Chapter 1

Data

Chapter 1 – Data

Chapter overview

This chapter describes some of the sources and varieties of data that are accumulating in health care and biomedical research settings, and the increasing ways in which data may be used.

Data provide the raw materials for reasoning and calculation. The informational value of data arises from the context in which they are placed, and how they relate to other data. The meaning, utility and value of data may be transformed as they appear within different contexts such as health care, research and public policy. Digitisation has allowed an escalating accumulation of data in health care and biomedical research settings, including:

- clinical care data (e.g. primary care and hospital records)
- data from clinical trials and observational studies
- patient-generated data (e.g. from 'life logging' or consumer genetic testing)
- laboratory data (e.g. from imaging, genome sequencing and other 'omics')
- administrative data or metadata

Advances in information technology (faster information storage, retrieval and processing) and data science (more powerful statistical techniques and algorithms) have created novel opportunities to derive insights from the analysis of big datasets, and particularly through the combination or linking of datasets. While these developments are not specific to biomedical research and health care, they are having a significant impact in these fields, with morally significant implications. They have led to the emergence of a new attitude towards data that sees them as exploitable raw materials, which can be put to use for a variety of purposes beyond those for which they were originally collected.

'Data initiatives' involve the re-using data in novel contexts and linking them with data from other sources. However, inconsistent data quality and peculiarities arising from the context of data collection can present technical difficulties in exploiting these opportunities. Furthermore, legal and ethical limitations placed on the re-use of data for secondary purposes limit the re-use of existing data sets.

Introduction

1.1 This chapter is about the accumulation and use of data in biomedical research and health care that has been enabled by developments in computing, biotechnologies, bioinformatics and professional practice since the last decade of the 20th Century. The dramatic growth in the volume and variety of these data and in our capacities for collecting, storing, combining, analysing and putting them to use, are the main advances that have given rise to this report. The Nuffield Council on Bioethics believes that these connected developments are significant and that the issues they raise are important not only for specialists, and in certain circumstances, but generally, and for all members of society.

1.2 Many of the sources of biological and health data described in this chapter are not new. We have been collecting and accumulating data in many areas of life since the advent of writing and the practice of analysing those data is as old as the practice of medicine.[1] Even those sources of data that involve the most advanced technologies have often taken some time to enter routine use.[2] The timeliness of this report rests on a claim that those innovations in data production, along with advances in the capture and analysis of data, have brought about a shift of emphasis in the way in which knowledge, well-being and public goods are pursued that has morally significant consequences.

Data and digitisation

1.3 'Data' means 'given things', i.e. things that are known or assumed as facts rather than deduced, inferred or imagined by us. Data produced by observation or measurement form the basis of reasoning and calculation.[3] For the purposes of this report we simply draw a distinction between *data*, which we treat as the raw materials for analysis, and their *informational value*, which is given by the relation in which they stand to other facts or conclusions within a particular context. It is through relational properties that data have in a particular context that they acquire real-world significance: whereas *data* are treated as simply *given*, *information* has *meaning*. Ethical questions arise from how we use data within a context that gives them a particular meaning. Such a context might be, for example, one created by a particular research question we are trying to answer or a decision with which we are faced.

1.4 When we use census data to calculate average lifespan within a population we are treating data as given (e.g. baby boys living in the most deprived areas in England in 2010-12 can expect to live 7.5 years less than those in least deprived areas).[4] When we investigate the accuracy of those data or question how they were collected we interrogate their informational value (e.g. what is the age range used to classify who falls into the category of a 'baby'?). When we investigate the social meaning of data, we begin to ask about assumptions and values underlying the information (e.g. what does 'deprived' mean in this context?). Changing the context in which data are presented can significantly alter their informational value, especially if there are unusual or atypical values in that the new context. For example, data about our individual biology collected to diagnose disease or predict disease risk may also serve to identify us or establish our relationship to others. Conversely, data that were not originally acquired for health purposes can become a valuable source of health information. For example, data about our lifestyles – our alcohol intake or the contents

---

1 For example, the Hippocratic corpus (Books I and III of *Epidemics*) contains forty two case histories. Hippocrates (1923) *Volume I: Ancient medicine. Airs, waters, places. Epidemics 1 and 3. The oath. Precepts. Nutriment* (London: William Heinemann), available at: https://archive.org/details/hippocrates01hippuoft.

2 For example, gene sequencing has been possible since the 1970s and has been in use in clinical practice for decades, (for example, in Down's syndrome screening, Philadelphia chromosome testing for leukaemia, and neonatal screening programmes. The sequencing method in use today was developed largely by Frederick Sanger in 1977; see: Sanger F, Nicklen S, and Coulson AR (1977) DNA sequencing with chain-terminating inhibitors *Proceedings of the National Academy of Sciences* **74(12)**: 5463-7, available at: http://www.pnas.org/content/74/12/5463. See also: Hutchison III CA (2007) DNA sequencing: bench to bedside and beyond *Nucleic Acids Research* **35(18)**: 6227-37, available at: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2094077/;Korf BF (2013) Integration of genomics into medical practice *Discovery Medicine* **16(89)**: 241-8, available at: http://www.discoverymedicine.com/Bruce-R-Korf/2013/11/08/integration-of-genomics-into-medical-practice/.

3 The Oxford dictionary gives one meaning of data as "things known or assumed as facts, making the basis of reasoning ". Stevenson, A and Waite, M (2011) *Concise Oxford English dictionary*, 12th edition (Oxford: Oxford University Press).

4 See: http://www.ons.gov.uk/ons/rel/subnational-health4/life-expec-at-birth-age-65/2006-08-to-2010-12/sty-life-expectancy-gap.html.

of our shopping baskets, our daily activities and exercise routines – can become 'health data' when framed by questions of mental health or disease risk in later life. Social data can help to predict the course of epidemics and inform public health responses to them.

1.5 The development that has allowed the collection and accumulation of unprecedented amounts of data is digitisation. The widespread use of electronic media means that data are generated at a rate that is difficult to imagine.[5] In less than a generation the recording of medical data has moved from 'doctors' notes' to computer-based records that capture standardised information, are accessible in a range of settings, and support a wide range of purposes in addition to clinical care of the individual patient (such as resource planning, cost effectiveness evaluations, etc.). A similar journey has taken place in biomedical and population health research: within the span of an academic career, many researchers have gone from using an edge-notched punch card system to digital data mining using cloud-based data services several orders of magnitude more powerful.[6]

Data, records and tissues

1.6 The association between data and the medium in which they are stored may create difficulties from the point of view of governance. Lloyd George who, as Chancellor of the Exchequer, introduced the National Insurance Act 1911, famously came into conflict with the medical profession over the question of who owned the new medical records he introduced. From that it emerged that the Secretary of State owned the paper, the doctor owned the writing on it, and the record would pass to the Government on the patient's death for statistical analysis.

1.7 A great deal of biological data, such as DNA sequence data is encoded within the tissues of the body, which enables it to serve so well as a biometric identifier for forensic purposes. Similarly, advances in synthetic biology have allowed DNA molecules to be used as a storage system that might conceivably be used in the future for archiving. (It has been claimed that world's total stock of information, 1.8 zettabytes at the time, could be stored in about four grams of DNA).[7] At present, retrieval is too slow and expensive to make this useful for computing purposes, although this limitation might be overcome in future. Developments in sequencing speed, for example, might eventually make it more cost effective to sequence patients' DNA as required rather than storing the information on more expensive magnetic or semiconductor memory.

1.8 The relationship between tissues and data has been subject to legal as well as technological displacement: in a significant case relating to DNA sample and profile

---

[5] IBM's website, for example, carries the claim that "Every day, we create 2.5 quintillion bytes of data — so much that 90 per cent of the data in the world today has been created in the last two years alone." See: www-01.ibm.com/software/sg/data/bigdata/.

[6] In a common example of the former system, information was recorded on a number of index cards, and holes were punched around the edges, each one representing a data point. The (binary) data value would be given by whether the hole was then notched to continue it to the outer edge of the card. This allowed the answer to a research question to be found by identifying the cards that fell from the stack (possibly over several iterations for complex Boolean questions) when something like a knitting needle was inserted into the appropriate holes around the edge. For larger data sets mechanical counter sorters could be used.

[7] Church GM, Gao Y and Kosuri S (2012) Next-generation digital information storage in DNA *Science* **337(6102)**: 1628, available at: http://arep.med.harvard.edu/pdf/Church_Science_12.pdf; Goldman N, Bertone P, Chen S *et al.* (2013) Towards practical, high-capacity, low-maintenance information storage in synthesized DNA *Nature* **494:** 77-80, available at: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3672958/pdf/emss-51823.pdf. This estimate is reported at http://www.computerworld.com/s/article/9230401/Harvard_stores_70_billion_books_using_DNA. A byte is a unit of digital information comprising 8 bits (each of which can have two values). A zettabye is 10^{21} bytes.

retention by the British police, the European Court of Human Rights has suggested that tissues containing DNA should be subject to the EU data protection regime.[8] The Article 29 Working Party (the European advisory body on data protection established under Article 29 of the European Data Protection Directive) acknowledged the need to attend carefully to the legal status of tissues and the range of data subjects' rights they engage.[9] The persistence of this question nevertheless exemplifies the fact that technological developments for extracting data – not restricted to DNA sequencing – have created complexities at the intersection of multiple regulatory and governance regimes for data and tissues. A case in point is the longstanding difficulty of determining what should and may be done with 'Guthrie' cards (records containing a blood spot sample routinely collected for neonatal health screening in many countries since the late 1960s).[10]

Observational data

Proposition 1

There is a growing accumulation of data, of increasing variety, about human biology, health, disease and functioning, derived ultimately from the study of people.

Clinical care data

1.9 One of the primary sources of data with which we shall be concerned is clinical care. From the moment of birth, each of us, in the developed world, is more likely to interact with health care professionals than almost any other public service. Since the introduction of the 'Lloyd George' record in 1911, information has been recorded routinely about all NHS patients.[11] Originally *aides-mémoires* that recorded the information an individual doctor judged to be useful in order to treat the same patient on subsequent occasions, or to refer them to a colleague, medical records have become increasingly standardised and multi-purpose.

1.10 The original paper records were vulnerable to physical deterioration, and to misfiling or being misplaced, and subject to increasing costs of storage.[12] The problems associated with paper record management combined with the need to gain rapid reimbursement led many GP practices to keep 'additional records' on computer despite the fact that the keeping of paper records remained mandatory in the UK until October

8 See *S. and Marper v United Kingdom* [2008] ECHR 1581.

9 Article 29 Data Protection Working Party (2004) *Working document on genetic data* (WP 91), available at: http://ec.europa.eu/justice/policies/privacy/docs/wpdocs/2004/wp91_en.pdf. See also Beyleveld D and Taylor M (2008) Patents for biotechnoloy and the data protection of biological samples and shared genetic data, in *The protection of medical data: challenges of the 21st century*, Herveg J (Editor) (Louvain-la-Neuve: Anthemis).

10 See: Laurie G, Hunter K, and Cunningham-Burley, S. (2013) *Storage, use and access to the Scottish Guthrie card collection: ethical, legal and social Issues* (The Scottish Government Social Research), available at: http://www.scotland.gov.uk/Publications/2014/01/7520.

11 The A5-sized record card envelope was introduced in 1911 by David Lloyd George when Minister for Health. The use of Lloyd George records was mandated until October 2000 and GP practices are still required to maintain extant paper records. A typical GP practice, with 6000 patients will house over 5,000,000 pages in Lloyd George envelopes.

12 It is notable that the Royal College of Physicians only approved standards for paper-based medical records in 2007, indicating the need for records to be interpretable outside their original source. Standards for electronic records were published by the RCP in 2013. See: https://www.rcplondon.ac.uk/resources/generic-medical-record-keeping-standards; www.rcplondon.ac.uk/resources/standards-clinical-structure-and-content-patient-records.

2000.[13] Computerisation has facilitated developments in medical practice, allowing multidisciplinary teams to work together across health care sites, specialties and agencies. It has also significantly enabled the possibility of research using health records.[14] Health care systems now record and standardise ever more data about people and their care, integrating data from other care providers (outside immediate health care), including 'plans' (e.g. care pathways) as well as outcomes, which include patient-reported outcome measures (PROMs).[15] Increasingly, substantial amounts of data are being recorded that are ancillary to the practice of medicine.[16] There is a growing expectation that more information about lifestyle and environmental factors will be recorded, as these are increasingly recognised as potentially modifiable determinants of health risk.

Clinical trials and observational studies

1.11 A significant amount of scientific data is collected during clinical trials for medicines, or other clinical research, in which researchers design the study, allocate different interventions to separate groups of people and attempt, as far as possible, to standardise other factors that could influence the outcomes. However these are limited in scale. In contrast, observational study data are collected alongside the provision of health care or periodically over time. Observational study data, either from disease-specific populations or from the public more generally, are the main resources for statistical analysis and modelling using epidemiological methods for public health research. Such data are also used in social science research to study the everyday behaviour of individuals or cultural groups.[17] Although clinical trials are often referred to as the 'gold standard' for investigating research hypotheses, both observational and clinical trial data have much in common in terms of the statistical methodology used to identify the relative importance of different characteristics or events, in other words, to invest the data with meaning and extract information.

1.12 Observational studies can involve a snapshot of a state of affairs at a particular time but longitudinal observational studies gather large amounts of data over long timescales (sometimes generational) during which contributory factors can be investigated.[18] Most prospective observational studies involve a recruitment and consent process in order to collect medical data, biological samples and other data, such as retrospective medical history or lifestyle data, via interviews or questionnaires. Studies vary considerably in the time commitment of participants and the possibilities of unintended harms whether physical, mental, emotional or informational for participants.[19] One of the most famous is the Framingham Heart Study, which has

[13] See chapter 6 (below) for a more extended discussion of health record systems.

[14] The business model for VAMP, an early GP computerised record system, was predicated not on sales of systems to GPs, but of sales of statistical data to pharmaceutical companies (see chapter 6 below).

[15] Although what patients feel and how they function have always been part of medical records, the codification of such data and the secondary uses that this enables are new.

[16] See NHS Confederation (2013) *Challenging bureaucracy*, available at http://www.nhsconfed.org/~/media/Confederation/Files/Publications/Documents/challenging-bureaucracy.pdf.

[17] For social science data collection, see: Lapan S, Quartaroli M and Riemer F (2012) *Qualitative research: an introduction to methods and designs* (San Francisco: Wiley).

[18] In the UK, the Office for National Statistics Longitudinal Study (LS, in England and Wales) and Scottish Longitudinal Study (SLS) link data for a sample of the population from administrative, 'vital events' and health data sets, starting with a sample from the 1971 and 1991 census returns, respectively. For ONS LS, see http://www.ons.gov.uk/ons/guide-method/user-guidance/longitudinal-study/index.html; for SLS see: http://sls.lscs.ac.uk/.

[19] This can range from the minimally intrusive (e.g. where data are collected in accordance with the participants' consent from health records, through periodic interviews (as in the case of the Avon Longitudinal Study of Parents and Children - ALSPAC) or sampling (UK Biobank) through to regular invasive sampling (for example, in the Harvard Biomarkers Core, which involves regularly taking a variety of biological samples from participants with Parkinson's disease and other

been running in Framingham, Massachusetts, since 1948. Observation of study participants has contributed significantly to understanding of the risk factors for cardiovascular (and other) disease, which were previously thought to be associated with natural ageing.[20]

1.13 Because, unlike clinical trials, the parameters of the study are not strictly controlled, scale is an important aspect of observational studies. Whereas the original Framingham study enrolled just over 5,000 adults in 1948, much larger observational studies have since been initiated. The UK 1958 birth cohort (and later ones) enrolled 98 per cent of the over 17,000 mothers giving birth in a particular week in England, Wales and Scotland, and follow-up of the children has continued at intervals ever since.[21] The UK's new Life Study will gather data on more than 80,000 babies to look principally at social and environmental determinants of development and health.[22] UK Biobank has enrolled a hundred times the number of volunteers in the original Framingham study in order to obtain sufficient numbers of cases of all the common diseases to facilitate a broad range of research investigations (see chapter 7 below). In the Million Women Study, over a million women were recruited through NHS Breast Screening Clinics between 1996 and 2001 and followed up for a range of health conditions including cancers, osteoporosis and cardiovascular disease.[23]

1.14 While certain practicalities of data capture and storage (participants' willingness, researchers' time and resources, and data storage technologies) limit observational studies, new monitoring devices or activity monitors, and wearable or implantable technologies (e.g. ambulatory heart rate monitoring devices), have made data collection more frequent (even continuous) and much less resource intensive, as well as socially acceptable.[24] Web-based questionnaires have also made the collection of other data from research participants much more efficient, and the rapid and widespread diffusion of mobile phone technology, allowing geospatial location and remote transmission, is providing innovative opportunities for collection of data relevant to real world scenarios.[25]

---

neurodegenerative diseases and a healthy control group, see: http://www.neurodiscovery.harvard.edu/research/biomarkers.html).

20 Mahmood SS, Levy D, Vasan RS and Wang TJ (2014) The Framingham Heart Study and the epidemiology of cardiovascular disease: a historical perspective *The Lancet* **383(9921)**: 999-1008.

21 Power C and Elliott J (2006) Cohort profile: 1958 british birth cohort (national child development study) *International Journal of Epidemiology* **35(1)**: 34-41, available at: http://ije.oxfordjournals.org/content/35/1/34.short. The Avon Longitudinal Study of Parents and Children (ALSPAC) studies a geographically local cohort, enrolling more than 14,000 pregnant women in the early 1990s and has consistently generated research findings about genetic and environmental aspects of health since then (see: http://www.bristol.ac.uk/alspac/).

22 See: http://www.lifestudy.ac.uk/homepage.

23 Research using data from the Million Women study has been influential in the development of clinical guidelines and public health, particularly for planning screening programmes and use of Hormone Replacement Therapy. See: http://www.millionwomenstudy.org/introduction/; http://www.ox.ac.uk/research/research-impact/million-women-study.

24 See: Pierleoni P, Pernini L, Belli A, and Palma L (2014) an android-based heart monitoring system for the elderly and for patients with heart disease *International Journal of Telemedicine and Applications*, available at: http://www.hindawi.com/journals/ijta/2014/625156/; Svagård I, Austad HO, Seeberg T, et al. (2014) A usability study of a mobile monitoring system for congestive heart failure patients *Studies in Health Technology and Informatics* **205**: 528-32, available at: http://ebooks.iospress.nl/publication/37543; Banos O, Villalonga C, Damas M, et al. (2014) PhysioDroid: combining wearable health sensors and mobile devices for a ubiquitous, continuous, and personal monitoring *Scientific World Journal*, available at: http://www.hindawi.com/journals/tswj/2014/490824/.

25 For example, Google's flu trends service, which 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, 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; http://www.ncbi.nlm.nih.gov/pubmed/24626916.

Lifestyle and social data innovations

1.15 A number of applications have emerged for tracking daily life ('life logging') in terms of inputs (e.g. food, air quality), states (e.g. mood, blood oxygen levels) and mental and physical performance. Such self-monitoring and self-sensing can combine wearable sensors (e.g. the 'fitbit') and computing (e.g. ECG, blood oxygen, steps taken).[26] The availability of screening devices such as continuous blood pressure recorders would seem to be quite widely used (in the UK) to supplement the blood pressure screening generally available through the National Health Service.[27] There is also an initially rather modest uptake of commercial genetic profiling, which provides genetic risk estimates to customers for a number of diseases and traits.[28] These technological innovations have had the effect of allowing non-specialists to develop their interests in research at both a personal and more public level.

1.16 The primary aim of collecting such data is for individuals to self-monitor as a means of improving health and fitness, often using apps that may involve uploading data to the Internet (members of the Quantified Self movement are enthusiastic sharers of lifestyle data through social networks), where it may be used to inform those with similar interests or taken up more widely into research.[29] There is also at least one platform that allows customers of direct-to-customer genetic tests to publish their results, to compare theirs with others and find information about their implications. There are also recreational family history services based around DNA testing.[30] Similar approaches using social networking platforms have also been adopted by patient groups that aim to generate data about conditions affecting the members. These data may then be made available to researchers to help in the development of more effective products, services and care. One of the best known is PatientsLikeMe (see chapter 7 below). Some companies who provide testing and interpretation (such as the genetic profiling company 23andMe) may themselves also carry out research using their customers' samples and information. The data they generate may be reported, although they are not usually made available for wider research use.

Laboratory data

Imaging

1.17 Imaging offers a way to understand complex biological phenomena by making use of human capacities for processing visual representations. Different wavelengths of energy are used, ranging from those for MRI (long wavelengths), through infrared

---

[26] An emerging technology is 'physiological computing': for example the Xbox One game console has a built-in camera that can monitor the heart rate of people in the room for the purpose of exercise games, but could be put to other uses. See: Fairclough S (2014) Physiological data must remain confidential *Nature* **505(7483)**: 263, available at: http://www.nature.com/polopoly_fs/1.14524!/menu/main/topColumns/topLeftColumn/pdf/505263a.pdf.

[27] Khattar RS, Swales JD, Banfield A, *et al*. (1999) Prediction of coronary and cerebrovascular morbidity and mortality by direct continuous ambulatory blood pressure monitoring in essential hypotension *Circulation* **100**: 1071–6, available at: http://circ.ahajournals.org/content/100/10/1071.short.

[28] One of the leading providers, 23andMe, encountered difficulties when the FDA halted some of its operations in the USA, although it has launched new services in other countries, including the UK, and claims to have 600,000 customers worldwide. See: Annas GJ and Sherman Elias S (2014) 23andMe and the FDA *New England Journal of Medicine* **370(11)**: 985-8, available at: http://www.nejm.org/doi/full/10.1056/NEJMp1316367; http://www.theguardian.com/technology/2014/dec/02/google-genetic-testing-23andme-uk-launch.

[29] For the quantified self movement, see: http://quantifiedself.com/. Examples of applications such as the Google Fit and Apple's Health Kit and iPhone and iPad Health app allow data to be used for app development and for further purposes, depending on the design and privacy settings.

[30] For publication of DNA profiles, see Greshake B, Bayer PE, Rausch H and Reda J (2014) openSNP–A Crowdsourced Web Resource for Personal Genomics *PLoS ONE* **9**(3): e89204, available at: http://www.plosone.org/article/infoper cent3Adoiper cent2F10.1371per cent2Fjournal.pone.0089204; for an example of a DNA genealogy service, see: http://dna.ancestry.com/.

thermal imaging, into the visual spectrum, to X-rays (very short wavelengths), as well as ultrasound. Most modern imaging applications related to health care generate digital data. In many cases, the image is used to summarise a complex set of quantitative data as 'picture elements', each of which encodes a value for the interaction of the imaging energy and object at a corresponding point in space. In practice, the raw data collected as a string of values have little intrinsic relationship to an image. The visual pattern or form is reconstructed from the full set of data by filtering it to enhance true signal against noise (e.g. scattered light or glare for visual images) and then assembling the most probable representation based on understanding of how the data were acquired. For imaging techniques like functional MRI (fMRI), the development of a functional image is explicitly probabilistic: hundreds of images are summed and statistically contrasted to estimate true signal changes associated with the changes in brain physiology that are linked with perception or thought.

1.18 Representations of brain activity associated with cognitive processes are becoming a tool for understanding brain functions in health and disease. Their popularization has provided a visual metaphor for 'thought' as the transfer of information in the brain. They have also suggested the possibility of brain imaging being used as a lie detector or even a 'mind reader'.[31] In fact, while the information content of images is high, all of the techniques are restricted to gathering limited dimensions of information. Brain imaging methods capture correlates of cognitive processes with limited spatial-temporal resolution. While correlates of different types of thoughts can be distinguished in a probabilistic way, the 'contents' of thought thus far cannot be captured in any general sense.[32]

1.19 A revolution has occurred in imaging as the very large datasets (often gigabytes even for a single complete set of brain MRI, for example) have been able to be manipulated and integrated with other datasets as well as easily searched for specific data using digital computational methods. This has allowed new kinds of features to be detected ('visualized'") and new measures to be defined as well as enabling easier storing and sharing of patient information among clinicians. Viewed in this way, imaging has become as much an extension of contemporary bioinformatics as the skilled use of a particular physical method.

Biomarkers

1.20 Scientific laboratory services have, for a long time, played an important part in the diagnosis of disease, initially through cellular and chemical evaluation of blood and tissues as well as molecular profiles. Pathology services interact with a variety of registries and tissue banks and, analogously to GP clinical records, information is managed in the UK through a range of Laboratory Information Management Systems (LIMS). Biomarkers (or biological markers) are measurable characteristics that can indicate an underlying biological state or condition, such as a disease state. The value of different biomarkers depends on the accuracy of the measurement, the association

[31] Some have even coupled this with notions of remote surveillance (e.g., satellite imaging) to conclude that there is a potential for *mass* mind-reading and, with it, the ultimate destruction of privacy. However, attempts at making fMRI work in lie detection have arguably been, to date, just as ineffective as the old-fashioned polygraph. See: Vrij A (2008) *Detecting lies and deceit: pitfalls and opportunities*, Volume two (Chichester: Wiley), p365ff.

[32] See also Nuffield Council on Bioethics (2013) *Novel neurotechnologies: intervening in the brain*, available at: http://nuffieldbioethics.org/project/neurotechnology/.

between what is measured and the underlying state of interest, and the relevance of this to the particular question to be addressed.

1.21 The use of biomarkers, and the role of laboratory services, has become increasingly widespread. Biomarkers may be able to identify disease prior to development of symptoms. Through the use of biomarkers, ostensibly similar clinical presentations have been revealed to be distinct, leading to more tailored therapeutic interventions (Personalised medicine). However, the identification and validation of biomarkers can be demanding, requiring large-scale data linking across large numbers of variables.[33]

Genome sequencing

1.22 Gene sequences have been used for decades as biomarkers to inform diagnosis, disease prediction and clinical management, but recent advances in sequencing technologies are changing practice. Next Generation Sequencing (NGS) technologies now in use are claimed to double the capacity to produce sequence data every year, outpacing Moore's Law.[34] The cost of a sequencing run has also decreased dramatically. Although estimates vary, a whole human genome – the full sequence of more than three billion base pairs comprising the DNA molecules contained in a human cell nucleus – can currently be sequenced for approximately $5,000 and this cost is expected to continue to drop.[35]

1.23 These factors have enabled researchers to produce enormous amounts of genomic data from humans, animals, plants, insects, fossils, bacteria and other organisms. In humans, over 3,500 Mendelian or single-gene disorders have been identified and now a variety of approaches, such as whole genome and exome sequencing and genome-wide association studies, are used to target rare variants.[36] Sequencing is also helping clinicians to understand disease better, for example to classify cancer tumour genomes to determine whether a certain drug or treatment will be more or less effective, and is now being used directly in clinical treatment.[37] Cancer tumour sequencing has been shown to be capable of producing results sufficiently quickly to allow a clinician to adjust a patient's treatment plan as a result of the sequence data.[38]

---

[33] See: Academy of Medical Sciences (2013) *Realising the potential of stratified medicine*, available at: http://www.acmedsci.ac.uk/viewFile/51e915f9f09fb.pdf.

[34] See: Illumina (2013) *An introduction to next-generation sequencing technology*, available at: http://res.illumina.com/documents/products/illumina_sequencing_introduction.pdf. Moore's law, first proposed in 1965, refers to the observation that the number of transistors able to be fitted on to an integrated circuit will grow constantly at an exponential rate, approximately doubling every two years.

[35] National Human Genome Research Institute (2013) DNA Sequencing Costs, available at: http://www.genome.gov/sequencingcosts/; Illumina claimed in early 2014 that its HiSeq X Ten sequencing system could reduce the cost of sequencing to as low as $1000 per whole human genome. See: http://www.nature.com/news/is-the-1-000-genome-for-real-1.14530. However, there are some indications that the rate of advance is not steady; the cost of sequencing actually increased by 12per cent between April 2012 and October 2012, although it then fell again. See Hall N (2013) After the gold rush *Genome Biology* **14(5)**: 115, available at: http://www.biomedcentral.com/content/pdf/gb-2013-14-5-115.pdf. For a discussion of the implications of the fall in cost, see: Stein LD (2010) The case for cloud computing in genome informatics *Genome Biology* **11(5)**: 207, available at: http://genomebiology.com/2010/11/5/207.

[36] See: Brunham LR and Hayden MR (2013) Hunting human disease genes: lessons from the past, challenges for the future *Human Genetics* **132(6)**: 603-17, available at: http://link.springer.com/article/10.1007/s00439-013-1286-3; on approaches used, see: Lee S, Abecasis GR, Boehnke M and Lin X (2014) Rare-variant association analysis: study designs and statistical tests *The American Journal of Human Genetics* **95**(1): 5-23, available at: http://www.sciencedirect.com/science/article/pii/S0002929714002717.

[37] Sekar D and Thirugnanasambantham K, Hairul Islam VI, and Saravanan S (2014) sequencing approaches in cancer treatment *Cell Proliferation* **47(5)**: 391-5; Roychowdhury S and Chinnaiyan AM (2014) Translating Genomics for Precision Cancer Medicine *Annual Review of Genomics and Human Genetics* **15**: 395-415.

[38] Welch, JS, Westervelt P, Ding L, et al. (2011) Use of whole-genome sequencing to diagnose a cryptic fusion oncogene *Journal of the American Medical Association* **305(15)**: 1577-84, available at: http://jama.jamanetwork.com/article.aspx?articleid=897152.

Other 'omics'

1.24 In research applications (and, it is likely, in some clinical applications in the near future), in addition to the increase in genomic data, data relating to other groups of biological molecules are increasingly being linked to genomic and other health data. These include proteomics (the study of the entire set of proteins expressed in a cell or tissue at a certain time); transcriptomics (the study of the set of RNA transcripts that that indicate the pattern of gene expression at any given time); metabolomics (small molecules such as sugars and fats in a biological cell, tissue, organ or organism, which are the end products of cellular processes); microbiomics (the microorganisms that inhabit the gut, genitalia, skin, lungs, etc.); and epigenomics (the reversible modifications of a cell's DNA or associated molecules that affect gene expression without altering the DNA sequence). Whereas a person's germline genome is relatively stable throughout their life, the other 'omics' listed above vary over time, yearly, daily and hourly, potentially providing a new insight into the interaction of an individual with their environment. For example, epigenomic studies have shown that, while monozygotic twins have an almost identical epigenomic profile during their early years, by middle age their profiles have diverged, which is likely to be due to different environmental exposures and may result in differing susceptibilities to disease.[39]

1.25 Findings such as these are already motivating research into how a person's clinical data, genome and other 'omic' profiles together determine personalised responses for health and disease. However, they require the substantial capacity and skills in the accumulation, management and analysis of large amounts of biological data.[40]

1.26 With the increasing amount of 'omic' data becoming available to use, there are renewed calls to improve the linking of those data with phenotypic data – an individual's observable or detectable traits and characteristics – in order to understand and catalogue variations within a population and, in turn, to improve the diagnosis and stratification of diseases.[41] Deviations from what is considered normal can be used to make a diagnosis and indicate treatment.[42] However, the precision of this kind of diagnosis is limited as a spectrum of phenotypic differences may be associated with any disease or condition and, conversely, a single phenotype may be associated with more than one disease. Knowing a patient has cancer, for example, or even breast cancer, leaves a clinician with a considerable range of options for treatment. Genotyping of breast tumours suggests that breast cancer should now be regarded as at least 10 distinct diseases that respond differently to different therapies.[43] To define subclasses of disease with a common biological basis, and therefore to discover and select the most appropriate care, more detailed ('deep') phenotyping is required.[44]

---

[39] Haque FN, Gottesman II and Wong AHC (2009) Not really identical: Epigenetic differences in monozygotic twins and implications for twin studies in psychiatry *American Journal of Medical Genetics Part C: Seminars in Medical Genetics* **151C(2)**: 136-41, available at: http://onlinelibrary.wiley.com/doi/10.1002/ajmg.c.30206/full.

[40] Costa FF (2014) Big data in biomedicine *Drug Discovery Today* **19(4)**: 433-40.

[41] Kohane IS (2014) Deeper, longer phenotyping to accelerate the discovery of the genetic architectures of diseases *Genome Biology* **15(5)**: 115, available at: http://www.biomedcentral.com/content/pdf/gb4175.pdf.

[42] Robinson PN (2012) Deep phenotyping for precision medicine *Human mutation* **33**(5): 777-80, available at: http://onlinelibrary.wiley.com/doi/10.1002/humu.22080/full.

[43] Curtis C, Shah SP, Chin S-F *et al.* (2012) The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups *Nature* **486**(7403): 346-52, available at: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3440846/.

[44] Robinson PN (2012) Deep phenotyping for precision medicine *Human mutation* **33**(5): 777-80.

Critically, this approach requires the use of computational informatics systems to manage the data and to analyse it along with other data, including genome and clinical data.

Administrative data or metadata

1.27 Health services routinely collect 'transactional data' in the same way as other sectors: billing information, activity data, etc., which may or may not reveal who was treated or what sort of treatment they had. Data are often collected to analyse relative performance, to identify good and bad practice, for anti-fraud measures, as well as general management information collected for operational purposes. Some data types may be peculiar and essential to health systems such as the English NHS, for example those needed for the purposes of the Quality Outcomes Framework (QOF).[45]

1.28 Some of these data may be 'personal data' or 'sensitive personal data' for legal purposes. For example, administrative data such as clinic diaries may indicate no more than that a person saw a particular clinician at a certain place and time. As we have argued, the significance of this information will depend on how it is framed in relation to the informational context and the interests of the subject and those who might have access to it: such information may be highly sensitive if it relates to a visit to an STI or fertility clinic, for example. Equally, missing data, such as non-attendance at a clinic, can be highly informative.

1.29 Clinic attendance records are a special case of metadata. These are data that describe the contents of substantive data files or records and the circumstances of their creation and processing, for example, the size of data files, the time or location at which they were processed, the identity of the author or processor, and various technical features of the data. Records of communications are another example: call logs of who called whom and when may reveal a highly sensitive patient relationship. Clinical computer systems have, for many years, recorded metadata: about the identity of the person accessing the record, the changes they made, the time it was accessed or altered, the transmission of data between systems, etc., generating substantial audit trails.

1.30 Metadata can be useful both for organising substantive data and as research data in their own right. The distinction between data and metadata may be increasingly difficult to sustain as metadata and measurement or observation data can be equally informative depending on the context: the fact of a communication between a patient and a consultant in a known specialism can reveal information about a patient's health; confirmation of an individual's presence at a specific time in a geographical location, whether the record of a mobile phone use or a photograph, can be equally informative.[46]

[45] The Quality Outcomes Framework is a voluntary incentive scheme for GP practices in England, rewarding them for how well they care for patients, requiring various indicators to measure performance (see also chapter 6 below.)

[46] The surveillance activities of the US National Security Agency that were brought to light in 2013 made more use of metadata than content as information about who called whom, when, and where, is often critical in unravelling criminal conspiracies or focussing investigations (see chapter 2 below).

Data science

Proposition 2

Advances in data technology and data science provide more ways, and potentially more powerful ways, to collect, manage, combine, analyse and understand data in biological research and health care.

1.31 The need to process complex sets of biological and health data has led to the development of specialist techniques and related fields of expertise including bioinformatics and health informatics. These fields contain the knowledge, skills and tools that are applied to produce, manage and analyse data in order to generate information for particular purposes. They typically involve the use of computing, statistical and mathematical sciences.

Big data

Proposition 3

Advances in data science and technology have given rise to a new attitude towards data that sees it as a valuable resource that may be reused indefinitely in other contexts, linked, combined or analysed together with data from different sources. These uses have both practical advantages and limitations.

Proposition 4

The opportunities arising from data linking and re-use are presented as both novel and significant in the way in which they bring about new relationships between data and theory ('data-driven' and 'big data' approaches to research) and between data and practice (data modelling for policy and clinical decision making).

1.32 The term 'big data' initially characterised a problem that gave birth to novel solutions: that the size of datasets was outstripping the ability of typical database software to capture, store, manage and analyse them.[47] Although there is no settled consensus as to the definition of 'big data', computational informatics professionals, who are concerned with the analysis of big data, initially gathered around a characterisation in terms of three 'V's: volume, variety and velocity.[48] To work with massive datasets, in particular those created as by-products of the electronic mediation of so many social

[47] "The term 'big data' is meant to capture the opportunities and challenges facing all biomedical researchers in accessing, managing, analyzing, and integrating datasets of diverse data types [e.g., imaging, phenotypic, molecular (including various '–omics'), exposure, health, behavioral, and the many other types of biological and biomedical and behavioral data] that are increasingly larger, more diverse, and more complex, and that exceed the abilities of currently used approaches to manage and analyze effectively." See: US National Institutes for Health http://bd2k.nih.gov/about_bd2k.html#bigdata.

[48] For the '3 Vs' definition, see: http://strata.oreilly.com/2012/01/what-is-big-data.html. Other commentators have embellished this basic characterisation with an arbitrary number of further Vs: veracity, validity, volatility, etc.

interactions, from public administration to Internet shopping, web-searching, and social networking, requires cost effective, high speed computing and high volume storage, as well as scalable computational frameworks for analysing the data.[49] It has also required the development of a variety of computationally intensive tools in order to extract insights from data (such as visualisation – see the discussion of 'imaging' above).[50] This understanding of big data presents the extraction of value from datasets as being essentially a technical challenge, for example, to integrate and exploit different sources of data, such as images, voice records and numerical databases.

1.33 In current usage, 'big data' therefore refers less to the size of datasets involved ('big' being a relative term) than to the approach to extracting information from them using analytical techniques successively described under the rubrics of 'statistics', 'artificial intelligence', 'data mining', 'knowledge discovery in databases' (KDD), 'analytics' and, more recently, 'data science'.[51] The common feature of these approaches is the interrogation of datasets to discover non-obvious patterns and phenomena through finding correlations within the dataset. This may be done with or without a prior hypothesis about the causal relationships involved. Because of the complexity of the datasets the interrogation of the data for this purpose involves the application of an automated procedure, an algorithm.

1.34 Advances in the fields of computational informatics and statistical data mining that characterise 'big data' initiatives have at least two kinds of significant implication. First, the possibility of increasing the useful information that can be extracted from given resources of data, in particular by combining or linking datasets, may lead to a substantial reconfiguration of human and other resources, having consequential impacts (such as the training of more analysts or, for example, hiring more analysts and fewer doctors). Second, the use of these techniques within biomedicine suggests the emergence of a new attitude to data held by researchers and health systems, namely, as a resource amenable to a wide variety of uses and in pursuit of an unbounded range of purposes. In short, health records and research data can be re-conceived as a kind of raw material, to which the image of 'data mining' is perfectly apt, rather than as existing to serve a circumscribed purpose or range of purposes.

1.35 Use of the term 'big data' therefore calls attention less to a technical achievement (or challenge) than to a change in perspective that entails associated changes in behaviour. Commentators point to emergent properties of data at a large scale and the advantages of big data approaches in dealing with 'noisy' or messy datasets. Some even speak in ideological terms about the virtues of liberation from hypothesis-guided inquiry. [52]

---

[49] Examples of such frameworks are Google's proprietary MapReduce data processing model or the open source Apache Hadoop model.

[50] The basic principles of this technique are not new, although the quantities and complex relationships between data involved require the use of substantial computing power. Early examples include Florence Nightingale's "rose charts" of mortality in the Crimean War (which showed that the numbers of soldiers dying as a result of combat injuries were far outweighed by the number of those dying from disease) and John Snow's plot of the 1854 Soho cholera outbreak (which narrowed the source to an infected water pump in Broad Street (now Broadwick Street) and helped to replace miasmic ('bad air') theory with modern understanding of cholera as a water borne disease).

[51] For a history of computational and data science from 1960 to 2009, see: The Royal Society (2012) *Science as an open enterprise* (figure 2.1, at page 15), available at: https://royalsociety.org/policy/projects/science-public-enterprise/Report/.

[52] Mayer-Schönberger V and Cukier K (2013) *Big data: a revolution that will transform how we live, work and think* (London: John Murray), at page 14.

Data quality

1.36 The ways in which data science uses information, however, have both advantages and limitations. Given uncertainties in the accuracy of data, data from a larger number of data points can, in theory, increase the statistical power of the analysis. If the data collection includes the whole population of interest ('*n*=all'), errors due to sampling are reduced. However, if the data are subject to ascertainment bias, then more such data may only exaggerate that bias.[53] Data quality therefore remains an issue, particularly with 'found' data, 'data exhaust', or data originally collected for different purposes.

1.37 The intrinsic precision of data can vary with their origin: alternative kinds of equipment may be used, there may be simple differences in the training of observers, external factors (such as stress on a patient when measuring blood pressure) which may not be ascertainable, and errors in transcribing or converting data can be introduced.[54] Both technology and methodology play a role in the generation of data: there will typically be differences in the genome sequence given for the same individual depending on which company has supplied the sequencer and associated informatics.[55] Data quality can also be affected by use of different terminologies or criteria for the use of specific terms in different data entry contexts.[56] For example, a GP may use different criteria for the diagnosis of depression to those used by a psychiatrist. It is important to be aware of these factors as users may place undue faith in a computer record, failing to appreciate that computers often store data collected by humans. The uncritical analysis of computer records can magnify any phenomena related to the human input. The complex technologies and procedures used to produce biomedical data (such as imaging or genome sequencing) may involve processing according to preset methodologies to clean data and impute the value of missing data before they are rendered amenable to analysis.[57]

1.38 Obstacles to the re-use of data have been a lack of widespread knowledge about what data are actually collected and held, lack of standardisation, and lack of tools and infrastructure to link, curate and analyse datasets. Ethical constraints and, of course, the constraints of data protection law and existing standards of good practice also limit data reuse. Many of the technical obstacles may be surmountable, although some limitations, especially the quality of data at the point of collection, will be less tractable and more persistent. Other constraints may represent important safeguards.

---

[53] Ascertainment bias is a systematic distortion in measuring the true occurrence of a phenomenon that results from the way in which the data are collected with the result that all relevant instances were not equally likely to have been recorded.

[54] See, for example, Kohn LT, Corrigan JM, and Donaldson MS (Editors) (2000) *To err is human: building a safer health system* (Institute of Medicine Committee on Quality of Health Care in America) (Washington: National Academies Press), available at: http://www.nap.edu/openbook.php?record_id=9728.

[55] Patel RK and Jain M (2012) NGS QC: a toolkit for quality control of next generation sequencing data PLoS ONE 7(2): e30619, available at: http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0030619#pone-0030619-g003.

[56] Fortier I, Burton PR, Robson PJ *et al.* (2010) Quality, quantity and harmony: the DataSHaPER approach to integrating data across bioclinical studies *International Journal of Epidemiology* **39(5)**: 1383-93, available at: http://ije.oxfordjournals.org/content/39/5/1383.short.

[57] An fMRI brain scanning experiment measuring the brain 'activity' of a dead salmon offers a sobering demonstration that methodologies commonly used in imaging can produce high false positive rates. See: Bennett C M, Baird AA, Miller MB, and Wolford GL (2009) Neural correlates of interspecies perspective taking in the post-mortem Atlantic Salmon: An argument for multiple comparisons correction (poster presentation), available at: http://prefrontal.org/files/posters/Bennett-Salmon-2009.pdf.

Data initiatives

1.39 The main ethical concerns that arise as a result of the production, accumulation and use of data that we have described are less about the size or detail of any one dataset in isolation but rather about the potential for extraction of information either directly, by the application of analytical tools, or by first linking or combining datasets. A particular source of concerns, although certainly not the only one, is the capacity for the data to reveal significant information about particular individuals in a way that they are either unaware of or unable to control.

1.40 Throughout this report we will refer to the kinds of activities that are of interest to us as 'data initiatives'. These may be large – at the scale of a national biobank, health system or international research collaboration – or small – on the scale of a discrete research project to examine co-incidence of cases in two data registries. Large or small, the essential feature of a data initiative is that it involves one or both of the following practices:

- Data collected or produced in one context or for one purpose are *re-used* in another context or for another purpose. This translation between contexts or transformation of purposes may mean that the data take on a different meaning and significance. (An example might be where medical records are used by the police to solve a crime such that 'markers of health and functioning' may become 'indicators of guilt'.) This may be described as 're-use', 'secondary use' or 'repurposing' of data.

- Data from one source are *linked* with data from a different source or many different sources. This may be in order to facilitate a purpose for which one of the datasets was produced, or for some further purpose, possibly unrelated to any of them. This might involve combining the datasets for the purpose of a single analysis or creating some durable (permanent or temporary) link between them. (An example might be where data from a disease registry are linked to data about the location of discharges of environmental pollutants to examine or monitor any link between them.)

1.41 There are several reasons to re-use data rather than collect it afresh. First and foremost, reusing data is efficient and allows the same data to do more work. Some of this work may be closely connected with the original purposes, such as allowing research results to be validated or allowing data from across research projects to be collated in meta-studies. Re-using data avoids the cost, inconvenience, and possibly the annoyance involved in having to approach people repeatedly to gather much the same data. Thus data collected in a clinical consultation may be used for health service planning and medical research, but is also potentially of interest as evidence for social policy making, actuarial purposes (e.g. insurance pricing), market research, product development, marketing, and many other purposes. It is not clear how often or how widely data, particularly non-standardised data, may be re-used as time and technology move on, generating novel sorts of questions and requiring new kinds of measurements (although if new data are needed there may nevertheless be benefit in linking them to earlier data). However, these limitations may be offset by increasingly sophisticated algorithms that allow data in existing datasets to be correlated and 'mined' for new insights. As a result, the limits of the potential utility of any given dataset are increasingly unforeseeable.

Conclusion

Proposition 5

Data collected in biomedical research and health care are not intrinsically more or less 'sensitive' than other data relating to individuals. However, they can be extremely 'sensitive' depending on the context in which they are used and how they are related to other information. The use of data in different contexts and for different purposes may influence how people are treated by others, including by public authorities, in ethically significant ways.

1.42 The description of any data as 'biological' or 'health' data is increasingly misleading. From the perspective of data science whether they are 'biological' or 'health' data depends on the use to which they are put as much as the source from which they were obtained or the purpose for which they were originally collected. Biomarker data may be used to inform someone's treatment, but they may also be used for the development of therapies, the allocation of costs, or the planning of services, moving variously between health care, research, financial and administrative contexts.

1.43 Nevertheless, data about individual biology and health are considered by many people to be somewhat more 'sensitive' than much other day-to-day information. Partly, this may be to do with social norms, and expectations about medical confidentiality and the importance attached to certain kinds of records: people may feel very differently about the use of data from their medical records than they might about the use of the same data taken from a research assessment, for example. Partly, this may have to do with the fact that the data may reveal stigmatising information, such as sexual and mental health states, though other personal data can be equally sensitive, depending on the context and circumstances.[58] The analysis, linking and use of certain kinds of data can also have critical implications for life and well-being.[59] This point can work towards both the need to protect confidentiality as well as the need to use the data wisely to improve safety and quality of health care.

1.44 The problem of pinning down data as 'health' data, or as 'sensitive' or 'personal' data is compounded by the fact that the relevant literatures are vexed by imprecise, inconsistent and sometimes conflicting terminology. It is a reflection of the novel and unsettled problems raised by the possibilities of data science that there is no universally accepted lexicon, although the lack of one is frequently bemoaned. That none exists may also bear witness to political tussles over the values embedded in such terms as 'data sharing' (which has connotations of beneficence and mutuality) and 'anonymisation' (which promises obscurity). [60]

[58] See, for example, Nagel T (2002) *Concealment and exposure* (New York: Oxford University Press), especially the title essay.

[59] Discussions on the draft European Union General Data Protection Regulation (to replace the existing EU Data Protection Directive) have sought to introduce 'genetic data' as a special class of data over and above mere health data, because of its potentially identifying and predictive nature. It is a consequence of the approach that we will develop in this report that such a categorisation misses the point. For the draft Regulation, see: http://ec.europa.eu/justice/data-protection/document/review2012/com_2012_11_en.pdf.

[60] Recognising this, the recent UK Information Governance Review recommended the adoption and use of a single set of terms and definitions relating to information governance that both staff and the public can understand. See: The Caldicott

1.45 Likewise, the developments in data use that have led to this report are of a general nature, and are not limited to the biological sciences and biomedicine, but are diffused across public administration, the provision of commercial and financial services, and other fields. Nevertheless, as a bioethics Council our principal interest is in the ethical use of data in relation to biology and medicine. Therefore, while conscious of this wider environment, in this report we shall nevertheless focus on data initiatives within medicine (or health care more broadly) and research in the biological, biomedical and clinical sciences.

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. In the absence of a satisfactory consensus, the way in which we use key terms in this report is described in the text and summarised in the Glossary.