83302.htm.

[107] Commercial data brokers such as Google or Microsoft keep data locked up in their data centres. They allow third parties to build apps on top of it, and monetise it through adverts or access charges.

Big data and the knowledge economy

2.30 The exploitation of data science and technologies has acquired a central role in the political narratives around revitalising the UK economy.[108] Especially since 2012, this field of investigation has been seen as one of the 'eight great technologies' around which UK research and industrial policy has been built.[109] Similarly, the *Europe 2020* growth strategy gives prominence to big data in the 'Digital Agenda' for 'smart growth'.[110] The revision of European data protection law (the replacement of the existing Directive 95/46/EC with a new Data Protection Regulation) was initially presented as a streamlining of rules to facilitate data movement and support commercial activities (as well as to secure citizens' rights in the face of technological advances and to harmonise implementation across the Community).

2.31 In the UK, health science and biotechnology is one of the main dimensions along which value is expected from big data.[111] This sector is seen as especially promising because of the existing academic and research base, favourable commercial conditions, and potentially exploitable national data collections. Substantial political energy and investment is being directed towards capacity building, investment in infrastructure, education and training, streamlining regulation, developing innovation pathways and creating a welcoming commercial environment. The Medical Research Council (along with nine other funders) has made substantial investment (including £20M capital funding) in the Farr Institute of Health Informatics Research and UK Health Informatics Research Network.[112] Similar developments in the use of administrative data have seen the establishment of a network of Administrative Data Research Centres (ARDCs) in each of the four home countries. Both systems provide safe havens for linkage of datasets and analysis by approved researchers.[113]

Data threats

> **Proposition 7**
>
> Decisions and actions informed by the use of biological and health data may have both beneficial and harmful effects on individuals or on broader groups of people (e.g. families, companies, social groups, communities or society in general).

[108] See, for example, the Cabinet Office Data Strategy Board (see: https://www.gov.uk/government/publications/data-strategy-board-and-public-data-group-terms-of-reference.

[109] See, for example, Willetts D (2013) Eight great technologies, available at: http://www.policyexchange.org.uk/images/publications/eightper cent20greatper cent20technologies.pdf.

[110] See: http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=COM:2010:2020:FIN:EN:PDF. It is worth noting, also, the persistent use of human genome sequencing as a trope for big data. See, for example: http://www.computerweekly.com/news/2240226052/Cameron-announces-300m-big-data-human-genome-database-project.

[111] The recognition of the political significance can be seen, for example, in the Prime Ministerial backing for the launch of the 100K Genomes project. See: http://www.genomicsengland.co.uk/uk-to-become-world-number-one-in-dna-testing-with-plan-to-revolutionise-fight-against-cancer-and-rare-diseases/.

[112] The Farr Institute is named after William Farr (1807-83), one of the 'founding fathers' of medical statistics. Centres ('nodes') have been established at University College, London, the University of Manchester, Swansea University, and the University of Dundee (http://www.farrinstitute.org/).

[113] ADRCs (http://www.adrn.ac.uk/) enable large numbers of academic and other researchers to analyse and link data from, for example, tax, benefit, and education systems. These developments make it possible that, for example, medical researchers will be able to study disease outcomes as a function of income or education level; equally, researchers interested in taxation policy or in mechanisms for reducing disability benefit claims will have access to large-scale medical data. See Department for Business, Innovation and Skills (2013*) Improving access for research and policy: the government response to the report of the administrative data taskforce*, available at: https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/206873/bis-13-920-government-response-administrative-data-taskforce.pdf.

Proposition 8

The potential benefits and harms that could arise from data use may be valued radically differently by different people and by the same people at different times.

Cyber security

2.32 Large datasets with multiple users and access points are more attractive targets for attack. Studies of reported data breaches across sectors by The Identity Theft Resource Center, a US non-profit organisation monitoring data theft, find an emerging trend of the health-care sector as the target for the largest share of attacks (a greater number than the business sector).[114] Attacks can involve technical penetration by outsiders, dishonesty by insiders, or subversion of the executives who control the system. In the NHS, attention has historically focussed on the first of these; the NHS network, for example, uses encryption and many users authenticate themselves with smartcards. While this may have forestalled possible attacks of the first kind it does offers no protection against threats that fall in the second two categories. Abuse by insiders has a long history and the NHS is often the single largest reporter of data breaches to the Information Commissioner's Office (ICO).[115]

2.33 In addition to direct compromise of data security there are also secondary threats from malicious inference from legitimately released data and statistics. These cannot be resolved by access control or improved security because they make use only of data available through legitimate interfaces. For example, considerable concern was expressed when it emerged that the consolidated Hospital Episode Statistics (HES) records of England and Wales, consisting of about a billion finished consultant episodes from 1998–2013, had been sold to a number of non-profit and for-profit researchers, some of whom had been granted commercial re-use licences that continue to allow them to re-sell the data.[116]

2.34 Distrust of centralised information systems, of the technical or human elements, as well as principled opposition to inadequately justified or governed data processing, has provoked responses from privacy and civil liberty perspectives,[117] and from powerful professional groups such as the British Medical Association (BMA).[118] This, in turn, has raised concerns that support for research, and the benefits that may follow from responsible data re-use, could be negatively affected by such a loss of confidence. (We discuss an example in more detail in chapter 6.)

[114] See http://www.idtheftcenter.org/ITRC-Surveys-Studies/2014databreaches.html.

[115] There is mandatory reporting of NHS Level 2 security breach incidents both to the Department of Health and to the Information Commissioner's Office. See: https://www.igt.hscic.gov.uk/Publications/IGper cent20SIRIper cent20Reportingper cent20Toolper cent20Publicationper cent20Statement_Final_V2per cent200.pdf. There is an automated tool within the NHS Information Governance Toolkit for this purpose. No other public or private body has the same degree of mandatory reporting. This, taken with the vast amount of data held on every citizen contributes to the fact that the NHS is often the single largest reporter of data breaches.

[116] See the 'Partridge Review' of data releases made by the NHS Information Centre (http://www.hscic.gov.uk/datareview). See also: http://www.telegraph.co.uk/health/healthnews/10656893/Hospital-records-of-all-NHS-patients-sold-to-insurers.html; http://www.computing.co.uk/ctg/analysis/2352497/nhs-data-governance-in-critical-condition.

[117] See, for example, http://www.medconfidential.org; http://www.no2id.net/.

[118] See: http://bma.org.uk/news-views-analysis/news/2014/march/caredata-confidentiality-concerns-cannot-be-ignored-say-doctors; http://bma.org.uk/practical-support-at-work/ethics/confidentiality-and-health-records/care-data.

State surveillance

2.35 The data protection movement arose in the 1960s out of concerns about states building 'data banks' which would enable them to exercise surveillance and control over their citizens.[119] Half a century later, in 2013, a contractor working for the US National Security Agency (NSA), fled the USA with information, which he subsequently made public, about the activities of the NSA and its PRISM project.[120] Edward Snowden's revelations, which were published in the UK by *The Guardian* newspaper, had a chilling effect on support for personal data systems sponsored by the state or big corporations. They contained (among many other things) evidence relating to the UK Government Communications Headquarters (GCHQ) having extensive access to sensitive personal information which it shared with the US National Security Agency.[121]

2.36 Snowden's disclosures have provoked a significant debate about the security of IT infrastructure in Europe, with Germany taking a lead in putting forward proposals to create a European communications network to keep data from passing via the US, where they are vulnerable to the NSA.[122] The European Commission was already promoting the idea of European cloud computing and the Snowden revelations gave this more impetus. Many legal and governance issues arise in respect of the use of data cloud resources: cloud computing is a good way of solving the practical problems of processing very large amounts of data and it allows researchers in different jurisdictions to access a large centralised resource without the need to transfer data.[123] However, most cloud computing facilities are either in the USA or run by US firms and thus open to US warrants and, in other cases, it is unclear who ultimately controls access to cloud data. (We discuss the use of cloud computing for data analysis and collaborative biomedical research in chapter 7.)

2.37 The timing of Snowden's disclosures cannot have failed to have an effect on the progress of the draft EU Data Protection Regulation, which was made significantly more restrictive by the lead parliamentary committee. The anticipated chilling effect on medical research of the Parliament's amendments has been viewed with concern, in turn, by medical researchers, research funders and some interest groups.[124]

---

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

[120] PRISM is the familiar name of a mass electronic surveillance programme run since 2007 by the US National Security Agency. It collects data on internet communications from providers of internet services pursuant to requests under the Foreign Intelligence Surveillance Amendments Act of 2008 and approved by the Foreign Intelligence Surveillance Court. The Snowden disclosures suggested that the scale of data collection went significantly beyond the scope intended by the legislation providing for it. For further information on the Snowden disclosures, see: http://www.theguardian.com/us-news/the-nsa-files.

[121] See: http://www.theguardian.com/uk/2013/jun/21/gchq-cables-secret-world-communications-nsa.

[122] See: http://www.bbc.co.uk/news/world-europe-26210053. The US "Uniting and Strengthening America by Providing Appropriate Tools Required to Intercept and Obstruct Terrorism Act of 2001" (otherwise known as the 'USA PATRIOT Act', see: http://www.justice.gov/archive/ll/highlights.htm) allows the US authorities, under prescribed circumstances connected with national security and crime, to gain access to data held by US companies (including data on non-US citizens). The Act itself was a consolidation of existing powers, but it has become emblematic of the proactive and wide ranging use of state powers in the name of national security post-September 2001.

[123] For example, since 2011 PA consulting has uploaded HES data obtained from the HSCIC to Google BigQuery in order to manipulate the data: see http://www.paconsulting.com/introducing-pas-media-site/releases/pa-consulting-group-statement-3-march-2014/ and http://www.hscic.gov.uk/article/3948/Statement-Use-of-data-by-PA-consulting.

[124] See the joint statements from non-commercial research organisations and academics (updated December 2014), scientific research organisations (May 2013) and Federation of the European Academies of Medicine (June 2012), all available at: http://www.wellcome.ac.uk/About-us/Policy/Spotlight-issues/Personal-information/Data-protection-legislation/index.htm.

Discrimination

2.38 People who regularly use Internet browsers and search engines will be familiar with targeted pop-ups and personalised recommendations. Those who use the Internet for shopping are likely to be aware of the recommendation services of the sort pioneered by online retailer Amazon.com. These functions rely on customer profiling, where an individual's online activities (or rather those associated with a specific IP address or linked via a cookie on a device) are linked to create a 'user profile'. This may be done via a single entity (e.g. Amazon) or through intermediary sites (e.g. Doubleclick) invoked during web-page transition on many sites where activity across many entities are gathered to give more comprehensive information about an individual's interests.[125] It is this latter 'surveillance' that raises concerns as it is hard for an individual to control this in any way. It may be that few people have problems with single entity uses, for example, to produce recommendations, although they may find some of the marketing pop-ups irritating or perhaps embarrassing, as they may inadvertently reveal that person's interests to other users of the IP address or device. These technologies are linked to 'risk profiling' and screening in health care, with the same concerns about trying to provide benefits to individuals without appearing to 'look over their shoulders'.

2.39 Similar approaches may be used to infer further attributes of the user (and associate them with their online profile). For example, it has been shown to be possible to infer gender and sexuality with a high degree of reliability from use of social networking sites.[126] The same may be possible for health conditions (see Google flu tracking[127]) or other private information. In other words, correlations found in large datasets can support inferences based on individual online behaviour that can 'create' personal data where none was provided directly. It is important to consider what may rest on such inferred information.

2.40 For the purposes of targeted advertising, a fairly high probability of statistical correlation supporting a correct inference is usually sufficient; but whereas receiving inappropriate advertisements may be a minor irritation in most cases, in some it may have more significant consequences. An infamous account concerns a father who complained about US retail company, Target, sending his school-aged daughter coupons for baby products, only to discover later that she had been correctly profiled as pregnant by the company's software, using her purchasing patterns as markers.[128] There is an interesting postscript to the story of Target's pregnancy divination. An academic researcher from Princeton University carried out an experiment to see whether it would be possible to hide her own pregnancy from big data analytics. Given that so many social and economic transactions are mediated by electronic devices linked to the Internet this presented a significant challenge. Her conclusion was that

[125] "Some generic work can be done with de-identified data that is related to anonymous purchase data, but better targeted marketing depends upon knowing at least some of the properties of the possible purchasers, and ideally their identity. The main concerns/fears are that people who are less fit/at higher risk of disease and/or who have functional impairment will be discriminated against if identifiable biomedical data about them is widely available outside clinical practice & academic research establishments." Consultation response by Ian Herbert, available at: www.nuffieldbioethics.org/project/biological-health-data/evidence-gathering/.

[126] See Kosinski M, Stillwell D and Graepel T (2013) Private traits and attributes are predictable from digital records of human behaviour *Proceedings of the National Academy of Sciences* **110(15)**: 5802-5, available at: http://www.pnas.org/content/110/15/5802.full?sid=8148a219-8733-4ad0-adfe-e1523cb5feba.

[127] Google's flu trends service aims to identify the spread of flu symptoms in near real time, based on search terms entered into its search engine and geolocation of searching, and thereby enabling timely public health measures to be taken in response. See: http://www.google.org/flutrends/. However, the approach has limitations that differ from those of traditional disease surveillance: see http://www.nature.com/news/when-google-got-flu-wrong-1.12413.

[128] See: http://www.nytimes.com/2012/02/19/magazine/shopping-habits.html?pagewanted=all&_r=0.

hiding from big data is so inconvenient and expensive that it would be a difficult lifestyle choice.[129]

2.41 Some applications, such as credit scoring and differential pricing (where individual consumers are charged different prices for the same product or service based on factors such as credit history), may have the effect of systematically compounding social inequalities.[130] Others may perpetuate and reinforce societal inequality and discrimination (even where they do not rely directly on information about, for example, race, ethnicity, religion, etc.).[131] Even the most sophisticated algorithms are imperfect predictors, and the more unusual the case they are applied to, the more unreliable they become. At the very least, computer profiling can be insensitive, and fail to respect individuality, changing preferences and caprice. It could even make options of interest invisible and inaccessible to individuals.

2.42 One of the key difficulties with regulating the use of algorithms is their opacity, often even to those who employ them, and their complexity, which makes them difficult to understand and therefore to combat. One response to this has been to call for enhanced regulation and public scrutiny, although the commercial value of algorithms means that there may be reluctance to disclose them and some may be protected as trade secrets.[132] The widespread adoption of profiling and algorithmic prediction has potentially significant social consequences and raises important questions of public ethics and regulation.

Misuse of data

2.43 We have referred to a change in attitude towards data, generated by the recognition of its secondary use value, and the need for shared access to these, brought about by the increasing complexity and data dependency of professional practices such as medicine. The change in emphasis can be traced in successive versions of the General Medical Council's guidance on confidentiality and the distance travelled between the two 'Caldicott reports' on information governance in the NHS.[133] The first Caldicott report (1997) set out conservative principles to ensure the maintenance of patient confidentiality in the complex data flows that followed the implementation of IT systems within the NHS, and their operation by staff unaccustomed to dealing with information governance issues.[134] It ushered in a new role in NHS institutions that quickly became known as the 'Caldicott guardian' who was usually a senior health professional

[129] See: http://thinkprogress.org/culture/2014/04/29/3432050/can-you-hide-from-big-data/.

[130] Differential pricing based on consumer profiles may result in those with poor credit histories being charged more for a product or services than the more well-off on the basis that they present a greater credit risk, thus compounding their disadvantage and exploiting their vulnerability. See: http://www.huffingtonpost.com/nathan-newman/how-big-data-enables-econ_b_5820202.html.

[131] Gandy OH Jnr (2010) Engaging rational discrimination: exploring reasons for placing regulatory constraints on decision support systems *Ethics and Information Technology* **12(1)**: 29-42.

[132] Danna A and Gandy OH Jnr (2002) All that glitters is not gold: digging beneath the surface of data mining *Journal of Business Ethics* **40(4)**: 373–86, available at http://web.asc.upenn.edu/usr/ogandy/DMper cent20published.pdf.

[133] After 2009, the section of the GMC guidance that emphasised, quite straightforwardly, the importance of medical confidentiality acquired an important qualification: "But appropriate information sharing is essential to the efficient provision of safe, effective care, both for the individual patient and for the wider community of patients". See: (pre-2009) http://www.gmc-uk.org/Withdrawn_core_guidance_watermarked.pdf_27014281.pdf and (post-2009): http://www.gmc-uk.org/static/documents/content/Confidentiality_-_English_0914.pdf, at page 6. The 2009 guidance also contains new elaborated sections on public interest and research.

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

responsible for the protection of patient information. The second Caldicott report, appearing after the National Programme for IT, sought to promote a culture of responsible data 'sharing', including for secondary uses, and identified the possibility that failures to use data effectively could compromise treatment of patients as surely as failures to protect data could harm them.

2.44 Evidence of direct harm arising from misuse of data, particularly from re-identification of individuals from de-identified or pseudonymised datasets, is difficult to find. Registers of data breaches and complaints to the data protection authorities exist, and many such breaches are reported in the media. However, despite much hearsay and anecdotal evidence, and a scattering of clearly described incidents in the literature, we found no systematic assessment of harms arising as a consequence of data misuse. Therefore, as part of the evidence gathering that informed our deliberations we commissioned, jointly with the Expert Advisory Group on Data Access (EAGDA), some independent research into this question.[135] The researchers developed an empirical typology of harms arising from the misuse ('abuse') of data from biomedical research and health care, which was related to the type of abuse that led to them and its root cause.

Box 2.2: Empirical typology of data abuses, their causes and resulting harms

Type of abuse (decreasing intentionality)	Causes of abuse (decreasing intentionality)	Harms caused by abuse (decreasing severity)
■ Fabrication or falsification of data ■ Theft of data ■ Unauthorised disclosure of or access to data ■ Non-secure disposal of data ■ Unauthorised retention of data ■ Technical security failures ■ Loss of data ■ Non-use of data	■ Abuse of data to meet NHS/organisational objectives ■ Abuse of data to protect professional reputation ■ Abuse of data for self-gain (e.g. monetary gain) ■ Abuse attributed to third parties (e.g. hackers) ■ Disclosure by the press or media ■ Unauthorised access without clinical or lawful justification (e.g. for curiosity) ■ Against the wishes/ objections of the individual ■ Abuse as a result of insufficient safeguards	■ Receipt of suboptimal care, resulting in detriment to health or death ■ Individual distress e.g. emotional, physical, etc. ■ Damage to individual reputation (e.g. societal, personal or professional) ■ Individual, financial loss ■ Damage to public interest (e.g. loss of faith in confidential health service, general loss of public trust in medical profession, delayed or stunted scientific progress etc.) ■ Damage to organisational reputation (e.g. to NHS)

[135] EAGDA was established by the Wellcome Trust, Cancer Research UK, the Economic and Social Research Council, and the Medical Research Council to provide strategic advice on the emerging scientific, legal and ethical issues associated with data access for human genetics research and cohort studies (see: http://www.wellcome.ac.uk/About-us/Policy/Spotlight-issues/Data-sharing/EAGDA/index.htm). The research was delivered by a multidisciplinary team of researchers from the Mason Institute at the University of Edinburgh's Law School (see: http://masoninstitute.org/) and the Farr Institute's CIPHER at Swansea University's College of Medicine (see: http://www.farrinstitute.org/centre/CIPHER/34_About.html). Their report, Laurie G, Jones KH, Stevens L, and Dobbs C (2014) *A review of evidence relating to harm resulting from uses of health and biomedical data* is available on our website at: www.nuffieldbioethics.org/project/biological-health-data/evidence-gathering/.

- Abuse arising out of a Freedom of Information request
- Abuse due to maladministration (e.g. failure to follow correct procedures)
- Abuse due to human error (e.g. sending a fax to the wrong recipient)
- Non-use due to misinterpretation of legal obligations
- Potential for harm to individual, organisation or the public interest in future
- No evidence of harm found due to lack of reported information

Source: Laurie G, Jones KH, Stevens L, and Dobbs C (2014) *A review of evidence relating to harm resulting from uses of health and biomedical data*, available on our website at: www.nuffieldbioethics.org/project/biological-health-data/evidence-gathering/.

2.45 Methodologies for identifying harm face a number of serious limitations. First, the definition of harm used by the ICO excludes many incidents that would be considered harmful by data subjects.[136] Second, for data from health care settings, there is a lack of central reporting in the NHS.[137] Third, there are obstacles to obtaining redress: in the UK (as opposed to the USA), costs shifting, whereby the loser of a civil case generally pays the winner's costs, constitutes a serious discouragement to civil action for breach of confidence, since it is thus extremely risky for a private individual with limited means to sue. Consequently, abuses do not show up in law reports. Furthermore, criminal prosecutions for breach of confidence appear not to be a priority for law enforcement and only become so in high profile cases, such as the News International sponsored 'phone hacking'.[138] There is also the added complication that the victim may be unaware, and may never become aware, of the 'harm' (for example, where they are unsuccessful in a job application owing to information illicitly in the possession of the would-be employer). Finally, it may very often not be in the interest of the victim to pursue relief for privacy harms given that privacy harms are likely to be compounded by any publicity. The scarcity of documented cases of harm does not, therefore, provide very much reassurance that they do not exist. Those that are well documented in the available literature probably represent the tip of a much larger iceberg, as figure 1 (below) suggests.[139]

[136] See Laurie G, Jones KH, Stevens L, and Dobbs C (2014) *A review of evidence relating to harm resulting from uses of health and biomedical data*, available on our website at: www.nuffieldbioethics.org/project/biological-health-data/evidence-gathering/.

[137] A Freedom of Information (FOI) request by a Working Party member to the HSCIC for information on data breaches relating to the Personal Demographics Services (PDS) drew a response that information was not collected centrally and it would be necessary to contact individual trusts to obtain the data. See: https://www.whatdotheyknow.com/request/pds_exploits_and_breaches.

[138] *R. v. Coulson and ors* (unrep.) 4 July 2014. See sentencing remarks of Mr Justice Saunders: http://www.judiciary.gov.uk/wp-content/uploads/2014/07/sentencing-remarks-mr-j-saunders-r-v-coulson-others.pdf.

[139] Figure 1 is not drawn from the commissioned report; it was produced by Peter Singleton, a former member of the Working Party.

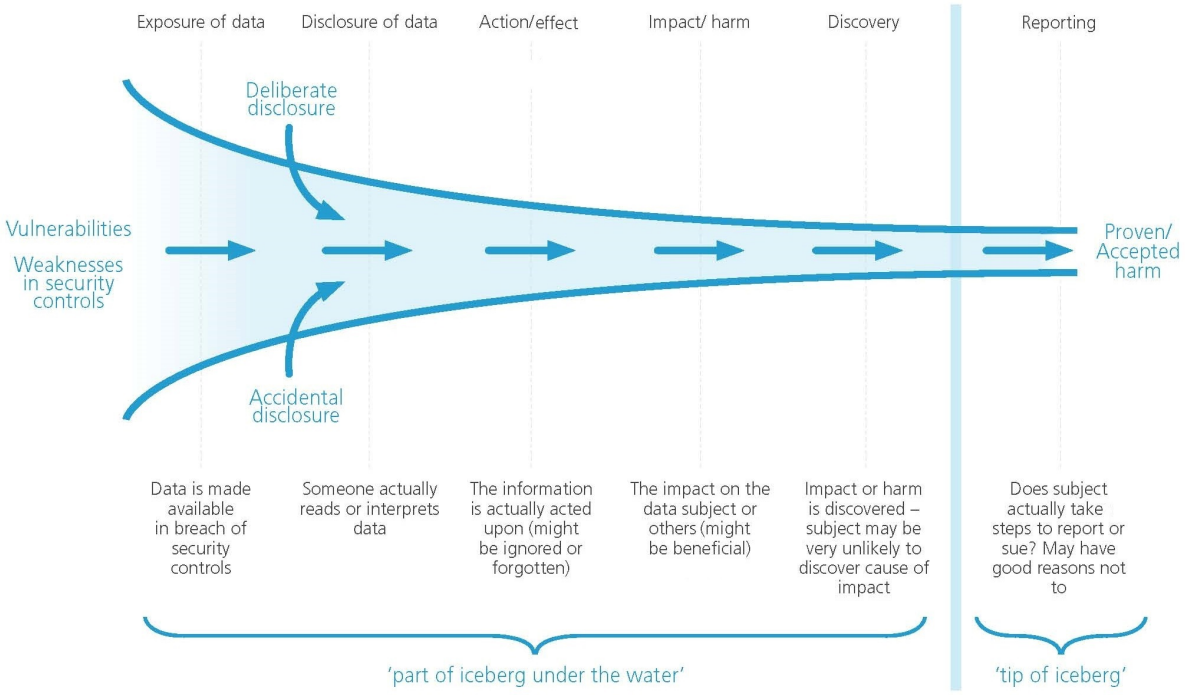

Figure 1: the 'confidentiality funnel'

2.46 Because of the demanding conditions for an adverse effect to be formally classified as a harm the commissioned research distinguished between 'harm', which could be recognised as a cause for action, and 'impact', which included non-actionable but nevertheless morally significant adverse effects. The research sought evidence of both 'harms' and 'impacts'. The research found that the broad category of 'maladministration' was the main cause of data abuse and therefore an important potential source of harm (although the majority of the evidence related to health care rather than research systems). One form of maladministration is simple human error.

Box 2.3: The case of Helen Wilkinson

While most national database collections are created with the best of intentions, it is important to recognise that they often use either identified data or at least identifiable data about patients' treatments, which entails privacy risks where information (whether true or, as in this case, false) may have consequences for an individual to whom it is connected.

In 2004 Helen Wilkinson was a GP Practice Manager who discovered that data submitted to the NHS-wide Clearing Service (NWCS) managed by McKesson for the NHS had a mis-coded record indicating that she had attended an alcohol advisory service in 1998 instead of a surgical procedure.[140]

Attempts to have this corrected were thwarted and led to her case being debated in Parliament in June 2005, including the fact that there were no facilities for patients to opt-out of their data being collected centrally.

Although the database concerned would not be used for actual medical treatment (or in

[140] The facts of this case are given in House of Commons Hansard, 16 Jun 2005, Col.495, where the case is reported as the subject of an adjournment debate.

> the case of screening invitations, would not use this particular data), Ms Wilkinson nevertheless suffered substantial personal embarrassment and distress as a result of the error and the difficulty in correcting it and, as a result, withdrew from the care of the NHS.

2.47 Errors of the kind described in Box 2.3 above may be particularly damaging if they are replicated across information systems and if they go undetected and inform the way in which individuals are treated (in the broadest sense). Where they are detected they can usually be corrected (although any dissemination that has already taken place through other systems may make this more difficult).

2.48 There are, however, abuses more intentionally damaging than simple errors of administration. For many years, NHS systems have been abused by private investigators, journalists and others to track down targets of investigation. A standard technique was the 'blag' or false-pretext telephone call, in which the caller phones one NHS organisation pretending to be from another NHS organisation and asks for information about a patient. [141] In 1996 the BMA issued guidance on how to detect and avoid such abuse: rather than simply handing out information over the phone, staff were advised to log all requests, consult a senior clinician for approval and call back only to numbers in the phone book rather than to a number given by the caller.[142] In that year, staff at the NW Yorkshire Health Authority, trained to follow this guidance, discovered several dozen false-pretext calls per week.[143] The system that is now the natural target for attacks is the NHS's Personal Demographics Service (PDS), which contains the private contact information of all NHS patients and is available to hundreds of thousands of NHS staff, who use it routinely to verify the names and dates of birth of patients presenting for treatment and look up NHS numbers so that records can be retrieved.

2.49 The broad conclusion of the research, which we endorse, was that relying on compliance with current legal requirements is insufficient to avert harm and that 'harm' as currently recognised by authorities (the ICO, tribunals and courts) failed to provide a complete picture of how harm resulting from abuse of data is perceived or experienced by individuals.[144]

> "This is not to suggest that groundless concerns or abstract fears should drive information governance practices. Rather... the range of considerations about what might be construed as harmful is far wider than the law alone recognises. As such, the lesson is that due attention should be paid to possible impacts when using health and biomedical data, and to ensuring that governance mechanisms and actors within

[141] This risk was highlighted in the case of Jacintha Saldanha, a night sister at King Edward VII Hospital in London, who committed suicide after transferring a hoax call from Australian radio station, 2Day FM, to a nurse caring for the pregnant Duchess of Cambridge, believing the call to be from the Queen and the Prince of Wales (http://www.theguardian.com/world/2014/sep/12/jacintha-saldanha-death-suicide-prank-call-dj-apologises).

[142] Anderson R (1996) Clinical system security – interim guidelines *British Media Journal* **312(7023)**:109–111, available at: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2349761/pdf/bmj00524-0047.pdf.

[143] Hassey A and Wells M (1997) Clinical systems security – implementing the BMA policy and guidelines, in *Personal medical information: security, engineering, and ethics*, Anderson R (Editor) (Berlin: Springer) pp 79–93.

[144] The research made a technical distinction between 'harms' from the hard evidence search (essentially those that satisfied a legal definition) from negative 'impacts' that were identified in the soft evidence.

them have the ability to assess and, where appropriate, respond to data subjects' expectations."[145]

2.50 The commissioned report was intended only as an initial scoping exercise. It has, however, sufficiently demonstrated the importance and urgency of carrying out more thoroughgoing research in order to form a realistic picture of the incidence and possible hazards of data abuse. Based on our examination of this area, and in the light of our deliberations, we make a number of recommendations below.

Recommendation 1

We recommend that relevant bodies, including public and private research funders and UK health departments, ensure that there is continued research into the potential harms associated with abuse of biological and health data, as well as its benefits. This research should be sustained as available data and data technologies evolve, maintaining vigilance for new harms that may emerge. Appropriate research that challenges current policy orientations should be particularly encouraged in order to identify and test the robustness of institutional assumptions.

Recommendation 2

We recommend that the Independent Information Governance Oversight Panel and the Health Research Authority supervise, respectively, the maintenance of comprehensive maps of UK health and research data flows and actively support both prospective and continuing evaluation of the risks or benefits of any policies, standards, or laws governing data used in biomedical research and health care.

Recommendation 3

We recommend that the Government make enforceable provisions to ensure that privacy breaches involving individual-level data that occur in health services and biomedical research projects are reported in a timely and appropriate fashion to the individual or individuals affected.

Recommendation 4

We recommend that the Health and Social Care Information Centre maintain prospective assessments to inform the most effective methods for preventing the inadvertent or fraudulent accessing of personal health care data by unauthorised individuals.

[145] Laurie G, Jones KH, Stevens L, and Dobbs C (2014) *A review of evidence relating to harm resulting from uses of health and biomedical data*, available on our website at: www.nuffieldbioethics.org/project/biological-health-data/evidence-gathering/, at page 161.

Recommendation 5

We recommend that the UK government legislate to introduce criminal penalties, comparable to those applicable for offences under the Computer Misuse Act 1990, for *deliberate* misuse of data *whether or not* it results in demonstrable harm to individuals.

Conclusion

2.51 The opportunities promised by advances in IT and data science, and the demands of wider industrial policy to develop the knowledge economy, have provoked a reorientation of policy on the use of biomedical and health data from care support towards a broader value extraction. This is the case in several major developed economies. In the UK, a particular focus has been on the development of genomic technologies and the exploitation of data collected by the NHS. The effect of policy decisions has been to promote – and, to an extent, to lock in – data-intensive initiatives as a generator of economic activity in the near term and to establish the conditions for improved and more cost-effective treatments and services in the long term. This nevertheless makes it difficult to disentangle the confusion of motives behind policies affecting the protection and exploitation of data in biomedicine and health care.