The collection, linking and use of data in biomedical research and health care:
ethical issues

NUFFIELD COUNCIL ON BIOETHICS

The collection, linking and use of data in biomedical research and health care: ethical issues

Nuffield Council on Bioethics

Professor Jonathan Montgomery (Chair)

Professor Simon Caney

Professor Bobbie Farsides*

Professor Peter Furness

Professor Ann Gallagher

Professor Robin Gill

Dr Andy Greenfield

Professor Erica Haimes

Professor Julian Hughes

Sir Roland Jackson

Professor Graeme Laurie

David K Lawrence

Professor Tim Lewens

Professor Ottoline Leyser (Deputy Chair)

Professor Anneke Lucassen

Professor Martin Richards**

Dr Tom Shakespeare

Dr Geoff Watts

Professor Robin A Weiss

Professor Heather Widdows

Adam Wishart

Dr Paquita de Zulueta

* co-opted member of the Council while chairing the Working Party on Children and Clinical Research

** co-opted member of the Council while chairing the Working Party on Health and Biological Data: Ethical Issues

Secretariat

Hugh Whittall (Director)

Peter Mills

Katharine Wright

Catherine Joynson

Sarah Walker-Robson

Seil Collins

Ranveig Svenning Berg

Tom Finnegan (until January 2014)

Kate Harvey

Laura Medhurst

Carol Perkins

Bettina Schmietow (from March 2014)

Anna Wilkinson

The terms of reference of the Council are:

1. to identify and define ethical questions raised by recent advances in biological and medical research in order to respond to, and to anticipate, public concern;

2. to make arrangements for examining and reporting on such questions with a view to promoting public understanding and discussion; this may lead, where needed, to the formulation of new guidelines by the appropriate regulatory or other body;

3. in the light of the outcome of its work, to publish reports; and to make representations, as the Council may judge appropriate.

The Nuffield Council on Bioethics is funded jointly by
the Medical Research Council, the Nuffield Foundation, and the Wellcome Trust

Acknowledgments

The Council would like to thank the Working Party for the generous contribution of their time, knowledge and expertise to this project.

We should also like to express our thanks to the large number of people, individually and from within institutions and other organisations, who responded to the open consultation that informed the development of the Working Party's understanding of and approach to the issues addressed in this report. As well as these, we are grateful to the knowledgeable contributors who attended our four 'fact-finding' meetings. A list of these people may be found at Appendix 2.

The Working Party commissioned two independent reports to assist their deliberations, which are published on the Council's website alongside this report. The first, commissioned jointly with the Expert Advisory Group on Data Access (EAGDA), explored evidence of the harms associated with misuse of data and the second the relationship between public and private sectors in the use of data in biomedical innovation. We would like to thank the authors of these reports, Professor Graeme Laurie, Ms Leslie Stevens, Dr Kerina H Jones and Dr Christine Dobbs (for the first report) and Professor Paul Martin and Dr Gregory Hollin (for the second). We should also like to thank the members and secretariat of EAGDA for their assistance in commissioning the first independent report and for giving us the opportunity to present and discuss with them the direction of our research at a very early stage.

We are also greatly appreciative of the value added to the report by the dozen external reviewers, representing a range of professional backgrounds and academic disciplines, who are listed in Appendix 2. In addition to these, members of the Council, Secretariat and Working Party also benefitted from discussions and correspondence with a number of other individuals at different stages of the project, including Professor Dame Fiona Caldicott, Dr Alan Hassey, Dr William Lowrance, Katherine Littler, Dr Natalie Banner, Dr Alfred Z Spector, Dr John Parkinson, Dr Mark J Taylor, Tim Kelsey, Kingsley Manning, Dr Stephen Pavis, Laura Riley and Adrienne Hunt. We also extend our thanks to Dr Anna Wilkinson of the Nuffield Council Secretariat for her careful proofreading and to Clare Wenham and Tom Burton for additional research and assistance.

Finally, we should like to acknowledge the contribution of two members of the original Working Party who, for supervening reasons, were unable to see the report through to completion: Lynn Molloy and Peter Singleton. While the Members of the Working Party listed below take responsibility for the final form of the report, we are very grateful for the valuable contribution of those colleagues to the deliberations.

Foreword

The germ of this report might, perhaps, be found in discussions leading to the Council's annual 'forward look' meeting in May 2011. Some time in 2010, with typical foresight and perspicacity, the Council identified a bundle of emerging issues to do with the application of information technology and data science to health records and biological information that had the potential to generate substantial public hopes and concerns. The discussion in May 2011 persuaded the Council that significant and under-explored ethical questions were at stake and that a Council report could offer a useful and timely contribution to debate.

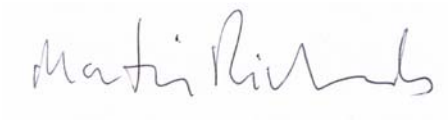
I doubt that many of those who, like me, attended the Council's initial scoping workshop in February 2012 would have predicted the shape this report would take or could have imagined just how wide-ranging, complex and passionate the debate would be. Recalling that the project began life under the rubric 'Genomics, Health Records, Database Linkage and Privacy' may shed light on the some of the preoccupations underlying the report that resurfaced repeatedly as it emerged. But when the Working Party set to work in March 2013, with the tide of enthusiasm for 'big data' reaching a high water mark, it seemed that the issues we had to consider were much more general than the use of particular 'kinds' of data or the linking of particular 'kinds' of records. This was confirmed when we began to gather evidence at a number of fact-finding meetings. It was further confirmed by the extraordinary richness of responses to our open consultation.

Our task was complicated by the fact that we built on shifting sands. The negotiations around the new European Data Protection Regulation, the well-publicised vicissitudes of the 'care.data' programme and the expectations built around the '100,000 Genomes' project: all these unfolded during our deliberations and remain largely unresolved as we go to press. Although the report addresses these and other examples explicitly, and will no doubt be read differently in the light of further developments, the reader should remember that our principal concerns are moral questions, not contingent questions of law and practice. We have tried to produce a report that adds insight to these but is not diminished by the fact that they may unfold in one way rather than another. We intend and expect, in other words, that readers will continue to find value in our report as initiatives, methodologies, and indeed laws, come and go.

Some of the matters that we have had to address – the challenging complexity of health service and research practices, the technical sophistication of data science, for example – were new to many of us and we relied heavily on our colleagues' expertise to bring these unfamiliar concepts into our common grasp. Nevertheless, we could not have reached our objective without the contribution of every member of the Working Party to our deliberations, since each brought a perspective that at times enriched, at others challenged, and often enriched through challenging, the understanding that developed among us. I should here express my gratitude to those colleagues for their hard work and commitment to our objective, and, though we have had some robust debates, for demonstrating a belief in the centripetal potential to produce something of value over the centrifugal force of our differences. Although, for purely contingent reasons, not all of the original Working Party were able to see the project to completion, the imprint of earlier contributions remains on the final document. As a group we are deeply indebted to the Nuffield Council and especially to the subgroup of Council members who reviewed and commented on our output at various stages of the project for their intellectual engagement, wise counsel and moral support throughout. Beyond these we have benefitted considerably from the contributions of numerous others, among them those who responded to our open consultation. Though we

have not engaged with specific responses explicitly in the report, these have been immensely valuable in informing our thinking, providing a common wealth of knowledge and questions for a diverse working party such as ours to draw on, to discuss and to pursue. We benefitted greatly from the contributions of those who attended our early fact finding meetings, who gave generously of their considerable knowledge and patiently addressed our questions, and of the authors of two independent commissioned reports. Our initial drafts were improved immeasurably thanks to our external reviewers who saw an early draft of this document and commented more in the spirit of collaborators than critics. Though we have not been able to reflect the totality of their advice in a single document, it has enriched and informed the report beyond anything that we might otherwise have achieved; any shortcomings that remain are, of course, our responsibility rather than theirs.

Finally, I should thank the members of the Nuffield Council Secretariat who supported our work. Peter Mills, not only organised the whole project, but also undertook the lengthy and arduous task of drafting and editing the report. Without his expertise, patience, and his dedication beyond the call of duty, our work would not have reached a successful conclusion. Peter was ably supported by our research officers, Tom Finnegan (to the end of 2013) and Bettina Schmietow (from March 2014) who also deserve our gratitude.



Martin Richards

Chair of the Working Party

Members of the Working Party

Professor Martin Richards
Emeritus Professor of Family Research, University of Cambridge (Chair)

Professor Ross Anderson
Professor of Security Engineering, University of Cambridge

Stephen Hinde
Former Head of Information Governance and Group Caldicott Guardian for the Bupa Group (retired December 2013)

Professor Jane Kaye
Professor of Health, Law and Policy, Nuffield Department of Population Health and Director of the Centre for Law, Health and Emerging Technologies at the University of Oxford

Professor Anneke Lucassen
Council Member, Consultant Clinical Geneticist and Professor of Clinical Genetics at the University of Southampton

Professor Paul Matthews
Edmond and Lily Safra Professor of Translational Neuroscience and Therapeutics and Head, Division of Brain Sciences, Department of Medicine, Imperial College London

Professor Michael Parker
Professor of Bioethics and Director of the Ethox Centre at the University of Oxford

Margaret Shotter
Member of advisory panels for medicines and research, including UK Biobank EGC

Dr Geoff Watts
Council Member, science writer and broadcaster

Dr Susan Wallace
Lecturer of Population and Public Health Sciences, University of Leicester

John Wise
Executive Director, Pistoia Alliance and Programme Coordinator, PRISME Forum

A full list of the relevant activities and interests of members of the Working Party can be found at Appendix 3.

Terms of reference

1. To identify developments in the collection, linking, use and exploitation of human biological and health data arising from advances in knowledge, technology, organisation and governance.

2. To identify, define and examine significant ethical questions raised by these developments.

3. To consider, in particular, the implications (including possible benefits and possible harms) of these developments, having regard to:

   a the meaning, importance and the practical exercise of privacy, autonomy, anonymity, identity, altruism, solidarity and citizenship;

   b ownership, control and interest in data, and the exercise of these via measures such as consent, authorisation, donation and sale;

   c the interaction between the interests of the individual data subject, other individuals, the public interest and commercial interests, particularly in cases in which these are not aligned;

   d the moral and legal duties of those involved in the collection, linking, use and exploitation of data;

   e the appeal to autonomy, rights, human dignity and common interest as justifications for processing data in different contexts.

4. To report on these matters and to make recommendations, as appropriate, for research, information governance and policy.

Table of Contents

Executive summary

Key findings and recommendations

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. Advances in information technology and data science provide more ways, and more powerful ways, to collect, manage, combine, analyse and derive insight from these data. The result is that data are now seen as a valuable resource with an indefinite range of potential uses.

2. There is a public interest in the responsible use of data to support advances in scientific knowledge, innovative treatments and improvements in health services. However, there is also a public interest in protecting the privacy of individuals: privacy is fundamentally important to individuals (and groups) in the establishment and maintenance of their identity, their relationships and their sense of personal well-being. In biomedical research and health care data initiatives, which link and re-use data, public and private interests are entangled in complex ways. Such data initiatives must address the following question:

   - what is the set of morally reasonable expectations about the use of data and what conditions are required to give sufficient confidence that those expectations will be satisfied?

3. Compliance with the law cannot guarantee that a use of data is morally acceptable. Faced with contemporary data science and the richness of the data environment, protection of privacy cannot reliably be secured merely by anonymisation of data or by using data in accordance with the consent from ‘data subjects’. Effective governance of the use of data is indispensable.

4. A set of morally reasonable expectations about the governance and use of data should be determined in accordance with four principles:

   - the principle of respect for persons
   - the principle of respect for established human rights
   - the principle of participation of those with morally relevant interests
   - the principle of accounting for decisions

5. Taking into account the current state of knowledge and practice, and the likely direction and pace of developments, and considering a number of specific data initiatives in biomedical research and health care, we recommend:

   - support for needed research into the potential harms associated with abuse of biological and health data, as well as the benefits of responsible data use
   - comprehensive mapping of UK health and research data use and the norms relevant to it
   - mandatory reporting of privacy breaches affecting individuals to the individuals affected
   - review of anti-blagging measures to protect data in health care systems and promulgation of best practice
   - criminal penalties, including imprisonment, for the deliberate misuse of data

- a public statement of expectations about who may be given access to health data and for what purposes, for each data initiative
- publication of all Health and Social Care Information Centre data sharing agreements and results of independent audits of compliance
- maintenance of an auditable record of all people given access to data held by the Health and Social Care Information Centre, that can be given to affected individuals
- a review of the appropriateness of public-private partnerships to secure public benefit from the research use of National Health Service records
- increased subject participation in design and governance of research projects
- wider use of explicit and flexible ethics and governance frameworks for research projects, including for international collaborative research
- restriction of access to research data to researchers (including international collaborators) who are subject to institutional oversight and effective sanction
- publication of policies on the use of cloud services by national bodies
- ethical and scientific appraisal to maximise the contribution of participant-led research to science while ensuring adequate protection of participants
- collaboration among all members of the research community to promote a more robust, explicit and candid foundation for extending access to data for research in the public interest.

Summary of the report

6. This report takes as its starting point the massive accumulation of data in biomedical research and health care, and the increasing power of data science to extract value by linking and re-using that data, for example in further health or population research. It examines the scientific, policy and economic drivers to exploit these opportunities, and the concerns and potential risks associated with doing so. The faltering ability of conventional information governance measures to keep pace with these developments is identified as a significant problem. The report therefore poses and addresses the following question:

    - how can we define a set of morally reasonable expectations about the use of data in any given data initiative and what conditions are required to give sufficient confidence that those expectations will be satisfied?

7. The report sets out a number of general recommendations, including four guiding principles for ethical design and governance of data initiatives. These help to identify specific examples of existing good practice and to make recommendations for improved practice in the use of data in the fields of health care (re-use of NHS records, clinical research, etc.) and population research (biobanks, epidemiological studies, etc.).

Data (chapter 1)

8. 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

9. 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.

10. 'Data initiatives' involve the re-use of 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.

Opportunities and threats (chapter 2)

11. The combination of advances in information technologies and in data science have generated considerable opportunities for economic activity. 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.

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

    - increasing efficiency and transforming service delivery through better informed decisions about resource allocation and greater involvement of patients through e-health care
    - generating improvements in medical treatment by building a stronger evidence base for prediction, prevention and treatment, and by using data to personalise treatment and care, linking phenotype and genotype data with lifestyle, environmental and social data
    - generating economic growth from the life sciences by using existing data in health systems with increased technological capacity and skills to invigorate the pharmaceutical and biotechnology industries.

13. To achieve these outcomes a number of policy orientations have been set in the UK and elsewhere, such as:

    - increasing IT intensity and introducing new infrastructure in health systems
    - establishing partnerships between the public and private sectors to promote resource exploitation and innovation
    - centralising data resources to facilitate analysis of linked data
    - promoting 'open data' and 'data sharing' to encourage the widest possible use of resources
    - promoting 'big data' and investing in the knowledge economy to foster development of new tools, methodologies and infrastructures.

14. However, the pursuit of opportunities must take account of the need to manage a number of threats to welfare. These threats take a number of forms, for example:

    - misuse of data leading to harms to individuals and institutions (ranging from detriment to health, loss of privacy, financial loss, reputational damage, stigmatisation and psychological distress)
    - discriminatory treatment, ranging from targeted advertising to differential pricing that compounds social disadvantage, to discrimination in insurance and employment
    - state surveillance of citizens, particularly in the light of revelations about the US National Security Agency, which is greatly facilitated by large databases and linked information systems

15. Independent research commissioned to inform our work found that 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 and obtaining redress is difficult.

Recommendations

1 Relevant bodies, including public and private research funders and UK health departments, should 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.

2 The Independent Information Governance Oversight Panel and the Health Research Authority should supervise, respectively, the maintenance of comprehensive maps of UK health and research data flows, and they should 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.

3 The Government should 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.

4 The Health and Social Care Information Centre should maintain prospective assessments to inform the most effective methods for preventing the inadvertent or fraudulent accessing of personal healthcare data by unauthorised individuals.

5 The UK government should 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.

Moral values and interests (chapter 3)

16. The concept of *privacy* and the distinction between public and private have evolved throughout history. Individual privacy is important in the formation of identity and the maintenance of personhood but privacy can also be attributed to families, and wider groups. Norms of information disclosure are important in the formation and maintenance of identities and relationships between individuals and groups, and different norms apply to different relationships.

17. An important class of privacy norms is enshrined in the rules and practices of *confidentiality*. These may exist as informal conventions among individuals but may be more formalised in professional relationships, contracts and laws. Medical confidentiality allows information sharing that might otherwise infringe privacy norms to take place for specific professional purposes.

18. At the same time *consent* provides a mechanism to make controlled exceptions to an existing privacy norm for specific purposes (for example, to permit a medical diagnosis or referral) without abolishing the underlying norm. However, consent does not itself ensure that all of the interests of the person giving consent are protected nor does it set aside the moral duty of care owed to that person by others who are given access to the information. On its own, consent is neither necessary nor sufficient for ethical extensions of data access.

19. While individuals have privacy interests in the use of data, they also share group interests in the wider use of data for health research. This broader *public interest*, which consists in securing objectives that are valued by society, may come into conflict with individual privacy. But the relationship between privacy and public interest in data is not simply one of opposition. The two are mutually implicated in each other: there are private interests in the achievement of common goals and a public interest in the protection of privacy that encourages cooperation. This complex relationship leads to a need to reconcile the articulation of the private within the public and the public within the private. A fundamental moral question facing data initiatives is therefore:

    - how may we define a set of morally reasonable expectations about how data will be used in a data initiative, giving proper attention to the morally relevant interests at stake?

20. Three sorts of considerations will be relevant to formulating an answer:

    - the norms of privacy and disclosure applicable among those who participate in a data initiative
    - the ways in which individual freedoms are respected, for example, the freedom to modify these norms by consent
    - the form of governance that will give acceptable assurance that the expectations will be met

Law, governance and security (chapter 4)

21. A number of overlapping legal measures exist to protect privacy, principally: formal privacy rights, which guarantee freedom from interference, albeit that they may be qualified by certain public interest considerations; rules of data protection, which control the ‘processing’ of various kinds of ‘personal data’; and duties of confidentiality, which protect against unauthorised or unreasonable breaches of confidence.

22. A number of technical measures may be applied to prevent the identification of individual subjects and reduce the risk of privacy infringements:

    - aggregation of data makes it harder to distinguish individual cases, although it is not wholly secure in the face of modern statistical techniques; it also makes further linking of data difficult

- anonymisation by the removal of identifiers also makes individuals difficult to re-identify, although re-identification may still be possible depending on what other data or information are available
- pseudonymisation, the replacement of identifiers with a code, enables linking of data where the correspondence between the code and the case is known, although data may still be vulnerable to inferential re-identification

23. While de-identification measures may help to protect privacy, re-identification may not be impossible and the risk of re-identification is both difficult to quantify and may become greater over time. To protect the privacy of data subjects, de-identification should therefore be combined with further controls on the access to and uses of data.

24. A standard control is to limit access to data in accordance with the consent of the 'data subject'. Broad consent allows data subjects to set certain parameters for the use of the data that are morally salient for them but the often far-reaching implications of data use may be obscure, and the scope of consent given when data are collected may become unclear, particularly in changing circumstances and in relation to novel uses of data. This is especially likely when data are held for long periods of time. Continuing involvement of subjects through 'dynamic' forms of consent can address this but is potentially demanding.

25. While seeking consent respects rights that individuals may have to make decisions about matters that may affect their interests, it cannot protect them from potentially harmful consequences of data use. Merely acting in accordance with consent cannot excuse data users from their moral duties towards data subjects, indeed towards all those who have a morally relevant interest the data initiative, whether they are people from whom the data were initially collected or others who stand to be affected by their use.

26. As neither anonymisation nor compliance with consent offer sufficient privacy protections in data initiatives, additional controls on the use of data – on who is permitted to access them, for what purposes, and how they must conduct themselves – are therefore required. These have administrative aspects (e.g. data access committees and agreements) and technical aspects (e.g. safe havens).

27. The need to meet two contradictory requirements at the same time places data initiatives in a double bind. In other words, they are required:

    - to generate, use and extend access to data (because doing so is expected to advance research and make public services more efficient); and, at the same time
    - to protect privacy as this is a similarly strong imperative, and a requirement of human rights law (and the more access is extended the greater the risks of abuse).

28. In order to meet this challenge the use of measures such as anonymisation, the mechanisms for respecting individuals' rights and interests (such as consent procedures), and the forms of governance that guide the conduct of professionals in the public interest need to be established coherently. These measures should be determined in relation to the underlying moral norms and values, and in relation to an agreed understanding of the hazards, benefits and uncertainties of data use in the context of particular data initiatives.

Ethical governance of data initiatives (chapter 5)

29. Data initiatives are practical activities that involve a number of actors (who might be individuals, groups, institutions, etc.) some of whom stand to benefit or lose from the outcomes. Tensions and potential conflicts between values and interests can arise at the level of the individual, of professions or of the public. The ethical formation of a data initiative is a matter of reconciling these values and interests in a coherent set of morally reasonable expectations.

30. A morally reasonable set of expectations should embody four principles.

Ethical principles for data initiatives

The use of data in biomedical research and health care should be in accordance with a publicly statable set of morally reasonable expectations and subject to appropriate governance.

- **The set of expectations about how data will be used in a data initiative should be grounded in the principle of respect for persons.** This includes recognition of a person's profound moral interest in controlling others' access to and disclosure of information relating to them held in circumstances they regard as confidential.
- **The set of expectations about how data will be used in a data initiative should be determined with regard to established human rights.** This will include limitations on the power of states and others to interfere with the privacy of individual citizens in the public interest (including to protect the interests of others).
- **The set of expectations about how data will be used (or re-used) in a data initiative, and the appropriate measures and procedures for ensuring that those expectations are met, should be determined with the participation of people with morally relevant interests.** This participation should involve giving and receiving public account of the reasons for establishing, conducting and participating in the initiative in a form that is accepted as reasonable by all. Where it is not feasible to engage all those with relevant interests – which will often be the case in practice – the full range of values and interests should be fairly represented.
- **A data initiative should be subject to effective systems of governance and accountability that are themselves morally justified.** This should include both structures of accountability that invoke legitimate judicial and political authority, and social accountability arising from engagement of people in a society. Maintaining effective accountability must include effective measures for communicating expectations and failures of governance, execution and control to people affected and to the society more widely.

31. The principle of respect for persons does not mean that individuals' interests may never be overridden, but that they may only be overridden where there is a legitimate reason to do so. As a principle of design of data initiatives, the principle of respect for human rights seeks to avoid potential rights conflicts and violations rather than leaving them to be dealt with retrospectively through judicial processes. The participation of people with morally relevant interests in the design and governance of data initiatives allows the identification of relevant privacy norms and the development of governance measures (such as design of consent and authorisation procedures) in relation to these norms; it allows preferences and interests to be expressed and transformed through practical

reasoning, and account to be given of how these interests are respected in decision making, helping to foster trust and cooperation. The principle of accounting for decisions ensures that expectations, as well as failures of governance and control, are communicated to people affected and to others more widely. It also ensures that data initiatives remain in touch with changing social norms.

Data initiatives in health systems (chapter 6)

32. Health-care IT systems were originally introduced to facilitate basic administrative tasks, such as managing patient records, issuing repeat prescriptions and tracking patients through different encounters with health care professionals. However, they developed to provide business intelligence for service improvement and support for observational research. These come together in the concept of a 'learning health system' which is seen as an inevitable response to increasing pressures on health services and the demand for new treatments.

33. These functions, together with the need to manage reimbursement and the appetite for data to inform health policy, combined to push data systems in the English NHS towards a centralised approach with electronic health records at its heart. The central collection of health care data in England is now managed by the Health and Social Care Information Centre (HSCIC). The HSCIC's model involves holding linked data centrally, publishing aggregate data and disclosing certain individual-level 'pseudonymised' data in controlled conditions. Debate around the 'care.data' programme to extract primary care data to the HSCIC focused attention on the assumptions made about the relationship between privacy norms relevant to NHS patients and the legal norms under which HSCIC operates. It highlighted the absence of governance arrangements to negotiate this difference, and raised questions about how the rights of individuals were respected. Failure to attend to these prospectively led to ad hoc policy changes and a damaging loss of public and professional trust.

34. The Scottish Informatics Programme to develop a research platform for health data involved initial public consultation to identify relevant social norms. On the basis of this it developed a model whereby data linkages are performed for specific purposes using a trusted third party, with analysis carried out on linked, pseudonymised datasets in a controlled environment. Data are not warehoused centrally and no individual-level data may be released from the safe haven. Direct access to the data is not available to commercial researchers. In addition to the role of the data custodians, authorisation for data use is provided by a risk-based, proportionate governance system that takes into account both privacy risk and public benefit, and refers to an explicit, potentially revisable, statement of guiding principles and best practices. The model demonstrates a number of features of good practice in relation both to its development and form that are consistent with the principles set out in this report.

35. The 100,000 Genomes project involves linking data from genome sequencing with individuals' NHS records for the investigation of cancers, rare diseases and some infectious diseases. Broad consent is obtained from individual subject participants (who do not expect direct therapeutic benefit). In operation the project will have an ethics committee and an explicit data access policy. Authorised researchers from all sectors may access a firewall-protected, pseudonymised dataset administered by a Government-owned company, Genomics England Ltd. No individual-level data may be released from this environment. The claimed public interest lies explicitly in securing economic as well as scientific and therapeutic benefits, by stimulating the commercial sector.

Recommendations

6 An independent, broadly representative group of participants should be convened to develop a public statement about how data held by the HSCIC should be used, to complement the Code of Practice on confidential information. This should clearly set out and justify what can reasonably be expected by those from whom data originate and be able to demonstrate that these expectations have been developed with the participation of people who have morally legitimate interests in the data held by the HSCIC, including data subjects, clinical professionals and public servants.

7 In addition to implementing the recommendations of Sir Nick Partridge's review, all Data Sharing Agreements held by the HSCIC should be published, along with the findings of a periodic independent audit of compliance those agreements.

8 HSCIC Data Sharing Agreements should include a requirement to maintain an auditable record of all individuals or other legal entities who have been given access to the data and of the purposes to which they are to be put; this should be made available to all data subjects or relevant authorities in a timely fashion on request.

9 Broader public consideration should be given to whether Genomics England Ltd provides the most appropriate model for the ethical use of genomic information generated in health services for public benefit before it becomes the *de facto* infrastructure for future projects.

Population research (chapter 7)

36. Biobanks are major resources of tissues and data that may be used for a variety of research purposes. They support the trend in life sciences research towards broader collaboration, larger datasets and greater varieties of data. UK Biobank is a large population biobank established to support the investigation of a range of common diseases occurring in the UK. Subject participants give broad consent to the use of data collected at recruitment, from their medical records and through supplementary data collections (e.g. the imaging study). The resource has a published Ethics and Governance Framework, compliance with which is overseen by an independent Ethics and Governance Council. Its design was foreshadowed by meaningful public engagement but the intention to establish a participant panel was not followed through. Subject participants' influence over the use of the data is limited to possible withdrawal from the resource on the basis of information published or communicated by the organisation. Pseudonymised data are released to researchers from recognised institutions for research that meets public interest criteria, and results are returned and published to support further research. There may be some need to review the set of expectations underlying the operation of UK Biobank in the light of changing circumstances (the evolving data environment, revaluation of data, etc.) One such area is feedback of information to subject participants; another is expectations about commercial access to the resource. Renewed engagement with public and participants is desirable in this context.

37. The UK10K Rare Genetic Variants in Health and Disease project was established to use existing research samples to characterise the genetic bases of rare diseases through comparison of genotypes of affected individuals with deeply phenotyped groups from cohort studies. This confronts the problem of controlled disclosure of highly specific individual-level data among different groups of researchers working on distinct studies. It is achieved through a common ethical framework of policies and guidelines developed with some patient interests represented. It places considerable reliance on ensuring that

only appropriately qualified researchers, bound by enforceable agreements, have access to data and on the role of principal investigators as data custodians.

38. International collaborative research initiatives such as the International Cancer Genome Consortium and the Psychiatric Genomics Consortium also rely on a common ethical framework operating across different research contexts. These need to accommodate differing local practices (e.g. different policies regarding the return of findings to subject participants, different standards of security, varying institutional sanctions) and tackle complex consent issues to do with re-use and international transfer of data. The use of cloud-based storage and processing services is becoming increasingly important but it raises issues such as third party access (for example, by security services).

39. The wide availability of information technology and social networking platforms have facilitated participant-led research, allowing individuals to group together to address research questions of interest to them and complement institutional research. The norms and social dynamics of patient-led research are different to more formal institutional research owing to the online medium, self-organising dynamics and the absence of formal review or oversight structures. They present different challenges of ensuring the protection of individual interests, of integration with institutional research, and of translation of findings into clinical products and practice.

40. Good practice is emerging in many population research initiatives but more needs to be done to protect the privacy interests of subject participants in order to secure the trust of current and future generations.

Recommendations

10 Appropriate mechanisms should be put in place to allow governance arrangements to evolve during the life of an initiative, through deliberation with morally relevant stakeholders including participants, the public, funders and the research community. Arrangements may include, e.g., representation of relevant stakeholder groups in the governance of the biobank; regular review of a public ethics and governance framework document legitimated through deliberation with interested parties that sets out the relationships of a biobank with participants, the research community, individual researchers, funders and the wider society. This may serve as an instrument to maintain alignment of the public interest in research with the privacy and other interests of the participants. Governance arrangements should, among other matters, outline policies for maintaining data security, the feedback of health-related findings to participants and for research access to the resource. In large scale and complex initiatives detailed diagrams of data flows should be available to support good governance. The responsibility to ensure appropriate governance arrangements are in place rests with funders.

11 Where broad consent is sought for the use of data additional, adaptive safeguards should be in place to secure the interests of participants over the life of a project. A possible model is provided by a publicly articulated, 'living' ethics and governance framework that reflects the expectations of participants and is subject to review and revision through mechanisms that involve representatives of the full range of interests of participants in the initiative.

12 Researchers should operate demonstrably within a local governance framework able to maintain reasonable surveillance in order to identify inappropriate data use and administer sanctions for misuse. Researchers should be members of a recognised research environment with appropriate arrangements in place to ensure their research meets ethical standards. They should provide undertakings regarding the confidential and secure use of data and that they will refrain from any attempt to

identify participants from whom data may have been derived.

13 All international collaborative data research initiatives should operate within an explicit, public ethics and governance framework that has agreement from the initiative's constituent partners. International collaborators should be able to demonstrate that they can fulfil recommendation 12 by applying equivalently strong governance standards (using legal and other mechanisms available in their national jurisdiction).

14 All partners in international collaborations should integrate the provisions of the ethics and governance framework (EGF) agreed by the initiative as far as possible at their local research site. The partner should ensure that they adhere to the EGF, for example by ensuring participants have given appropriate consent for the use of data and samples in the initiative and that they are informed of potential transfer across borders.

15 National bodies should publish their policies on the use of cloud services in health data settings so that data initiatives can include this in their decision making and interactions with publics and participants.

16 Biomedical researchers should give consideration to arrangements that will maximise the potential of participant-driven research to contribute to generalisable health knowledge and secure public benefits while providing adequate protection of those involved through continuing ethical and scientific appraisal. Key stakeholders are citizen patient researchers, biomedical research bodies, research funders and journal publishers. All stakeholders should encourage optimal use of human studies for improved health outcomes

17 The research community, including all academic and commercial partners, data controllers and custodians, public services and government agencies should actively foster opportunities to create a more explicit and reflective foundation for extending data access in the public interest. We urge all stakeholders in the medical research enterprise to continue to develop robust and comprehensive, yet efficient privacy protecting rules, guidelines and measures. Among other things these should aim at:

- Providing greater clarity for members of the public about ways that their biomedical data are used, and may be used in the future, along with a realistic acknowledgement that no system can guarantee privacy and confidentiality in all circumstances.
- Securing commitments from data controllers to a responsible approach to the extension of data access as part of their core mission statement; they must publish information about their approach to data access, transparency and accountability, and whether, and on what terms, they will consider extending access to data.
- Demonstrable and continual improvement of collection, storage and data access procedures against explicit standards for accuracy, reliability and security

Reflections (chapter 8)

41. Consideration of the state and direction of travel of information technology, data science, research and governance described in the report, and reflections on examples in health care and population research, lead to some practical precepts for professionals involved in data initiatives. In particular they should:

- identify prospectively the morally relevant values and interests in any data initiative.
- take special care to identify those people whose interests may be especially at risk, and interests that arise from diverse values

- not rely simply on compliance with the law as a way of securing that data use is morally appropriate, as the law does not always fully reflect moral norms
- identify the existing privacy norms in relation to the contemplated uses of data
- involve a range of those with morally relevant interests in the design of the data initiative in order to arrive at a publicly statable set of expectations about how data will be used
- state explicitly the set of morally reasonable expectations about the use of data in the initiative.
- involve a range of those with morally relevant interests in the continuing governance and review of the data initiative.

Introduction

The Council's terms of reference charge it, among other things, 'to identify and define ethical questions raised by recent advances in biological and medical research.' The developments with which this report is concerned are not peculiar to biomedicine although their impact on biomedicine raises significant and distinctive issues. The relevant 'recent advances' that the Working Party is responding to are principally two, and they are closely linked:

- first, the unprecedented quantity and variety of data collected through technologies of biological measurement (e.g. genomic and imaging data) and the accumulation of these stored data (e.g. medical records, biobanks), increasingly in machine readable formats;
- second, the development of more powerful ways of transferring, linking and manipulating data afforded by information technologies, and of analysing these data afforded by data science.

The report contains 17 propositions relating to these developments, four principles for ethical governance of data initiatives, and 17 specific recommendations for action. It has a broadly tripartite structure. The first part is largely descriptive: it describes some of the relevant advances in data collection and use, and the conditions and influences that are pushing it in particular directions. It describes the limitations of existing security, legal and governance measures that have been applied to the new uses of data and the challenges that therefore arise. The middle part explores the nature of the morally relevant values and interests at stake in data initiatives and develops a way of understanding data initiatives as social practices. Through a consideration of how the values are incorporated through the social processes involved in their formation and governance, it proposes a way of securing morally desirable outcomes. In the light of this, the third part of the report draws examples of good practice and identifies areas for improvement in selected data initiatives taking shape in health care and institutional research contexts. The intention is that, in addition to making specific recommendations, the report will provide an enduring resource and a support for constructive engagement with questions about the ethical use of data in health care and biomedical research.

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.

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.

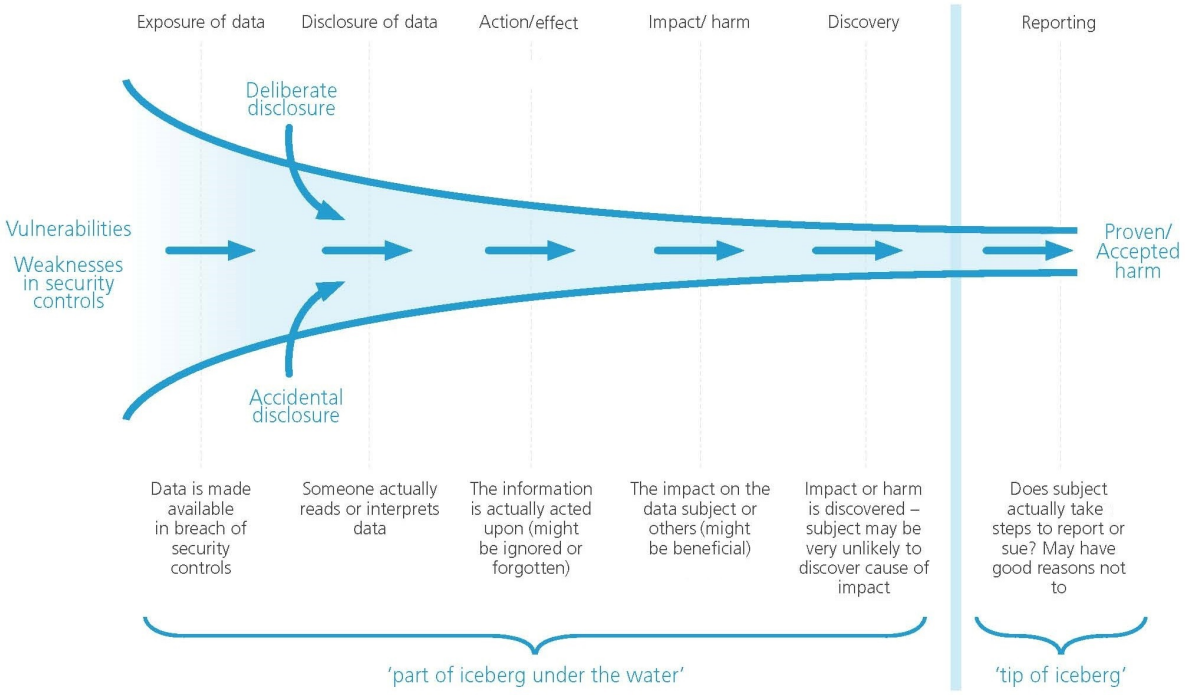


Figure 1: the 'confidentiality funnel'

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.

Chapter 3

Values and interests in data initiatives

Chapter 3 – Moral values and interests in data initiatives

Chapter overview

This chapter discusses the significance and nature of information privacy norms and the relationship between privacy and public interests.

The concept of privacy and the distinction between public and private have evolved throughout history. Privacy is important in the formation and maintenance of individual and group identities. Norms of confidentiality and information sharing characterise different relationships between people and groups. Medical confidentiality allows information sharing that might otherwise infringe privacy norms to take place for specific professional purposes.

Consent provides a mechanism to make controlled exceptions to an existing privacy norm for specific purposes. However, consent does not itself ensure that all of the interests of the person giving consent are protected nor does it set aside the moral duty of care owed to that person by others who are given access to the information. On its own, consent is not always necessary, nor always sufficient for ethical extensions of data access.

While individuals have privacy interests in the use of data, they also share group interests in the wider use of data for health research. The broader public interest may come into conflict with individual privacy but the relationship is usually complex. This complex relationship leads to a need to reconcile the articulation of the private within the public and the public within the private. A fundamental moral question facing data initiatives is therefore: 'How may we define a set of morally reasonable expectations about how data will be used in a data initiative, giving proper attention to the morally relevant interests at stake?'

Introduction

3.1 In this chapter we consider the morally relevant values and interests engaged by the use of data in biomedical research and health care. Our aim will be to understand what is at stake when claims are made about whether it is right or wrong to allow or to make particular disclosures of information. While data initiatives involving computerised data analysis have developed mainly in the late 20th and early 21st Century, many of the core moral questions they raise have been debated in some form for at least two-and-a-half millennia. We shall nevertheless try to formulate as clearly as possible the questions that must be addressed when ethical concerns are raised about current and future data initiatives.

The value of privacy

Proposition 9

Privacy is fundamentally important to individuals (and groups) in the establishment and maintenance of their identity, their relationships and their sense of personal well-being.

The public and private spheres

3.2 Human beings, as the philosopher Aristotle observed, have the capacity to form political communities and do so almost everywhere they exist.[146] The benefits of cooperative action are obvious: people working together can achieve what one person alone might never achieve. Nevertheless, just as human beings are born into, and drawn into, communities that advance their individual and common aims, they simultaneously value and preserve a sphere of 'private' thought and action.

3.3 The concept of privacy has a long and evolving history in Western social and political philosophy. In the fourth century BCE, Aristotle described the distinction between the spheres of private life – the life of the household – and public life – the free life of citizens in the *polis*, the Greek city state.[147] The private household was rigidly organised in order to supply the necessary conditions of life efficiently. In contrast, the public or political sphere was characterised not by necessity and toil but by freedom and discourse.[148]

3.4 The public/private distinction, as it would have been understood by the ancient Greeks, is transformed in modernity when labouring activities formerly constrained to the private realm of the household are transferred to the public sphere and organised through economic cooperation within societies. In the modern age, private life became increasingly important for the flourishing of individuality, for personal development and the cultivation of intimate relationships, both inside and outside the home.[149]

3.5 The vigorous defence of a sphere of free individual action to which society had no claim was a central preoccupation of modern liberal philosophers such as John Stuart Mill. "The only part of the conduct of any one, for which he is amenable to society," wrote Mill in *On Liberty*, "is that which concerns others. In the part which merely concerns himself, his independence is, of right, absolute. Over himself, over his own body and mind, the individual is sovereign."[150] This conception of a protected sphere was important, however, not only as a defence from society, but as a source of free moral action. According to Isaiah Berlin's famous distinction, freedom can be thought as having negative ('freedom from') and positive ('freedom to') aspects.[151] The first is the protection of a sphere of action from interference by others; the second concerns

[146] Aristotle, *Politics*, book I, available at: http://classics.mit.edu/Aristotle/politics.html.

[147] Ibid. In the ancient Greek *polis*, privacy is connected with the body, with manual labour necessary for the maintenance of life and with reproductive labour necessary for the continuation of the species.

[148] The freedom of the *polis* meant an equal freedom for citizens, although this has little in common with our contemporary understanding of equality: it presupposed the existence of (an inevitable majority of) 'unequals' – women, children, servants and slaves, whose labour was subordinated to the freedom of the male head of the household. A more recent critique in the feminist tradition argues that the defence of privacy can provide cover for the abuse and degradation of women and others. For a critical survey of some feminist writing on privacy, see Gavison R (1992) Feminism and the public/private dimension *Stanford Law Review* **45(1)**: 1-45.

[149] The political theorist Hannah Arendt argued that modern society inverts the norms of behaviour characteristic of the ancient Greek political realm, so that public behaviour becomes highly regulated, with the home life becoming a private refuge of 'intimacy'. Arendt H (1958) *The human condition*, 2nd Edition (Chicago: University of Chicago Press). See also Habermas J (1991 [1962]) *The structural transformation of the public sphere: an inquiry into a category of bourgeois society* (Cambridge MA: MIT Press) for an extensive analysis of the concept of the public sphere.

[150] Mill JS (1859) *On liberty*, available at: http://www.gutenberg.org/files/34901/34901-h/34901-h.htm, at page 18. This thought is embodied in Mill's 'harm principle'.

[151] See: Berlin, I (1969 [1958]) Two concepts of liberty, in Berlin I (1969) *Four essays on liberty* (Oxford: Oxford University Press), available at: http://spot.colorado.edu/~pasnau/seminar/berlin.pdf.

the freedom expressed by being the 'author' of one's actions, and an active agent in the formation of one's social world.

Informational privacy

Proposition 10

Control of certain information is generally viewed by individuals as an important aspect of maintaining their privacy; access to or disclosure of information contrary to their wishes can affect individuals' well-being and infringe their rights.

Proposition 11

Not all personal information is private and some personal information is legitimately public. Privacy norms depend on the nature of the relationship between individuals or between individuals and institutions, including the state.

3.6 Some philosophers argue that having the opportunity to be free from observation by others is essential to the formation and maintenance of individual identity and 'personhood'. A variety of arguments, all tending to this general conclusion, are offered in the literature. For example, privacy is claimed to be psychologically essential for personhood because the possibility of withdrawing from observation is necessary in order to assimilate and reflect on life experiences, and to identify one's unique individuality.[152] Privacy is said to be practically necessary because observation by others inevitably transforms the conditions in which the person chooses and acts, and, by placing external constraints on their moral choices, denies respect for them as a moral agent.[153] Privacy is said to be necessary, furthermore, because it is through the ritual of respecting privacy that the social group recognises the entitlement of an individual to their own moral existence.[154] Finally, privacy is necessary for intimacy, which is nurtured by a process through which people progressively share hidden aspects of themselves.

3.7 Disclosure and withholding of information between people has an important function in establishing the structure of social relationships, as a means by which particular people are included or excluded. Family, group, community – even national – identities may be formed and confirmed by norms of information sharing. The fact that someone shares intimate information with one person and not with others, can function as a token of friendship, cement social bonds, promote trust, and encourage reciprocal sharing, all of which lay the foundation for future cooperation.[155] The sharing of information has been said to establish 'moral capital' that is a currency for interpersonal relationships.[156] Breaching such norms (reading someone's private diary without their permission, for

[152] See, for example, Van Manen M and Levering B (1996), *Childhood's secrets: intimacy, privacy, and the self reconsidered* (New York: Teachers College Press), available at: https://archive.org/details/childhoodssecret00vanm.
[153] Benn SI, Privacy, freedom, and respect for persons, in Schoeman FD (Editor) (1984) *Philosophical dimensions of privacy: an anthology* (Cambridge: Cambridge University Press), pp223-44.
[154] Reiman JH (1976) Privacy, intimacy, and personhood *Philosophy and Public Affairs* **6(1)**: 26-44.
[155] See: Fried C (1970) *An anatomy of values: problems of personal and social choice* (Cambridge MA: Harvard University Press).
[156] Fried C (1970) *An anatomy of values: problems of personal and social choice* (Cambridge MA: Cambridge University Press). See also Gavison R (1992) Feminism and the public/private dimension *Stanford Law Review* **45(1)**: 1-45.

example, or 'hacking' into their telephone messages) both has the practical consequence of undermining trust and exhibits a moral attitude of lack of respect for them as a person. On the other hand, enforcing non-disclosure norms (such as 'keep it in the family' in cases of domestic abuse) can be equally morally unacceptable. We therefore have to ask whether the norms themselves are morally appropriate, as well as who has the authority and the power to modify or transgress them.[157]

3.8 Different norms of disclosure and withholding of information will apply to different kinds of relationship. A disclosure (e.g. infection with a stigmatising disease) that might be expected in one context (e.g. between close family members or clinical professionals) might be surprising in another (e.g. among work colleagues). A person may participate simultaneously in many different relationships, governed by different norms, for example, distinct personal, social and professional networks. The norms governing these relationships, along with the membership of the networks themselves, may change over time.[158] In the contemporary world, control of access to and disclosure of information has become increasingly significant in measure with the role played by information exchanges in the conduct of life. The presence of information technology – giving the capacity to store, replicate and communicate data indefinitely – has acted as an exponent to this. People may consider public buildings (swimming baths, for example) public places where anyone might observe or overhear them but they might consider the presence of CCTV or a 'webcam' an 'invasion of privacy'.[159] To answer the question whether such behaviours should be considered an 'invasion' of privacy, or whether disclosures of personal confidences are a 'breach' of privacy, we must define not only the nature of the expectations that have been frustrated, but what people are entitled to expect in these circumstances.

Confidentiality and consent

Proposition 12

Expectations of privacy relating to norms of access to and disclosure of information may be formalised and enforced, for example, through rules of confidentiality. These rules and expectations may be modified by individuals in specific cases, for example through explicit consent procedures.

[157] Ruth Gavison, for example, argues for the importance of a critique of the deployment of a public/private distinction rather than of the distinction itself. Gavison R (1992) Feminism and the public/private distinction *Stanford Law Review* **45(1)**: 1-45.

[158] Mark Taylor uses the term 'norms of exclusivity' to describe how the conditions of information access are deployed between different people in different social contexts: "Privacy is established by norms regulating access to individuals or groups of individuals: it represents a relevant state of separation defined and mediated by particular standards. In order to capture more fully the idea that relevant separation can only be assessed according to particular norms, I suggest that privacy concerns 'norms of exclusivity'." Taylor M (2012) *Genetic data and the law: a critical perspective on privacy protection* (Cambridge: Cambridge University Press), at page 25. See also Helen Nissenbaum, who posits 'contextual integrity' as a benchmark for privacy 'to capture the nature of challenges posed by information technologies': "Contexts, or spheres, offer a platform for a normative account of privacy in terms of contextual integrity. [...] contexts are partly constituted by norms, which determine and govern key aspects such as roles, expectations, behaviours, and limits. There are numerous possible sources of contextual norms, including history, culture, law, convention, etc. Among the norms present in most contexts are ones that govern information, and, most relevant to our discussion, information about the people involved in the contexts. I posit two types of informational norms: norms of appropriateness, and norms of flow or distribution. Contextual integrity is maintained when both types of norms are upheld, and it is violated when either of the norms is violated." Nissenbaum H (2004) Privacy as contextual integrity *Washington Law Review* **79(1)**: 119-58, at page 119.

[159] See, for example, the House of Commons Home Affairs Committee (2008) HC 58-I *A surveillance society?* (fifth report of session 2007–08), available at http://www.publications.parliament.uk/pa/cm200708/cmselect/cmhaff/58/58i.pdf. See paragraph 4.6.

3.9 An important class of privacy norms is embodied in the principles and practice of confidentiality. Whereas privacy may be about access to a number of different things (such as access to one's body, one's home or one's possessions) confidentiality is exclusively about information. However, confidentiality is not simply synonymous with informational privacy.

Box 3.1: 'Privacy' and 'confidentiality'

The terms 'privacy' and 'confidentiality' are sometimes used imprecisely. They are often used in conjunction (as in 'privacy and confidentiality issues') which may contribute to an elision of the distinct concepts, and in casual discussion they are occasionally used interchangeably. It is therefore worthwhile clarifying how we understand their distinct meanings.

- **Privacy** concerns the interest people have in others' access to themselves, their homes and property, or to information about them. What counts as 'private' can change depending on social norms, the specific context, and the relationship between the person concerned and those who might enjoy access. Informational privacy is maintained by selectively withholding or allowing access or through establishing limits on acceptable behaviour by others (e.g. proscribing voyeurism).

- **Confidentiality** concerns the assurance that information provided by a person (or by another body) will not be further disclosed without their permission (except in accordance with certain established laws, norms or expectations about when confidentiality obligations may be set aside). Duties of confidence are created by well-established expectations that attach to certain relationships (for example, between a doctor and a patient or between a lawyer and a client) or may be agreed between parties in a specific context (for example, parties to a commercial contract or contract of employment). They allow information to be made available for the purposes of that relationship (and perhaps also to others whose involvement is necessary to achieve those purposes), but for no other purpose. In short, confidentiality is one – but only one – of the tools used to achieve and maintain privacy.

3.10 Moral duties of confidence exist among individuals (friends sharing a secret, for example) but some duties of confidence are made enforceable through contractual and legal instruments or through the reasonable expectations that patients, for example, have of their doctors. Medical confidentiality exists because information that doctors might need in order to diagnose or treat a patient might be information that is non-obvious and of a type that the patient might not otherwise (other than for the purpose of obtaining diagnosis and treatment) want to disclose (including to the doctor concerned) or publish more widely. Similar considerations apply in the case of research. Medical confidentiality protects patients from harm in two ways: it both encourages them to disclose information essential to their treatment, so that they do not suffer the harm of untreated disease, and it provides assurance against any harm that may occur to them from a more general disclosure of the information. Over time, respecting confidence helps to foster trust.

3.11 Privacy norms may be modified informally by individuals simply through their behaviour (in the way that they may impart private information to others), particularly

where they trust the individual to understand and observe the appropriate norms and not to disclose the imparted information any further. Privacy norms may also be modified through formal mechanisms, such as giving consent, particularly where a formal structure exists to provide assurance that those norms will be complied with, such as obtains in the case of medical treatment or legal advice.

Box 3.2: Consent

Consent sets aside norms and standards, such as the expectation of confidentiality, in specific ways for specific purposes.[160] Consent does not abolish the underlying norm but modifies its application by creating a specific exception. (We discuss the operation and limitations of consent procedures further in chapter 4.)

Valid consent is consent that is freely (autonomously) given: for it to count as valid it cannot be obtained by coercion or deception. Furthermore, the person consenting should be aware of the morally relevant implications of giving consent. This does not mean that the consenting person needs to be aware of every last detail and consequence of the use of the data (so-called 'broad' consent may count as valid consent) but they should be aware of those details about the proposed use and the reasonably foreseeable implications that are morally significant to them. There is clearly room for considerable debate about how much information and understanding is necessary for consent to be valid, when different forms of argument and encouragement may undermine freedom and when the limits of previously given consent (perhaps one given in very different circumstances) are reached.

3.12 It should be noted that, while observing the terms of consent respects the interests of the person giving it in a limited way, the fact that information is only disclosed in accordance with the terms of the consent does not in itself protect the person from any harm resulting from the use of the information. Thus, consent should not be thought of as shifting the liability for any privacy infringements from the user of data to the 'consenting' person, and simply obtaining consent does not exhaust the moral 'duty of care' owed by the user of the data. This is consistent with the structure of consent, which implies permission to use the data but no obligation to do so, particularly where doing so would infringe the subject's interests. The fact that someone's consent must be sought is not, however, necessarily or straightforwardly empowering for the person giving it, particularly where the options available to them are highly constrained.[161] Consent is often, in fact, a rather blunt tool, allowing only a binary 'yes' or 'no' response. Genuine respect for the autonomy of individuals as 'world forming' is likely to be better realised through a richer involvement in the formation of norms and options than simply accepting or refusing options presented by others.

3.13 Consent is neither always necessary (since not all norms would otherwise prohibit data access and disclosure) nor sufficient (since it does not set aside the moral duty of the user of data with respect to others) for ethical use of data. However, where there is a

[160] The normative function of consent in creating a conditional waiver of pre-existing rights is discussed in Manson N and O'Neill O (2007) *Rethinking Informed Consent* (Cambridge: Cambridge University Press).

[161] This is the case, for example, where use of social networking software requires that the user accept the terms of an agreement that allows the provider to extract their use data and mine it for their own purposes or release it for others to use.

reasonable expectation that disclosure of information may infringe a well-grounded entitlement to privacy, consent may play an important role in enabling that disclosure.

Community and solidarity

3.14 Conventional forms of Western morality treat the person – usually a living human individual – as the fundamental unit of moral agency and value. Nevertheless, the assumption that privacy relates only – or primarily – to personal forms of identity is not universal, or not necessarily applicable to all forms of information. There may be different indices for privacy norms, some relating, for example, to the family group, tribe or community, even when the information in question pertains most obviously to a single individual among them. These norms may differ significantly between distinct cultures or contexts.[162]

3.15 In some recent bioethical writing on data access, a concept of solidarity has been promoted as a reaction to what many bioethicists see as the over-privileging of individual autonomy at the expense of wider public interest. The concept of solidarity describes social cohesion as a result of the homogeneity or interdependence of individuals making up a community.[163] The concept has been applied particularly in relation to genetic and genomic information (our 'shared genetic heritage'), but also in relation to biobanks and biomedical research more generally.[164] This concept of solidarity may be embodied and formalised in institutions.[165]

3.16 The arguments for solidarity as a moral basis for extending data access in relevant cases have, generally, been cautious rather than revolutionary. In the literature, solidarity tends to be proposed as a default social norm from which individuals retain the entitlement to withdraw, rather than as a moral obligation from which they may be released only exceptionally. While they change the emphasis, in ways that may bear on decisions about appropriate forms of governance (we discuss concrete examples in chapters 6 and 7), the arguments for solidarity rarely seek to overturn the primacy of individual autonomy. Implicitly, the shift in disposition comes about as the result of increased opportunities available to derive public benefit from personal data where the privacy risks to individuals are well managed, as well as the increasing practical difficulties of maintaining individual level controls in complex data flows.[166] In practice it amounts to little more than a justification for 'broad' models of consent: for example, the replacement of specific individual consent for research uses of data (to the extent

[162] For a specific case, see guidance on ethics in relation to the Canadian First Nations, Inuit and Metis communities: http://www.pre.ethics.gc.ca/eng/policy-politique/initiatives/tcps2-eptc2/chapter9-chapitre9/.

[163] See the distinction between 'mechanical' and 'organic' solidarity in Durkheim É (1984 [1893]) *The division of labour in society* (Basingstoke: Macmillan), book I, chapter II and III. The concept is also important in ethics and political philosophy: for a discussion of the recent emergence of the concept in bioethics see Prainsack B and Buyx A (2011) *Solidarity: reflections on an emerging concept in bioethics*, available at: http://nuffieldbioethics.org/project/solidarity/.

[164] See the concept of 'genetic solidarity' in: Human Genetics Commission (2002) *Inside information: balancing interests in the use of personal genetic data* (London: HMSO), available at http://webarchive.nationalarchives.gov.uk/20061023110946/http://www.hgc.gov.uk/UploadDocs/DocPub/Document/insideinformation_summary.pdf. In relation to biobanks, see, for example, Stjernschantz Forsberg J, Hansson MG and Eriksson S (2009) Changing perspectives in biobank research: from individual rights to concerns about public health regarding the return of results *European Journal of Human Genetics* **17(12)**: 1544-9; Buyx A and Prainsack B (2013) A solidarity-based approach to the governance of research biobanks *Medical Law Review* **21(1)**: 71-91.

[165] See the 'third tier' solidarity in Prainsack B and Buyx A (2011) *Solidarity: reflections on an emerging concept in bioethics*, available at: http://nuffieldbioethics.org/project/solidarity/.

[166] See Stjernschantz Forsberg J, Hansson MG and Eriksson S (2009) Changing perspectives in biobank research: from individual rights to concerns about public health regarding the return of results *European Journal of Human Genetics* **17**(12): 1544-9.

that they should be re-contacted to consent to novel uses) with general 'participation agreements'.[167]

3.17 It is relevant to observe, moreover, that just as the (re)emergence of the concept of solidarity can be seen as a reaction to the constraints of autonomy and individualism, the assertion of autonomy in bioethics was itself partly a reaction to extreme derogation of human rights that took place in Europe in the middle of the 20th Century and continued in totalitarian regimes during the latter part of that century.[168] Partly, also, it was an expression of growing resistance to institutionalised paternalism in fields such as medicine.[169] Although these political conditions may have been largely extinguished in Europe, it is easy to conceive that the pendulum may swing back towards autonomy as a consequence of large-scale privacy intrusions by states, for example, though the activities of state organisations such as the US National Security Agency (see chapter 2).

Public interest

3.18 The 'public interest' is not the opposite of private interests although it is sometimes contrasted with them. The 'public interest' can be thought of as securing objectives that are valued by society.[170] There are two questions that we must address when we consider the relationship between the public and private interests. The first question is about the *content* of the public interest (i.e. its objects) and how this relates to the aims and interests of individuals. This leads us to questions about legitimate procedures for making collective decisions. The second question is about the *force* that should be given to the public interest, most importantly where it is in tension with private interests. This takes us to juridical questions about when it is legitimate to limit or even override private interests in the name of the public interest. Clearly, these questions are interconnected: the nature of the objects of the public interest and the value assigned to them will relate to the force that public interest claims have.

The objects of the public interest

3.19 Identifying the proper objects of the public interest is not straightforward and political philosophers have argued about the merits of different approaches. A general distinction can be drawn between those that rely on abstract reasoning from premises (such as propositions about the moral nature of human beings) and those that employ empirical methods to discover actual preferences (such as, voting or deliberative decision making).

---

[167] Prainsack B and Buyx A (2013) A solidarity-based approach to the governance of research biobanks *Medical Law Review* **21(1)**: 71-91.

[168] Chadwick R and Berg K (2001) Solidarity and equity: new ethical frameworks for genetic databases *Nature Reviews Genetics* **2(4)**: 318-21.

[169] See Katz J (1984) *The silent world of doctor and patient* (New York: Free Press).

[170] We use the terms 'community' and 'society' to indicate, respectively, associations in which there is a unity of values and a common will among the members, and associations in which their shared project represents a compromise for the sake of self interest. The distinction was made by sociologist Ferdinand Tönnies (using the terms *Gemeinschaft* and *Gesellschaft*, respectively, to denote these forms of association). Tönnies F (2001 [1887]) *Community and civil society* (Cambridge: Cambridge University Press), book I. A community founded on kinship ties is the example of the first, whereas a trading group is an example of the second. All social groups are, in reality, a mixture of the two so the terms are used to emphasise the nature of a given association rather than to describe it. Feintuck uses definition of public interest within the field of regulation: "[...] the concept of public interest as a justification for regulatory intervention into private activity, limiting the exercise of private power, in pursuit of objectives valued by the community." Feintuck M (2004) '*The public interest' in regulation* (Oxford: Oxford University Press), at page 6.

3.20 The 'common good' theories of medieval Christianity take up Aristotle's understanding of the relation between particular goods and the good of all as being the good of human beings in view of their nature. In supposedly pious societies the object of the common good was implicitly common to all (or ought to be so, once people were enabled to understand the vanity of their temporal concerns and carnal appetites).[171] The identification of the public good as the set of goods that all people share is appealing, but it offers a much poorer prospect of guiding policy or action in diverse societies where there may be disagreements over priorities and where individuals have strongly developed and diverse private interests. In these circumstances the objectives common to all are likely to be rather abstract: things like 'security', 'health' and 'prosperity'.

3.21 As the exclusive pursuit of conflicting private interests is likely to be destructive (or, at least, sub-optimal), a way of limiting conflict and securing cooperation and the provision of public goods is desirable.[172] One way of doing this is to envisage the terms of an implicit contract that specifies what limitations to the free pursuit of their interests people ought to accept and, in return, what the legitimate role of the state will be in securing public goods. The idea of a 'social contract' of this kind was formulated by thinkers of the European Enlightenment (Locke, Hume, Rousseau) and had an influence on the founding fathers of the United States of America (Jefferson, Madison).[173]

3.22 An alternative approach that attempts to draw out the public interest by aggregating individual private preferences is offered by utilitarianism, as formulated in the 18th Century by, for example, Jeremy Bentham.[174] Utilitarianism assumes that if members of the community 'vote' to maximise their own happiness or 'utility' the aggregation of their interests will indicate the outcome that will maximise absolute utility. This both recognises and accepts that people may have different ideas about their preferred outcome but aims to find a resolution that they should all accept if they agree in advance that the voting procedure is a fair way of resolving them.

3.23 Any approach to determining where the public interest lies will have advantages and disadvantages. Any approach that derives rules of action from abstract principles must make assumptions about what it is people *should* value, notwithstanding what their actual subjective preferences may be. It will then have to account for how this can be consistent with respect for individuals as free moral agents. Aggregative approaches have the virtues of clarity and simplicity but, pursued mechanically, they can lead to perverse outcomes (for example, in a three voter system where two voters vote to kill the third and appropriate her property). There are many historical examples of

---

[171] Aquinas T, *Summa contra gentiles*, III.17.6, available at: http://dhspriory.org/thomas/ContraGentiles.htm.

[172] The concept of "public goods" is used loosely here to mean goods that are provided for public benefit: in economics, public goods are those that are non-rivalrous (my use of it does not deprive you of the opportunity to use it) or non-excludable (it cannot be made available to me without also making it available to you), or both. For those reasons public goods are typically provided by the state rather than by the market, since it is difficult to make commercial profit and public goods are vulnerable to 'free riding' (people consuming the goods without paying for them). Examples include policing and street lighting, and public health initiatives.

[173] The idea of the social contract predates the Enlightenment as it is usually described (i.e. beginning in the late 17th Century and lasting until the rise of romanticism in the late 18th). An important precursor was Thomas Hobbes's *Leviathan* (1651) written during the English Civil War. The social contract tradition endures in the US notably in the work of John Rawls. See Rawls J (1999) *A theory of justice* (Cambridge, MA: Belknap Press of Harvard University Press).

[174] See Bentham J (1789) *An introduction to the principles of morals and legislation*, available at: http://www.earlymoderntexts.com/pdfs/bentham1780.pdf. Classical utilitarianism identifies the public interest with the preponderance of interests within a political community, which is arrived at simply by aggregating the interests of individual members. The chief appeal of an approach of this kind is its procedural fairness: "everybody to count for one, nobody for more than one." This dictum was attributed to Bentham by John Stuart Mill in *Utilitarianism* (1861), available at: http://www.earlymoderntexts.com/pdfs/mill1863.pdf, at page 44.

individual interests being disregarded in the name of the common good. For this reason, where utility calculations are used as the basis of public decision making (e.g. in the form of cost-benefit analysis), the outcomes are usually qualified by some distributive principle (to ensure fairness) or other limiting factor (such as rights to non-interference) in order to prevent the interests of some being sacrificed for the good of others.

3.24 The best approach may well lie in some combination of both principle (to foreclose intuitively unacceptable outcomes) and practical reasoning (to allow expression to a plurality of legitimate interests). The approaches discussed above are not the only ones available and we will return to this discussion when we tackle the question of finding ethically appropriate forms of governance for specific data initiatives in chapter 5.

The force of the public interest

3.25 In relation to the data initiatives with which we are concerned, the public interest will be an important legal and regulatory concept. This is particularly so where these initiatives are public initiatives, carried out with the involvement of the public sector or with public funding, or are otherwise aimed at delivering public goods. The claim that the object of any data initiative is an important public good may offer a justification for modifying the usual privacy norms (as, for example, some contagious disease reporting has been made mandatory to avert epidemics). However, it is not only the state that can appeal to public interest as a justification for normative action. Private actors may appeal to public interest as a justification for interfering with others' privacy, as when newspapers publish 'private' information about public figures. Individuals may claim a breach of confidentiality norms is justified by the public interest, for example when 'whistleblowers' make public interest disclosures.

3.26 Whereas public interest plays an important regulatory function in keeping private interests in check, or may justify overriding them in certain circumstances, there is a concern that it might be used to justify unacceptable levels of paternalism or state intrusion into private affairs. A significant bulwark against the intrusion of the state into the lives of its citizens, and of individuals into each others', is provided by human rights instruments.[175] (We will discuss human rights law in the next chapter.) Human rights law guarantees the protection of private life against interference except where this is necessary for an overriding public interest. A key concept in determining when a particular interference is justifiable, developed in the jurisprudence, is that of 'proportionality'. In other words, the interference, which must be necessary in order to achieve a legitimate aim, must be proportionate (sufficient but not excessive) to the achievement of that aim. Furthermore, it must be knowable, in a way that informs individuals' expectations and allows them to modify their actions accordingly. Thus, uses of data that might, at face value, interfere with privacy interests can be justified so long as these conditions are met. However, for the time being we are not concerned with situations in which rights come into conflict, but with interests and preferences, and the production of moral norms.

[175] "The European Court of Human Rights (ECtHR) makes important contributions to how the UK should conceptualise the notion of privacy and concomitantly protect against prospective violations of individuals' Article 8 rights, outwith the more narrow confines of the Data Protection Act 1998." Laurie G, Jones K, Stevens L, and Dobbs C (2014) *A review of evidence relating to harm resulting from uses of health and biomedical data*, at page 11 available at: www.nuffieldbioethics.org/project/biological-health-data/evidence-gathering/.

The mutual implication of public and private interests

3.27 Individuals are embedded in communities in complex ways: each has a private interest in protecting their privacy but also in contributing to the public good because, as a member of the public, they and those they care about benefit from the good that they bring about through cooperation with others in society. Likewise each knows that there is a public interest in respecting their privacy because the good of the community depends on their willingness to enter into voluntary cooperation with others under conditions in which they must share private information with confidence.

3.28 Consequently, our problem is not finding a 'balance' between privacy and public interest for a data initiative, but resolving a double articulation, between the private interest in protecting privacy and promoting the public good, and the public interest in protecting privacy and promoting the public good.[176] We all have interests on both sides, private and public, as individuals, members of families, groups, communities and nations. Navigating among these different relationships with other individuals, professionals and institutions requires a subtle negotiation of many different norms of information access and disclosure, of when and how they may be modified and where hard and fast limits should be drawn.

Proposition 13

A fundamental question to be addressed in relation to the ethical design and conduct of data initiatives is:

How may we define a set of morally reasonable expectations about how data will be used in a data initiative, giving proper attention to the morally relevant interests at stake?

3.29 The question proposed above (proposition 13) is the one that we will be mainly concerned with answering, along with examining how data initiatives have addressed it, explicitly or implicitly. Formulating the question in this way recognises that people's interests may be mutually limiting or mutually reinforcing. What it is morally 'reasonable' to expect depends upon an assessment of the moral claims of those interests. When we consider whose interests are relevant it is important to remember that these include the expectations not only of those to whom data relate, but of those making use of the data, and those who have an interest in the aims or outcomes of a data initiative. It is also important to acknowledge that those whose privacy interests are engaged may also have interests in securing the individual and public benefits of data use. Three sorts of consideration will be relevant in formulating an answer to this question:

- the identification of the norms of privacy and disclosure that are applicable in relation to a specific data initiative. (Some, but not all of these, may be encoded in laws and professional rules of conduct.)
- how respect is shown to persons, especially where their individual preferences do not coincide with these norms. (This may often be for good reason, e.g. because their

[176] In *X v. Y* [1988] 2 All E.R. 648, for example, the law was understood as requiring a balance of two public interests (in maintaining confidentiality and in disclosing information of public interest).

contextual vulnerabilities, or even arbitrary preferences, may deserve to be respected.)

- the form of governance (for example, the regulation of professional conduct) that will give acceptable assurance that the expectations will be met.

Conclusion

3.30 In this chapter, we have examined the kinds of interests that are at stake in data initiatives and why these are morally relevant, and arrived at the formulation of a key moral question that faces data initiatives. Different approaches to resolving complex sets of interests may have advantages and disadvantages but it is likely that an appropriate approach will involve a mixture of ethical principles and empirical methods. This will be the subject of chapter 5. First, however, we will consider the problems with which conventional governance approaches are faced as a result of the developments in data science, information technology and data policy that we discussed in the previous two chapters.

Chapter 4

Law, governance and security

Chapter 4 – Law, governance and security

Chapter overview

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

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

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

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

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

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

Introduction

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

Legal framework for use of biological and health data

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

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

A legal right to privacy

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

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

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

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

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

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

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

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

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

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

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

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

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

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

[186] See paragraph 3.8.

Data protection law

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

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

---

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

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

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

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

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

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

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

Common law

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

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

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

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

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

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

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

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

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

Security of data

Operations designed to prevent the identification of data subjects

Aggregation

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

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

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

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

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

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

Anonymisation

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

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

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

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

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

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

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

Box 4.1: Re-identification: some examples

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

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

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

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

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

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

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

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

Pseudonymisation

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

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

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

---

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

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

[211] See paragraph 2.43.

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

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

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

Weaknesses of de-identification

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

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

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

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

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

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

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

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

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

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

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

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

Box 4.2: Technical anonymity

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Controlling data access and use

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

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

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

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

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

Consent

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

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

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

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

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

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

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

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

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

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

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

---

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

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

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

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

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

Difficulties for consent in data initiatives

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

The limited role of consent

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

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

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

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

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

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

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

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

Governance and security

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

Authorisation of data access

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

---

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

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

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

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

Limiting data access

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

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

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

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

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

Limitations to data use

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

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

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

Conclusion

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

---

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

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

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

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

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

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

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

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

Chapter 5

Ethical governance

Chapter 5 – Ethical governance of data initiatives

Chapter overview

This chapter develops an ethical approach to the design and governance of data initiatives and sets out some principles for guidance.

Data initiatives are practical activities that involve a number of actors (who might be individuals, groups, institutions, etc.) some of whom stand to benefit or lose from the outcomes. Tensions and potential conflicts between values and interests can arise at the level of the individual, of professions or of the public. The ethical formation of a data initiative is a matter of reconciling these values and interests in a coherent set of morally reasonable expectations.

A morally reasonable set of expectations should embody four principles:

- the principle of respect for persons
- the principle of respect for established human rights
- the principle of participation of those with morally relevant interests
- the principle of accounting for decisions

The principle of **respect for persons** does not mean that individuals' interests may never be overridden, but that they may only be overridden where there is a legitimate reason to do so. As a principle of design of data initiatives, the principle of respect for **human rights** seeks to avoid potential rights conflicts and violations rather than leaving them to be dealt with retrospectively through judicial processes. The **participation** of people with morally relevant interests allows the identification of relevant privacy norms and the development of governance measures (such as design of consent and authorisation procedures) in relation to these norms; it allows preferences and interests to be expressed and transformed through practical reasoning, and account to be given of how these interests are respected in decision making, helping to foster trust and cooperation. The principle of **accounting for decisions** ensures that expectations, as well as failures of governance and control, are communicated to people affected and to others more widely. It also ensures that data initiatives remain in touch with changing social norms.

Introduction

5.1 In chapter 3 we examined the moral values and interests engaged by the collection, retention and use of data in biological research and health care. We saw that private and public interests in these activities are interrelated, often in complex ways, and may be – but are not automatically – in tension. At the end of that chapter we proposed a question to structure reflection on the moral acceptability of data initiatives. They should, we argued, define a set of morally reasonable expectations about how data will be used in the data initiative, giving proper attention to the morally relevant interests at stake. We suggested that an answer should take into account three sorts of considerations: the underlying norms of data access and disclosure, the respect for people in terms of their individual values and interests, and the governance of professional conduct in the public interest. In the previous chapter we examined different legal frameworks and concluded that the minimal conditions they offered did not exhaust or always correspond to morally relevant norms for specific data initiatives. We discussed different consent procedures and concluded that obtaining consent from the 'subjects' of data was not sufficient (or always necessary) to make the use of data

morally acceptable. And we discussed tools and procedures of governance and concluded that they gave only a partial answer to questions of moral accountability.

5.2 In this chapter we will take up the question posed at the end of chapter 3 and propose a way of moving, usually in conditions of some uncertainty, from the complex of often poorly articulated and possibly inconsistent values and interests that are engaged by data initiatives to a more coherent, shared and publicly articulated solution. Our approach is based on an understanding of the establishment and conduct of data initiatives as an activity that requires cooperation between people whose interests they engage. We examine how engaging in this activity can inform and develop the relationship between moral norms, individual values and interests, and governance in the public interest, and help to define appropriate governance arrangements. In the course of this we will propose an ethical framework comprising four elements (which are set out in Box 5.1 at the end of the chapter).

Morally relevant interests

5.3 Interests are not abstract ideas, existing independently of their bearers and outside time, or unreflective desires demanding satisfaction. They are tied to the people whose interests they are, to a particular material context and orientated towards specific future goals. The contexts in which interests are expressed may involve a number of different people, professions and practices. Indeed, the potential of a data initiative to extract value, and the novel features that give rise to ethical questions, typically result from converging developments in a number of fields of endeavour (for example, the application of computational methods to human biology) rather than a tipping point in the development of any one field. Although no list can be exhaustive, data initiatives in biomedical research and health care might involve:

- Information governance professionals
- Clinicians and other health care practitioners who hold or use data
- Biomedical scientists and researchers (including pathologists, imaging specialists, geneticists, epidemiologists, etc.)
- Social and behavioural scientists
- Bioinformaticians, statisticians and data scientists
- Information technology developers
- Research funders (who may be the public as national taxpayers)
- Commercial firms
- Public policy makers and administrators (service commissioners, etc.)
- Independent advisors (lawyers, bioethicists, etc.)
- Regulators
- Patients or research study participants whose data are included in the initiative (who might also be members of any of the forgoing categories)
- The wider 'public' (or 'publics')[258]

5.4 Along with their skills and resources (including data), each of those involved in a data initiative will bring a particular set of interests and expectations.[259] These may be more

[258] 'The public' is, of course, a controversial category: see discussion in Nuffield Council on Bioethics (2012) *Emerging Biotechnologies: technology, choice and the public good* (esp. Chapter 5), available at: http://www.nuffieldbioethics.org/emerging-biotechnologies.

[259] Interests may be in either maintaining or altering these norms, either in the specific case of an initiative or in general.

or less shared and more or less stable within professional or disciplinary groups. (They may be established through professional codes of ethics of good practice guidelines within membership organisations, for example.) But in other cases they may be contested within a given field or profession, perhaps as different 'schools' or 'movements'. They may be further complicated by historical peculiarities, political differences, as well as national traditions and legal contexts (particularly in large international collaborations). The public interest in a data initiative may also be more complex and far-reaching than the immediate aims of the initiative. As well as the immediate aims of the initiative, there may be a public interest in supporting national research or production capacity (so that it can support the development of other products) or even in generating economic activity more generally.[260] Two further sets of potentially, but not necessarily, conflicting interests will be those of the public (or of that portion of the public that the data initiative aims to benefit) and those – who may overlap with the first group – from whom the data were collected.[261]

5.5 As well as alignments and tensions between professional groups and among individuals within those groups, the interests of individuals themselves can be inconsistent, contrary, and changeable. Research in psychology, behavioural economics, and other social sciences has shown, for example, that the behaviour of individuals in regulating access to and disclosure of private information may not follow rational or predictable patterns.[262] One example of this is the so-called 'privacy paradox', which refers to the dissonance between individuals' stated and revealed preferences (for example, people's stated preferences for privacy and their behaviour using public online social networks).[263]

5.6 The formation of a data initiative is therefore a complex social practice where tensions and potential conflicts of interests exist at many scales: at the level of the individual, of professions, and of the public. Thinking about data initiatives in this way avoids placing different interests (public and private, researchers and subjects, science and society, etc.) in simple opposition. (See chapter 3 where we drew attention to the mutual implication of public and private interests.) It focuses attention, instead, on how initiatives are formed by those with relevant interests and how, within this context, those involved may collectively develop their moral 'craft' through shared

[260] A common feature of discourse around innovation is equivocation between the scientific, therapeutic and broader strategic and economic aims (for example, between improving treatments for everyone and beating international competitors in a race to develop those treatments). See chapter 2 (above) for a discussion of this confusion of public interests.

[261] Health research is generally in the public interest on the basis that any healthy member of the public may be affected by ill health. This is less true for rare hereditary diseases, for example, where the existence of solidarity relations between those at risk and other members of the community comes into the question.

[262] Irrational behaviour has a number of explanations including the presence of 'framing effects' that distort the appraisal of evidence; preferential modes of reasoning identified by moral psychology that function especially when relevant information exceeds available cognitive capacity; and prevailing social norms. On 'framing effects' see Tversky A and Kahneman D (1981) The framing of decisions and the psychology of choice *Science* **211(4481)**: 453-8; on 'moral psychology' see Haidt J (2012) *The righteous mind: why good people are divided by politics and religion* (New York: Pantheon); on 'prevailing social norms' see Utz S and Krämer N (2009) The privacy paradox on social network sites revisited: the role of individual characteristics and group norms *Cyberpsychology: Journal of Psychosocial Research on Cyberspace* **3(2)**, article 1, available at: http://cyberpsychology.eu/view.php?cisloclanku=2009111001&article=1.

[263] For a survey of evidence about the relationship between what people say and what they do online, see Acquisti A, John LK and Loewenstein G (2013) What is privacy worth? *The Journal of Legal Studies* **42(2)**: 249-74, available at: http://www.heinz.cmu.edu/~acquisti/papers/acquisti-ISR-worth.pdf. While, for example, the promiscuity of younger generations in online social networking is a popular trope, it is also argued that they are actually more cautious and adept at simultaneously managing multiple interactions governed by different privacy norms. See, for example, Marwick AE and Boyd D (2014) Networked privacy: how teenagers negotiate context in social media *New Media and Society* **17(6)**: 1051-67.

understanding of 'good practice'.[264] It also suggests that the elucidation of relevant interests should be an important initial step in the formation of a data initiative.

Morally reasonable expectations

5.7 The interests that apply to different data initiatives will vary according to the initiative in question. Nevertheless, there is a general need to find a way of reaching decisions about the use of data that command respect, particularly among those who may feel that their own preferences have not prevailed. If the decision-making process lacks moral legitimacy, they may feel that their interests have been disregarded by others, especially to the advantage of those with greater political, economic or social power.

Moral reasonableness

5.8 The question with which we concluded chapter 3 concerned what it might be reasonable for those who participate in data initiatives to expect concerning the use and control of the data. There are broadly two ways in which something might be argued to be 'morally reasonable'. First, a proposition might be morally reasonable if it conforms to an objective moral standard or principle. What establishes that standard and who judges conformity with it are therefore important second-order questions. Second, a proposition might be judged to be reasonable where it is the outcome of a legitimate procedure, for instance democratic decision making.[265] In this case, who participates in this procedure and how it is conducted are equally important secondary questions.

5.9 The weakness of approaches based on substantive principles is that if they are too abstract they leave open a wide margin of interpretation concerning how they should be applied. If they are too prescriptive they may proscribe solutions that can optimise ethical data use according to legitimate and possibly diverse values. Purely procedural approaches, on the other hand, can result in morally perverse outcomes if they are not constrained or guided by some principle (as we noted in chapter 3). Procedural approaches therefore generally include an appeal to objective moral standards that both legitimise and place some limitations upon the relevant procedures. The approach we propose here has a strong procedural dimension – emphasising both participation and accountability – but is grounded in and constrained by a strong commitment to 'respect for persons'.

> **Principle 1 – Respect for persons**
>
> **The set of expectations about how data will be used in a data initiative should be grounded in the principle of respect for persons.** This includes recognition of a person's profound moral interest in controlling others' access to and disclosure of information relating to them held in circumstances they regard as confidential.

[264] See Parker M (2012) *Ethical problems and genetics practice* (esp. chapter on 'moral craft' pp. 112-30) Cambridge: Cambridge University Press).

[265] An example of this approach is outlined by Normal Daniels under the rubric 'accounting for reasonableness': see Daniels N (2000) Accountability for reasonableness: establishing a fair process for priority setting is easier than agreeing on principles *British Medical Journal* **321(7272)**: 1300-01, available at: http://www.bmj.com/content/321/7272/1300?ijkey=e1c0e7705033bda924d2556bec2d6af8da87175d&keytype2=tf_ipsecsha.

5.10 The principle of respect for persons is the principle that all persons have a special moral status that means they are owed respect simply in virtue of being persons and not because of any contingent characteristics, individual merit or social position.[266] Respecting persons takes the practical form of treating persons in ways that have regard to their own interests, not merely treating them as tools to secure our own ends or gratification.[267]

5.11 The principle of respect for persons, with regard to information access and disclosure, implies that consideration should be given to how their wishes about certain uses of information should be taken into account. This does not imply simply doing what people want. In many cases the things that different people want are incompatible. Data initiatives are collective activities that require cooperation. What is 'reasonable' in the context of a data initiative must, therefore, pay appropriate respect to all those with morally relevant interests. It means, furthermore, that the initiative should not assume that their interests can be respected simply by taking account only of the interests of a family, tribe, community, or nation to which they belong.

5.12 There may be rare cases in which a data initiative may depend on people accepting something (the disclosure of certain data relating to them, for example) that they would prefer not to happen. This is not always incompatible with respect for persons, even in the face of their active and specific objections. There are two main cases in which the argument for mandatory inclusion of individual data in a dataset by appeal to necessary and proportionate interference is made (see paragraph 4.5). The first case is where an aim of paramount public interest can only be achieved by either comprehensive participation (or could not be achieved by a level of participation expected under non-compulsory conditions) or can only be achieved by the inclusion of particular individuals. Such cases as these arise (although not without controversy) in the domain of public health, where individual objections to state intrusion into private life are sometimes overruled in the public interest. Limitations of this sort underpinned a series of Vaccination Acts in the UK in the 19th Century (for the eradication of smallpox) and may be invoked in contemporary public health emergencies.[268] A relatively uncontroversial example is the mandatory reporting by doctors of 'notifiable' diseases.[269] The second case is one in which full participation is not necessary, but where it can be argued that 'free riding' (*i.e.* benefitting from a public good that others have borne the cost of providing) is regarded as morally unacceptable. Some have sought to apply this argument to national health data initiatives to develop or improve medical treatments or care.[270]

---

[266] Philosophers have argued about the criteria for being a 'person' and whether it applies to 'marginal cases' (e.g. neonates, people with severe cognitive impairments, or cognitively advanced higher primates or computers). Such considerations may become operationally relevant when decisions about data relating to particular subjects fall to be made.

[267] Philosophical support for the principle is usually derived from the work of the Prussian Enlightenment philosopher, Immanuel Kant. See: Kant I (1998 [1785]) *Groundwork of the metaphysics of morals* (Cambridge: Cambridge University Press).

[268] The 1853 Act, for example mandated vaccination of every infant whose state of health would stand it and registration of the fact with the registrar of births. See: Porter D and Porter R (1988) The politics of prevention: Anti-vaccination and public health in 19th century England *Medical History* **32**: 231-52, available at: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1139881/pdf/medhist00062-0007.pdf.

[269] A duty is placed on a doctor to notify the relevant local authority officer if they suspect that a patient has a 'notifiable disease' under the Public Health (Control of Disease) Act 1984 (http://www.legislation.gov.uk/ukpga/1984/22) and Public Health (Infectious Diseases) Regulations 1988 (http://www.legislation.gov.uk/uksi/1988/1546/contents/made). For the legal basis of information sharing in relation to health, see: Department of Health (2007) *NHS information governance. Guidance on legal and professional obligations*, available at: http://systems.hscic.gov.uk/infogov/codes/lglobligat.pdf.

[270] See: Chan, Sarah and Harris, John. 2009. Free riders and pious sons – why science research remains obligatory *Bioethics* **23(2)**: 161-171, available at: http://onlinelibrary.wiley.com/doi/10.1111/j.1467-8519.2008.00648.x/full; Stjernschantz Forsberg

5.13 The fact that some data initiatives involve relationships between individuals and the state, which, as is often said, has a monopoly on the legitimate use of force to require compliance where it is not freely given, gives rise to our second requirement, concerning human rights.

> **Principle 2 – human rights**
>
> **The set of expectations about how data will be used in a data initiative should be determined with regard to established human rights.** This will include limitations on the power of states and others to interfere with the privacy of individual citizens in the public interest (including to protect the interests of others).

5.14 Some data initiatives have been challenged in the courts because they have been seen as breaching privacy rights enshrined in UK and European law.[271] The mechanics of the judicial process, however, mean that the acceptability of a practice may not be tested until there is a victim. It is often the case that the evolution of the law lags behind the invention of new initiatives and in some cases behind the evolution of social norms (although in some cases it can encourage such evolutions, as with the recognition of the rights of certain minorities). For reasons that we discussed in chapter 2, in the case of data abuses there can be a long interval between the cause of the harm and the effect, and abuses may continue for years before a harm is detected and addressed. Furthermore, as our commissioned research showed, it is necessary to consider a broader range of morally relevant effects than might meet the standards of harm required to engage legal rights.[272] A final consideration against merely relying on the judicial process is that seeking relief might actually compound the harm by drawing attention to it, so 'victims' may be unwilling to take action.

5.15 The purpose of promoting human rights as a formative principle for data initiatives is therefore to encourage a prospective consideration of how conflicts might arise and how they might be resolved (particularly where legal provisions are inadequate, inaccessible, unclear or conflicting), or avoided altogether. The aim is also to encourage the foundation of data initiatives on the moral rights that underpin legal systems, rather than to focus on simply satisfying the requirements of positive law, possibly on the construction most favourable to the aims of the initiative (and not necessarily to the interests of all those affected by it).

5.16 The principle of respect for persons and the requirement to respect human rights set the criteria for what expectations about the use of data may qualify as morally reasonable. Together, they provide the substantive 'guide rails' for the formation of morally acceptable data initiatives, without prescribing what specific measures should be adopted in the context of any particular data initiative. Within these bounds we still need to determine how the relationship between the relevant norms, the interests of

J, Hansson MG, and Eriksson S (2014) Why participating in (certain) scientific research is a moral duty *Journal of Medical Ethics* **40**: 325-8, available at: http://jme.bmj.com/content/40/5/325.short.

[271] An example of an initiative subject to a successful claim is the UK National DNA Database (see *S. and Marper v. The United Kingdom;* http://www.bailii.org/eu/cases/ECHR/2008/1581.html*).*

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

individuals and the external controls of governance should be resolved in any given concrete situation.

Moral reasoning

5.17 While individual interests may sometimes be overridden in the public interest, more usually, however, people will accept outcomes that they would not necessarily prefer because they are an indissociable part of a 'bundle' of goods that they value. This may be the case, for example, with freely provided Internet services, where users 'trade off' some unfavourable 'terms and conditions' in order to gain access to a service.[273] This, however, assumes that people are mere consumers, reduced to accepting or refusing an option presented to them, or choosing between several available 'offerings'. Where the question is about the design of a data initiative rather than selection from among a number of available options, it is no longer a matter of evaluating the different tradeoffs as a consumer but instead about negotiating between the moral interests of different participants (where one 'participant' may represent the 'public interest'). There are strong reasons to believe that, in the case of data initiatives in which questions of public interest are at stake, involving those with interests in the design and conduct of the initiative is preferable to simply offering pre-determined options, not least because there is a public interest in the optimisation of outcomes for all. Rather than treating norms and values as fixed or imponderable, such an approach may offer a way of bringing these into play in order to produce a new equilibrium within a particular governed context.

5.18 One way in which decision makers have sought to understand prevailing norms in society is through research into public opinion. A number of qualitative and quantitative exercises, including surveys, consultations and public dialogue events, have been carried out into the use of data for biomedical research and other purposes. These have mainly been sponsored by research funders, who typically want to understand how use may legitimately be made of available data resources and to promote public trust in research. A body of received wisdom that claims support from these findings has built up in the UK about the use of stored information. This suggests that there is a broad majority of public support for information to be used for a range of secondary purposes (including in biomedical research and health service improvement) *so long as* people are asked about this. However, this support is said to fall away to a significant extent where they are not asked, or where the research involves private companies operating for profit.[274] The current reality of medical research is that it relies upon

[273] There is a well-known apothegm in information technology that if someone is not paying for the service, 'they're not the consumer, they're the product', meaning that the reason they are able to obtain the service without payment is that the service provider is able to make a return by selling data provided (when registering for or using the service) or somehow monetise it thereafter. See: http://blogs.law.harvard.edu/futureoftheinternet/2012/03/21/meme-patrol-when-something-online-is-free-youre-not-the-customer-youre-the-product/.

[274] The published evidence comes from a mixture of commissioned market research and academic social science. See Hill EM, Turner EL, Martin RM, and Donovan JL (2013) "Let's get the best quality research we can": public awareness and acceptance of consent to use existing data in health research: a systematic review and qualitative study *BMC Medical Research Methodology* **13(1)**: 72, available at: http://www.biomedcentral.com/1471-2288/13/72. The Consultation response by Ian Herbert (available at: www.nuffieldbioethics.org/project/biological-health-data/evidence-gathering/) also provides a very helpful list of relevant research, drawn from his 2012 report *Fair shares for all: sharing and protecting electronic patient healthcare data,* available at: http://phcsg.org/publications/fair-shares-for-all-final/. Subsequent relevant research includes Ipsos MORI *Dialogue on data: exploring the public's views on using administrative data for research purposes*, available at: http://www.esrc.ac.uk/_images/Dialogue_on_Data_report_tcm8-30270.pdf; CM Insight research for Wellcome Trust (2013) *Summary report of qualitative research into public attitudes to personal data and linking personal data,* available at: http://www.wellcome.ac.uk/About-us/Publications/Reports/Public-engagement/WTP053206.htm; Sciencewise ERC (2014) *Big data: public views on the collection, sharing and use of personal data by government and companies*, available at: http://www.sciencewiseerc.org.uk/cms/assets/Uploads/SocialIntelligenceBigData.pdf?utm_medium=email&utm_source=Ricardo-AEA+Ltd&utm_campaign=4132283_SWper cent2fMonthly_digestper cent2fNAOper cent2fED57482500_May+2014&dm_i=DA4,2GKHN,1SUKGU,8Y5R6,1); Ipsos MORI's research on trust in data and

clinical and commercial research collaborations and partnerships to develop innovations for the health care system (see chapter 2 above). Difficult issues therefore arise, for example, if data are collected initially through the health care system or academic institutions and then access is given to a pharmaceutical company as a part of collaborations.[275] It is not clear, however, that these details are considered when people give their opinions to researchers.[276] This evolving area would benefit from further research. As some of the academic papers acknowledge, the research may suffer from a sample bias that favours participants who are positive about research.[277]

5.19 While public opinion research gives a valuable indication of some relevant norms its limitations as a support for decision making must be understood. It presents decision makers with a number of difficulties, not least the value that should be given to opinions from the interested and the disinterested, those informed by morally relevant interests or those informed by merely prudential or even immoral ones, those that are top-of-the-head and those that result from earnest and prolonged deliberation.[278]

5.20 Public opinion research typically provides evidence in decision making where the actual decisions are taken elsewhere, in contexts to which access is restricted and through procedures that are often obscure.[279] In the case of data initiatives there are reasons to give particular attention to people whose morally relevant interests are engaged, not simply as a source of evidence of the norms that must be managed, but as collaborators in the elaboration of the whole system. This means not only the professionals who deliver it and those who stand to benefit but also – and perhaps most importantly – people whose privacy and welfare are at stake. This can be prudent because their decisions about whether to participate or not can enable or frustrate the initiative once it is established. But it also expresses respect for them as persons who have morally significant interests and the capacity to contribute positively to the shaping of the social world.

5.21 In chapter 3 we discussed a number of ways of resolving problems of collective action. We mentioned common good, social contract and utilitarian approaches to these problems and noted that all had advantages and disadvantages.[280] If individuals are to be included there are two main ways in which their interests may be brought to bear

attitudes toward data use/data sharing for the Royal Statistical Society (2014), available at: http://www.statslife.org.uk/news/1672-new-rss-research-finds-data-trust-deficit-with-lessons-for-policymakers.

[275] See: Ipsos MORI for MRC (2007) *The use of personal health information in medical research*, available at: http://www.mrc.ac.uk/documents/pdf/the-use-of-personal-health-information-in-medical-research-june-2007, which found pharmaceutical companies to be among the least trusted organisations where personal health information is concerned (trusted by just 6 per cent of the population, confirming long-standing trends research on general trust in professions).

[276] Clemence M, Gilby N, Shah J, *et al*. (2013) *Wellcome Trust monitor wave 2: tracking public views on science, research and science education*, available at: http://www.wellcome.ac.uk/stellent/groups/corporatesite/@msh_grants/documents/web_document/wtp053113.pdf.

[277] For such an acknowledgement, see Hill E, Turner E, Martin R and Donovan J (2013) "Let's get the best quality research we can": public awareness and acceptance of consent to use existing data in health research: a systematic review and qualitative study *BMC Medical Research Methodology* **13**(1): 72, available at http://www.biomedcentral.com/1471-2288/13/72.

[278] On the problems and paradoxes of public engagement, see Nuffield Council on Bioethics (2012) *Emerging biotechnologies: technology, choice and the public good*, especially chapter 5 ('Public perspectives'), available at: http://www.nuffieldbioethics.org/emerging-biotechnologies.

[279] This is self-consciously true of government which reserves a 'safe and protected space' for policy making. See also: Nuffield Council on Bioethics Nuff' said blog (4 February 2013) *Engagement in open policy making; or how to train your academic*, available at: http://nuffieldbioethics.org/blog/2013/engagement-in-open-policy-making-or-how-to/.

[280] See paragraph 3.19ff.

fairly, namely through aggregation or deliberation.[281] The basic difference is that aggregative approaches assume that people have stable or rational preferences from which the preferences of the group can be deduced, whereas deliberative approaches place value on the fact that individual moral interests in collective outcomes may be transformed as they encounter each other in a context of reasoned argument and discussion.[282]

5.22 A deliberative approach may address the complex problem of resolving varied, confused, and possibly conflicting norms and interests into a coherent and mutually acceptable set of common aims and expectations.[283] This demonstrates respect for persons because (and insofar as) it arises from the face-to-face encounter between moral agents who recognise and treat each other as such.[284] It also recognises that the elements of any solution (the norms, mechanisms to account for diverse values and forms of governance) are interrelated and co-dependent.

Principle 3 – Participation

The set of expectations about how data will be used (or re-used) in a data initiative, and the appropriate measures and procedures for ensuring that those expectations are met, should be determined with the participation of people with morally relevant interests. This participation should involve giving and receiving public accounts of the reasons for establishing, conducting and participating in the initiative in a form that is accepted as reasonable by all. Where it is not feasible to engage all those with relevant interests – which will often be the case in practice – the full range of relevant values and interests should nevertheless be fairly represented.

5.23 The principle of participation requires decision makers not merely to imagine how people with morally relevant interests *ought* to expect data to be used but to take steps to discover how they do, *in fact*, expect data to be used and to engage with those expectations. The participation of people with interests at stake in the design of data initiatives gives decisions a strong claim to legitimacy. Independently of the outcome, participants, and the wider public, are more likely to accept the process as being a fair and respectful way of resolving any differences between them with regard to decisions that may affect them all.[285]

5.24 The outcome of such a process is by its nature provisional, a 'working solution'. Circumstances or expectations may change; they may prove unrealistic; there may be

[281] Both aggregation and deliberation have a claim to procedural fairness. See also: Knight J and Johnson J (1994) Aggregation and deliberation: on the possibility of democratic legitimacy *Political Theory* **22(2)**: 277-96; Gutmann, A and Thompson D (2004) *Why deliberative democracy?* (Princeton: Princeton University Press).

[282] Daniels and Sabin note that moral values are not simply like tastes or preferences: aggregation "seems insensitive to how we would ideally like to evolve moral disputes, namely through argument and deliberation." Daniels N and Sabin J (1997) Limits to health care: fair procedures, democratic deliberation and the legitimacy problem for insurers *Philosophy and Public Affairs* **26(4):** 303-50, at page 338.

[283] There are a number of well-known advantages and disadvantages of deliberative approaches. Some of these are discussed in Parker M (2007) Deliberative bioethics, in Ashcroft RE, Dawson A, Draper H, and McMillan JR (Editors) *Principles of health care ethics* (Chichester: John Wiley& Sons), pp185-91.

[284] Among the implications of the principle of respect for persons is not only a respect for the things that persons value but also a recognition that they themselves can, when they are enabled to do so, take responsibility with others for actively creating the conditions that, for example, manage their privacy and promote the common interest. Furthermore, it is an implication of respect for persons that this is morally preferable (*i.e.* more respectful) when this is done by those people themselves or through their nominated representatives.

[285] See also guidance outlined in the recent report for Sciencewise, *Data policy and the public: shaping a deeper conversation*, available at: http://www.sciencewise-erc.org.uk/cms/assets/Uploads/Data-policy-and-the-publicJan-2015.pdf.

improvements or failures. Consequently, though the process may be provisionally concluded with the production of a publicly statable set of expectations about how data will be used and governed, there is often a need for continuing reflection and review. The principle of participation therefore applies equally to the establishment of a data initiative and to its continuing governance.

Accounting for decisions

5.25 Deliberation is a social activity that requires participants to engage in a common 'public' discourse through which they can account to each other for the positions they take.[286] This accounting, given through a face-to-face encounter between moral agents with a common purpose, can build trust among people with different interests, allow them to discover where trust may be placed intelligently and design the terms of a data initiative accordingly.[287] Like any social activity, however, deliberation is vulnerable to abuse, domination or capture by those with power.[288] It is also at risk from cognitive and social effects such as framing and 'groupthink'.[289] There are two main antidotes to these effects: the internal commitment to fair conduct, and openness to external scrutiny and revision. It is important, therefore, that the initiative is embedded in broader system of accountability that allows for challenge and dispute resolution. The notion of accountability emphasises the extension of the processes of moral deliberation, namely the requirement to 'give an account' that is intelligible and acceptable to the person to whom it is given. It also emphasises the function of 'holding to account', namely the imposition of a judicial power with legitimate authority.

> **Principle 4 – Accounting for decisions**
>
> **A data initiative should be subject to effective systems of governance and accountability that are themselves morally justified.** This should include both structures of accountability that invoke legitimate judicial and political authority, and social accountability arising from engagement of people in a society. Maintaining effective accountability must include effective measures for communicating expectations and failures of governance, execution and control to people affected and to the society more widely.

5.26 The principle of 'accounting for decisions' emphasises two forms of accountability that face in notionally divergent directions. The first is formal accountability, through regulatory, judicial and political procedures. In a democratic society this should be

[286] This notion of the public use of reason that claims the right of reason to challenge authority draws on a modern tradition from Kant (*What is enlightenment?*) that has found different contemporary expressions in European thinkers such as Jürgen Habermas (*Moral consciousness and communicative action*) and proponents of deliberative democracy (e.g. Amy Gutmann; Norman Daniels).

[287] For a discussion of the significance of trustworthiness and the placing of trust, see: O'Neill O (2002) *Autonomy and trust in bioethics* (Gifford Lectures 2001) (Cambridge: Cambridge University Press). It seems preferable to construct an initiative in conditions that foster trust, and that allow the interrogation of and making provision for the limits of trustworthiness, rather than simply to treat participants as consumers in a free market of more or less trustworthy initiatives.

[288] Those able to exploit differentials of knowledge, articulacy, economic and political power, for example.

[289] On framing see Tversky A and Kahneman D (1981) The framing of decisions and the psychology of choice *Science* **211(4481)**: 453-8; on 'groupthink' see Turner ME and Pratkanis AR (1998) Twenty-five years of groupthink theory and research: lessons from the evaluation of a theory *Organizational Behaviour and Human Decision Processes* **73(2/3)**: 105-15, available at: http://homepages.se.edu/cvonbergen/files/2013/01/Twenty-Five-Years-of-Groupthink-Theory-and-Research_Lessons-from-the-Evaluation-of-a-Theory.pdf.

accessible to those with a relevant moral interest (potentially up to the level of the whole political community where the data initiative in question raises issues of public policy). The second is accountability to the broader mass of moral stakeholders who, perhaps for practical reasons, cannot participate directly in the formation or governance of the initiative. Periodic engagement with a broader public, for example, provides a way of ensuring that they are fairly represented and that governance is not 'captured' by partial interests. Formal and social accountability are closely linked, as a system that loses its social mandate will come under pressure politically. However for either to work, people (and in particular the dispersed and potentially vulnerable participants whose information is used) need effective means of learning what has happened to their data. This inevitably requires a careful consideration of the design not just of the systems that a given data initiative calls into being, but of the institutional structures in which they are embedded, to ensure that there is sufficient transparency and incentives to report abuses and to rectify them.[290]

Conclusion

5.27 The question we posed at the end of chapter 3 was: 'How may we define a set of morally reasonable expectations about how data will be used in a data initiative, giving proper attention to the morally relevant interests at stake?' Different data initiatives will have different objects, engage different moral values and interests, and give rise to different sets of expectations. We have suggested that a good answer may be given through a procedure of moral reasoning that is bounded by respect for persons and human rights, involves the participation of those representing the range of morally relevant interests at stake, and is embedded in institutional and social procedures for accountability.

5.28 With these principles in mind, in the following chapters we consider a number of concrete data initiatives in order to highlight instances of good practice and areas where attending to these principles may offer better solutions than those that have been found.

Box 5.1: Summary – ethical principles for data initiatives

The use of data in biomedical research and health care should be in accordance with a publicly statable set of morally reasonable expectations and subject to appropriate governance.

- **The set of expectations about how data will be used in a data initiative should be grounded in the principle of respect for persons.** This includes recognition of a person's profound moral interest in controlling others' access to and disclosure of information relating to them held in circumstances they regard as confidential.
- **The set of expectations about how data will be used in a data initiative should be determined with regard to established human rights.** This will include limitations on the power of states and others to interfere with the privacy of individual citizens in the public interest (including to protect the interests of others).
- **The set of expectations about how data will be used (or re-used) in a data initiative, and the appropriate measures and procedures for ensuring that those expectations are met, should be determined with the participation of people with**

[290] See recommendation 3.

morally relevant interests. This participation should involve giving and receiving public account of the reasons for establishing, conducting and participating in the initiative in a form that is accepted as reasonable by all. Where it is not feasible to engage all those with relevant interests – which will often be the case in practice – the full range of values and interests should nevertheless be fairly represented.

- **A data initiative should be subject to effective systems of governance and accountability that are themselves morally justified.** This should include both structures of accountability that invoke legitimate judicial and political authority, and social accountability arising from engagement of people in a society. Maintaining effective accountability must include effective measures for communicating expectations and failures of governance, execution and control to people affected and to the society more widely.

Chapter 6

Data initiatives in health systems

Chapter 6 – Data initiatives in health systems

Chapter overview

This chapter discusses developments in the functions and purposes of health information systems and draws lessons from some specific initiatives.

Health-care IT systems were originally introduced to facilitate basic administrative tasks but have evolved to provide business intelligence for service improvement and to support observational research. These come together in the concept of a 'learning health system'.

Information systems in the English NHS have moved towards a centralised approach now overseen by the Health and Social Care Information Centre (HSCIC). Debate around the 'care.data' programme focused attention on assumptions about the relationship between privacy norms relevant to NHS patients and the legal norms under which HSCIC operates. It highlighted the absence of reflection on this difference and of a capacity to address it, and raised questions about how the rights of individuals were respected, resulting in a damaging loss of public and professional trust.

The Scottish Informatics Programme involved initial public consultation to identify relevant social norms. It developed a model of bespoke data linkage using a safe haven, subject to proportionate governance that takes into account both privacy risk and public benefit. Governance refers to an explicit, potentially revisable, statement of guiding principles and best practices that takes account of the findings of public engagement.

The 100,000 Genomes project involves linking data from genome sequencing with individuals' NHS records to investigate some cancers and other diseases. Authorised researchers from all sectors may access a firewall-protected, pseudonymised dataset with the broad consent of patients. The dataset is administered by a Government-owned company, Genomics England Ltd. The claimed public interest lies explicitly in securing economic as well as scientific and therapeutic benefits, by stimulating the commercial sector.

A number of recommendations are made in relation to defining reasonable expectations of data use, accounting for data use and delivering outcomes in the public interest.

Introduction

6.1 This chapter and the following one examine concrete contexts in which data initiatives have taken shape. In this chapter we will show how the introduction of information technology has wrought a transformation in our understanding of whole health care systems, particularly the NHS in the UK. From being a system focussed on the delivery of health care to those in immediate need, the introduction of IT in the health services has broadened and blurred the horizons of the system.

6.2 In the context of health care, perhaps more than elsewhere, the potential uses of data have influenced choices about information technology infrastructure, which is only partly about the delivery of health care. Other functions include the delivery of research insight, better resource allocation, support for innovation and evidence to inform policy. Data and IT initiatives also seek to extend the boundaries of health care beyond responding to illness, for example, to predicting who might get ill before they experience any symptoms and to understanding the rich combination of medical, behavioural and environmental factors relevant to health conditions.

6.3 The health services in the UK offer a number of examples of data initiatives at different scales, from the macro (e.g. the National Programme for IT – NpfIT), through the meso (e.g. the '100,000 Genomes' project) to the micro (e.g. a clinical evaluation of a particular intervention). These have not always gone smoothly. While their possible failings as infrastructure projects are of moral relevance (when they miss opportunities to deliver benefits, or when they use public resources inefficiently, or undermine public trust) our interest will be primarily to identify good practice and areas for improvement in how they manage the relationship between underlying norms of access and disclosure, respect for individual values and interests, and governance in the public interest. Drawing on our discussion so far, the critical decisions for each initiative will be those that place them at different points on the following critical axes:

- The arrangements for storage (whether data are retained close to the point of collection or gathered together in safe havens or in a single, central repository)
- The arrangements for data disclosure/access (whether data are published, subject to controlled disclosure, controlled access or mediated access)
- The role given to individual patients (from explicit individual consent, through implicit consent with opt-out, to no individual authorisation)
- The range of users and purposes approved for access/disclosure (from restricting access to particular classes of users, such as academic researchers, or an expectation that the broader public interest will be served by any responsible use of data, including by commercial users).

6.4 The optimum relationship between norms, private freedoms and public objectives may be found through consideration of the interests and values at stake in the practical context. Health services in the UK have a long and complicated history, and giving attention to the way in which information use has developed will help us to understand the interests and drivers involved.

IT innovation and developing information requirements

6.5 As in other sectors, health care has a number of basic information needs that are increasingly met by computerised systems. The way in which these systems were introduced to health care paralleled other sectors in many respects, usually starting with routine administrative tasks.

Tracking patients

6.6 General Practitioners (GPs) were relatively quick to adopt computers to manage patient medical records, primarily to facilitate repeat prescriptions and later to manage complete patient records. Some of the earliest systems in the UK were developed on microcomputers in the late 1970s by technically adept GPs themselves.[291] The Department of Health (DH) introduced standards ('requirements for accreditation') in the late 1990s, along with government subsidies for compliant systems. These ensured that public administrative requirements would be incorporated alongside clinical requirements. In 2000 GPs were officially allowed to stop keeping paper records and

[291] One of the first GP computer systems, the Integrated General Practice system was developed by an Essex GP in collaboration with IT staff from a local shoe factory (an example of skills clustering leading to innovation) and successfully marketed to other practices via a venture capital funded company (Value Added Medical Products, or VAMP). For a brief history, see: Health Service Journal (12 August 1999) *VAMP comes alive*.

move to a paperless system although many had converted to a fully electronic system before then, albeit without necessarily having transferred legacy data from the retained paper files. In 2003 the new GP contract transferred effective ownership of GP systems away from GPs to Primary Care Organisations, which thus became the software vendors' customer.[292]

6.7 Hospitals adopted electronic patient administration systems from the late 1960s to manage patient appointments and track patients across the hospital, but these were not primarily for managing medical data, which often continued to be held separately in paper notes. Patients were usually identified by an assigned 'hospital ID' and, where available, their NHS Number.[293] Hospitals might also have local clinical or laboratory systems to capture information directly from different types of machine, such as analytical instruments, or from clinicians. These vary from simple spreadsheets of results to complex scientific systems. However, they would often be only loosely integrated with the central patient administration system, perhaps having the patient's hospital ID as the only common feature. Regional health care systems often developed on the back of national requirements to facilitate collection of datasets in the days before universal use of the Internet (or a Virtual Private Network such as NHSNet). The data collected could then be used for regional planning and commissioning.

6.8 The English NHS Number used to be managed electronically through the National Strategic Tracing Service (NSTS), introduced in 1999 to allow NHS organisations to discover or check the NHS Number for individuals. This followed the recommendation of the first Caldicott Committee to increase the use of the NHS Number so that simple anonymisation could be used for administrative analyses of patient data, minimising the need to transmit data with full identifiers in order to link and de-duplicate datasets.[294] There were a number of restrictions on access to NSTS to prevent it being used to re-identify anonymised records or to trace individuals. Unfortunately, these were so cumbersome or intrusive that many health regions set up their own 'master patient index' to provide a more rapid service locally, while using the NHS number only for those patients coming from outside their region.[295] In Scotland a separate Community Health Index (CHI) Number is used to identify patients nationally, which must be used for each health care episode.[296] Some countries, such as the USA, have no such universal health number, and it is often argued that this makes national collation of data difficult.[297] This argument is somewhat suspect, however: universal naming systems are often sold as a means of improving data quality but in practice there are many difficulties and they rarely fulfil their potential.[298] Similarly, while some people oppose universal numbering on privacy or even religious grounds, most patients are easily identified even by traditional methods such as name and date of birth.

---

[292] Primary Care Organisations (PCOs) are NHS England in England, Health and Social Services Boards in Northern Ireland, Local health boards in Wales, and primary care divisions within area health boards in Scotland. See: http://www.rcgp.org.uk/training-exams/gp-curriculum-overview/~/media/Files/GP-training-and-exams/Curriculum%20previous%20versions%20as%20at%20July%202012/curr_archive_4_2_IMT_v1_0_mar06.ashx.

[293] The old-style NHS Number (XXXXX NNN) originally appeared on a baby's birth certificate; the new NHS Number (NNN-NNN-NNNN) is allocated from birth (see: http://systems.hscic.gov.uk/nhsnumber/patients/yournumber).

[294] 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.

[295] See, for example: http://www.tamesidehospital.nhs.uk/documents/EnsuringAccuratePatientInformationPolicy.pdf.

[296] Those in the Border regions may have both, as their GP may be in England but their nearest hospital in Scotland. http://www.ehealth.scot.nhs.uk/support-documentation/document-holder2/.

[297] See, for example, http://www.cardiosource.org/en/Advocacy/Issues/Health-Information-Technology/ACC-Policies-and-Activities/Unique-Patient-Identifier-Principles.aspx.

[298] See "Naming" in Anderson R (2008) Security engineering – a guide to building dependable distributed systems (Wiley), available at: http://www.cl.cam.ac.uk/~rja14/book.html.

Observational research

6.9 Various GP system providers have set up research databases based on proprietary extracts from their systems. First was VAMP (now InPractice Systems) whose original business model was to provide free PCs to GPs and then sell access of the aggregated data to pharmaceutical companies. This facility was donated to the Department of Health in 1994, where it became the GP Research Database. Most GP systems support a facility, MIQUEST, introduced in 1996, that allows queries to be run by participating GP systems at the discretion of the GP practice and return aggregated data for research or other purposes. Exceptionally, individual-level data could be requested, but only age and postal area (not full postcode) would be returned to limit identifiability. In 2012, the Clinical Practice Research Datalink (CPRD) was established. This combines the activities of the MHRA's General Practice Research Database (GPRD) and the Department of Health's NIHR Research Capability Programme.[299] The CPRD is accessible to researchers for a variety of research purposes, including observational research and planning interventional trials and can support clinical decision making by providing clinicians with relevant, real-world data to inform their consultations with patients.[300]

6.10 A number of condition-specific registries have been developed over the years to help to understand the effect of those conditions and provide integrated services to those affected. For example, cancer registries (of which there are 11 run by eight regional organisations, also confusingly called 'registries').[301] These cancer registries link with the Office of National Statistics death register to identify when patients on the various registries die (so mortality statistics can be generated – and compared with other countries' experience).[302] Recruitment to registries used to be seen by recruiters as part of clinical care with the possibility for patient opt-out rather than specific opt-in consent. It was a challenge to this presumption (from some wording in the GMC Confidentiality guidance in 2001) that led to the introduction of Section 60 of the Health & Social Care Act 2001 (now Section 251 of the NHS Act 2006) and the Health Service (Control of Patient Information) Regulations 2002 to enable the use of this information for research without specific consent under some conditions (see chapter 4). The path established by cancer registries subsequently provided a template for other specialties.

Performance evaluation and improvement

6.11 Organisations need feedback in order to improve quality and ensure safety. For this purpose, health systems generally, health organisations separately, clinical teams as well as individual clinicians compare their performance with the norm or with others' performance. If there is a public interest in health care, there is also a public interest in supporting good practice and identifying and eliminating poor provision. In the UK, health care intelligence providers such as Dr Foster have created a business in monitoring and analysing outcomes so that patients and funders can see which

[299] See: http://www.cprd.com/home/.
[300] Ibid.
[301] Some people reserve the term 'register' for the database and 'registry' for the organisation where registering takes place.
[302] One of our reviewers drew to our attention the fact that delays in death registration in England, Wales and Northern Ireland can have the effect of undermining the evidence-base for epidemic monitoring, record-linkage research and policy development. The Royal Statistical Society has recommended that registration of the fact of death (pending determination of cause of death) should take place in a timely fashion (http://www.publications.parliament.uk/pa/cm201213/cmselect/cmpubadm/406/406we08.htm).

hospital departments have the lowest mortality rates.[303] On the one hand, this can be useful in detecting failing institutions; on the other, if not interpreted with reference to the context, it can penalise hospitals that tackle more complex, high-risk cases turned down by others. Nevertheless, the trend for increased public monitoring of medical performance now appears to be well established. In the USA, websites like ZocDoc and Vitals are establishing patient feedback as a norm, as TripAdvisor has done for hotels.[304]

6.12 The distinction between striving for continual improvements in productivity through efficiency and innovation, and improving health through developing better patient information and treatment – that is, through research – begins to disappear in the concept of a so-called 'learning health care system'. While these two senses of 'service improvement' (business productivity and improved care) have always been related, they have been the subject of separate information and governance systems. The integration of these systems represents a more recent innovation.

6.13 Three moral justifications have been suggested for this integration: the need for a just system, for high quality care, and for a system that supports economic well-being. [305] These are said to entail moral requirements on both clinicians and patients to participate in research aimed at service improvement ('learning').[306] The key claim is that there is no 'do nothing' option because of the growing pressure of circumstances: at an individual level, people are getting ill (healthy people becoming ill, ill people getting more ill, and new illnesses appearing); at a system level, resources to meet these demands are more or less tightly constrained (although the level of constraint is, of course, a result of political decisions). Proponents of learning health systems may accept that the risk of data abuses increases as a result of a learning activity, and that such activities may impose additional burdens on patients (such as extra visits to clinics). However, they believe that these can be minimised through appropriate controls and that the residual risk is justified.[307]

Public administration and service delivery

6.14 Over the last 25 years, there has been continuous pressure for more administrative access to records for purposes such as clinical audit, service planning and cost control.[308] Indeed, these may often have been the real drivers behind centralisation efforts, with 'research' promoted as a desirable further purpose and often as the public rationale. The biggest single driver of centralisation in the UK was, however, the

---

[303] See: www.drfoster.com.

[304] The Economist (26 July 2014) *DocAdvisor*, available at: http://www.economist.com/news/international/21608767-patients-around-world-are-starting-give-doctors-piece-their-mind-result/.

[305] See Faden RR, Kass NE, Goodman SN, *et al*. (2013) An ethics framework for a learning health care system: a departure from traditional research ethics and clinical ethics *Hastings Center Report* **43(s1)**: S16-S27, available at: http://onlinelibrary.wiley.com/doi/10.1002/hast.134/full.

[306] Faden et al. (op.cit) offer a Rawlsian account of common good to justify a 'norm of common purpose'. This is based on "the reciprocal obligation that arises among strangers who occupy the role of patient over time" (at page 23). In their system, the first 4 principles they advance are designed to protect individual rights and freedoms, and limit the claims that can be made of any individual participant.

[307] Ibid.

[308] The collection of UK national statistics was a recommendation of the Körner Committee, which carried out a major review of health service information between 1980 and 1984. In the 1980s the NHS Exeter system was developed to bring together data sets from across England with 21 regional centres collating data and transmitting them to the national centre at Exeter (building on an earlier initiative to develop a national PAS system there). The data covered a range of topics from GP registrations, organ donor registration, national screening programmes, to GP capitation payments. The Exeter system is now known as the National Health Applications and Infrastructure Services (NHAIS).

purchaser/provider split, introduced in 1991.[309] The fact that medical procedures performed in hospitals had to be paid for meant that information about the procedure, the patient, and the cost had to flow on a large scale. This led to the establishment of the Hospital Episodes Statistics (HES) database, amongst others, to track activity across the NHS as a whole. [310] The HES database was largely populated using data from the NHS-wide Clearing Service which handled payments. Otherwise data was only available on a piecemeal basis, gleaned from individual audits or research studies.[311]

6.15 In 1992, the Department of Health published the Information Management and Technology (IM&T) Strategy, whose vision was a single electronic health record (EHR) for each patient, accessible to everyone working within the NHS.[312] The British Medical Association objected that making patient records available beyond the teams responsible for a patient's direct care would compromise both safety and privacy.[313] This led to the first Caldicott review and the creation of 'Caldicott Guardians'.[314] In 1998, NHS England released a new IT strategy, *Information for Health*, promoting the adoption of both EPRs that would be held at a particular care provider and EHRs to be shared across all providers.

6.16 In 2002, with high-profile backing from the (then) Prime Minister, the Department of Health announced a major infrastructure initiative, the National Programme for IT (NPfIT). This was to be implemented by Connecting for Health, a new agency established to drive forward 'ruthless standardisation' across the NHS in England.[315] (Wales and Scotland had separate initiatives taking rather different approaches.) The aspiration was to deliver the EPR systems promised in *Information for Health*, in a time frame of about three years.[316]

6.17 The NPfIT supported a number of features. Some, like the Picture Archiving and Communication System, which provided an accessible electronic archive for radiology images, proved to be both useful and successful.[317] Others, like the 'Choose and Book'

---

[309] NHS and Community Care Act 1990, available at: http://www.legislation.gov.uk/ukpga/1990/19/contents. The internal market was brought to an end in 1997 but the split continued through the introduction of Primary Care Groups (then Primary Care Trusts).

[310] Initially data were collected annually, then quarterly. At the time of writing they are collected on a monthly basis (http://www.hscic.gov.uk/hes).

[311] It is notable that it was only in 1999 that Professor Sir Brian Jarman at Imperial College used HES data to produce comparative mortality statistics, thus creating quite a storm about why these figures were unreliable, though few suggestions on how to improve on them and reduce adverse incidents. See: Jarman B, Gault S, Alves B, et al. (1999) Explaining differences in English hospital death rates using routinely collected data British Medical Journal 318: 1515-20, available at: http://www.bmj.com/content/318/7197/1515.full.

[312] Department of Health (1992) Getting better with information: an IM&T Strategy for the NHS in England (London: HMSO). See: http://www.intosaiitaudit.org/9899371.pdf.

[313] R Anderson (1996) Security in clinical information systems, British Medical Association, available at: http://www.cl.cam.ac.uk/~rja14/policy11/policy11.html.

[314] See paragraph 2.43.

[315] Department of Health (2002) *Delivering 21st century IT support for the NHS*, national strategic programme, available at: http://webarchive.nationalarchives.gov.uk/20130107105354/http://www.dh.gov.uk/prod_consum_dh/groups/dh_digitalassets/@dh/@en/documents/digitalasset/dh_4071684.pdf.

[316] Computer Weekly (18 Feb 2008) *Secret papers reveal Blair's rushed NpfIT plans*, Computer Weekly, available at: http://www.computerweekly.com/blogs/public-sector/2008/02/secret-papers-reveal-blairs-ru.html; *BBC News* (25 October 2007) *NHS IT time-frame 'ludicrously tight'*, available at: http://news.bbc.co.uk/1/hi/health/7061590.stm; Campion-Awwad O, Hayton A, Smith L and Vuaran M (2014, MPhil Public Policy 2014, University of Cambridge) The national programme for IT in the NHS – a case history, available at: http://www.cl.cam.ac.uk/~rja14/Papers/npfit-mpp-2014-case-history.pdf, at page 13.

[317] Picture Archiving and Communications System (PACS), though not part of the original specification for NPfIT, moved radiology images from film-based processing to electronic recording. The systems were generally set up on a regional basis as a shared service to hospitals (with feeds to GP practices as well) for economies of scale. In this case, the technology was well developed if not widely adopted: a few leading sites (mainly in the USA) had taken on such facilities and proved them to

electronic booking system to enable a GP to book a specialist appointment for a patient while the patient was still in the surgery, were much less so. One feature strongly promoted by Connecting for Health was the Summary Care Record (SCR), a nationwide system containing a GP record summary (initially, current prescriptions and allergies) that would facilitate out-of-hours care and could also enable patients to view their own records via a mechanism called HealthSpace.

Box 6.1: The Summary Care Record

The first national care record to be implemented was the Scottish Emergency Care Summary (ECS), starting in 2002 and achieving near complete coverage by 2007.[318] The role of the ECS was simply to allow the most immediate medical details to be available in an emergency, by out-of-hours GPs, clinicians at Accident and Emergency (A&E) departments in hospitals, or call-centre staff at NHS24.

It was limited to basic allergies, adverse drug reactions, and current (<12 months) medications data. Patients were able to opt out and would be asked for 'consent to view' the record by clinicians treating them.[319]

The introduction of the Summary Care Record (SCR) in England was more ambitious, carried out on a reduced timescale (2½ years) as the first stage in a move to full electronic health records and involved a wholesale replacement of almost all NHS systems as part of the National Programme for IT. Initially it would contain a summary of information from the patient's GP system and referral and discharge correspondence (though not clinical details), and, later, Common Assessment Framework (CAF) care plan documents to permit 'joined up' care delivery between health and social care services.

The decision to provide access to the system through a browser-based stand-alone application raised security and confidentiality concerns as it separated the access to the SCR from any local governance process. It also created some confusion between the SCR as a record and the SCR 'application', and how each might be accessed.

The SCR has been bedevilled by technical and policy difficulties that derived largely from the scale of the ambition and the number of variables, including arguments about what options individuals might exercise (which information could be uploaded from detailed care records, including, for example, 'sealed'/'sealed and locked' envelopes), the legitimacy of default options (the RCGP and BMA argued for an 'opt in' approach, for example), and how individuals' choices could be given effect through the technical architecture of the system.

be effective. The technology also required relatively little change to clinical practice (calling up an image on screen was often easier than obtaining the film) but it enabled a number of further enhancements. It raised the possibility of rationalising radiology reporting services across hospitals, rather than requiring each hospital to have its own staff; teams could be merged, relocated, or even outsourced if necessary, though this led to important debates about patient understanding of how their images might be shared and whether reduction of clinician-radiologist interaction mattered.

[318] Although this was the first national record, the Hampshire and Isle of Wight Care Record was implemented even earlier. While all 14 Health Boards participated, in 2007 some 7 GP practices had chosen not to participate, so their patients' records where not present. It is not clear if this was for reasons of principle or practicalities, such as incompatible systems.

[319] Nevertheless the system was the subject of a high-profile confidentiality failure when a doctor accessed the records of a number of public figures. See: "Medical records of Gordon Brown and Alex Salmond hacked" Daily Record, 1 March 2009, available at: http://www.dailyrecord.co.uk/news/scottish-news/medical-records-of-gordon-brown-and-alex-1011941.

6.18 However, a review of the SCR in 2010 found little evidence of benefit. It found that even staff in Accident and Emergency services did not often access it and that patients also made little use of HealthSpace.[320] A significant number of people also opted out of the SCR owing to concerns about data would be handling.[321] (Criticism that the patient information circulated in pilot areas made opting out difficult led the campaign 'The Big Opt Out' to produce an opt-out letter, which was extensively downloaded.) The NPfIT was officially dismantled in 2013 with major objectives unachieved and amid public criticism of management failure and cost overrun.[322]

Box 6.2: Connecting for Health – an assessment

NPfIT has been described as 'the largest ever civilian IT project failure in human history'.[323] It has been documented by many articles in the computing and health IT press and by successive parliamentary committee inquiries. A recent case history classifies the problems according to three main themes:[324]

- *Haste.* Politicians and programme managers rushed policy making, procurement and implementation processes, allowing insufficient time for consultation with key stakeholders and failed to deal with confidentiality concerns;
- *Design.* The government pursued an overambitious and unwieldy centralised model, without giving consideration to how this would impact user satisfaction and confidentiality issues; and
- *Culture and skills.* NPfIT lacked clear direction, project management and an exit strategy, meaning that the inevitable setbacks of pursuing such an ambitious programme quickly turned into system-wide failures. Furthermore, the culture within the Department of Health and government in general was not conducive to swift identification and rectification of strategic or technical errors.

One further problem with the centralised commissioning and ownership of health IT is that IT suppliers respond to the priorities of the customer rather than the clinical user of the system, which allowed the collection of data for central use to compete with the objective of improved care or efficient local service delivery.[325]

[320] Greenhalgh T, Stramer K, Bratan T, *et al*. (2010) *The devil's in the detail: final report of the independent evaluation of the Summary Care Record and HealthSpace programmes*, available at: http://www.ucl.ac.uk/news/scriefullreport.pdf. This was equally true of patient-centred health record offerings from Google and Microsoft: the former has been closed and the latter developed into a more conventional 'white label' product. For a discussion of online personal health records, see Nuffield Council on Bioethics (2010) *Medical profiling and online medicine: the ethics of 'personalised healthcare'* in a consumer age, available at: http://nuffieldbioethics.org/project/personalised-healthcare-0/.

[321] As of 18 December 2014 the number of people who had opted out was 528,034 according to the HSCIC (see: http://www.hscic.gov.uk/article/2220/FOI-disclosure-log).

[322] House of Commons Public Accounts Select Committee (2013) The dismantled National Programme for IT in the NHS (Nineteenth Report of Session 2013–14) HC 294 (London: TSO), available at: http://www.parliament.uk/business/committees/committees-a-z/commons-select/public-accounts-committee/news/npfit-report/.

[323] See: The NHS's national programme for information technology: a dossiers of concerns (2010), available at: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.304.9601&rep=rep1&type=pdf; House of Commons Public Accounts Committee, The national programme for IT in the NHS: progress since 2006, 2nd report of session 2008-09, HC 153, available at: http://www.publications.parliament.uk/pa/cm200809/cmselect/cmpubacc/153/153.pdf; Computer Weekly (22 September 2011) *The world's biggest civilian IT project finally looks to have failed but is the NHS IT failure a surprise?*, available at: http://www.computerweekly.com/blogs/outsourcing/2011/09/the-worlds-biggest-civilian-it-project-finally-looks-to-have-failed-but-it-is-no-surprise.html.

[324] A recent case history classifies the problems according to three main themes. See: Campion-Awwad O, Hayton A, Smith L and Vuaran M (2014, MPhil Public Policy 2014, University of Cambridge) The national programme for IT in the NHS – a case history, available at: http://www.cl.cam.ac.uk/~rja14/Papers/npfit-mpp-2014-case-history.pdf.

[325] Foundation for Information Policy Research, Evidence submitted by the Foundation for Information Policy Resarch (EPR 61), Select Committee on Health, 15 March 2007; at http://www.publications.parliament.uk/pa/cm200607/cmselect/cmhealth/422/422we22.htm.

6.19 The experience of the NPfIT may, nevertheless, be salutary for health care data initiatives more generally because it highlight the risks of external drivers overtaking the establishment of data initiatives (see chapter 2) and of lack of involvement or imbalance of key interests, and the need adequately to address values and norms relating to confidentiality (chapter 5). These lessons are important, moreover, because the SCR service, along with many of the other projects previously overseen by Connecting for Health, have themselves continued, with responsibility for their delivery having been passed to the Health and Social Care Information Centre (HSCIC).

The Health and Social Care Information Centre

6.20 The Health and Social Care Information Centre (HSCIC) has replaced Connecting for Health as the Department of Health's agency for driving the implementation and use of health IT for business intelligence and to equip the NHS to be a learning health system.

Box 6.3: The Health and Social Care Information Centre

The Health and Social Care Information Centre (HSCIC) is an executive non-departmental public body that took its current form following the Health and Social Care Act 2012 (HSCA). The HSCIC was created with the intention of being a national focal point for information collection across health and social care that is responsible for collecting, transporting, storing, analysing and disseminating the nation's health and social care data.[326] Continuing the work of its predecessor, the NHS Information Centre, and absorbing continuing elements of the Connecting for Health programme (e.g. the NHS Spine, the National Back Office and SUS), a key aim of the HSCIC is to provide a trusted 'safe haven' for health data.[327] NHS health data comprises approximately a quarter of this data; the majority is social care data.

The stated aim of the HSCIC is to 'revolutionise' the ability to unlock NHS and care data. It has been given the legal and administrative power and responsibility to:

- Collect information from health and social care bodies
- Hold that information within a secure environment
- Make that information readily available for others to turn into "actionable business intelligence"[328]

At present, the HSCIC collects a range of information, including (monthly) Hospital Episodes Statistics (HES) which relate to in- and out-patient appointments, and accident and emergency admissions. This is intended to be supplemented by primary care data through NHS England's 'care.data' programme (see below).

6.21 The intention is that the HSCIC will hold comprehensive and integrated information about the care patients receive from all parts of the health service, including hospitals and GP practices. It is hoped that by collecting this information and analysing it,

[326] The responsibility to provide information under the Health and Social Care Act 2012 was analysed in Grace J and Taylor MJ (2013) Disclosure of confidential patient information and the duty to consult: the role of the health and social care information centre *Medical Law Review* **21(3)**: 415-47.

[327] Health and Social Care Information Centre (2014) A strategy for the Health and Social Care Information Centre 2013-2015, available at: http://www.hscic.gov.uk/media/13557/A-strategy-for-the-Health-and-Social-Care-Information-Centre-2013-2015/pdf/hscic-strategy-2014.pdf.

[328] Health and Social Care Information Centre, 2012. Exploiting the potential of health and care data, available at: http://www.hscic.gov.uk/ourrole.

researchers can compare the safety of different NHS hospitals, monitor trends in different diseases and treatments and use the data to plan new health services.

6.22 There are two levels of information release relevant to the HSCIC: general publication (i.e. open data) and limited access.[329] Data disclosed for limited access may include potentially identifying individual-level data or, with the subjects' consent or if it is in the public interest, identifying data. Disclosures are governed by agreements about purposes for which the data will be used, who will have access, and whether they may be sold on to third parties.[330] (In theory, the HSCIC could audit compliance with these agreements but in practice it does not and this does not seem to be budgeted for in the costs recovered through the charges made for data extracts.) The HSCIC also offers to link its own data to data supplied by the customer (see Box 6.5 below).

Box 6.4: Data available from HSCIC

Product	Description
Tabulation	A statistical table of aggregate data.
Bespoke extract - pseudonymised	A one-off extract tailored to the customer's requirements of specified data fields containing patient identifiable data, sensitive data items or both.
Standard extract	Cumulative data for the financial year to date, delivered on a monthly basis via a subscription service. Users sign up to receive a year's worth of data, delivered in monthly increments
Bespoke data linkage	A bespoke service linking one or more datasets held by the HSCIC to data supplied by the customer.
Patient status and/or tracking	Products designed to enable customers to receive one-off or on-going notifications of mortality and morbidity events affecting a specified patient cohort.
List cleaning	Validating demographic data to ensure it is accurate and improve linkage outcomes.

Source: Data Linkage and Extract Service: Service Charges 2013/14[331]

6.23 While HSCIC will not exclude particular organisations from using their data, access to datasets will only be granted if the data will be used broadly for purposes beneficial to health or health care.[332] In providing data, the HSCIC operates on a 'cost recovery' basis: it does not charge for data itself but applies charges to cover the costs of processing and delivering its service. The cost is determined by the amount and type of data required (between a few hundred and a few thousand pounds). The HSCIC takes

[329] The Health and Social Care Act 2012 (s.260) imposes a general duty on the HSCIC to publish data that it collects, although identifying data are exempt from this requirement.
[330] 'commercial reuse licenses' enable firms to resell personal health information to third parties. See also the HSCIC register of approved data releases: http://www.hscic.gov.uk/dataregister.
[331] Available at: www.hscic.gov.uk/media/12443/data-linkage-service-charges-2013-2014-updated/pdf/dles_service_charges__2013_14_V10_050913.pdf.
[332] This was clarified in the Care Act 2014, s.122; which amends s.261 of the Health and Social Care Act 2012 to specify that the HSCIC may only further disseminate information for the purposes of 'the provision of health care or adult social care' or 'the promotion of health'. The breadth allowed of these purposes was debated during the passage of the Act.

advice on data access from its Data Access Advisory Group (DAAG), which makes available details of all approved projects that utilised HSCIC's datasets containing 'sensitive' data. Sensitive data may include a patient's NHS number, postcode, date of birth and/or death, physical and mental health, etc. This arrangement was later fortified, but not before the HSCIC had been exposed to considerable political turmoil.

6.24 As we have argued in this report, questions about the terms under which information may be collected and disclosed by the HSCIC need to be answered by first establishing the norms of access and disclosure that govern the kinds of information transactions involved. These questions found a public focus in the debate that arose around the programme to extract GP data for inclusion in the HSCIC database. This brought many GPs and civil society groups into conflict with NHS England, resulting in delayed implementation and redesign of the initiative.

Box 6.5: The NHS England 'care.data' programme

Care.data is a programme commissioned by NHS England and promoted with the caption 'better information means better care'. Its purpose is to bring together routinely collected information from NHS organisations within the HSCIC. Care.data has been particularly focussed on the extraction of data from primary care records (GP data) and widening the range of data collected from secondary care (hospital data), to add to the hospital episode statistics (HES) data already routinely collected by HSCIC.[333]

The collection of GP data is through a new General Practice Extraction Service (GPES), which extracts information stored electronically within GP practices and sends it to the HSCIC.[334] This information does not include patient names but does include each individual's NHS number, date of birth and full post code, as well as the patient's history of diagnoses, prescriptions, vaccinations, etc. At present, particularly sensitive information (e.g. pregnancy termination) and handwritten or 'free text' notes will not be extracted.

In the future, it is anticipated that care.data will join up with other NHS projects, for example, allowing phenotypic data to be linked to the genomic data produced by the 100k Genome Project being delivered by Genomics England Ltd (GeL).[335]

---

[333] In Scotland, the mandated use of the CHI Number in relation to all health episodes, the comprehensive computerisation of routine clinical data and the centralisation of NHS data has allowed the Scottish Informatics Programme (SHIP) to develop a platform for EPR research involving the linkage of health and other records (now based at the Farr Institute @ Scotland (http://www.farrinstitute.org/centre/Scotland/3_About.html). Though different (and clearer) in its aims from care.data, SHIP has had a less troubled development than care.data, involving more academic reflection and a programme of public engagement in relation to its governance arrangements.

[334] See: http://www.hscic.gov.uk/gpes. This was originally due to begin in Spring 2014 but was delayed owing to opposition from GPs and civil society groups, with a pilot finally announced in late 2014. A function of GPES is to feed the Quality and Outcomes Framework (QOF) used to calculate a significant part of GPs' remuneration. QOF payments are made for meeting multiple targets for activities from screening through immunisation, and were previously claimed using paper forms. The proposal that QOF payments be computed automatically on live patient data that would be subjected to bulk upload without the practical possibility of an opt-out for either GPs or patients led to widespread protest. GPES is owned and managed for the NHS by the HSCIC and has an Independent Advisory Group (IAG), which considers applications to extract data from GP clinical systems from other organisations (e.g. NHS England, Public Health England) according to a defined approvals process (see: http://www.hscic.gov.uk/article/3472/Customer-requirements).

[335] Indeed, Sir John Chisholm, Executive Chair of GeL, has recently become a Non-Executive Director at HSCIC. GeL state that it is intended that the findings of the 100k Genome Project will be linked with identifiable data from primary care and hospital records and that this can be linked to relevant clinical data. See: 'On the progress and outcomes from the NHS England Genomics Strategy Board and the genetics lab reconfiguration', http://www.england.nhs.uk/wp-content/uploads/2013/07/180713-item16.pdf. See also: Tim Kelsey, The Guardian (5 November 2013) *Five ways to enable transparency in the NHS*, available at: http://www.theguardian.com/healthcare-network/2013/nov/05/transparency-operating-principle-nhs.

6.25 A number of arguments have been deployed to establish the moral reasonableness of the care.data programme. The example is highly instructive because the sites of these arguments have moved from the extension of implicit norms, through attempts to rebalance these in legislation, to a public account of political decision making.

The moral justification of the 'care.data' programme

6.26 When care.data was initially proposed it was framed as an extension of previous programmes, such as the Secondary Uses Service (SUS) initiated by Connecting for Health.[336] Analogies can also be drawn to other programmes such as the Clinical Practice Research Datalink (see paragraph 6.9 above), which already links a number of the same data sources that will be linked by the HSCIC and which has been implemented without critical attention. A first argument, then, is that if there is no morally significant distinction between these initiatives, and others have been accepted, then to object to care.data would be inconsistent.

6.27 Any argument by analogy to previous initiatives is, however, only as robust as the justification for those previous initiatives. The identification of an inconsistency in attitudes between care.data and SUS or CPRD (if it is an inconsistency) does not indicate which (if either) of the inconsistent beliefs is correct. Care.data attracted a high level of media attention whereas SUS did not: however, it might be that people would have objected to SUS more strongly or in greater numbers if there had been a greater level of awareness of it.[337] This raises important questions about how those norms were established (questions that we suggest may be addressed by our third and fourth principles of participation and accounting for decisions– see chapter 5). It has since become common ground that there was insufficient communication and consultation with key stakeholders prior to the planned introduction of the care.data programme.[338]

6.28 A second potential weakness of this argument is the robustness of the analogy. That is, whether the HSCIC's proposed activities are constrained within the scope of those previously accepted (in the case of SUS, for example) and established in data protection law (see below) or, on the other hand, whether they transgress previously accepted moral and, arguably, legal norms of privacy and disclosure. Crucial in this respect is the question of who might have access to data or might receive extracted data from HSCIC, and for what purposes (which has been taken up through legislative action to re-establish a legal norm and associated mechanisms).[339]

6.29 A second site of justification, then, concerns the legal norms applicable to care.data, specifically those of data protection and data sharing. The cornerstone of data protection law, that data processing should be fair, requires reasonable steps to be taken to give notice to the data subject of the purposes for which their data will be used.[340] In August 2013 NHS England maintained that it was the responsibility of GPs,

---

[336] See http://www.connectingforhealth.nhs.uk/systemsandservices/sus. It is interesting that care.data is widely seen as controversial where SUS was not. Indeed, it has been argued that care.data became controversial only because the Secretary of State, in undertaking to honour the choice to 'opt out', actually gave prominence to the question of choice, and that the media became widely engaged with the issue.

[337] In this case, the argument turns in the opposite direction and becomes a criticism of previous failings to inform the public.

[338] See: http://www.england.nhs.uk/2014/02/19/response-info-share/; http://www.england.nhs.uk/wp-content/uploads/2014/04/cd-stakeholder-lett.pdf.

[339] See the Care Act 2014, discussed below.

[340] DPA 1998, Sched.1, Part II. See also: Grace J and Taylor MJ (2013) Disclosure of confidential patient information and the duty to consult: the role of the Health and Social Care Information Centre *Medical Law Review* **21(3)**: 415-47.

as controllers of primary care patient data, to inform their patients about the data extractions. Owing to the extraordinary demands this would place on GPs' resources, an attempt was made to inform patients through a NHS England leaflet, 'Better information means better care', which was sent to households in England.[341] This was, however, widely criticised for its uninformative design, its limited circulation and its resulting lack of success in generating awareness; it did little to persuade many GPs that patients had been adequately informed.[342] This information would be all the more important if patients were to be able meaningfully to exercise the opt-out they had been promised.[343] This notwithstanding, however, the Health and Social Care Act 2012, in effect, mandated the submission of data to the HSCIC. The effect of the two Acts placed GPs at the intersection of potentially inconsistent requirements – not to submit data without giving adequate notice to the data subjects and to submit data to the HSCIC – that each invoked norms of privacy and public interest.[344] On the eve of the first extractions, pressure from GP bodies and civil society groups caused a postponement of the data extraction to allow time to 'build understanding' of the aims and benefits of care.data.[345]

6.30 A third site of justification is the public account given of the conditions to which care.data was the proposed response, conditions that implicitly entailed the recognition of new norms in the light of which the care.data approach would be justified. Alongside considerations of individual privacy, there is a public interest in making use of data if doing so results in more efficient health service planning and delivery, better treatment and the development of scientific knowledge. Other things being equal, there is

---

[341] This was a *volte-face* from the position in August 2013, when the NHSE maintained that it was the responsibility of GPs to inform their patients about care.data. For the leaflet, see: http://www.england.nhs.uk/wp-content/uploads/2014/01/cd-leaflet-01-14.pdf. The leaflet did not go to households that have registered with the Royal Mail's 'door to door opt-out'. However, it was delivered to households using the Mail Preference Service (see: https://www.whatdotheyknow.com/cy/request/royal_mail_contract_for_caredata).

[342] In relation to its design, it was criticised for biased presentation of the benefits and risks of care.data, for failing to provide adequate information about how to opt out, and for not taking advice (Dame Fiona Caldicott, speaking to the BBC Radio 4 PM programme, said that NHS England had gone ahead without waiting for her IIGOP committee's advice on the leaflet (see: http://www.bbc.co.uk/news/health-27069553). It was also criticised in terms of its circulation (for example, an ICM survey for the BBC found that only 29% of 860 adults polled recalled receiving the leaflet and about 45% of people remain unaware of the plan to share some data from GP medical records; see: http://www.bbc.co.uk/news/health-26187980). Further criticisms included that the leaflet was not accessible for people with visual impairments and some others, and that it did not include the term 'care.data' anywhere.

[343] Patients were promised the opportunity to opt out of the programme by the Secretary of State for Health in the wake of the publication of the 2012 Information Governance Review report. As the HSCIC privacy impact assessment acknowledges, "patients have a right to object to personal information about them being collected and used by the Health and Social Care Information Centre" (http://www.hscic.gov.uk/media/12931/Privacy-Impact-Assessment/pdf/privacy_impact_assessment_2013.pdf, at page 16). In law, this 'opt out' is constituted by a right to object to the use of data (with the ICO as an arbiter), which, pursuant to the Secretary of State's undertaking, would be automatically upheld. Even in June 2014, however, an Ipsos MORI poll of 1,958 UK adults commissioned by the Joseph Rowntree Reform Trust found that 51% of respondents said they had never heard of the scheme and 13% had heard of it but did not know what it was. On the substantive issues, a small majority were opposed to the 'opt out' basis on which care.data is proceeding, with 40% saying that their GP should only be allowed to share their data with explicit consent, and 13% saying that said data shouldn't be shared by GPs under any circumstances. (This compares with 27% who said GPs should be able to share patient records if they had been informed and given a chance to opt out and 10% who said that they should be shared even without informing them.) See http://www.jrrt.org.uk/sites/jrrt.org.uk/files/documents/IpsosJRRTprivacypollMay2014full.pdf.

[344] It is important to appreciate, as we have argued, that norms of privacy and public interest are implied in both the applicable Acts (i.e. it is not the case that one defends individual privacy and the other the claims of public good), although arguably to different effect. This is also the case with the Care Act 2014. This is not inconsistent with the proposition that legislative action in this case is attempting to affect moral norms rather than merely to reflect them.

[345] In a letter to NHS England on the eve of the first intended extractions, the RCGP said: "While we recognise the substantial programme of activity and materials that has already been developed to communicate care.data, we believe that there is a deficit of awareness and understanding regarding the scheme amongst many members of the public and professionals." (Letter from RCGP Honorary Secretary, Professor Nigel Mathers, to NHS England (18 February 2014), quoted in RCGP news release; see: http://www.rcgp.org.uk/news/2014/february/rcgp-calls-for-reassurances-before-controversial-data-scheme-goes-ahead.aspx). See Pulse (17 February 2014) *GPC calls for urgent talks over public awareness of care.data scheme*, available at: http://www.pulsetoday.co.uk/your-practice/practice-topics/it/gpc-calls-for-urgent-talks-over-public-awareness-of-caredata-scheme/1/20005884.article?&PageNo=1&SortOrder=dateadded&PageSize=20#.VLZvx3uj9Oc.

likewise a public interest in using health data to generate economic growth by stimulating economic activity around it, and generating revenue or saving costs for the health service. The proponents of care.data have argued that there is such an overwhelming public interest in the programme that privacy interests should be qualified proportionately.[346] This argument has been stated in terms no less than that 'the future survival of the NHS depends upon it.'[347]

6.31 The cogency of this argument depends on accepting its premises. These include (1) that changing conditions require such a measure to improve the efficiency of health care so that it can continue to be provided as a public good; (2) that the proposed approach is likely to contribute to this objective in the way intended[348]; and (3) that the proposed approach is preferable to all alternatives.[349] These premises rest on a number of further factual, political and speculative claims. It is empirically true, for example, that, with ageing populations and increasing technological intensity, inflation in health care costs has outstripped the rise in other prices worldwide. Nevertheless, UK health care is largely funded through taxation and the adequacy of resources is substantially a matter of political decision rather than a hard constraint. It is likely, therefore, that the NHS will continue to be caught in major political arguments about funding and resource allocation, with or without the business intelligence and advances in treatment promised by the HSCIC.[350]

Moving the arguments forward

6.32 With moral and, arguably, also legal norms unresolved a second layer of debate concerned how individual freedoms should be respected. The default setting (choice of opt-out or opt-in) and the level of public information became a new focus of debate. The justification of default inclusion might be seen as extending the norm of social solidarity that is conventionally assumed to underwrite the provision of public health care. However, the content of those norms is not well developed or widely discussed (as the previous section shows). For example, the 'imagined community' to which these norms are referred is often taken to exclude commercial firms. However, disclosure of data to commercial firms is within the intentions of the HSCIC, so long as their objects are consistent with the purposes specified in law.[351]

---

[346] See ECHR, Art.8(2).

[347] See evidence to House of Commons Health Select Committee, February 2014. Tim Kelsey: "My view is that it is one of the most important public debates we are having and it is about the future of the health service." (Oral Evidence: Care.data database, HC 1105, 25 February 2014, available at: http://data.parliament.uk/writtenevidence/committeeevidence.svc/evidencedocument/health-committee/handling-of-nhs-patient-data/oral/6788.pdf, at page 34).

[348] We noted in chapter 2 that the policy discourse has been consistently ambitious. For example, a 2014 report from a big data solutions company claims that better use of data analytics could free £16.5 billion and £66 billion worth of NHS capacity.

[349] Some have argued that if the same project were set on a voluntary basis people would decline to participate in such numbers that the data would be significantly less valuable. A major concern of those who opposed an opt-in approach is that it would disproportionately disadvantage those who are most vulnerable and in need of its benefits (i.e. those who have low social power and status, and are already disadvantaged, including the elderly and those with long-term illness and disability). The second premise has been challenged, however: it can be argued that a certain level of non-participation (whether through 'opting out' or choosing not to 'opt in') would be tolerable and still allow the aims of the programme to be met. (See: evidence to HC Health SC, February 2014, http://www.parliament.uk/business/committees/committees-a-z/commons-select/health-committee/news/14-02-25-cdd-ev/ and http://www.parliament.uk/briefing-papers/SN06781/caredata. On mandatory inclusion and the (un)acceptability of 'free riding', see paragraph 5.12.

[350] See Nuffield Council on Bioethics (2014) *Funding pressures in the NHS: an ethical response* (Forward Look background paper), available at: http://nuffieldbioethics.org/wp-content/uploads/NHS_resource_pressures_final_FL_paper1.pdf.

[351] On 'imagined communities', see Busby H and Martin P (2006) Biobanks, national identity and imagined genetic communities: the case of UK biobank *Science as Culture* **15(3)**: 237-251. As we noted, the research and innovation system is complex, so

6.33 The question of who might have access to the data and to whom it might be disclosed in turn raised the question of governance of data use by the HSCIC. This turned the spotlight on the organisation's decision making procedures and their (or their predecessor, the NHS information Centre's) management of data sharing agreements. HSCIC provided little clarity on how data would be used in the run up to the initial launch date of the care.data extraction programme and, indeed, had not at that time published a code of practice.[352] Civil society groups opposed to the proposed introduction of care.data and the media began to draw attention to cases in which individual-level health data had been disclosed widely (including by the HSCIC's predecessor body, the NHS Information Centre) in ways that presented re-identification risks.[353]

6.34 In these ways, the key moral elements of a data initiative that we have discussed in this report, namely, the relationship between the underlying moral norms (including their relationship to legal norms), the way of respecting individual moral interests and values, and the way in which decisions are made and accounted for, finally became explicit and the subject of public and political discourse.

Solutions for the HSCIC

6.35 Care.data has raised to public prominence a debate about the justification, beyond the purely legal basis, of the HSCIC's collection and release of data. According to our approach, outlined in chapter 5, the moral question that confronts the HSCIC is to define a public set of morally reasonable expectations about the use of data generated by health and social care services. This should be done within a framework of principles that takes the mutual implication of public and private interests that we discussed in chapter 3 into account.

6.36 The first task should be to identify the relevant norms. There may have been an assumption that there were no novel features and therefore that the norms of social and health care, of the 'presumed broad consent' of patients to the use of data within the NHS, would apply unaltered to the HSCIC initiative. However, the HSCIC initiative appeared to want to keep the scope of potential uses broad, in the spirit of treating data as a resource with multiple and undefined potential uses. It is not at all clear that all these possible uses would fall within the expectations of patients, and the reaction of GPs, civil society groups and the media demonstrated this. Indeed, findings from public opinion research suggest, as we noted in chapter 5, that while there is broad public support for some further uses of care data, such as biomedical research, many individuals still want to be asked about this, and there are other uses, such as commercial uses, that are viewed with suspicion.[354] It may be that these views would

---

that even if medical research is carried out in English universities on an entirely non-profit basis, the commercialisation of any drugs discovered will almost inevitably involve the pharmaceutical industry.

[352] A code of practice has now been published by the HSCIC whilst this report was being finalised: HSCIC (2014) *Code of practice on confidential information*, available at: http://systems.hscic.gov.uk/infogov/codes/cop/code.pdf.

[353] Complaints were made to the ICO regarding the publicised 'Staple Inn incident' where individual-level data were apparently released to insurance actuaries. See: Information rights and wrongs blog (24 February 2014) *Hospital records sold to insurance companies – in breach of the data protection act?*, available at: http://informationrightsandwrongs.com/2014/02/24/hospital-records-sold-to-insurance-companies-in-breach-of-the-data-protection-act/. A review of data releases made by the NHS IC was subsequently carried out by a non-executive director of the HSCIC, Sir Nick Partridge. The review identified a number of incidents relating to the previous body and also to HSCIC; some refer to ongoing practices at HSCIC and made recommendations, which have been accepted by the HSCIC. See: http://www.hscic.gov.uk/datareview.

[354] Hill EM, Turner EL, Martin RM, and Donovan, JL (2013) "Let's get the best quality research we can": public awareness and acceptance of consent to use existing data in health research: a systematic review and qualitative study BMC Medical Research Methodology 13: 72, available at: http://www.biomedcentral.com/1471-2288/13/72.

not withstand further reasoning: we noted in chapter 2 that the involvement of commercial firms is a feature of the innovation system on which the medical developments that people want rely. It may be that increasingly broad use of data will play an important role in a sustainable future for the NHS. Such arguments gesture to the way in which private interests may be implicated in the broader public interest. But there was no open debate of these issues and arguments in public. One place in which these questions were opened up to deliberation, albeit belatedly, was through the 'autonomous' advisory group, which was set up when the planned GPES extractions were postponed and which includes a number of NGOs and other stakeholders.[355] Broader consultation was also promised which may help at least to identify norms specifically relating to the proposed uses, as will the 'pathfinder' exercises in four areas, which will examine ways of communicating with patients to make the ability to opt out meaningfully exercisable.[356]

6.37 As we have noted in this report, the existence of a framework of legal norms offered by various legal instruments (governing data protection, the duties of public health care services, and human rights – discussed in chapter 4) does not entail complete moral freedom to act within it. While these offer a starting point, they are clearly not sufficient and may even come into conflict.[357] (No more can moral norms be shifted simply by passing new legislation: the law and public morality must correspond with each other, procedurally as well as structurally.) Though the Care Act 2014 was intended to clarify the purposes for which HSCIC data may be used, 'the provision of health care or adult social care, or the promotion of health' merely sets apparently more limited but by no means more precise criteria for decision making. [358] What the Care Act 2014 (and currently awaited secondary legislation) does recognise is the need for further moral guidance on data access for which the HSCIC, as an executive agency, was not itself resourced. The solution it provides is to take advice from the HRA Confidentiality Advisory Group (CAG), which, through its predecessors, the Patient Information Advisory Group (PIAG) and the Ethics and Confidentiality Committee of the National Information Governance Board (ECC), has experience with the questions of data access, moral norms and reasonable expectations, with which the HSCIC is faced.[359] The elusive – and possibly illusory – significance of the difference between individual-level data and identifying data, as we have argued in this report (see chapter 4), makes the reliance on the Confidentiality Advisory Group (CAG) for advice on disclosures a natural link in the scheme of governance as it presently exists. This is because, in effect, the question that must be posed for all disclosures of individual-level data (and a lot of aggregate data too) is the one that CAG is implicitly established to address, namely, whether the proposed disclosures would fall within the scope of reasonable expectations of disclosure. This advisory system is also buttressed by the appointment

[355] See: http://www.england.nhs.uk/ourwork/tsd/ad-grp/.
[356] See: http://www.england.nhs.uk/wp-content/uploads/2014/10/caredata-pathfinder-proposal.pdf.
[357] The position of GPs with regard to the DPA v. HSCIC, see paragraph 6.29.
[358] Care Act 2014, s.122(3), which amends section 261 of the Health and Social Care Act 2012. The scope of this amendment was debated during the passage of the Act: "Here the current wording of the Commons Amendment, although well meant—we are pleased to have it—leaves open too many questions for interpretation. Their amendment suggests that use of patients' data may be allowed for "the promotion of health". This leaves us open to two types of possible interpretation that may be undesirable. For example, "promotion" could be taken to mean that food manufacturers could use data in their marketing campaigns for so-called healthy foods. That may or may not be desirable but it would put many off if it appeared that their data were being used for commercial gain in a competitive market." per Lord Turnberg, HL Hansard,(7 May 2014) c1514.
[359] New Economics Foundation (NEF) research found public opposition to the use of s.251: see recommendation 5 of http://b.3cdn.net/nefoundation/2cb17ab59382fe7c67_bfm6bdoas.pdf.

of Dame Fiona Caldicott as the first National Data Guardian for health and care, a role described as 'the patients' champion on security of personal medical information'.[360]

6.38 While this entails that particular decisions of the HSCIC will receive support, it does not make any more explicit the scope of what are reasonable expectations, nor does it require an engagement with interested participants (although CAG is very likely to take account of social research to inform its own reasoning). Rather it suggests that the norms will be established and elaborated through developing expertise and precedent. This has the advantage of flexibility to respond to the evolving debate. However, it fails to make any clearer to the public what expectations they may have about who will have access to individual-level NHS data or for what purposes, and therefore from what they might be opting out when they consider whether to do so.

Recommendation 6

We recommend that an independent group of participant representatives is convened to develop a public statement about how data held by the Health and Social Care Information Centre should be used, to complement the Code of Practice on confidential information. This should clearly set out and justify what can reasonably be expected by those from whom data originate and be able to demonstrate that these expectations have been developed with the participation of people who have morally legitimate interests in the data held by the HSCIC, including data subjects, clinical professionals and public servants.

Recommendation 7

We recommend that, in addition to implementing the recommendations of Sir Nick Partridge's review, all Data Sharing Agreements held by the HSCIC should be published, along with the findings of a periodic independent audit of compliance with those agreements.[361]

Recommendation 8

We recommend that HSCIC Data Sharing Agreements should include a requirement to maintain an auditable record of all individuals or other legal entities who have been given access to the data and of the purposes to which it has been put; this should be made available to all data subjects or relevant authorities in a timely fashion on request.

[360] See: https://www.gov.uk/government/news/national-data-guardian-appointed-to-safeguard-patients-healthcare-information.
[361] Sir Nick Partridge (2014), Review of data releases by the NHS Information Centre, available at: http://www.hscic.gov.uk/datareview. The recommendations included that HSCIC develops one Data Sharing Agreement to be used for all releases of data, and which includes clear sanctions for any breaches.

The Scottish Informatics Programme and the Farr Institute

6.39 In many ways the Scottish approach to the question of the reuse of data from health care has been the inverse of the English experience. The Scottish authorities began with public engagement to determine acceptability (rather than being forced to engage after the fact) and developed an integral governance approach along with the informatics infrastructure (rather than having to invent one to fill a vacuum between the legal framework and operational decision making), building public trust rather than undermining it. The Scottish Health service is, of course, much smaller than the English NHS, and organised and run differently. Data linkage is facilitated by the almost universal use of the CHI in Scotland, and the Scottish experience of moving to electronic patient records (EPRs) has been both more measured in ambition and smoother in execution than in England.

6.40 The Scottish Informatics Programme (SHIP – formerly the Scottish Health Informatics Programme) was initiated to develop a research platform to support more systematic collection, governance and research use of linked EPRs, and to establish a research arm within the Information Services Division (ISD) of NHS National Services Scotland and to support the efficient functioning of public services more generally. It provides for linkage not only of care data but also data that have been gathered in cohort studies and other forms of publicly-held administrative data.[362]

Box 6.6: Health record linkage in Scotland

SHIP began in 2008 as a collaboration between the universities of Dundee, Edinburgh, Glasgow and St Andrews and the Information Services Division (ISD) of NHS Scotland. It was established with a £3.6 M grant from the Wellcome Trust, the Medical Research Council and the Economic and Social Research Council. The development of the infrastructure involved three workstreams:

- Public engagement to ascertain acceptability
- Legal review and development of proportionate governance approach
- Development of data linking methodology and associated IT infrastructure

In 2014 the programme was moved to the Farr Institute @ Scotland, one of four Farr centres across the UK (see paragraph 2.31 above).[363]

A central element of the infrastructure is a National Safe Haven.[364] Unlike in the English HSCIC, data are not collected and held centrally; centralisation of data was found to be less acceptable during the prior public consultation. Instead, locally held datasets are linked within the safe haven for the purposes of specific analyses. The safe haven comprises three elements, provided by three separate organisations to increase privacy protection:

- An **indexing service** provided by National Records Scotland, which establishes links

[362] SHIP (2012) A blueprint for health records research in Scotland, available at: http://www.scot-ship.ac.uk/sites/default/files/Reports/SHIP_BLUEPRINT_DOCUMENT_final_100712.pdf. These requirements should apply more specifically in the areas of clinical trials, pharmacovigilance, diabetes epidemiology, and research resulting from the linkage of EPRs to socioeconomic and environmental data.

[363] There is a Farr Institute – CIPHER – in Wales. For reasons of space we have not included a discussion of the situation in Wales, but the conditions in Wales are more akin to those in Scotland than those in England.

[364] The SHIP Blueprint cites the 2008 Data sharing review report as an inspiration and SHIP's use of safe havens develops proposals within that report. See: The SHIP Blueprint 2012, available at: http://www.scot-ship.ac.uk/sites/default/files/Reports/SHIP_BLUEPRINT_DOCUMENT_final_100712.pdf.

between cases in different datasets using a CHI look-up table or probabilistic linkage
- An **analytic platform** housed in the University of Edinburgh's Advanced Computing Facility, which holds the research dataset (without direct personal identifiers) and is accessed by researchers to conduct analyses
- A **researcher Advice and Disclosure Control Service (eDRIS)** provided by National Services Scotland NHS, which provides advice to researchers and statistical disclosure control.

Data linking and governance

6.41 Unlike in England where the HSCIC extracts data from local sources to a central data warehouse, in Scotland different datasets are held in distributed collections, each of which is overseen by a data custodian who must agree to the release of data for the specific purposes of the research. Also, unlike in England, Scotland (as does Wales) applies precautionary de-identification measures by splitting and encrypting demographic and clinical data prior to linking. This is achieved in the following way. Individual-level data are sent by each contributing data source to an indexing service where identifiers are removed and a code assigned. A different code is assigned to each individual case for each data source, but an index of corresponding codes is produced by the indexing service, which links the different codes assigned to the same individual where they occur in different data sources. The coded data are then returned to the source and sent on to the national safe haven for analysis by authorised researchers. The researchers are sent the index of corresponding codes by the indexing service so that they can match up cases from each source. It is therefore only within the safe haven that the data from the multiple sources can be linked, as a result of knowing the corresponding codes assigned to cases in the different sources. Data from different sources (e.g. primary care, administrative, etc.) are linked centrally only for the purposes of a specific analysis; the linked data are held securely in the safe haven and destroyed as soon as practicable after the results have been obtained.[365] This is apparently more complex than the HSCIC approach, but represents a balance between efficiency and acceptability (as determined with reference to prior public engagement).

6.42 One of the sources of data will be primary care records, although GPs in Scotland exercise more control over data extractions than is proposed under NHS England's care.data programme. The Scottish Primary Care Information Resource (SPIRE) will be built from Quality Outcomes Framework (QOF) data about GP activity, as in England (see note 336 to Box 6.6 above), a standardised set of data about patients using primary care services that GPs may choose to provide (via an opt-in mechanism), and other bespoke extracts obtained with the permission of GPs.[366] As in England, GPs are in a crucial position and bear significant responsibility to protect their patients' interests alongside their own, although GPs in Scotland have not had to fight, as have English GPs, to assert their powers and responsibilities as data controllers.

Proportionate governance

6.43 The SHIP approach begins with an assumption that the public and medical doctors are in most cases expecting health data to be used for socially beneficial research, an

[365] See also: http://www.isdscotland.org/Products-and-Services/Health-and-Social-Care-Integration/docs/IF_Framework_HSC_Integration_Route_Map_V1.7.pdf.
[366] See http://www.spire.scot.nhs.uk/. SPIRE is expected to be operational from March 2015.

assumption supported by a dedicated research programme to review and build on existing evidence to this effect.[367] The governance arrangements for data services are built around the concept of 'proportionate governance' and a stated set of guiding principles and related best practices.[368] Among the guiding principles are protection of privacy and promotion of the public interest.[369]

6.44 'Proportionate governance' denotes an approach to information governance in which the balance of risks and benefits and appropriateness of means to ends are central.[370] It aims to improve on existing approaches by including assessment of the relative merits of different governance mechanisms, the selection of appropriate governance pathways, and a choice of different governance tools appropriate to any given research application. It does not place reliance on any particular combination of anonymisation, consent or authorisation (see chapter 4). Furthermore, the approach is not focussed solely on the 'kind' of data in play but also, importantly, on the context in which it is in play. Rather than applying a 'one-size-fits-all' solution, risk assessment is used to determine what tools and pathways are appropriate in a particular case. Risk assessment has two functions in SHIP: one with respect to data protection and the risk of individual identification, and another with respect to authorisation of research in the public interest.

6.45 The use of risk assessment explicitly orientates SHIP initiatives towards 'proportionate' governance as distinct from 'precautionary' governance.[371] This approach is possible where the context is one that is reliably controlled in such a way as to minimise risk ('safe data, safe people, safe environment').[372] A particular difficulty of assessing risk in this area is lack of evidence. It makes good sense to ground risk assessment in evidence because it offers an explicit, public reference point rather than remaining a matter of private opinion or judgement. However, there are two problems in this case, relating to uncertainties about the future conditions and about present facts (inductive and epistemological uncertainties). The first is a standard problem of induction, which points to the difficulty of using past evidence to make judgements about the future *when conditions or circumstances are changing significantly*.[373] That is because it is the regularity of the circumstances that underwrites the inference about the future. We

[367] See: SHIP public engagement work-stream, http://www.scot-ship.ac.uk/c4.html and SHIP (2011) SHIP public engagement: summary of focus group findings, available at: http://www.scot-ship.ac.uk/sites/default/files/Reports/Focus_Group_Findings_Briefing_Paper.pdf.

[368] The 'good governance framework' has four key elements: (1) guiding principles and best practice, (2) proportionate governance, (3) roles and responsibilities of data controllers and (4) researcher training. See: Sethi N and Laurie G (2013) Delivering proportionate governance in the era of eHealth: making linkage and privacy work together *Medical Law International* **13(2-3)**: 168-204; Laurie G and Sethi N (2013) Towards principles-based approaches to governance of health-related research using personal data *European Journal of Risk Regulation* **1**: 43-57.

[369] See the *SHIP blueprint*, Appendix 7, available at http://www.scot-ship.ac.uk/publications.html.

[370] "Proportionality is the overarching principle that ties the varying components of good governance together and should be the ultimate benchmark against which to assess the appropriateness of conduct – both at the level of individual linkage decisions and the choice of what counts as appropriate governance over those linkages." Laurie G and Sethi N (2012) Laurie G and Sethi N (2012) Information governance of use of health-related data in medical research in Scotland: towards a good governance framework, University of Edinburgh, School of Law Research Paper Series No 2012/13, available at: http://www.scot-ship.ac.uk/sites/default/files/Reports/Working_Paper_2.pdf, at page 12.

[371] A precautionary approach is often thought to be appropriate in conditions of significant uncertainty, where there is a potential for widespread and/or irreversible harm. A precautionary approach might be thought to entail that where there is any risk of re-identification, for example, then data should be treated as 'personal data' and, reliance on other grounds being insecure, the consent requirements of the Data Protection Act duly engaged.

[372] Laurie G and Sethi N (2012) Information governance of use of health-related data in medical research in Scotland: towards a good governance framework, University of Edinburgh, School of Law Research Paper Series No 2012/13, available at: http://www.scot-ship.ac.uk/sites/default/files/Reports/Working_Paper_2.pdf.

[373] Hume D (1975 [1748]) *Enquiries concerning the human understanding and concerning the principles of morals* (3rd Edition) (Oxford: Oxford University Press), section IV.

have already discussed the significant problem of rapid developments in data science and the accumulation of data; indeed it is this scale and pace of these developments that provoked this report. The SHIP approach of not retaining linked data and not disclosing raw data, limits the dimension of uncertainty that arises from holding linked data for indefinite time periods in changing circumstances. In such a context, a negligible risk of re-identification may obviate the need for additional consent from patients where it would otherwise be required (albeit that, in line with the findings of public consultation, the burden falls on the researchers to demonstrate why consent is either unnecessary or inappropriate). A second problem with an evidence-based approach to risk is the fact that, while some relevant evidence might exist, there are reasons to think that undesirable outcomes are significantly under-reported and intrinsically difficult to find, rather than simply absent. This is strongly suggested by our commissioned research into harms associated with data abuse (see chapter 2).[374]

6.46 Even if it were reliably determinable, however, risk alone cannot be a sufficient basis for governance. This is because the nature of the questions with which we are concerned, which are partly moral questions, are not tractable solely by the application of evidence, no matter how much evidence is available. Their resolution requires, additionally, a form of reasoning that reveals and resolves the values and tolerances associated with different possibilities and consequences.[375] Whereas robust approaches to data handling may minimise the risk of re-identification of individual patients, they do not in themselves address questions of privacy and potential harms to privacy.

Authorisation, decision making and accountability

6.47 Under the SHIP model a system of 'authorisation' of research operates alongside and somewhat independently of any requirement for consent and measures to de-identify the data, addressing instead the duty of care owed by professionals to the public out of respect for them as moral agents. (This is consistent with our conclusion, at which the SHIP analysis also explicitly arrives, that consent is neither necessary nor sufficient to protect the interests of those involved.)[376] Whether or not consent is judged to be required, the second function of risk assessment is to support the authorisation process in considering the 'relative risk' of a research initiative, taking account of the full range of interests at stake and the balance of likely hazards and benefits (not merely the risk of re-identification). This is to recognise that, from a public policy perspective, there are risks of both using and not using data, and that these may be distributed differently among the range of those with interests in the initiative. (Thus, the minimisation of risks for some, taken to its logical conclusion of simply locking down data and preventing all re-use, is not without consequences in terms of benefits foregone, both for those people and for others.)

6.48 When one moves from the simple principle of minimising risk to the principle of optimising the balance and distribution of risks and potential benefits, the essentially political question of how this optimum is determined (and by whom, and in whose interest) becomes salient. This is, in effect, the most fundamental question of privacy: how we construct the norms that enable cooperation in pursuit of common goals but

[374] One response to this is to pursue further research as we have recommended (see recommendation 2).
[375] See our third and fourth principles (participation and accounting for decisions).
[376] Laurie G and Sethi N (2012) Information governance of use of health-related data in medical research in Scotland: towards a good governance framework, University of Edinburgh, School of Law Research Paper Series 2012/13, available at: http://www.scot-ship.ac.uk/sites/default/files/Reports/Working_Paper_2.pdf.

keep the unreasonable demands of others in check. Of course, this question is, in reality, always present and is only held in abeyance by a restricted framing of the question as if it were simply about *this* particular risk and how to manage it (the risk of re-identification of given individuals, for example); in reality, reducing the exposure to risks for one person almost always increases it for others. It is for this reason that, in chapter 5, we presented data initiatives as *social* practices, embedded in the wider life of the political society, which require those with interests at stake to reason together regarding their resolution.

6.49 As we have done in this report, SHIP sets out principles that provide a reference point for deliberation and decision making without prescribing what should be done in any particular set of circumstances.[377] Though the principles are not the same, they demonstrate a clear and explicit engagement with the question of the relationship between public and private interests that led us to posit our principles of respect for persons and human rights (see chapter 5). They provide a common language and frame of reference to consider what morally relevant interests are at stake, to reason through the issues, and to ground justifications offered for particular outcomes. It is important, therefore, to ask who should participate in this process and to whom and by what means their decisions are accounted for.

6.50 In Scotland, the authorisation for release of data is acknowledged as being initially in the hands of GPs and other data custodians of the contributory datasets (e.g. NSS, NHS or within Health Boards). (The English system, by contrast, seems designed to shift these powers as far as possible to the centre. The fact that the GPES extraction of primary care data, rather than HES data, for example, became the main battleground owes much to the involvement of GPs as interested participants – as participants, of course, with their own interests in the possible data initiatives, alongside researchers, firms, patients and others.) Further authorisation for use of data is required (by law) from the data custodians and may be sought, additionally, from a body such as the Scottish Privacy Advisory Committee (PAC).[378] In addition to the legal question of whether data *may* be released, the SHIP approach explicitly requires consideration of the moral question of whether data *should* be released. It is here that the question of 'balance of risks' is engaged.

6.51 The SHIP approach connects to the social basis of data sharing through a commitment to public engagement as an input to governance.[379] The commitment to public engagement has been taken on by the Farr Institute @ Scotland, to raise awareness of

[377] The SHIP principles contain references to human rights and respect for individuals as moral agents. They are nevertheless more operational (and more numerous) than the higher-level principles set out in this report, which are intended to serve a broader range of applications. For the SHIP principles, see: http://www.scot-ship.ac.uk/sites/default/files/Reports/Guiding_Principles_and_Best_Practices_221010.pdf.

[378] PACs play a role somewhat analogous to the one now marked out for CAG in the English HSCIC system, although they have a formal approval function where CAG remains advisory. See: http://www.nhsnss.org/pages/corporate/privacy_advisory_committee.php.

[379] "13. Public and stakeholder engagement. **Principles** (1) Public and stakeholder engagement is an integral part of good governance. As far as possible, account should be taken of the full range of stakeholder positions in the development and implementation of governance arrangements. (2) The interests of one (or a few) stakeholder(s) should not dominate use/linkages or the conditions of the same, especially where this might be at the expense of other stakeholder interests. Robust justifications must be given for any departure from this principle. **Best Practice** (1) Stakeholder interests and expectations should be monitored over time by an appropriate body or individuals with appropriate expertise for the task. Where necessary, governance arrangements should be adapted to take account of shifting stakeholder needs and expectations. (2) Active engagement exercises should be developed and implemented over time to monitor and respond to stakeholder interests." (http://www.scot-ship.ac.uk/sites/default/files/Reports/Guiding_Principles_and_Best_Practices_221010.pdf, at page 16).

medical research and its benefits and to foster trust and to allow two-way communication between professional and public participants. As well as *ad hoc* activities, part of this commitment involves supporting a constituted Public Panel (currently with 20 members drawn from a cross-section of the Scottish public) that meets twice a year.[380] (This is consistent with recommendations we make in relation to research in the next chapter.)

6.52 One upshot of this has been the prior identification of expectations with regard to the involvement of the commercial sector. The stipulation that research should be in the public interest does not, in principle, bar private companies (e.g. pharmaceutical companies) using the resource. However, a key feature of the way the SHIP approach has been developed by the Farr Institute @ Scotland is that it takes seriously public guardedness about engaging with commercially motivated researchers by not allowing commercial users direct and un-chaperoned access to the data. At the same time, the approach acknowledges that in the science and innovation ecosystem as it currently exists, commercial actors can have an important role to play in promoting the public interest. In order to make use of the resource, therefore, private sector researchers must demonstrate that the research is in the public interest and form a partnership with NHS or academic researchers. It is these NHS or academic researchers who will undertake the analysis and have direct access to the raw data.[381] If a private company is involved, they will have access to the results of the research although results will only be released following approval from the PAC to ensure that no identifying data will be released. An explicit consideration is the reputational risk and the impact on public trust that commercial involvement may represent in any particular case.

6.53 The SHIP initiative demonstrates a number of elements of good practice according to our analysis and principles. It pays regard to context rather than simply the 'type' of data in use; it acknowledges the importance of responsible behaviour on the part of professionals over and above the duty to respect the consent of patients, even where data with a low risk of re-identification are used; it aims to resolve the 'double articulation' of public and private interests that we described in chapter 3, partly through a commitment to public engagement; and it takes seriously the need for trust and concerns about the involvement of commercial interests (which we consider further, in another context, in the next section).

The '100,000 Genomes' Project

6.54 The rich phenotypic and, increasingly, laboratory data held by the NHS and other health services, and the NHS's continuing relationship with patients, offer scientifically and politically attractive opportunities to carry out biomedical research alongside treatment. One such initiative is the UK 100,000 Genomes Project.[382] This project brings into conjunction more explicitly than any other current initiative, the research, policy, and national economic drivers that we discussed in Chapter 2, and embodies a commitment to the prospects of genomic medicine and to the idea that the NHS should be their proving ground. This has led to a notable two-stranded rhetoric around the

[380] See www.farrinstitute.org/centre/Scotland/21_Public-Engagement.html.
[381] The electronic Data Research and Innovation Service (eDRIS) provides a single point of contact for researchers. See: www.isdscotland.org/Products-and-Services/eDRIS/.
[382] A member of our Working Party, Professor Michael Parker, chaired the initial CMO's ethics working group leading up to the launch of the 100k Genomes Project. He is currently a Non-executive Director of GeL and chair of its Ethics Advisory Committee.

project, which freely mixes the objectives of advancing science to improve human health with ambitious 'techno-nationalism'[383]:

> "...the race is on. The benefits to human health (better and earlier diagnoses as well as personalised care) are so enormous that everyone will want to be in the game. Even so, the insights we can unlock are so numerous, there's enough potential reward for all players. But when it comes to building the critical mass of data needed to tackle some of our most serious healthcare challenges, there will be one winner, and that will be Britain."[384]

Box 6.7: The UK 100K Genomes Project and Genomics England Ltd

The 100K Genome Project, announced in a speech by the Prime Minister in December 2012, and launched formally on 1 August 2014, is a project to generate 100,000 whole genome sequences from NHS patients in England.[385] The project gives shape to the ambition to realise the benefits of genomic medicine in the NHS.[386] Its focus is on generating sequences from cancer patients and their tumours, patients with rare diseases (those affecting <1:1500 people) and their parents, and those with infectious diseases (HIV, tuberculosis, and hepatitis C) and antibiotic resistance. The project design was initially informed by reports from three working groups (on strategic priorities, ethics and data) established by the Chief Medical Officer for England.[387] Following these, the decision was taken to establish a private limited company, Genomics England Limited (GeL), wholly owned by the UK Government, to deliver the project and to manage the extraction of value from it.[388] The project has an embedded ethics team and an ethics advisory group. The aims of the project are stated as follows:

- to bring benefit to patients
- to create an ethical and transparent programme based on consent
- to enable new scientific discovery and medical insights
- to kickstart the development of a UK genomics industry[389]

Suitable NHS patients will be invited to participate by their health professionals and complete consent forms outlining the aims of the initiative, the possible uses of their data and the mechanisms for governing this. No immediate therapeutic benefits are promised to those taking part but in some cases the information may be used to inform their

[383] On 'techno-nationalism' see Edgerton DEH (2007) The contradictions of techno-nationalism and techno-globalism: a historical perspective *New Global Studies* **1(1)**: 1-32, available at: https://workspace.imperial.ac.uk/historyofscience/Public/files/c_contradictions_of_technoglobalism.pdf.

[384] Sir John Chisholm, Executive Chair of GeL and also a non-executive director of HSCIC. See: http://www.nesta.org.uk/news/14-predictions-2014/great-whole-genome-race. For a discussion of the link between biotech R&D investment and national economic growth, see Nuffield Council on Bioethics (2012) *Emerging biotechnologies: technology, choice and the public good* (esp. Chapter 7), available at: http://www.nuffieldbioethics.org/emerging-biotechnologies.

[385] The actual number of patients will be about 75,000 as the remaining sequences will be tumour sequences (see: http://www.genomicsengland.co.uk/the-100000-genomes-project/faqs/).

[386] Human Genomics Strategy Group (2012) *Building on our inheritance: genomic technology in healthcare. a report by the Human Genomics Strategy Group*: "We are currently on the cusp of a revolution in healthcare: genomic medicine – patient diagnosis and treatment based on information about a person's entire DNA sequence, or 'genome' – becoming part of mainstream healthcare practice.", available at: https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/213705/dh_132382.pdf, at page 12.

[387] See https://www.gov.uk/government/publications/mapping-100000-genomes-strategic-priorities-data-and-ethics.

[388] "Although this [GeL] is a company, it is only formed as a company so it can move more quickly to do these things [help patients, the NHS], to bring maximum benefit at the fastest speed." Mark Caulfield, cited in Martin P and Hollin G (2014) *A new model of innovation in biomedicine? A review of evidence relating to the changing relationship between the private and public sector in the use of human genomics and personal medical information,* available at: www.nuffieldbioethics.org/project/biological-health-data/evidence-gathering/, at page 14.

[389] See: http://www.genomicsengland.co.uk/about-genomics-england/.

treatment. A specific set of known genetic predispositions will be looked for and it is likely that the protocol will allow this information to be fed back to participants with their consent.[390] Patients who do not wish to enrol will continue to receive the best NHS treatment currently available for their condition.

The intention is to make data available to researchers in a secure setting where they can study it without extracting or storing it on a different infrastructure, so that their interactions with the data may be tracked and audited.[391] It will also contribute to making the NHS a 'learning health system' (see above) through Genomics England Clinical Interpretation Partnerships (GeCIP).

6.55 The decision to deliver the 100,000 Genomes project through a limited company (owned and seeded with £100M investment by the Department of Health, but with ambition to seek substantial additional investment) was to allow it to operate flexibly and responsively, to enter into contracts and relationships with businesses, and to seize opportunities as they appeared. GeL's main asset will be the 100,000 genomes database, from which it expects to make a financial return, either through direct payment for data access, or through royalty sharing or joint venture schemes with other companies.[392] Although GeL has responsibility for ownership and delivery of the project it will contract other UK companies, universities and NHS institutions to carry out sample collection, sequencing, annotation and storage. The company has a number of parameters for how it will work which include that it will "ensure the benefits of the investment flows from the company to a large range of companies and contractors including SMEs" and "use any surplus to benefit the public health community."[393] Those benefitting are likely to be software developers, sequencing and annotation providers, and the life sciences industry, which will use the knowledge generated by the project to develop new products.[394]

Information governance

6.56 Although GeL's corporate governance arrangements follow a commercial model, information governance follows a pattern more familiar from biomedical research. A formal data access process will be established with advice from the Ethics Advisory Committee and wider consultation. The procedure will include a data access committee that will examine and make decisions about data access applications, and a formal data access agreement to be signed by researchers.

6.57 Access to information with linked patient identifiers will be restricted to the clinicians working with patients (and, indeed, findings relevant to treatment of individual patients will be fed back to clinicians). Other users will only have access to de-identified data. All data will be held behind an NHS firewall, a model that has been compared to a 'reference library': unlike HSCIC, GeL will not provide extracts of individual-level data. GeL's customers (including academic and industry researchers) will have controlled access to that environment to carry out data analyses on linked, de-identified clinical

[390] The final protocol was not available at the time of writing.
[391] In the pilot phase, data is released to commercial organisations and annotation of the genome sequences is conducted within the annotators' infrastructure, outside the NHS firewall.
[392] See: http://www.genomicsengland.co.uk/wp-content/uploads/2013/11/Annotation-Supplier-Event-November-5-2013.pdf.
[393] See: http://www.genomicsengland.co.uk/about-genomics-england/how-we-work/.
[394] GeL peremptorily signed contracts with the sequencing provider, Illumina, to generate the genome sequences, thereby inflating a bubble in UK-based sequencing capacity and associated infrastructure. 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/.

data and genome sequences. All operations within the environment will be traceable so that, for example, it will be possible to see who has reviewed what data, when and for how long and GeL will carry out regular 'penetration testing' to assess security.[395] When their analyses are complete, users will only be able to export their results (and not the raw data) across the NHS firewall.[396] There is also an intention to develop a function that allows users to build 'apps' within the environment to interrogate data.[397]

Consent

6.58 The project will seek patients' active and explicit consent to retrieve samples and generate genome sequences and to authorise clinical, research and commercial use of the data, clinical feedback, re-contact through the patient's clinician and lifelong access to patients' medical records to allow the continuing updating of the GeL database. In other words, there will be a 'broad' initial consent procedure (for which UK Biobank has been widely cited as a model – see chapter 7, below). The richness of whole genome sequence and other -omic data, its association with one individual, and the possibility of it revealing predictive information about biological relatives, raises familiar privacy concerns as well as important treatment/benefit options for family members.[398] GeL states clearly that it cannot promise study participants that it will not be possible for users of the data to identify them, despite the security measures in place to prevent this.

6.59 Because many participants are patients or relatives of patients with serious disease the structure of their interests and motivations may be different than, for example, with an administrative or prospective research database involving apparently healthy people. Although they may have strong incentives (both self- and other-regarding) or dispositional vulnerabilities (owing to their status as patients) the thorough consent process suggests that their preferences can be respected, even if the options available do not correspond to their personal interests: even if, for example, the prospect of commercial use of the resource is a personal disincentive, it need not be morally unreasonable to offer people an option that allows them to trade this off against positive benefits as they see them. In other words, the circumstances of consent do not include coercion or unreasonable inducement, although there is potentially some uncertainty about the possible personal therapeutic benefit. [399] In clinical genetic practice it has sometimes been the case that the boundary between clinicians as physicians and as researchers has become somewhat blurred. So, for example, having collected samples from patients for genetic aetiological research clinicians may have later been in a position to provide clinical benefit through feedback of information about genetic risks for patients and family members. In the 100,000 Genomes project clinicians will both recruit patients to the project and later may carry responsibilities for the feedback of clinically significant information to their patients or to use this to inform their treatment. Clearly, in such a situation, it is important that the patients being

[395] See: Notes of the meeting on 24th March, 2014 between patient advocates and Professor Michael Parker and Ms. Vivienne Parry of Genomics England, available at: http://independentcancerpatientsvoice.org.uk/consultations/.
[396] See http://www.genomicsengland.co.uk/town-hall-engagement-event/. The arrangements for the pilot phase are different from the main project – see: http://www.genomicsengland.co.uk/annotation-supplier-event/.
[397] Described by Professor Tim Hubbard at Progress Educational Trust/GeL event "Genetic Conditions: how should your DNA be used in the 100,000 Genomes Project?" See: http://www.progress.org.uk/geneticconditions.
[398] Lucassen A and Parker M (2010) Confidentiality and sharing genetic information with relatives *The Lancet* **375(9725)**: 1507-09.
[399] On the 'therapeutic misconception', see paragraph 4.37.

recruited have clear and realistic expectations about the clinical benefits that may or may not accrue to them through participation in the project.

6.60 Patients and patient groups were, in fact, consulted and involved in the design of the consent process and materials. Although the consent given by participants is not open to additional conditions, and therefore does not offer an opportunity to express preferences that might shape the design of the resource or its use beyond their simple participation or non-participation, it is clear that candidate patients' receipt of the best currently available treatment is not dependent on their participation in the 100,000 Genomes project.

6.61 The long term nature of the resource is also relevant. It is possible that clinically significant findings might emerge at any point throughout the life time of the project. However, families also change over time and the salience of findings may change for family members (for example, as they face reproductive choices). Some mechanism for managing these uncertainties, such as setting a time limit for feedback of information, will therefore be of great importance. Many of the most significant ethical questions relate to how the use of the resource will also develop in the future and what this will mean in changing circumstances. The mechanism for responding to such changes relies on the governance provided through the institution, in which the Ethics Advisory Committee plays an important role, and the option for participants to withdraw from the study.

Elements of an ethical approach

6.62 Looking at the 100,000 Genomes project through the lens of the approach we set out in chapter 5 focuses attention on two aspects in particular: the design of the project (and how this incorporates the public and private interests involved) and the governance of the project (and how well this respects the interests of participants in changing circumstances).

6.63 Though there were a number of 'Town Hall' meetings in the development phase of the project, and the consent materials were developed through interviews and focus groups with patients and members of the public of diverse ages and backgrounds, there is no evidence of broader social engagement around questions of project design.[400] Like NPfIT, the policy process was carried out in haste with a strong political (indeed, Prime Ministerial) impetus, and many of its key parameters and elements of infrastructure were locked in prior to determination of the governance systems. Without social accountability there is a possibility that political, commercial and health drivers may conflict (or, at least, that their relationship may appear ambiguous, with potentially adverse consequences for broader public trust). Especially given the context of ambitious public rhetoric about realising the promise of genomic medicine in the UK, special care needs to be taken to avoid overselling the prospects of therapeutic benefit to participants.

6.64 We know, from previous public engagement, that the issue of commercial involvement excites particular preferences for many, who treat this as a morally salient, rather than morally irrelevant, feature.[401] This may be due to considerations of justice (equitable sharing of benefits), trust (belief that commercial involvement represents greater privacy risk) or other reasons. Although the choice to participate is not coerced, the

[400] See, however: http://www.genomicsengland.co.uk/town-hall-engagement-event/.
[401] See paragraph 5.18.

question arises whether the design of the project could have offered a morally preferable set of options than those offered by GeL. This is not a trivial question since, though a government-owned company, GeL exists to promote the public interest. Of course, the public interest may also be served by generating national income though the commercial activities stimulated by the resource. Furthermore, commercial input is an increasingly important part of contemporary academic research and ruling this out completely (as the 1958 birth cohort does, for example) may mean that the capacity for potentially desirable academic research is limited.[402] Given that there are other ways to achieve this, as the example offered by SHIP shows (see above), the case needs to be made out publicly that GeL represents the politically, scientifically and morally optimum resolution of the public and private interests at stake.

6.65 The second question relates to the possible implications of as-yet-undefined uses of data in what may be a technologically and informationally very different future environment, and how these will be governed for public benefit and the protection of individual interests. Given the broad consent model and the breadth of potential uses and users, once the 100,000 genomes resource is established, considerable reliance will be placed on the governance system. One way of ensuring broader accountability would be to set out transparently the set of morally reasonable expectations about data use within the powers that are available to decision makers and to use this as a focus for both governance and for more inclusive decision making about the future of this public resource. The key document will be the data access policy.[403] Such mechanisms, should not limit the flexibility and ability to seize opportunities that were seen as advantages of constituting GeL as a private sector actor, but would provide greater clarity and assurance with regard to policy.

6.66 Genomic testing is likely to become more routine in health systems and GeL may form a bridgehead for new forms of data linking in the NHS. The first 100,000 genomes could well be merely the vanguard for a more substantial genomic and phenotypic database, especially as the value of such databases increases significantly in relation to their size. Whether the model offered by GeL represents the most appropriate model for securing the public interest in the ethical use of genomic information in health services is therefore an important question since this was not extensively or publicly debated prior to the initiation of the project.

Recommendation 9

We recommend that broader public consideration should be given to whether GeL provides the most appropriate model for the ethical use of genomic information generated in health services for public benefit before it becomes the *de facto* infrastructure for future projects.

[402] For information on the '1958 Birth Cohort' (the National Child Development Study), see: http://www2.le.ac.uk/projects/birthcohort/1958bc.
[403] At the time of writing this is still under development.

Conclusion

6.67 In this chapter we have looked more closely at the formation of a number of data initiatives that represent different approaches to the linking and re-use of data in health systems. We have looked at this formation not only structurally, but also in terms of how they came about, what incentives and drivers pulled and pushed them in different directions, and in the light of our contention that data initiatives where public and private interests are at stake are social and political practices, as well as moral and scientific ones.

6.68 Each of the data initiatives resolves questions about centralisation and distribution of resources, how data are disclosed or accessed, the range of users and purposes, and how control is exercised in different ways. Thus the HSCIC and GeL models are more centralised than SHIP; the HSCIC model allows some disclosure of individual-level data whereas SHIP and GeL will only disclose the results of analyses carried out within their infrastructure; HSCIC and GeL allow direct access to data by commercial companies whereas SHIP is more guarded; GeL requires explicit, though broad, patient consent, whereas HSCIC and SHIP have (at least initially) rather different authorisation procedures in which individual preferences and values may figure (through a rather blunt opt-out or through a more constitutive participation); etc. Though there are lessons to be shared amongst these various experiences, perhaps the most salient lessons relate to our principles of participation and accounting for decisions through formal governance and wider social engagement. These strongly suggest that there are serious consequences for public trust and for the viability of data initiatives if they do not first take steps to identify the applicable moral norms that they must negotiate and put in place, in relation to these, well-supported measures to respect the interests engaged, supported by credible justification.

6.69 As we have argued a key question that faces data initiatives and the health systems as a whole is what uses of data should be ‘expected’ as part of delivering national health care with quality, safety, and cost-effectiveness with ongoing improvement in the standards of care. It is becoming increasingly evident that there are commercial drivers behind many high-profile initiatives that have been proposed in recent years and, as empirical studies show, this is of significant concern for the public. The issues go beyond individual privacy, especially as datasets start to be linked together, and there is a need to have governance structures in which all interests are enabled to participate and that involve continuing review and reflection on the societal implications of such initiatives. Unless there are trustworthy governance systems in place that can engage with and reflect reasonable expectations in continuously evolving circumstances, initiatives that could have wide public benefits may continue to be challenged and fail to secure public confidence.

Chapter 7

Population research data initiatives

Chapter 7 – Population research data initiatives

Chapter overview

This chapter describes research data initiatives in three clusters (biobanks, international collaborative research projects, and participant-driven research), identifies examples of good practice and draws lessons from some specific initiatives.

Biobanks are major resources of tissues and data that may be used for a variety of research purposes. UK Biobank is a large population biobank established to support the investigation of a range of common diseases occurring in the UK. Key features of this are the broad consent model, its Ethics and Governance Framework and the independent Ethics and Governance Council. Questions arise for which it is necessary to review the set of expectations underlying its operation, including feedback of findings and commercial access to the resource.

The UK10K Rare Genetic Variants in Health and Disease project confronts the problem of controlled disclosure of highly specific individual-level data among different groups of researchers working on distinct studies. It achieves this through a common ethical framework of policies and through guidelines that place considerable reliance on institutional sanctions and on the role of principal investigators.

International collaborative research initiatives such as the International Cancer Genome Consortium and the Psychiatric Genomics Consortium need to accommodate differing local practices and tackle complex consent issues to do with re-use and international transfer of data. The use of cloud-based storage and processing services is becoming increasingly important but it raises issues such as third party access (for example, by security services).

The wide availability of social networking platforms has facilitated participant-led research with norms and social dynamics that differ from more formal institutional research. They present distinct challenges of ensuring the protection of individual interests, of integration with institutional research, and of translation of findings into clinical products and practice.

A number of recommendations are made in relation to a greater role for subject participants, accounting for governance through explicit frameworks, and the use of institutional measures.

Introduction

7.1 Increased capability for linking databases, along with technological and IT innovations, have accelerated the pace of data acquisition for large-scale population studies. Typically, population research initiatives of this type collect and store information from identifiable individuals, the study participants. They are supported by the increasing pace and falling cost of automated collection and laboratory analysis of biomedical samples and information, including genetic analysis. Many of the research methodologies in use are well established and build on earlier developments in population epidemiology and data science, but life sciences research is turning towards increasingly large resource platforms for use by international teams of researchers studying a wide range of health-related conditions, and using data from many thousands of individuals. Another feature of these initiatives is the richness of the data available for each participant, made possible by linking existing medical databases, the inclusion of data derived from genomic analysis, as well as ‘deep phenotyping’ (see

chapter 1) including lifestyle and behavioural data, data from imaging technologies and data derived through sensing devices.

7.2 Biomedical research raises significant ethical and governance issues including recruitment of participants, how their morally relevant values and interests are respected (for example, through the choice of consent mechanisms), data security, decisions on feedback of medical information to participants, data access arrangements for researchers, data linking and movement trans-nationally, governance of the resource in the public interest, and strategies for disseminating the outcomes of research. In this chapter we consider a number of data initiatives in three broad clusters: biobanks, international collaborative research projects, and participant-driven research.

Biobanking

7.3 The term 'biobank' has become a catch-all phrase for many types of collection of biological samples and related data.[404] Here we are concerned with collections that are established as prospective research resources comprising material and data from many participants. Our examples are two flagship British resources, UK Biobank and UK10K.

UK Biobank

7.4 The UK Biobank initiative is a major resource designed to support a range of research 'to improve the prevention, diagnosis and treatment of a wide range of serious and life-threatening illnesses' and the promotion of health throughout society.[405] Through invitations sent to patients on geographically selected general practice lists, more than 500,000 people between the ages of 40 and 69 were recruited. The intention is to follow them for at least 25 years through their GP and NHS hospital records as well as through periodic collection of data directly from the subject participants themselves.[406] In 2012 the resource opened for use by researchers worldwide. Applications to use the resource are screened to ensure that projects meet established ethical and scientific standards, and to consider how the research meets the criterion of being in the public interest. Researchers using the resource give undertakings to abide by certain conditions of use and to treat data confidentially. They pay a modest fee calculated on a cost recovery approach for providing the data or samples. UK Biobank carries out extensive work validating and cleaning data and preparing it for use but does not undertake the research itself.

7.5 An International Scientific Advisory Board advises on scientific and policy matters and the research community is invited to contribute to shaping the direction of future data

[404] See: Kaye J, Gibbons SMC, Heeney C, and Parker M (2012) *Governing Biobanks: understanding the interplay between law and practice* (Oxford: Hart Publishing); Shaw DM, Elger BS, and Colledge F (2014) What is a biobank? Differing definitions among biobank stakeholders *Clinical Genetics* **85(3)**: 223-7.

[405] See: http://www.ukbiobank.ac.uk/about-biobank-uk/.

[406] At the time of recruitment, participants give consent for UK Biobank to have long-term access to their existing and future NHS medical records and other health-related records. The consent form states:"I give permission for access to my medical and other health-related records, and for long-term storage and use of this and other information about me, for health-related research purposes (even after my incapacity or death)." Participants are able to withdraw from the study at any time on the basis of either no further contact, no further access to their data or no further use of their data. See http://www.ukbiobank.ac.uk/wp-content/uploads/2011/06/Consent_form.pdf and http://www.ukbiobank.ac.uk/faqs/can-i-withdraw-from-uk-biobank/.

acquisition. Right from the start it was recognised that such a large multi-purpose biobank resource, designed to collect data and samples prospectively to facilitate research, would raise ethical and governance challenges both at recruitment and in later decades. Decisions were taken to put in place two distinctive measures, namely, the UK Biobank Ethics and Governance Framework and an independent Ethics and Governance Council to advise the funders and UK Biobank.[407]

Recruitment

7.6 There is a long history of population studies in Britain, including the 1946 and 1958 Birth Cohorts studies up to the Life Study begun in 2012.[408] It was recently estimated that 2.2 million people are currently taking part in large population studies in the UK (approximately one in thirty of the general population).[409]

7.7 UK Biobank sought to recruit as widely as possible across England, Wales and mainland Scotland (but not in Northern Ireland). Participants completed an automated questionnaire at local assessment centres and were interviewed about lifestyle, medical history and diet. In addition, basic assessments including weight, body mass index (BMI), heart function, blood pressure and bone density were made. Blood and urine samples were taken to be assessed for biomarkers, and DNA extracted for genomic analysis.[410]

7.8 The overall UK Biobank volunteer rate was approximately 5.5 per cent of those approached.[411] The locations of Assessment Centres, and the surrounding GP practices where the potential participants were registered, were selected with the aim of creating a generalisable population sample, 'so that research may ultimately benefit a wide diversity of people'.[412] It is well known that it is harder (and more expensive) to recruit to population studies those with low income and of lower social status, those with poorer health and/or chronic conditions, and those from ethnic minority and rural communities.[413] This proved to be the case for the UK Biobank sample. It is unclear how far strategies to ensure greater representation from at least the larger UK ethnic minority groups were pursued, but it seems that a compromise was reached between the costs of ensuring participation from hard to reach communities and accomplishing timely recruitment of 500,000 people. Regardless, some researchers have been critical of the lack of representativeness of the sample for the UK population.[414] The UK Biobank recruitment strategies were justified on the grounds that nested case-control studies do not require a (near) representative sample. This may, however, limit the

---

[407] UK Biobank has been funded (about £62 million by the time participant recruitment was completed) by the Medical Research Council, the Wellcome Trust and the Department of Health together with a number of other public and charitable resources. For the EGF and EGC, see: http://www.ukbiobank.ac.uk/ethics/.

[408] See the Cohort and Longitudinal Studies Enhancement Resources (CLOSER*)* for details of British birth cohort studies: http://www.closer.ac.uk/.

[409] Pell J, Valentine J and Inskip H (2014) One in 30 people in the UK take part in cohort studies *The Lancet* **383(9922)**: 1015-6 and http://www.mrc.ac.uk/news-events/publications/maximising-the-value-of-uk-population-cohorts/.

[410] This was followed by an enhanced baseline assessment including eye measure, additional blood collection for RNA analysis and saliva sample collection. See http://www.ukbiobank.ac.uk/data-showcase-timeline.

[411] Allen N. Sudlow C, Downey P, et al. (2012) UK Biobank: current status and what it means for epidemiology *Health Policy and Technology* **1(3)**: 123-6; see also Swanson JM (2012) The UK Biobank and selection bias *The Lancet* **380(9837)**: 110.

[412] See UK Biobank Ethics and Governance Framework, http://www.ukbiobank.ac.uk/wp-content/uploads/2011/05/EGF20082.pdf, at page 4.

[413] Ridgeway JL, Han LC, Olson JE *et al.* (2013) Potential bias in the bank: what distinguishes refusers, nonresponders and participants in a clinic-based biobank? *Public Health Genomics* **16(3)**: 118-26; for ethnic minorities, issues include distrust of the medical profession, lack of awareness and economic burden. See Paskett ED, Reeves KW, McLaughlin JM *et al.* (2008) Recruitment of minority and underserved populations in the United States: the centers for population health and health disparities experience *Contemporary Clinical Trials* **29(6)**: 847-61.

[414] Swanson JM (2012) The UK Biobank and selection bias *The Lancet* **380(9837)**: 110.

range of research studies that use UK Biobank since those requiring a (near) representative sample may lack sufficient scientific validity. It also follows that the ambition for research outcomes to benefit a 'wide diversity of people' might fail to be achieved equally.

Consent and governance

7.9 Population-based biobanks generally create a resource for the use of researchers for future research that is unspecified at the time of collection. Thus, at the time of recruitment it is not possible to tell would-be participants in detail what research may be carried out with the data and samples they may donate. A 'broad consent' model is therefore adopted by many biobanks and was adopted by UK Biobank. Participants were given information about what data and samples would be collected and how the project would be governed but the scope of future research was defined only in general terms, as health-related research in the public interest. Participants are able to withdraw at any time from the project to different degrees: they may choose no further contact, no further access to their data or no further use of their data.[415] As we noted in chapter 4, the broad consent model is not uncontroversial and these controversies have found a focus in relation to biobanks.[416] As we argue in chapter 5, however, consent, as a way of respecting morally relevant individual values and interests, is only ever part of the story – it is the relationship between norms of privacy, the way of respecting individual preferences (in this case, through broad consent), and mode of governance regulating the public and private interests in play, that determines whether a data initiative can find an ethically acceptable form. While broad consent has been a practical solution to the difficulties of obtaining informed prospective consent for a large number of diverse research projects from thousands of participants, the need for complementary governance structures is generally accepted. For many population studies, an ethics oversight committee is established, but for UK Biobank an independent Ethics and Governance Council (EGC) acts as the guardian of a dedicated and detailed Ethics and Governance Framework (EGF) (see Box 7.1.)

Box 7.1: UK Biobank Ethics and Governance Council and Framework

The UK Biobank Ethics and Governance Council (EGC) is an advisory body with members appointed by the funders independently of UK Biobank. It has no formal regulatory role but rather advises UK Biobank in the manner of a 'critical friend'.[417]

The Ethics and Governance Framework (EGF) sets out the relationship between UK

[415] See http://www.ukbiobank.ac.uk/faqs/can-i-withdraw-from-uk-biobank/. They are unable, however, to have their data expunged entirely from the biobank's systems. In fact, in June 2007, UKBB was obliged to amend its advice regarding the 'no further use' option offered as part of the right to withdraw as a consequence of a technical feature of the data archiving system, after some participants had been recruited. The advice was amended to indicate that while information would be made unavailable to researchers it would nevertheless be retained for archival and audit purposes. Furthermore, UK Biobank advises that while it would destroy all biological samples, 'it may not be possible to trace all distributed sample remnants' and data could not be removed from completed analyses. See: 'No further use' withdrawal option: February 2008' at http://www.ukbiobank.ac.uk/resources/.

[416] For criticism of broad consent as inadequately informed, see: Greely, HT (2007). The uneasy ethical and legal underpinnings of large-scale genomic biobanks *Annual Review of Genomics and Human Genetics* **8(1)**: 343-364; for a defence of broad consent for larger-scale research infrastructure projects, see: Knoppers, BM (2005) Consent revisited: points to consider *Health Law Review* **13(2-3)**: 33-8; Hansson MG, Dillner J, Bartram CR, *et al*. (2006) Should donors be allowed to give broad consent to future biobank research? *The Lancet Oncology* **7(3)**: 266-9. See also: Corrigan O, McMillan J, Liddell K, *et al*. (Editors) (2009) *The limits of consent* (Oxford: Oxford University Press).

[417] Laurie, G (2011) Reflexive governance in biobanking: on the value of policy led approaches and the need to recognise the limits of law *Human Genetics* **130(3)**: 347-56.

> Biobank and participants, research communities, individual researchers and society. This twenty page document was widely circulated and discussed in the research and bioethics communities, and at public meetings, before a version was agreed by the funders during the planning stage of the project. The EGF may be seen as an instrument, legitimised through wide discussion, which serves to align the public interests in research and the privacy and other interests of participants, as well as engendering trust. It is a 'living' document that is intended to be responsive to changing needs and circumstances. The EGC is charged with monitoring and reporting publicly on the conformity of UK Biobank with the EGF and to advise more generally on the interests of research participants and the general public in relation to UK Biobank.

7.10 While participants cannot know in advance for exactly what purposes their data will be used, UK Biobank has attempted to keep them informed about the developments of the resource, the research being carried out using it and the results of that research. It does this via newsletters and information on the website, as has been the case with other biobanks.[418] It is a way of encouraging the continued support and motivation of participants on which the long-term value of the resource depends.[419] But it is also important in order to enable participants meaningfully to exercise the option to withdraw from the resource if they consider that morally salient features of its use, as approved under the governance system in place, depart from the scope of their expectations.[420] Aside from this, however, these means of communication are mainly one-way, an opportunity for the biobank to deliver information, rather than a deliberate opportunity to hear the views of participants or involve them in the development of the resource.

7.11 The EGF states, among other things, that 'further consent will be sought for any proposed activity that does not fall within the existing consent'. As circumstances change it may not be clear whether new proposals raise issues that fall outwith the scope envisaged by the initial broad consent. However, relying on renewed consent may limit activities that are in the interests of participants and in the public interest. This is because seeking renewed consent from the whole body of participants may be impractical, very expensive and likely to damage the resource as many participants, especially those who are hard to reach, might drop out by default. Thus activities beyond the obvious scope of initial consent are likely to be avoided. Possible examples here include data linking to other biobanks or to administrative data that are not directly health related.

7.12 Revising the EGF is one potential response to these changing circumstances and the changing horizons of public and private interests. However, revising the EGF simply by agreement with the Board of Directors falls short of the standard of engagement set by UK Biobank's creation and legitimisation, which was characterised by wide-ranging and inclusive discussions, pursued through public meetings. Despite time and resource implications there is growing support for the view that long-term studies should attempt to involve participants in a meaningful way and some initiatives have moved in this

[418] McCarty CA, Garber A, Reeser JC, and Fost NC on behalf of ; the Personalized Medicine Research Project Community Advisory Group and Ethics and Security Advisory Board (2011) Study newsletters, community and ethics advisory boards, and focus group discussions provide ongoing feedback for a large biobank *American Journal of Medical Genetics* **155A(4)**: 737-41.

[419] Gottweis H, Gaskell, and Starkbaum J (2011) Connecting the public with biobank research: reciprocity matters *Nature Reviews Genetics* **12(11)**: 738-9.

[420] Melham K, Briceno Moraia L, Mitchell C, *et al*. (2014) The evolution of withdrawal: negotiating research relationships in biobanking *Life Sciences, Society and Policy* **10**:16, available at: http://link.springer.com/article/10.1186/s40504-014-0016-5.

direction.[421] However, whereas the UK Biobank EGF proposes a representative participants' panel, this has not been pursued to date.[422] Given its central role in the regulation of the relationship between UK Biobank and its participants, there is a strong argument both that the process of revision and development of the EGF would benefit significantly from including participants' views and interests, and, as we argue in chapter 5, there is a substantive moral reason for them to participate.[423] Furthermore, the involvement of 'publics' (e.g. participants, policy makers, researchers, future beneficiaries and other publics) in open discussion, particularly in relation to issues where practical decisions are to be made, can help build, maintain and develop a trusted governance structure.[424]

Data security, access and linkage

7.13 Permanently de-identified data (which the data controller cannot link back to the individual case) is little use in the context of continuing longitudinal research. Data and samples are therefore assigned a pseudonym or code (see chapter 4) so that they can be linked back to the same index case over time. The obvious personal identifiers (names, addresses, etc.) are separated and stored in a protected file store. Coded data are then made available to researchers, who are usually bound by undertakings to not try to re-identify participants. UK Biobank has an Access Committee (a subcommittee of the UK Biobank Board) that takes decisions about research access in the light of advice from UK Biobank managers and external ethics advisors. Access requirements for users focus on the three areas described below in order to protect the confidentiality of participant data, as well as to promote the trustworthiness of the project.

7.14 First, researchers are checked to see if they are 'bona fide' (acting 'with good faith') and from recognised institutions (and so governed by ethical codes of practice). If there is a breach of use by a researcher, there are a number of ways in which penalties may be imposed.[425] Although there may be no legal basis for UK Biobank to reprimand individuals, beyond refusing further access to the resource, there exist ways that those using data could be penalised by their institution and ways that some institutions can be penalised by funders if one of their staff breaks the rules, which may be given effect through contracts and legal agreements.[426] Second, proposals must meet criteria set by the biobank for the use of the resources. Access committees need to review requests for samples to ensure that the research is scientifically valid and the use falls

[421] UK Biobank Ethics and Governance Council (2009) Workshop report: involving publics in biobank research and governance, available at: http://www.egcukbiobank.org.uk/sites/default/files/meetings/EGCper cent20workshopper cent20report.pdf.

[422] UK Biobank Ethics and Governance Framework, http://www.ukbiobank.ac.uk/wp-content/uploads/2011/05/EGF20082.pdf, at page 8.

[423] As it happens, some members of the EGC have also been participants. However, there is no requirement that there should be participants on the Council.

[424] Burgess MM (2014) From 'trust us' to participatory governance: deliberative publics and science *policy Public Understanding of Science* **23(1)**: 48-52. See also Armstrong V, Barnett J, Cooper H, et al. (2007) Public attitudes to research governance. On the Governance of biomedical research: a qualitative study in a deliberative context, available at: ghttp://www.wellcome.ac.uk/stellent/groups/corporatesite/@policy_communications/documents/web_document/wtx038443.pdf.

[425] Joly Y, Zeps N, and Knoppers BM (2011) Genomic databases access agreements: legal validity and possible sanctions. *Human Genetics* **130(3)**: 441-9.

[426] The 2008 Thomas and Walport Data Sharing Review Report points, with approval, to the application of legal penalties through the Statistics and Registration Service Act 2007: "The Board may extend access to researchers from various organisations, including academic institutions, public bodies and nongovernmental organisations. These researchers are then bound by a strict code, which prevents disclosure of any personal identifying information. Any deliberate or negligent breach of data security by the approved researcher would entail criminal liability and the prospect of a custodial sentence up to a maximum of two years." Thomas R and Walport M (2008) *Data sharing review report*, available at: http://systems.hscic.gov.uk/infogov/links/datasharingreview.pdf/view. See also our recommendation 5 above.

within the broad categories given in the participant consent. Third, researchers must agree to certain undertakings regarding the confidentiality of the data and handling of samples. Material and Data Transfer Agreements are used to bind researchers through their institutions.[427] Researchers are required, for example, to provide a secure environment for the samples and data, to ensure that the data are not used other than for the agreed purposes and not to attempt to re-identify individuals from data. UK Biobank also asks researchers to return results of their work, which may be added to the resource, and for research results to be published, which promotes outcomes for public benefit and thereby demonstrates respect for the altruistic motivations of participants.[428]

7.15 One of the core aims of UK Biobank is to link data provided by participants with other health-related and administrative records in order to track the emergence and/or progression of disease and to collect data to support epidemiological research. The complex process of linking UK Biobank records to both NHS hospital and GP records is currently underway. A recent MRC strategy review of population cohort studies saw increasing opportunities for cohort studies not only to link to NHS records but to also to link more widely to cross-sector administrative and environmental information. It noted a number of initiatives that will improve secure access to data.[429] However, the resource could potentially have much broader uses, some of which may challenge the initial health-related purposes or address them in unusual and unanticipated ways. Examples already exist of biobanks seeking to link their data with criminal convictions and cautions, as well as financial benefits, earnings and employment data.[430]

7.16 There are also opportunities to use commercial data (such as geospatial location data from mobile phones) in research. It has been suggested that purchasing data from supermarkets might be used to infer the effects of diet on the health of research participants. With researchers' growing interests in 'deeper' phenotyping (see chapter 1) the appetite for data from a wider range of sources is likely to increase. As we noted at the outset, when these are linked in the context of health-related research, this data may be informatively 'health-related' but such linkages may well fall outside the expectations participants had when they signed up to participate.[431] To determine this, it may be necessary to reason in relation not only to the nature of the research proposal within the developing field of science and the initial expectations of participants but also the norms that apply at the time.[432] As further opportunities emerge, enhanced participation may prove to be an important way of demonstrating respect for subject participants and undertaking the moral reasoning necessary to relate the scope of reasonable expectations to the scope of potential uses.

---

[427] Fortin I, Pathmasiri S, Grintuch R, and Deschênes M (2011) 'Access arrangements' for biobanks: a fine line between facilitating and hindering collaboration *Public Health Genomics* **14(2)**:104-14.

[428] UK Biobank Return of Results Data: Guidance Note for Approved Projects, available at: http://www.ukbiobank.ac.uk/wp-content/uploads/2011/06/Return-of-Results_Guidance-Note_v2.pdf. Where research results may be seen as controversial, UK Biobank can ask for sight of papers before publication.

[429] Medical Research Council (2014) Maximising the value of UK population cohorts. MRC strategic review of the largest UK population cohort studies, available at: http://www.mrc.ac.uk/news-events/publications/maximising-the-value-of-uk-population-cohorts/.

[430] See for example, the Avon Longitudinal Study of Parents and Children data linkage information. http://www.bristol.ac.uk/alspac/researchers/resources-available/data-details/linkage/.

[431] See footnote 407 above.

[432] See Laurie, G (2009) Role of the UK Biobank Ethics and Governance Council *The Lancet* **374(9702)**: 1676, available at: http://www.thelancet.com/journals/lancet/article/PIIS0140-6736per cent2809per cent2961989-9/fulltext. Whether proposed activities fall within the scope of the original consent "depends on what is proposed scientifically, expectations of participants, and social mores at the time of an application."

Overlap between research and medical care

7.17 UK Biobank states clearly that participants are not likely to benefit directly from participation; the intention is instead that research discoveries will benefit future generations and be in the broad public interest. This appeal to altruistic participation mirrors appeals to participate in blood donation programmes or early phase clinical trial research. But despite no lack of clarity in the message from UK Biobank there remains some ambiguity at its reception. For example, follow-up research in relation to other biobanks has shown that participants may regard the initial assessment as a 'health check'.[433]

7.18 This tension is particularly keen when looking at the potential to feed back health-related findings that may have significant benefit for individuals. At the initial assessment, UK Biobank participants are given results from some of the measurements taken, such as their blood pressure and weight. If staff notices abnormalities, such as elevated blood pressure or a suspicious mole, they may advise the participants to see their GP.[434] However, participants are told at the time of recruitment that no further personal feedback would be offered either from analyses carried out by UK Biobank or by researchers using the resource. This blanket 'no feedback' policy was typical for many population and cohort studies that had been established prior to UK Biobank.[435] Since then, however, such a policy has become more contentious in the light of discoveries that may come about through further data collection from participants or as a result of the analysis of data. It is now often argued that there is a moral obligation to consider feedback of health-related findings, including in the case of whole genome sequencing.[436]

Box 7.2: UK Biobank imaging study

UK Biobank is seeking to enhance the resource through the addition of an imaging study. In the pilot study, launched in May 2014, existing participants are invited for an integrated series of imaging studies of the brain, heart, abdomen, bones and carotid arteries.[437] The aim is to use these data as part of an increasingly detailed ('deep') phenotyping of participants to enhance the potential to deliver research objectives of UK Biobank.

In discussion with the EGC, the UK Biobank International Scientific Advisory Board, the project's funders and others (though not subject participants), it was agreed that during the pilot phase participants and their GPs would receive feedback on any serious health-related finding that might be observed during the collection of the imaging data.[438] Those receiving this feedback will be followed up so the consequences of being offered such

[433] Halverson CME and Ross LF (2012) Incidental findings of therapeutic misconception in biobank-based research *Genetics in Medicine* **14(6)**: 611-5, available at: http://www.nature.com/gim/journal/v14/n6/abs/gim201150a.html.

[434] See UK Biobank Ethics and Governance Framework, http://www.ukbiobank.ac.uk/wp-content/uploads/2011/05/EGF20082.pdf, at page 7.

[435] Wallace SE and Kent A (2011) Population biobanks and returning individual research results: mission impossible or new directions? *Human Genetics* **130(3)**: 393-401.

[436] Wolf SM, Crock BN, Van Ness B, *et al*. (2012) Managing incidental findings and research results in genomic research involving biobanks and archived datasets Genetics in Medicine 14(4): 361-84. See also Medical Research Council and Wellcome Trust (2014) Framework on the feedback of health-related findings in research, available at: http://www.wellcome.ac.uk/stellent/groups/corporatesite/@policy_communications/documents/web_document/wtp056059.pdf.

[437] See http://www.ukbiobank.ac.uk/2014/05/uk-biobank-imaging-study-launched/.

[438] Some participants were invited to comment on the information materials for the imaging study but the direction of the study was set, effectively, by the groups mentioned.

feedback can be assessed. This will provide vital evidence of the implications and consequences of receiving feedback, an area that is remains very much debated worldwide.[439]

7.19 One of the assumptions behind offering findings to participants is that it is likely to prove of benefit through, for example, earlier diagnosis and therefore more effective treatment of a condition. There is also an acknowledgement that some people may be harmed, for example by anxiety caused by notification about a serious abnormality that may be revealed later to be of little or no consequence to their well-being. However, given the lack of evidence about the consequences of providing such feedback it is unclear what the benefits or disadvantages may be. Some have argued that it is not the right time for biobanks to institute such a policy.[440] However, if a project considers revising a current, general 'no feedback' policy, this may be considered as outside the terms of the initial consent given by participants.[441]

Commercialisation

7.20 One specific use of UK Biobank data that has stimulated discussion is the use by commercial entities. While research suggests that academic researchers are generally trusted by the public, industry is viewed with more suspicion owing to its supposedly more mixed motives (as we note in chapter 5).[442] The debate focuses on whether the outcomes of the research will be shared and benefits returned to the public domain instead of boosting profits of commercial entities. There may also be fears that private interests will restrict the benefits available to the public, through, for example, commercial pricing of products.[443] It is possible that public discussion about the commercial use of the resource may have led to some people choosing not to join the study and so depressing the uptake rate for the project.

An ethical framework

7.21 A review of the conditions under which UK Biobank has been established suggests that it does provide a secure moral basis for the proposed uses in most respects. There was initial consultation with the public and other stakeholders in the development of the plans for the project and its governance. In order to respect the range of interests involved while at the same time acknowledge uncertainty about the specific future uses of the resource it uses a model of broad consent, together with a governance framework for the use of data that includes a criterion of public interest to ensure conformity with the expectations of the stakeholders. The EGF provides an explicit

[439] See Johnson KJ and Gehlert S (2014) Return of results from genomic sequencing: A policy discussion of secondary findings for cancer predisposition *Journal of Cancer Policy* **2(3)**: 75-80; Hallowell N, Hall A, Alberg C and Zimmern R (2014) Revealing the results of whole-genome sequencing and whole-exome sequencing in research and clinical investigations: some ethical issues *Journal of Medical Ethics* (Online First), available at: http://jme.bmj.com/content/early/2014/07/18/medethics-2013-101996.long#responses.

[440] Viberg J, Hansson MG, Langenskiold S and Segerdahl P (2014) Incidental findings: the time is not yet ripe for a policy for biobanks *European Journal of Human Genetics* **22(4)**: 437-41, available at: http://www.nature.com/ejhg/journal/v22/n4/abs/ejhg2013217a.html.

[441] Trinidad SB, Fullerton SM, Ludman EJ *et al.* (2011) Research practice and participant preferences: the growing gulf *Science* **331(6015)**: 287-8, available at: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3044500/; Stjernschantz Forsberg J, Hansson MG, and Eriksson S (2011) The risks and benefits of re-consent *Science* **332(6027)**: 306.

[442] Clemence M, Gilby N, Shah J, *et al*. (2013) *Wellcome Trust monitor wave 2: tracking public views on science, research and science education*, available at: http://www.wellcome.ac.uk/stellent/groups/corporatesite/@msh_grants/documents/web_document/wtp053113.pdf.

[443] Huzair F and Papaioannou T (2012) UK Biobank: consequences for commons and innovation *Science and Public Policy* **39(4)**: 500-12; Caulfield T, Borry P, and Gottweis H (2014) Industry involvement in publicly funded biobanks *Nature Reviews Genetics* **15**: 220.

correlative for the expectations of participants and, as a 'living' element of governance, provides a focus for reflective governance. The challenge to UK Biobank is to ensure that this ambition is realised in practice, in response to developments, such as developments in the potential uses of data we describe in this report.

7.22 Because of the developments of the initiative (such as the introduction of imaging) and new possibilities for using the data, as well as increasing knowledge (for example, in human genetics), this process of reflection and engagement with participants, should be maintained throughout the life of the initiative. The norms and expectations, such as 'no feedback', on which initial consents were premised, and corresponding views about the level of duty of care owed to participants, may alter in relation to new information. The EGC has responded to these but there is no embedded process through which others' views can be engaged in relation to such matters as part of the revision of the ethics and governance framework and the development of practice and governance more generally.[444] There is scope for bringing the views of the subject participants, the research users of the resource and the greater research community into this reflection so that both the promotion of research in the public interest and the privacy and other interests of all participants in the process are enhanced.

UK10K Rare Genetic Variants in Health and Disease

7.23 The UK10K project, established in 2010, was concerned with using genome sequencing to illuminate the genetic contribution to disease in a research culture where open access to data had been the norm. The premise of open access to genetic sequence data was established by the Human Genome Project (1999-2004) and the key principles were set out in a series of international agreements.[445] This has been widely accepted and endorsed by the research community and open access with DNA sequence data deposited online became the accepted practice. This was based on the assumption that there would be no risk of re-identification of research participants who had given biological samples for sequencing. However, this assumption was overturned by a study showing that data from individuals could be distinguished in genome-wide association study (GWAS) data using only summary statistics.[446] A later study demonstrated that male participants could be re-identified by linking individual mutations (single nucleotide polymorphisms, or SNPs) on the Y chromosome with data found in publicly available datasets on the Internet (see chapter 4).[447] Policies then changed: some datasets have been removed from the web, and models of managed or conditional access to data have developed, of which the UK10K project is one example. However, some initiatives in the spirit of citizen science, for example the Personal Genome Project, have continued to offer open access (see paragraph 7.41ff.)

---

[444] See, for example, Bjugn R and Casati B (2012) Stakeholder Analysis: A Useful Tool for Biobank Planning *Biopreservation and Biobanking* **10(3)**: 239-44; Lemke AA, Wu JT, Waudby C *et al.* (2010) Community engagement in biobanking: experiences from the eMERGE Network *Genomics, Society, and Policy* **6(3)**: 35-52, available at: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3434453/.

[445] Muddyman D, Smee C, Griffin H, Kaye J, and UK 10K Project (2013) Implementing a successful data management framework: the UK10K managed access model *Genome Medicine* **5(11)**: 100, available at: http://link.springer.com/article/10.1186/gm504.

[446] Homer N, Szelinger S, Redman M *et al.* (2008) Resolving Individuals Contributing Trace Amounts of DNA to Highly Complex Mixtures Using High-Density SNP Genotyping Microarrays *PLoS Genet* **4(8)**: e1000167, available at: http://www.plosgenetics.org/article/infoper cent3Adoiper cent2F10.1371per cent2Fjournal.pgen.1000167#pgen-1000167-g003.

[447] Gymrek M, McGuire A, Golan D, Halperin E and Erlich Y (2013) Identifying personal genomes by surname inference *Science* **339(6117)**: 321-4.

7.24 The objectives of the UK10K project are to apply genome-wide sequencing to existing research collections of patients from the UK and abroad with specific diseases (some 5,500 individuals) using comparisons with some 4,000 deeply phenotyped participants from the Twins UK and ALSPAC longitudinal cohort studies. This allows the identification of genetic sequence variants that may be associated with specific (and usually rare) diseases recorded in the phenotype data with the aim of characterising the genetic basis of diseases. In addition, a long-lasting research resource for UK and global genetic research is being established through rapid data release to the European Genome-phenome Archive, which is held by the European Bioinformatics Institute.[448]

7.25 The key instrument of governance for the initiative is, like UK Biobank, an Ethical Governance Framework (EGF),[449] based on informed consent and approval from an appropriate ethics committee. This was drafted by the UK10K Ethical Advisory Board with independent advice and international (but not public) review. The Board includes members representing the interests of some patients through the patient interest group Genetic Alliance UK and cohort study participants.

7.26 The aim of the UK10K EGF is to enable the UK10K project to operate as a federated system. This means that the projects can work together under a common ethical framework, which can acknowledge the nuances of particular studies while still allowing them to be part of a common endeavour. At the same time, it strives to ensure that there can be sufficient harmonisation so that these very different studies can participate in an ethically-coherent project that maximises the research benefit, while acknowledging the responsibilities and obligations that are owed to research participants.

7.27 An important precept is acknowledging and respecting the role of the principal investigator (PI) of each collaborating study, the person (usually a senior researcher) who has management responsibility for each study. Many of the principal investigators have collected samples from research participants themselves and may have a continuing clinical and/or research relationship with research participants. This means that they may be in a good position to develop an understanding of the interests of research participants, through consultation with participants themselves and others.

7.28 The EGF describes both policies, by which all project members agree to abide, and guidelines, which represent best practice as it is currently understood. It deals with aspects of the project including the feedback to patients of pertinent and incidental clinically significant findings, and management pathways. Feedback may take place when a clinician believes it to be appropriate, the patient has consented and the finding has been validated to clinical standards.[450] There is a diagram illustrating the process of data flow through the project and the necessary approvals.[451] Data access policy is described in a Data Sharing Policy Document.[452] All this information is publicly available. Access for the research community to sequence data held in the European Genome-phenome Archive is overseen by an independent Access Committee, which

[448] See https://www.ebi.ac.uk/ega/dataproviders/EGAO00000000079.
[449] See http://www.uk10k.org/ethics.html.
[450] Kaye J, Hurles M, Griffin H, et al. (2014) Managing clinically significant findings in research: the UK10K example *European Journal of Human Genetics* **22(9)**: 1100-4, available at: http://www.nature.com/ejhg/journal/v22/n9/full/ejhg2013290a.html.
[451] Muddyman D, Smee C, Griffin H, Kaye J, and UK 10K Project (2013) Implementing a successful data management framework: the UK10K managed access model *Genome Medicine* **5(11)**: 100, available at: http://link.springer.com/article/10.1186/gm504, at page 3.
[452] See http://www.uk10k.org/data_access.html.

will only approve applications from 'appropriately qualified' researchers who sign a legally binding agreement, making a number of undertakings that include protecting data confidentiality, providing appropriate data security and not attempting to identify individual participants.

7.29 The UK10K project is more modest in resources and limited in purposes and research methods than UK Biobank, and aims to leverage existing resources with genomic sequencing to generate new knowledge. Nevertheless it shares many of its governance principles with UK Biobank. Both make use of explicit Ethics and Governance Frameworks and place considerable reliance on institutional academic regulation to ensure the probity of individual researchers. Both foreground the role of consent and recognise the challenges of interpreting it in different and changing circumstances (UK Biobank through the EGC, UK10K placing significant emphasis on the role of the principal investigator to interpret the interests and expectations of participants). Reflecting on these different approaches we make a number of recommendations below with regard to governance that are relevant to biobanks.

Recommendation 10

We recommend that appropriate mechanisms should be put in place to allow governance arrangements to evolve during the life of an initiative, through deliberation with morally relevant stakeholders including participants, the public, funders and the research community. Arrangements may include, e.g., representation of relevant stakeholder groups in the governance of the biobank; regular review of a public ethics and governance framework document legitimated through deliberation with interested parties that sets out the relationships of a biobank with participants, the research community, individual researchers, funders and the wider society. This may serve as an instrument to maintain alignment of the public interest in research with the privacy and other interests of the participants. Governance arrangements should, among other matters, outline policies for maintaining data security, the feedback of health-related findings to participants and for research access to the resource. In large scale and complex initiatives detailed diagrams of data flows should be available to support good governance. The responsibility to ensure appropriate governance arrangements are in place rests with funders.

Recommendation 11

Where broad consent is sought for the use of data additional, adaptive safeguards should be in place to secure the interests of participants over the life of a project. A possible model is provided by a publicly articulated, 'living' ethics and governance framework that reflects the expectations of participants and is subject to review and revision through mechanisms that involve representatives of the full range of interests of participants in the initiative.

Recommendation 12

We recommend that researchers should operate demonstrably within a local governance framework able to maintain reasonable surveillance in order to

identify inappropriate data use and administer sanctions for misuse. Researchers should be members of a recognised research environment with appropriate arrangements in place to ensure their research meets ethical standards. They should provide undertakings regarding the confidential and secure use of data and that they will refrain from any attempt to identify participants from whom data may have been derived.

International collaborative research

7.30 As noted earlier, science is becoming an increasingly global enterprise, and as the ease with which research groups can communicate, share knowledge and carry out research collectively increases, more international collaborations will be formed. International collaborative initiatives can allow the sharing of knowledge and best practice and spread research expertise and funding across both well- and less-well-supported countries.[453] While the benefits can be great, there are significant difficulties to be faced. For example, scientists in one country cannot 'police' the activities of those in another country as there may be differing national laws and governance frameworks that may prevent single policies being imposed.[454] This requires conducting science in a way that provides accountability both at the local and consortium level, while respecting local legal, ethical and cultural norms.

Box 7.3: Examples of international collaborative research involving genetic data

International Cancer Genome Consortium

The International Cancer Genome Consortium (ICGC) coordinates large-scale cancer genome studies in tumours from 50 different cancer types and/or subtypes that are of clinical and societal importance across the globe.[455] As of May 2014, 74 projects representing over 17 countries and jurisdictions had sequenced over 25,000 cancer tumour genomes. Samples are held by each member project, while data is deposited in a central repository located in Toronto, Ontario. The project distinguishes two 'types' of data. Open access data, which does not contain obvious personal identifiers, is available from the ICGC Data Portal.[456] Controlled access data, which is more readily identifying, is available to authorised researchers for approved research through the ICGC Data Compliance Office.[457] After approval the researcher is able to download the data onto their own system for analysis.

The Psychiatric Genomics Consortium

The Psychiatric Genomics Consortium (PGC) is an international initiative with over 500 investigators from over 80 institutions in 25 countries. Its purpose is to conduct mega-analyses (individual-level data meta-studies) of genome-wide genetic data for

[453] For example, the Global Alliance for Health and Genomics has created a Framework for Responsible Sharing of Genomic and Health-Related Data, http://genomicsandhealth.org/; the Human Heredity and Health in Africa (H3Africa) Initiative is a cross-continental initiative to support African researchers and improve African health, http://h3africa.org/.

[454] Romeo-Casabona C, Nicolás P, Knoppers BM, *et al*. (2012) Legal aspects of genetic databases for international biomedical research: the example of the International Cancer Genome Consortium (ICGC) *Revista de Derecho y Genoma Humano/ Law and the Human Genome Review* **37**: 15-34.

[455] The International Cancer Genome Consortium (2010) International network of cancer genome projects *Nature* **464(7291)**: 993-8. See: https://icgc.org/.

[456] Zhang J, Baran J, Cros A *et al.* (2011) International Cancer Genome Consortium Data Portal—a one-stop shop for cancer genomics data *Database* **2011**: bar026, available at: http://database.oxfordjournals.org/content/2011/bar026.long.

[457] Joly Y, Dove ES, Knoppers BM, Bobrow M and Chalmers B (2012) Data sharing in the post-genomic world: the experience of the International Cancer Genome Consortium (ICGC) Data Access Compliance Office (DACO) *PLoS Computational Biology* **8(7)**: e1002549, available at: http://www.ploscompbiol.org/article/infoper cent3Adoiper cent2F10.1371per cent2Fjournal.pcbi.1002549.

psychiatric disorders. It is the largest biological experiment in the history of psychiatry.[458] The PGC data repository is located in the Netherlands. All phenotype and genotype data is stored there and all analyses of the data are carried out on its Genetic Cluster Computer.[459]

7.31 Similarly to biobanks, because of the developing knowledge environment, flexible and continually reviewed governance mechanisms are needed to guide the science while protecting the interests of participants. However, because of the diversity of their membership such consortia have to rely on agreements among members, peer pressure and the limited sanctions that can be imposed at the local level, such as cessation of funding. In the early days, ICGC members agreed to a set of overarching policies, together with flexible guidance that could be followed if desired.[460] It was agreed that projects, which must obtain participants' consent for whole genome sequencing, could be flexible regarding the return of individual health-related findings.[461] This contrasts with the more formal ethics and governance frameworks previously discussed and highlights the reluctance of researchers to impose a single governance structure over many different national and regional groups, a reluctance not shown in other sectors (finance, for example).

7.32 While ICGC was, for many, a prospective study, the PGC Schizophrenia Working Group uses existing data from a number of previous studies that were carried out in different counties to identify the genetic variants which may confer genetic risk for individuals.[462] Retrospective studies such as this present an obvious challenge to conventional research governance arrangements because the data were originally collected from participants in a number of different countries, and in varying circumstances. Furthermore, the extensive repurposing, data linking and analysis are carried out by a research collective whose members are themselves based in institutions in a number of countries. Gaining consent for international extensions of data access, or even disclosure outside a single institution, is actually a relatively recent circumstance for researchers to consider, let alone subject participants, and it is unclear whether the original consents would have been informed by foresight of such extensive re-use.

7.33 It is especially difficult to elicit participants' expectations, because deliberative engagement is extremely difficult at the international consortium level. Even if such procedures were undertaken for each local project, it would be very difficult to generalise results across sites. Yet it is important that participants and publics are able to obtain information about what is being done with the data they have provided. Specific investigations are therefore needed to assess the impact of cross-border research and how best to verify consent for extended uses, as well as how to disseminate the results of research in as transparent and accessible but secure a way as possible. Similarly, policy bodies and funding agencies need to be clear under what circumstances data can be re-used.

[458] http://www.med.unc.edu/pgc.
[459] http://www.med.unc.edu/pgc/documents.
[460] See https://icgc.org/icgc/goals-structure-policies-guidelines.
[461] Wallace SE and Knoppers BM (2011) Harmonised consent in international research consortia: an impossible dream? *Genomics, Society and Policy* **7**: 35-46, available at: http://www.lsspjournal.com/content/pdf/1746-5354-7-1-35.pdf.
[462] Schizophrenia Working Group of the Psychiatric Genomics Consortium (2014) Biological insights from 108 schizophrenia-associated genetic loci *Nature* **511(7510)**: 421-7, available at: http://www.nature.com/nature/journal/v511/n7510/full/nature13595.html.

7.34 Another concern is the security of data, as data protection regimes are not the same across all countries.[463] Global commerce has been dealing with this issue for many years, and it has been suggested that the research and clinical communities can learn from their experience.[464] For example, the adequacy test asks if the level of protection in the jurisdiction that receives the data is comparable to that of the origin of the data with the implication that if it is, personal data may be transferred with confidence.[465]

Cloud storage and computing

7.35 As more and more data become available, larger data analysis projects are being undertaken. These use the Internet to access appropriate technologies and operating power to transfer files, provide storage and drive analyses. Although commonly used in corporate settings, cloud computing is still a relatively new and potentially confusing concept to many in the research setting. It raises fears of loss of privacy due, for example, to a lack of clarity about responsibilities for data protection. Cloud services can be layered, with one provider being responsible for software while another is responsible for infrastructure.[466] With competing responsibilities there is a fear that adequate protections may not be in place for secure data processing. As already mentioned, if a provider is located in a different country, the data may be subject to a different data protection scheme, possibly one of lesser stringency.

7.36 The location of data repositories is now an issue of potential concern. This led, for example, to some countries declining to participate in the Type 1 Diabetes Genome Consortium, owing to study requirements for the processing of samples at network laboratories and/or final deposition of samples in US-based Central Repositories.[467] Of particular concern is that many companies offering cloud computing facilities are based in the USA and therefore all data is subject to the homeland security legislation that allows access to data by the NSA (see chapter 2 above). The use of cloud computing in research highlights the difficulties of balancing the desire to analyse large datasets from around the world to advance scientific discovery with the risks of the potential loss of the confidentiality of data.

Box 7.4: International Cancer Genome Consortium PanCancer Analyses of Whole Genomes (ICGC PCAWG)

The ICGC PCAWG will study the whole genome sequence from tumours and matched samples (usually blood) from an estimated 2,000 patients internationally who have been recruited to ICGC member projects. Demographic, clinical and pathology data will be available for all 2,000 matched samples. The dataset on which the analyses will run is expected to exceed one petabyte of data. Examining and comparing data across cancers internationally is now possible due to the large number of cancer genome

[463] See, for example, The Organisation for Economic Co-operation and Development (2013) *The OECD privacy framework*, available at: http://www.oecd.org/sti/ieconomy/oecd_privacy_framework.pdf.

[464] Kosseim P, Dove E, Baggaley C *et al.* (2014) Building a data sharing model for global genomic research *Genome Biology* **15(8)**: 430, available at: http://genomebiology.com/2014/15/8/430.

[465] For this purpose the US is deemed to be a 'safe harbour' by the EU as US organisations have voluntarily self-certified that they will comply with mutually agreed-upon data protection principles. See: http://ec.europa.eu/justice/policies/privacy/thridcountries/adequacy-faq1_en.htm. However, see chapters 2 and 4 regarding concerns and evolving case law about jurisdictional differences.

[466] Information Commissioner's Office (2012) *Guidance on the use of cloud computing*, available at: http://ico.org.uk/for_organisations/guidance_index/~/media/documents/library/Data_Protection/Practical_application/cloud_computing_guidance_for_organisations.ashx.

[467] Hilner JE, Perdue LH, Sides EG *et al.* (2010) Designing and implementing sample and data collection for an international genetics study: the Type 1 Diabetes Genetics Consortium (T1DGC) *Clinical Trials* **7**(1 suppl): S5-S32, available at: http://ctj.sagepub.com/content/7/1_suppl/S5.short.

> tumours that have been sequenced and made available in accessible form. As no one site will have sufficient capacity to host the project, using a cloud environment is being explored, if it can be shown to be consistent with the ethical and legal requirements of the ICGC.
>
> Several cloud providers, rather than one, may be used, based in different countries in North America, Europe and Asia. Annai Systems, an academic cloud based in the US, was approved for use by the ICGC Executive in November 2014 and ICGC data is already being mirrored, so far for five projects in its cloud. Any provider's Terms of Service will be reviewed and agreed by the ICGC. A small number of ICGC Portal staff and PCAWG working group members will align and annotate the data to create a uniform dataset. Only ICGC PCAWG team members who have received approval will be allowed access to the dataset. The dataset will be removed from the cloud providers after the analysis is completed and archived at the EGA.

7.37 At least four benefits to using cloud services have been identified for international collaborative biomedical research: lower costs, as one 'rents' space rather than purchases it; better data security, as such providers have the money to invest in state-of-the-art security mechanisms; increased data storage capacity; and lower environmental impact, as a resource is being reused rather than newly constructed.[468] However, the terms of service of many of these providers have not necessarily been developed with specific attention to the needs and sensitivities of biomedical research.[469] There may therefore be a gap to close through the research community (including participants) working with providers to agree control of the data, security measures (such as appropriate encryption), and access to the data.

7.38 Cross-border data access and transmission and the use of cloud services should provoke research studies to review the ethical and legal implications, particularly where they are introduced to existing projects. For example, using cloud providers was not considered at the beginning of the ICGC and is not included specifically in consent documents. Seeking specific consent for this use from the 2,000 participants from multiple countries would be unfeasible. Through its oversight committees, the ICGC has approached this problem by working with cloud suppliers who will design systems that will provide for the needs of the scientific community.

7.39 One US-based market intelligence firm has predicted that by 2020 80 per cent of all health care data will pass through a cloud provider at some point and that cloud-based products will increasingly be used to manage costs and enable the analysis of the increasing amounts of health-related data becoming available.[470] This could be simply one more standard technology that will be commonly used. However, it is not clear that there is a high level of understanding amongst the general public, and indeed researchers and health care administrators, of the implications and, therefore, the moral relevance of cloud technologies.

7.40 Details of how data initiatives use cloud systems need to be disseminated and discussed in the public arena, allowing any misconceptions to be explored and facts

[468] Dove Es, Joly Y, Tassé A, et al. (2014) Genomic cloud computing: legal and ethical points to consider *European Journal of Human Genetics* (advance online publication), doi:10.1038/ejhg.2014.196.
[469] Ibid.
[470] http://www.idc.com/getdoc.jsp?containerId=prUS25262514.

explained so that this technology can be used transparently and with appropriate safeguards. Any research-based data initiative seeking to use such technologies should discuss this with study partners and, if possible, potential participants for acceptability. This would allow prospective initiatives to include details of cloud use in consent materials as well as the governance framework. Any agreements with providers will need to be tailored to ensure that data will be kept secure from breaches of privacy and reviewed regularly. As norms in research practice change, it may be that the use of cloud providers will no longer be seen as contentious. But this will only happen with detailed examination of the issues and public debate, which will help us recognise morally reasonable expectations and formulate appropriate governance and oversight mechanisms.

Recommendation 13

We recommend that all international collaborative data research initiatives should operate within an explicit, public ethics and governance framework that has agreement from the initiative's constituent partners. International collaborators should be able to demonstrate that they can fulfil recommendation 12 by applying equivalently strong governance standards (using legal and other mechanisms available in their national jurisdiction).

Recommendation 14

We recommend that all partners in international collaborations integrate the provisions of the ethics and governance framework (EGF) agreed by the initiative as far as possible at their local research site. The partner should ensure that they adhere to the EGF, for example by ensuring participants have given appropriate consent for the use of data and samples in the initiative and that they are informed of potential transfer across borders.

Recommendation 15

We recommend that national bodies publish their policies on the use of cloud services in health data settings so that data initiatives can include this in their decision making and interactions with publics and participants.

Open data

7.41 There are a number of projects that involve uploading individual genomic data and other data to the world wide web so that it becomes freely available to anyone to use for any purpose. The best known such initiative is the Personal Genome Project. This was initiated by the prominent Harvard University genomics researcher, George M. Church in 2005.[471] The Personal Genome Project is a long-term cohort resource that aims to publish the genome sequence, medical records and various other measures

[471] Church GM (2005) The Personal Genome Project *Molecular Systems Biology* **1**: 2005.0030, available at: http://msb.embopress.org/content/1/1/2005.0030; Lunshof JE, Bobe J, Aach J, et al. (2010) Personal genomes in progress: from the Human Genome Project to the Personal Genome Project *Dialogues in Clinical Neuroscience* **12(1)**: 47-60, available at: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3181947/.

such as MRI images of 100,000 volunteers so that the data are freely available to anyone who chooses to use them and to enable citizen science. George Church himself, together with other prominent figures in the biotechnology industry, genome science, and science policy made up the initial 10 participants, the 'PGP-10'. (The Harvard University Medical School Review Board that considered the project had requested that the first group of volunteers included Church himself and other stakeholders in genomic science.) Today more than 3,500 volunteers have joined the USA study and additional studies have been established in Canada (2012) and UK (2013), with others planned. There is a long waiting list of potential volunteers, with over 10,000 people registered in the UK within three months of the project launch, although there is a significant lag, due to funding constraints, in generating and uploading sequence data.

7.42 Participants in the PGP go through a different recruitment process to many conventional research projects. Firstly, there is no promise that the identity of individuals will remain anonymous as the whole purpose of the project is to make sequence data freely available. To be accepted, volunteers must be over 21 and pass an examination to test whether they are aware of the potential risks to participation – including possible discrimination by insurers and employers. If accepted for inclusion in the study, participants are required to contribute a sample from which a genome sequence will be produced and encouraged to upload other kinds of medical information. Before the sequence information is deposited on the project website, they have thirty days to review the data and make a decision whether they want it to be made public. If they decide to withdraw from the project during this period their data will not be publicly released. However, if data are put on the web and participants later decide to withdraw from the study, already released information will remain publicly available and only future information will not be released. Participants are asked to report any discrimination or harm that they experience as a result of participation in the project. There is a continuing relationship and engagement between participants and the project. Participants in PGP (USA) were required initially to do this on a quarterly basis, but this was reduced to six-monthly as it was felt to be too onerous for participants who had nothing to report. Participants in the PGP may be viewed as 'information altruists' who are prepared to allow their genome sequence to be made public.[472] Although this level of openness is not for everyone, the positive response to the launch of the UK and the Canadian arms of the project suggests that such projects do have public appeal.

7.43 The open publication of data, as exemplified by the Personal Genome Project, is a limit case for the governance of data for research. Nevertheless, it is not meaningless to ask what the morally reasonable expectations of participants may be. In terms of what they may expect the limits to data use to be, the answer will depend on public norms rather than those maintained in the context of a specific data initiative, and on governance by law and the conventions of public morality. But while the expectations may not be bounded (indeed, participants are urged to contemplate the worst that can reasonably be imagined) they may nevertheless have some positive content. Subject participants may expect, for example, that their supposedly altruistic gesture should be answered by a commitment on the part of the PGP organisation actively to secure the best use of the data to advance scientific knowledge (for example, by ensuring the

[472] See also Kohane IS and Altman RB (2005) Health-information altruists – a potentially critical resource *New England Journal of Medicine* **353**: 2074-77.

quality, accessibility and interoperability of the data published). Beyond that, the PGP implicitly poses a challenge to societies to affirm as a norm that the relevant rights of altruistic subject participants will be protected. To do this is to abandon the 'arms race' of developing ever stronger data security measures and rely instead on regulating the conduct of data users, not purely within the context of 'bona fide' research, subject to institutional codes and penalties, but generally, under public morality and the rule of law.[473]

Citizen science and participant-driven research

7.44 Increasing access to digital technologies and the rise of online social networks has facilitated the formation of communities of people engaged in establishing and conducting health research including self experimentation, self surveillance, analysis of genomic data and genome-wide association studies.[474]

Box 7.5: PatientsLikeMe

Founded in 2004, the largest participant-driven research network, PatientsLikeMe (PLM) has more than a quarter of a million members representing over 2,000 health conditions. Through this company ('Live better, together') people connect with others who may have the same disease or condition, and track and share their own experiences. In doing so they generate data about the real world nature of disease that can help researchers, pharmaceutical companies, regulators and health providers develop more effective products, services and care. PLM allows members to contribute their own data about their conditions (treatment, history, side effects, hospital episodes, symptoms, function scores, weight, mood, quality of life, etc.) on a continuing basis. The resulting longitudinal record is organised into charts and graphs that allow members to identify patterns, gain insight and place their experiences in context, as well as to see what treatments may have helped other patients like themselves. The website also gives members lists of relevant clinical trials and they can search the site for trials for which they may be eligible. The company also offers a commercial service to actively message potential participants for specific clinical trials.

PLM describe their four core values as follows[475]:

- Honour the trust patients put in us – patients trust the company to protect their health data and to use it to advance knowledge of their disease.
- Transparency. The company aims for 'no surprises'. It discloses its business partnerships, what it does with patient's data and how the company makes money.
- Openness. The company believes that sharing health information openly has potential to benefit patients.
- Create 'wow'. This is a goal for what patients should feel when they visit the website.

The company has a team of in-house researchers who produce many (peer reviewed) papers and also a number of collaborative partnerships with academic research

[473] There are two routes to this: general data protection legislation and anti-discrimination legislation. A number of legal instruments give protection against discrimination and the existence of measures comprising the 'welfare state' offers some practical insurance against the effects of discrimination. Although the Equality Act 2010 prohibits discrimination on the ground of health status, it does not explicitly include genetic status as a 'protected characteristic'; in response to the green paper that foreshadowed the Act, many, including the UK Human Genetics Commission, argued that it should. See http://webarchive.nationalarchives.gov.uk/20100419143351/http://hgc.gov.uk/client/document.asp?DocId=134&CAtegoryId=4.

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

[475] https://support.patientslikeme.com/hc/en-us/articles/201245710-What-are-the-company-s-core-values-.

groups.[476] It is run as a for-profit company that makes money by selling data uploaded by patients to other companies.

7.45 An early and influential example of PLM's research was their amyotropic lateral sclerosis (ALS) lithium study. ALS is a progressive and incurable disease. In 2008 a small Italian study suggested that lithium carbonate could slow the progression of ALS. In response to this, many members of PLM began taking the drug. Two members with advanced stage ALS (from Brazil and the USA) initiated a study using self-generated data from members on the platform to test these findings. (Both died before the study was completed). The nine-month study indicated that lithium did not slow the progression of the disease, a result that was later confirmed in four randomised controlled trials.[477]

7.46 Patient-led and participant-driven research (PLR and PDR) is gaining wider recognition as a potential source of generalisable health knowledge that benefits both participants and society more widely, and that can realise the values of solidarity among communities of patients suffering from a common disease. It can complement conventional research on conditions, or on aspects of them, that may have been neglected. The researchers involved have claimed that it can speed up clinical discovery, and could potentially maximise it, setting a stage for better trials with more engaged participants.[478] However, this may require new governance arrangements.

7.47 Like any clinical research, PDR can involve the risk of harms to participants or their relatives, including children. Self-experimentation can lead to participants taking excessive risks. Furthermore, the existence of a strongly solidaristic patient community may create or allow undue peer pressure or even exploitation. Conventional research has both scientific and ethical oversight, which facilitates the production of generalisable health knowledge that can be used by participants and society more widely. Research conducted outside the conventional academic and commercial institutions may not be subject to such oversight and study reports may not meet the basic acceptance criteria for peer-reviewed journals. While some participant-driven research may involve collaborators within the conventional system, thus bringing it within its ambit, much does not. However, trying to force this research into the conventional mould may stifle the very features that could make it so valuable.

7.48 While all forms of scientific research involving human participants should be subject to ethical as well as scientific appraisal, the appropriate standards for ethical oversight need to be adapted to the distinctive features of PDR. There have been calls for a broad dialogue to address the issues and to generate consensus on best practice as well as warnings that a failure to do this may pose threats of harms to participants,

[476] http://www.patientslikeme.com/research/publications.

[477] Wicks P, Vaughan TE, Massagli MP and Heywood J (2011) Accelerated clinical discovery using self-reported patient data collected online and a patient-matching algorithm *Nature Biotechnology* **29(5)**: 411-4, available at: http://www.nature.com/nbt/journal/v29/n5/abs/nbt.1837.html; Wicks P, Vaughan TE, and Heywood J (2014) Subjects no more: what happens when trial participants realize they hold the power? *British Medical Journal* **348**: g368, available at: http://www.bmj.com/content/348/bmj.g368.

[478] Wicks P, Vaughan TE, and Heywood J (2014) Subjects no more: what happens when trial participants realize they hold the power? *British Medical Journal* **348**: g368, available at: http://www.bmj.com/content/348/bmj.g368, at page 2.

risks of undermining the credibility of PDR, and may provoke a backlash of over-regulation.[479]

Recommendation 16

We recommend that biomedical researchers give consideration to arrangements that will maximise the potential of participant-driven research to contribute to generalisable health knowledge and secure public benefits while providing adequate protection of those involved through continuing ethical and scientific appraisal. Key stakeholders are citizen patient researchers, biomedical research bodies, research funders and journal publishers. All stakeholders should encourage optimal use of human studies for improved health outcomes.

Conclusion

7.49 In view of the rapidly increasing importance to the research community of extending access to data, and the benefits that such research can bring to the public at large, developing best practice for the collection, governance and use, and extension of access to data in biological research and health care should be a very high priority across both research and clinical settings. Some work has been done and is continuing by international organisations such as the Public Population Project in Genomics and Society (P^3G) which have brought together best practice regarding population research and the Global Alliance for Genomics and Health which has created a Framework for Responsible Sharing of Genomic and Health-Related Data.[480] But more needs to be done at the level of individual patients and research participants and respect for their circumstances and protection of privacy must be at the centre of such systems. There are many stakeholders in the collective enterprise of health promotion and medical treatment but to marginalise those individuals who provide data for research will be to risk the trust of current and future generations.

Recommendation 17

We recommend that the research community, including all academic and commercial partners, data controllers and custodians, public services and government agencies, actively foster opportunities to create a more explicit and reflective foundation for extending data access in the public interest. We urge all stakeholders in the medical research enterprise to continue to develop robust and comprehensive, yet efficient privacy protecting rules, guidelines and measures. Among other things these should aim at:

- Providing greater clarity for members of the public about ways that their biomedical data are, and may be used in the future, along with a realistic acknowledgement that no system can guarantee privacy and confidentiality in all circumstances.
- Securing commitments from data controllers to a responsible approach to the extension of data access as part of their core mission statement; they must publish information about their approach to data access, transparency and accountability, and whether, and on what terms, they will consider extending access to data.

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

[480] www.p3g.org; http://genomicsandhealth.org/about-the-global-alliance/key-documents/framework-responsible-sharing-genomic-and-health-related-data.

- Demonstrable and continual improvement of collection, storage and data access procedures against explicit standards for accuracy, reliability and security.

Chapter 8

Reflections and conclusions

Chapter 8 – Reflections and conclusions

Chapter overview

This chapter reviews the state and direction of travel of information technology, data science, research and governance described in the report before drawing together the elements of the ethical argument. It concludes by setting out some practical precepts for professionals involved in data initiatives.

Introduction

8.1 In this final chapter we briefly reflect on the issues we have identified, the ethical argument that runs through this report and some of the conclusions to which our deliberations have led. Our hope is that our approach will prove useful to those proposing to extend the use of data in biomedical research and health care. We therefore conclude by proposing a number of practical precepts for those involved in the establishment or governance of data initiatives.

The state of the art

8.2 We began by setting out a number of propositions, which describe the area of interest and ethical issues that arise within it. The first two of these propositions describe the conditions from which the subsequent propositions follow, namely the accumulation of data from people in health care and biomedical research, and advances in information technology and data science that allow those data to be put to use. We recognise that these underlying technical advances are not specific to the fields of health care and biomedical research, but their impact in healthcare and biomedical research is profound and raise issues of special ethical significance. One reason for this is that the opportunities to which these advances give rise invite us to think about data as a resource with broadly exploitable potential rather than as an output bound to the intentions that motivated its original collection. This way of thinking is, in general terms, very different to the way in which information governance-conscious clinicians and researchers have, until now, been expected to think about data.

8.3 The principal ways of realising this new potential involve reframing the data within a novel context, created by a novel ‘research question’ or by linking them with other data, either from a different source or collected especially for the purpose. This led us to observe that the moral significance of data is therefore related to the kinds of questions that the data can help to answer and when or by whom those questions are addressed.

8.4 Data collected in health care and biomedical research contexts are not intrinsically more or less ‘sensitive’ than other data relating to individuals, but the medical context in which they are acquired (and in which they may be used) will often mean that they touch important personal interests. On the other hand, we draw attention to the fact that there is a strong public interest in the responsible use of data in research to support the development of knowledge and innovation intended to improve the well-being of all by enabling advances in healthcare. In fact, the use of data can have both beneficial and harmful effects on individuals or groups. These effects may be distributed in different ways: benefits for one group may entail welfare restrictions for another. Furthermore, different people may value different consequences in different

ways: something that might be profoundly troubling for one person might be a matter of indifference for another. It is principally these features – the potential for both beneficial and harmful consequences of data use, the possibility that they may be distributed differently among people, and the diverse ways that different people may value them – that constitute the problematic moral terrain of this report.

8.5 Negotiating this terrain is made difficult by the fact that so much about the personal and social consequences of data use is unknown, partly because there is a lack of existing evidence, but mainly because we have to consider an indefinite future in which these data will persist and in which the potential for data use and its impacts could be transformed in unanticipated ways. The digital world of data is growing rapidly and the ways in which datasets can be related and information from them derived are changing constantly. Making decisions about how data are best managed is complicated further by changing and powerful scientific, economic and political interests. In some cases this has led to the terms of publicly significant data initiatives being established the terms for many data initiatives *prior to any wider public debate*. These factors make it reasonable to expect that norms will shift in unpredictable ways over time. It is likely that well-established social norms of privacy and data access that apply today will no longer be applicable even in the near future as the actions of business, major institutions or government seek to *impose* new norms independently of social processes. A possible example is where using patient data offers opportunities meaningfully to inform health care service design, this becomes accepted as 'necessary' or is then legally mandated.[481] Meanwhile, the bulwarks that have hitherto protected a satisfactory and workable accommodation of interests, principally, the de-identification of data and the 'informed' consent of data 'subjects', have been substantially weakened in a hyper-connected (or potentially hyper-connectable) 'big data' world.

8.6 The morally relevant issues here are not merely to do with the re-identification of individuals: there also are social choices about the terms on which data are used that have moral consequences both because they determine how specific individuals might be treated (they may underwrite discrimination, for example) and because they may have a broader social impact (they may be used to inform political decisions). The challenge recognised in this report is for us as a broader society to get this right, to use data responsibly to promote the public interest, in a way that and best reconciles the morally relevant interests of individuals and groups, and respects their fundamental rights.

Ethical approach

8.7 Our ethical approach takes the perspective that the collection and use of data, and the determination of the circumstances in which these take place, are social activities that involve and affect people, individually or as members of groups, through time. Our focus has not been on identifying particular kinds of data as being of special concern (as almost all data can be 'sensitive' or 'personal', depending on the context), but on the human relationships that variously facilitate or restrict the use of data, or which may be created or affected by particular uses of data, and which change through time. Our

[481] See, for example, the arguments about the 'need' to use patient data to support health care service design and resource allocation that was put forward by NHS England in support of its 'care.data' programme, and the legislative action to facilitate data sharing a range of health-related purposes through the Care Act 2014.

aim has been to identify from among the influences and drivers shaping these relationships the values and interests that are *morally relevant* and how they should be respected accordingly.

8.8 *Privacy* is important to people for a number of reasons relating to their ability to maintain their identity, relationships and well-being. Respecting people's privacy can be seen as an aspect of showing respect for them as persons. The *public interest* is an interest that people share as members of a society, e.g. the promotion of commonly valued conditions like security, physical and mental health and material prosperity. People are simultaneously both individuals and members of wider groups with shared values and interests: they thus have interests both in allowing other people to access data that relates to them and in guarding against this to preserve their privacy, just as they have interests both in access to data about others and in their privacy. *Private and public interests are fundamentally entwined*: there is both a private and public interest in maintaining acceptable levels of privacy, and a private and public interest in making responsible use of data compatible with this. Data initiatives therefore have to perform a *'double articulation'* that seeks to reconcile the private and public interest in using data, and of the private and public interest in protecting privacy, rather than simply 'balancing' privacy interests *against* public interest.

8.9 Recognising the complex interrelation of morally relevant interests at stake leads to a more nuanced ethical approach than simply that of distinguishing the morally acceptable from the morally unacceptable. This is not to ignore that there might be unacceptable outcomes: those that do not respect persons or that violate their human rights are unacceptable, a point reinforced by our first two 'substantive' principles of *respect for persons* and *human rights*.

- Asserting the principle of respect for persons requires that the terms of a data initiative are set as a result of moral reasoning that takes the complex interrelationship of public and private interests into account. Enabling those with morally relevant interests to assert their own interests and offering them a reasoned account of decisions regarding data use that recognises those interests as being morally relevant are ways in which data initiatives may demonstrate respect for persons.
- Asserting the principle of respect for human rights entails that people should be free to exercise, and that others should respect, rights derived from people's core, morally relevant interests (among which is the right to protection of private and family life and personal correspondence). It also entails that this freedom may only be restricted for weighty reasons, where is it necessary to achieve an end that the person is expected, through their membership of the society, to find reasonable and compelling, and in a way that is proportionate to achieving this aim.

8.10 However, the ethical approach also recognises that that job of moral reasoning should not cease once the threshold of acceptability is passed but should continue throughout the process of establishing and governing a data initiative, and permeate it at every point. Opportunities for ethical reflection should therefore be built into data initiatives. Moral reasoning thereby assumes a constructive role: rather than that of the external conscience poised to say 'no' to certain practices that step over a notional line of acceptability, the recognition that there are ethical arguments on both sides of any question about data use allows them to be harnessed in the search of good and better solutions, not merely the delineation of acceptable ones. Hence the notion of what is

morally reasonable is not merely about satisfying some formal standard of reasonableness but rather about the outcome of a process of moral reasoning in which values and interests confront and challenge each other in a concrete situation.[482] In the report we offer two further principles to guide this positive search for a set of morally reasonable expectations.

- Following the principle of *participation* of those with morally relevant interests in a deliberative procedure can *optimise* the relationship between public and private interests because it allows values and interests to be transformed and reconciled through dynamic interaction (rather than assuming that they are fixed and immutable). This is in contrast to approaches that simply dictate terms of an initiative to fulfil particular interests and invite others to take part. Participation demonstrates respect for persons by involving them in the design of data initiatives (it enables them to engage in forming the conditions of a future in which they have a direct interest rather than merely responding to it) and is more likely to produce outcomes that secure their commitment and build *trust*.
- Following the principle of *accounting for decisions* is a necessary complement to the principle of participation, since not all interests can be represented through participation and not all interests may be satisfied with any outcome. This ensures not only that a decision can be 'accounted for' in a community, but also that there is an opportunity to challenge and even to re-evaluate the decisions, through formal structures (e.g. regulation or appeal to a legitimate authority) and broader social processes (e.g. open and continuing debate). It follows that the set of morally reasonable expectations must be a *publicly statable* in a way that allows an account to be given to all those with morally relevant interests of how their interests have been respected. The principle recognises the necessarily provisional nature of decisions about data management and governance, since the horizon of possibilities – and the values and interests invested in them – are constantly changing as the social, political, technological and information environments evolve.

8.11 Together, we believe that these principles offer the best chance of producing, for any particular data initiative, a morally reasonable set of expectations capable of being satisfied in practice. Such a set of expectations must incorporate the principles of respect for persons and human rights; it must include, in other words, expectations about how respect for diverse values and interests will be shown and about how moral conduct of others will be assured, while at the same time resolving the 'double articulation' of public and private interests through a process of moral reasoning. We found that there are always three essential elements to the set of morally reasonable expectations, and that the content of these expectations will be strongly interrelated in any data initiative.

- Identifying applicable norms: mere compliance with the law is inadequate to ensure that data use is morally reasonable. This is because law both stands in a broadly derivative relationship with respect to morality and because it provides only a minimal framework for action rather than full determination for moral action. It is therefore important to identify the *moral* norms of privacy and data access applicable in the use context.

[482] The central moral question facing data initiatives, as we formulate it in chapter 3 is: "How may we define a set of morally reasonable expectations about how data will be used in a data initiative...?".

- Respecting individual moral freedoms: similarly, consent is often relied upon as an important way of respecting individual moral agency but it is not sufficient on its own to resolve the morally relevant interests at stake, nor is it always necessary (for example, where the applicable privacy norms do not require it). An appropriate way of respecting individual freedoms must be found in relation to the applicable norms and governance for any particular initiative, which may involve different forms of consent (broad, explicit, etc.) or legitimate authorisation.
- Assuring moral conduct by others: individuals are entitled to have expectations of others using data (particularly professionals involved in data initiatives), including expectations of who these others will be, and how their conduct will be governed. Furthermore, there is a public interest in ensuring that those involved in data initiatives discharge a moral duty of care owed to others, a duty that is not exhausted simply by complying with subjects' consent.

8.12 It is these three elements – the content of expectations, how they were defined and the way in which they relate to each other in the context of specific data initiatives – that we considered when we looked for examples of good practice in specific initiatives in chapters six and seven.

Some practical precepts for data initiatives

8.13 The key to acting ethically with personal health information in a world of Big Data will be to maintain the engagement of, and oversight by, patients and other affected people not just as a new initiative is being developed, but as it evolves over time. It is natural for the evolution of a system to be driven by its heaviest users, and so an initiative that was initially acceptable to both patients and researchers may within a few years have a quite different balance. The promoters and operators of data initiatives using health and biomedical data must therefore give careful thought not just to how they secure moral acceptability and provide adequate transparency at the beginning, but also how this is to be maintained as the system evolves. Failure to maintain a workable reconciliation of moral, legal, social and professional norms, just as much as a failure to produce it in the first place, can lead to loss of public trust and compromise both the respect for private interests and the attainment of public benefits.

8.14 How, then, does our ethical approach translate into practical actions? What steps might someone approaching a data initiative take, perhaps as a principal investigator in a research project, a lead policy official or a commissioner of services? Clearly, the appropriate measures that may be taken will vary according to a number of factors including with the nature and size of the initiative. Nevertheless, from our examination of this area we might distil a number of useful precepts.

- **Identify prospectively the morally relevant values and interests** in any data initiative. Some process of stakeholder mapping and reflection on this will be essential as an initial step to understand where these interests are located and what informs them.[483] These will include private interests but may also include economic and political interests, for example. Explicating their moral content may allow them to be set in the same light as other moral interests. This critical reflection may very often reveal that what appear to be 'hard constraints' or 'strategic imperatives' rest on moral assumptions or prior value commitments that ought themselves to be brought into question.

[483] See recommendation 2 at paragraph 2.50 above (regarding mapping data flows) although the interests in a data initiative are not only those of people and groups at the terminal points of data flows.

- **Take special care to identify those interests that may be especially at risk or that arise from diverse values**. Identifying situational vulnerabilities (i.e. why the consequences of a particular data initiative might disproportionately affect certain individuals or groups) and understanding how different people value the potential benefits and hazards of data initiatives is essential to explore what forms of respect for individual freedoms (e.g. consent) and forms of governance may be required.
- **Do not rely simply on compliance with the law to secure that data use is morally** appropriate, particularly where it does not fully reflect moral norms. The norms enshrined in legal instruments, while they determine how data *may* be used (and, in certain cases, how it must be used) are insufficient to determine how they *should* be used. It should never be assumed that compliance with the requirements of law will be sufficient to ensure that a particular use of data is morally reasonable.
- **Establish what existing privacy norms are engaged** by the contemplated uses of data. These will have a number of different sources, including social conventions, value and belief systems, and needs of individuals, groups and communities. This might include, for example, norms of professional confidentiality, of data sharing within families or social groups, or of wider acceptance of data use. Findings from consultation or public opinion research will be informative at this stage (but caution should be exercised when relying on existing research as the circumstances, values and interests may differ from one data initiative to another). Resistance among the public to the involvement of profit-seeking commercial actors may be an important phenomenon in this context. If private sector organisations are going to play a role in the delivery of public services and public goods, this must be engaged with in formulating reasonable expectations. Attempts to shift norms or impose new norms without engagement risks undermining trust and therefore the objectives of the initiative.
- **Involve a range of those with morally relevant interests in the design of data initiatives** in order to arrive at a publicly statable set of expectations about how data will be used.[484] Participation helps to ensure both that different values and interests may be represented and that expectations are statable in a way that is intelligible from different perspectives. It also helps ensure that an account is given of how morally relevant values and interests are respected. Structured public dialogue or other forms of deliberative engagement, including direct participation of representatives in the initiative, will often be valuable.
- **State explicitly the set of morally reasonable expectations** about the use of data in the initiative. These are likely to include who will have access to data and for what purposes, the way in which disclosures will be authorised (including the form of any relevant consent procedures) and how the conduct of those with access to data will be regulated or accounted for.[485] This statement might take the form, for example, of a written and published ethics and governance framework document that can be accessed easily, with explicit arrangements for it to be reviewed.
- **Involve a range of those with morally relevant interests in the continuing governance and review of data initiatives.** What constitutes morally reasonable expectations may alter over time as new opportunities and threats emerge and as norms shift. Measures such as monitoring relevant social research, periodic consultation or a standing reference panel of participants are desirable.[486]

[484] See recommendations 6 and 7 (which are specifically relevant to the HSCIC)
[485] See recommendation 7 (which is specifically relevant to the HSCIC but covers the publication of data sharing agreements) and recommendations 11 and 13 (with regard to research using broad consent models).
[486] See recommendation 10 (with specific relevance to biobanks).

Appendices

Appendix 1: Method of working

Background

The Nuffield Council on Bioethics established the Working Party on *The collection, linking and use of data in biomedical research and health care: ethical issues* in March 2013. The Working Party met ten times over a period of 18 months. In order to inform its deliberations, it held a public consultation and a series of 'fact-finding meetings' with external stakeholders and experts. It also commissioned two reports on topics relevant to the work of the project and received comments on a draft of the report from twelve external reviewers. Further details of each of these aspects of the Working Party's work are given below and in Appendix 2. The Working Party would like to express its gratitude to all those involved, and the invaluable contribution they made to the development of the final report.

Consultation document

The Working Party launched a consultation in October 2013. The consultation ran until January 2014. 51 responses were received, of which 22 were submitted by individuals and 29 on behalf of organisations. Those responding to the consultation included researchers, interest groups and professional organisations. A full list of those responding (excluding those who asked not to be listed) is set out in Appendix 2. A summary of the responses is accessible on the Council's website. Copies of individual responses will also be made available on the website in those instances where the Council has permission from respondents to do so.

Fact-finding

As part of its work, the Working Party held a series of 'fact-finding' meetings. Invited guests gave brief presentations and then participated in discussion with Working Party members and other guests.

Big data: 19 July 2013

- Francine Bennett, Mastodon C
- Fiona Cunningham, EBI
- Tim Hubbard, Sanger Centre
- Martin Landray, University of Oxford

Patient/participant choices and privacy solutions: 12 September 2013

- John Bowman, Ministry of Justice
- Ian Brown, Professor of Information Security and Privacy and Associate Director (Cyber Security Centre), Oxford Internet Institute
- Toto Ann Gronlund , Head of Patient and Public Partnerships, NHS CFH
- Alastair Kent, Director, Genetic Alliance UK
- John Loder, Young Foundation
- Sam Smith, Privacy International
- David Townend, Professor of Law and Legal Philosophy in Health, Medicine and Life Sciences, Maastricht University
- Effy Vayena, Senior Research Fellow, University of Zurich
- Tim Williams, Director of myClinicalOutcomes

Screening and risk profiling: 14 November 2013

- Kerry Bailey-Jones, Health Lead, We Predict
- Ramona Liberoff, Senior Vice President (Innovation Analytics, Nielsen UK
- Grigorios Loukides, Lecturer in Computer Science & Informatics, Cardiff University
- Monique Mackenzie, Consultant in DMP Stats and Statistics Lecturer, University of St Andrews
- Anne Mackie, Director, National Screening Committee
- Nora Pashayan, Senior Clinical Lecturer in Applied Health Research, UCL
- Matt Sperrin, Lecturer in Health Data Science, University of Manchester
- Paul Taylor, Reader in Health Informatics, UCL

Biomedical data in research and clinical practice across jurisdictional boundaries: 8 January 2014

- Ruth Boardman, Bird and Bird
- Marc Dautlich, Pinsent Masons
- Paul Flicek, EMBL-EBI
- Dennis Keho, AIMES Grid Services
- Katherine Littler, Wellcome Trust
- Ioannis Pandis, Imperial College London
- Becky Purvis, AMRC
- Judith Rauhofer, University of Edinburgh
- Jane Reichel, Uppsala University
- Jonathan Sellors, UK Biobank

Evidence reviews

In order to inform its deliberations, the Working Party commissioned two reports from external academics.[487]

The terms of the reviews are set out below.

Review 1: Actual harms resulting from security breaches or infringements of privacy involving sensitive personal biomedical and health data.

Purpose: to assist the Working Party in understanding:

(a) The nature of the actual harms resulting from security breaches involving sensitive personal biomedical and health data (i.e. misuse of personal data in terms of both system security, breaches of confidentiality and the potential to re-identify individuals from anonymised data). This applies both to the research and healthcare domain, and translation between them;

(b) the incidence and prevalence of such harms and the appropriate context in which to assess them;

[487] Review 1 was commissioned jointly with the Expert Advisory Group on Data Access (EAGDA).

(c) relevant definitions (e.g. meaning of 'sensitive personal data' in the regulatory context);

(d) the effectiveness of mechanisms of redress in documented known cases;

(e) possible alternative governance options for controlling data and their likely consequences; and

(f) the nature and significance of any drivers or conditions favouring misuse of data (e.g. a 'black market' in personal data).

The review was carried out by Professor Graeme Laurie, University of Edinburgh, Ms Leslie Stevens, University of Edinburgh, Dr Kerina H.Jones, Swansea University, and Dr Christine Dobbs, Swansea University. It is available at: www.nuffieldbioethics.org/project/biological-health-data/evidence-gathering/.

Review 2: Evidence relating to the history of the relationship between the private and public sector in the field of human genomics

Purpose: to assist the Working Party in understanding:

(a) The nature of the relationship between the public and private sectors in the development and execution of biological and health research;

(b) whether there is an identifiable change in the nature of that relationship e.g. from competition (Human Genome Project; BRCA identification) to collaboration (Genomics England);

(c) if such a change can be identified, what relationship, if any, does it have with the current conjunction of open data/open policy making?; and

(d) if the nature of public sector involvement in research has changed.

The review was carried out by Professor Paul Martin, University of Sheffield and Dr Greg Hollin, University of Nottingham and is available at: www.nuffieldbioethics.org/project/biological-health-data/evidence-gathering/.

External review

An earlier version of the report was reviewed by twelve individuals with expertise in disciplines relevant to the project. These individuals were:

- Professor Carol Brayne
- Dr Deborah Peel
- Professor Douwe Korff
- Dr Eerke Boiten
- Dr Effy Vayena
- Harry Cayton OBE
- Leila El Hadjam
- Dr Mark Taylor
- Dr Neil Manson
- Professor Sheila M Bird OBE FRSE
- Professor Steve Yearley
- Dr Vitaly Shmatikov

The Working Party deeply appreciates the time and thought the reviewers brought to this investigation and thanks them for their helpful contributions.

The views expressed within this Report are those of the Working Party and the Council and do not necessarily reflect the views of any participants in the various activities undertaken by the Working Party in connection with this Report.

Appendix 2: Wider consultation for the Report

The aim of the consultation was to obtain views from as wide a range of organisations and individuals interested in the area as possible. The consultation document was published online (available in hard copy on request). Individuals and organisations known by the Working Party to be interested were also directly alerted by email and encouraged to respond. The document was divided into two parts, with the first one outlining three areas of development relevant to the topic of the report and the second asking for more detailed responses to the following questions, for each of which potentially relevant aspects were highlighted:

- Do biomedical data have special significance?
- What are the new privacy issues?
- What is the impact of developments in data science and information technology?
- What are the opportunities for, and the impacts of, the use of linked biomedical data in research?
- What are the opportunities for, and the impacts of, data kinking in medical practice?
- What are the opportunities for, and the impacts of, using biomedical data outside biomedical research and health care
- What legal and governance mechanisms might support the ethical linking of biomedical data?

Respondents were encouraged to answer as many, or as few, as they wished. Fifty-one responses were received, 22 from individuals and 29 from organisations.

All the responses were circulated to Working Party members and a summary of responses was considered in detail at a subsequent Working Party meeting.

A summary of the responses received, together with the original consultation paper, is available on the Council's website. Individual responses will also be published in full on the website, where respondents have granted permission for the Council to do so. The responses received played an important role in shaping the Working Party's thinking, and the Working Party is grateful to all those who contributed.

Anonymous

Three respondents wished to remain unlisted in the report.

Individuals

- Professor Sheila M Bird OBE FRSE
- Martin Bobrow
- Dr Jo Bowen
- Professor Carol Brayne
- Shawneequa Callier
- Patrick Finlay PhD
- Jane Halliday
- Ian Herbert
- Julian Hitchcock
- Atina Krajewska and Ruth Chadwick
- Professor Neil Lawrence, Department of Computer Science and Sheffield Institute for Translational Neuroscience, University of Sheffield
- Nadine Levin, University of Exeter

- Pauline McCormack, Simon Woods and Jackie Leach-Scully, PEALS Research Centre, Newcastle University
- Sylwia Maria Olejarz
- Dr John Saunders, Chair of the Royal College of Physicians (RCP)
- Bettina Schmietow, European School of Molecular Medicine and University of Milan
- Professor Tim Spector, KCL
- Dr Mark J Taylor, The University of Sheffield
- Professor Martin Widschwendter and Dr Daniel Reisel, UCL Department of Women's Cancer's, Institute of Women's Health, University College London

Organisations

- Association of Medical Research Charities (AMRC)
- British Dental Association
- British Medical Association
- Cancer Research UK
- CARE
- Centre for Longitudinal Studies (ESRC Resource Centre), Institute of Education, University of London
- Christian Medical Fellowship
- Clinical Trial Service Unit & Epidemiological Studies Unit (CTSU), Nuffield Department of Population Health, University of Oxford
- eHealth Research Group, University of Leeds Institute of Health Sciences
- EMBL - European Bioinformatics Institute
- Exeter Centre for the Study of the Life Sciences (Egenis)
- Farr Institute @CIPHER, with input from the Innovative Governance Group of the Farr Institute for Health Informatics Research
- GeneWatch UK
- Information Commissioner's Office
- Medical Research Council (MRC)
- National Bioethics Commission of Mexico
- Nowgen
- PHG Foundation
- Privacy Advisory Committee, Northern Ireland
- Progress Educational Trust
- Royal Academy of Engineering
- The Human Fertilisation and Embryology Authority
- The Mason Institute for Medicine, Life Sciences and the Law, University of Edinburgh
- The Physiological Society
- The Wellcome Trust
- U.S. Presidential Commission for the Study of Bioethical Issues
- UCL Centre for Health Informatics and Multiprofessional Education
- Wellcome Trust Sanger Institute
- World Medical Association (WMA)

Appendix 3: The Working Party

Professor Martin Richards (Chair)

Martin Richards is Emeritus Professor of Family Research at the University of Cambridge. Until his retirement (2005) he was Director of the Centre for Family Research where he continues to researched parent-child relationships, family life and genetic and reproductive technologies. He is a member of the Cambridge University United Hospitals NHS Foundation Trust Tissue Management Committee. Until recently he was Vice-Chair of the UK Biobank Ethics and Governance Council and he previously served on the Human Genetics Commission and the Ethics and Law Committee of the Human Fertilisation and Embryology Authority. His books include, *The Troubled Helix* (edited with T. Marteau, Cambridge University Press, 1997), *The Limits of Consent* (edited with O. Corrigan and others, Oxford University Press, 2009), *Reproductive Donation: Practice, Policy and Bioethics* (edited with J. Appleby and G. Pennings, Cambridge University Press, 2012), *Relatedness in Assisted Reproduction* (edited with T. Freeman, S. Graham and F. Ebtehaj, Cambridge University Press, 2014). He is currently working on a history of assisted reproduction.

Professor Ross Anderson

Ross Anderson is Professor of Security Engineering at Cambridge University. He is a Fellow of the Royal Society, the Royal Academy of Engineering, the IET and the IMA, and wrote the textbook *Security Engineering – A Guide to Building Dependable Distributed Systems.* He has a long-standing interest in health care IT, having advised the BMA on the safety and privacy of clinical information systems in the 1990s and more recently having been a special advisor to the House of Commons' Health Committee. He was an author of *Database State*, the Joseph Rowntree Reform Trust report which led to the cancellation of the ContactPoint children's database. He also chairs the Foundation for Information Policy Research and is on the advisory council of the Electronic Privacy Information Centre.

Stephen Hinde

Stephen Hinde was the Head of Information Governance and Caldicott Guardian for the Bupa Group for eighteen years. He retired in December 2013 and now advises a number of Christian charities on data protection and confidentiality. He is a member of the UK Council of Caldicott Guardians, having served as its inaugural chairman. He also sits on the Wales Information Governance Board. He was Deputy Chairman of the National Information Governance Board for Health and Social Care (NIGB) and was a member of its Ethics and Confidentiality Committee, which advised on s251 requests.

Professor Jane Kaye

Jane Kaye is Professor of Health, Law and Policy and Director of the Centre for Law, Health and Emerging Technologies (HeLEX) at the University of Oxford. Her research involves investigating the relationships between law, ethics and practice in the area of emerging technologies in health. Her main focus is on genomics with an emphasis on biobanks, privacy, data-sharing frameworks, public engagement, global governance and translational research. She is leading the ELSI 2.0 initiative and is on a number of Advisory Boards for large scientific projects, as well as journal editorial boards.

Professor Anneke Lucassen

Anneke Lucassen is Professor of Clinical Genetics and Honorary Consultant Clinical Geneticist, University of Southampton Cancer Sciences Division and The Wessex Clinical Genetics Service. Her clinical expertise is in cancer and cardiac genetics and she leads the clinical ethics and law unit at Southampton faculty of medicine (CELS) which researchers the social, ethical and legal aspects of genomic developments in clinical practice. She is chair of the Southampton University Hospitals NHS Trust Clinical Ethics Committee and of the British Society of Genetic medicine's Ethics and Policy committee. She is cofounder and organiser of the UK Genethics Club. She is a former member of the human genetics commission and current member of the ethics advisory committee for Genomics England.

Professor Paul Matthews

Paul Matthews is the Edmond and Lily Safra Professor of Translational Neuroscience and Therapeutics and Head of the Division of Brain Sciences at Imperial College, London, where he serves as the Medicine representative on the Data Science Institute Research Board. He is also a Fellow by Special Election of St. Edmund Hall, Oxford and holds a number of other honorary academic appointments. He received his training in neurology at Oxford, Stanford and McGill. He chairs the Imaging Enhancement Working Group and sits on the Steering Committee of UK Biobank. He also chairs the Imaging Network and is an ad hoc Executive Team member of the Dementias Platform UK. His personal research focuses on the characterising relationships between microglial activation, adaptive plasticity and neuroaxonal loss in the progression of in people with multiple sclerosis.

Professor Michael Parker

Michael Parker is Professor of Bioethics and Director of the Ethox Centre at the University of Oxford. His main research interest is in the ethics of collaborative global health research. He leads the ethics programme of the Malaria Genomic Epidemiology Network (MalariaGEN) and, together with partners at the Wellcome Trust Major Overseas Programmes in Africa and South East Asia, he co-ordinates the Global Health Bioethics Network. Since 2001, he has been one of the co-ordinators of the Genethics Club, a national ethics forum for health professionals and researchers in the UK to discuss ethical issues arising in their practice. The forum's work was published in the book Ethical Problems and Genetics Practice. Michael is currently Chair of the Ethics Advisory Committee of the 100,000 Genomes Project, Chair of the Data Access Committee of the Wellcome Trust Case-Control Consortium and a member of the Medical Research Council's Ethics, Regulation and Public Involvement Committee.

Margaret Shotter

Margaret Shotter fulfils a number of advisory roles relating to ethics and governance issues for medical research following a career in biostatistics with a special interest in the research methodology of clinical trials and then as senior manager for research ethics at UBC, Vancouver. There she had overall responsibility for the ethical reviews of human subject research across the university, including its affiliated hospitals and research institutes. Since 2008, she has been a member of the UK Biobank Ethics and Governance Council. She is a lay member of Expert Advisory Groups reporting to the Commission on Human Medicine, and serves on several other advisory panels relating to medical research.

Dr Geoff Watts

Geoff Watts spent five years in research before becoming a science and medical writer and broadcaster. He presented BBC Radio 4's Medicine Now and, more recently, its science programme Leading Edge. He was a founder member of, and served for six years on, the Human Genetics Commission. Geoff chaired the Council's Working Group on mitochondrial donation.

Dr Susan Wallace

Susan E. Wallace is Lecturer of Population and Public Health Sciences, Department of Health Sciences, University of Leicester. Her research interests include the legal and policy implications of population-based and disease-based longitudinal cohort studies and biobanks, ethical issues in biomedical research, research ethics review, and public health genomics. Currently, she is a member of the International Cancer Genome Consortium (ICGC) Ethics and Policy Committee, the UK ICGC Prostate Project Oversight Group and the Canadian Partnership for Tomorrow Project ELSI Task Force. She conducts policy research in collaboration with the UK National Child Development Study (1958 Birth Cohort) and is involved in the BioSHaRE-eu (FP7) project which focuses on the development and evaluation of tools and methods for accessing and exploiting data from biobanks and cohort studies.

John Wise

John Wise is the Executive Director of the Pistoia Alliance and the Programme Co-ordinator for the PRISME Forum. The Pistoia Alliance is a not-for-profit, cross-company organisation committed to lowering the barriers to innovation in Life science R&D. The PRISME Forum is a not-for-profit Pharma R&D IT leadership group focussed on the identification and palliation of "hot topics", and the sharing of industry best practices. John has worked in life science R&D informatics in a variety of organizations including academia, the pharmaceutical Industry, a cancer research charity as well as in the technology supply side of the industry. This has provided him with direct, hands-on experience of delivering computer-based services across the life science R&D value chain. John graduated in physiology before obtaining a post-graduate certificate in education.

Glossary

Algorithm: an effective method for calculating a function, expressed as a finite list of well-defined instructions. In the report, this term is used with particular reference to statistical data-mining.

Anonymisation: the removal of the names, addresses and other identifying particulars of data subjects – in the report with particular reference to their medical records – with a view to making their re-identification more difficult.

Article 29 Working Party: the European advisory body on data protection established under Article 29 of the European Data Protection Directive. It consists of all the DP Supervisors of EU member states, the European Data Protection Supervisor, and a representative of the European Commission. It publishes opinions which are influential, but which explain the law rather than having legal force of themselves.

Ascertainment bias: (also sampling bias) a systematic distortion in measuring the true occurrence of a phenomenon resulting from the way in which data are collected, where all relevant instances were not equally likely to have been recorded.

Barnardisation: a procedure of randomly altering data values prior to publication of statistics to make identification more difficult, named after mathematician George Alfred Barnard.

Bespoke data linkage: a service designed for a customer linking one or more data sets to data supplied by that customer.

Bespoke extract – pseudonymised: a one-off extract tailored to the customer's requirements of specified data fields containing patient identifiable data, sensitive data items or both, and linked to an existing scheme of pseudonyms.

Big data: a term used to describe large and rapidly growing datasets in all areas of life and accompanying technologies and developments to analyse and re-use these, especially with the ambition to release latent or unanticipated value. *See also Knowledge discovery in databases*; *Data mining*; *Data science*; *Machine learning*.

Biobank: a repository which collects, processes, stores and distributes tissue and data for biomedical, genomic or other research purposes.

Bioinformatics: an interdisciplinary field that develops methods and software tools for understanding biological data.

Biomarker: a measurable characteristic that can indicate an underlying biological state or condition, such as a disease state or pharmacologic response.

Biomedicine: medical research which applies principles of biology, biochemistry, etc., to medical research or practice.

Biometrics: the use of metrics of unique human characteristics or traits for identification and surveillance purposes.

Blagging: to obtain or disclose personal data or information by impersonation or other forms of deception and as such an offence under the UK Data Protection Act (1998).

Caldicott Guardian: a senior person responsible for protecting the confidentiality of patient and service-user information and enabling appropriate information-sharing at NHS and social care organisations as well as voluntary and independent sector organisations. The role was created following a recommendation in the "Report on the Review of Patient-Identifiable Information" (1997) chaired by Dame Fiona Caldicott.

Citizen science: scientific work undertaken by members of the general public, often in collaboration with or under the direction of professional scientists and scientific institutions. In the report, mainly with reference to publics providing data about themselves for biomedical research. See *Patient-driven research*.

Clinical care data: data collected about individuals in the context of a clinical intervention.

Clinical trial: a study using human subjects to investigate the efficacy and/or safety of a medicine or other clinical intervention. Traditionally, 'Phase I' trials involve a small group of healthy volunteers or patients with the disease to evaluate the effect of a range of doses and identify potential side effects. In 'Phase II' trials the drug or invervention is given to a larger group of patients to evaluate effectiveness and assess safety. 'Phase III' trials investigate effectiveness and safety, normally in comparison to existing treatments, in preparation for wide-scale use.

Cloud computing: a term for shared computing used to refer to the outsourcing of computation to centralised shared resources, typically in remote data centres.

Coded data and samples: biological samples and associated data labelled with at least one specific code, which allows traceability to a given individual, the ability to perform clinical monitoring, subject follow-up, or addition of new data.

Cohort study: a form of observational, longitudinal study in which a group (cohort) linked by some characteristic is followed over time to study the effects of, for example, particular risk factors.

Confidentiality: a societal norm or legal duty relating to the disclosure of information obtained in contexts (often, but not exclusively, professional ones) where expectations underpinning such norms are reasonable i.e. 'in confidence'. In specific professional contexts, there may be established codes of practice to safeguard confidentiality. See also *Hippocratic oath*.

Consent: the voluntary, informed and competent agreement of an individual to any action that would otherwise constitute an infringement of fundamental personal interests or rights.Consent is an important ethical mechanism in medical treatment, research participation and processing of sensitive data.

Correlation: a broad class of statistical relationships involving formal dependence between two random variables or two sets of data.

Data: Literally 'given things', i.e. evidence from observation or measurement, or facts that are assumed as the basis for further analysis, calculation or reasoning. Cf. *Information*.

Data abuse: in this report, a broad category of insecure, inadequate or unethical uses of data that have been empirically observed, including fabrication or falsification of data; data theft; unauthorised disclosure of or access to data; non-secure disposal of data; unauthorised retention of data; technical security failures and data loss. Cf. empirical typology of data abuses (box 2.2).

Data access committee: a governance infrastructure within (research) institutions or associated with research studies scrutinising or authorising data specific applications or for specific data collections. See also Appendix 3.

Data breach: the wrongful release of personal or sensitive information, whether as a result of accident, negligence or malice. See also *Data abuse*

Data initiatives: purposive activities in which data collected for one purpose are used for a new purpose, often involving linking with other data sources.

Data mining: the computational process of discovering patterns in big data sets. See also *Big data*; *KDD; Machine learning*.

Data science: the extraction of knowledge from data using techniques drawn from computer science, mathematics, and statistics.

Data sharing: extending access to data to data users who were not intended users at the time of data collection, usually for the purpose of further research or analysis; the term is common in research and policy contexts and is suggestive of the social benefits of data re-use.

Deep phenotyping: comprehensive analysis of phenotypic characteristics or abnormalities in which the individual components of the *Phenotype* are observed and described, particularly in *Precision medicine* and using computational and imaging applications.

De-identification: see *Anonymisation*.

Digitisation: converting analogue signals such as images, sounds and documents into digital ones.

DNA: deoxyribonucleic acid; the chemical that carries a person's genetic information. Most cells of a person's body contain a copy of that information. A DNA molecule consists of a long chain of nucleotides whose sequence codes for the production of proteins in cells.

Dynamic consent: interfaces for patients or research participants that aim to enable them to give or withdraw consent for their information to be used in specific research projects, thereby it is argued overcoming the limitations of all-or-nothing consent regimes.

Electronic health record: a computerised record of a patient's medical history, such as medication, allergies, results of health tests, lifestyle and personal information that can be used in different health care settings. See *Medical Record*

Epidemiology: a field of study investigating incidence, prevalence, causes and effects of disease in a defined population, as well as appropriate prevention or treatments.

Epigenomics: a field of study investigating chemical tags on the genome that control the activities of genes in contrast to *Genomics*.

Evidence-based medicine (EBM): the conscientious, explicit and judicious use of current best evidence in making decisions about the care of individual patients. This involves integrating individual clinical expertise with the best available external clinical evidence from systematic research.

Genome: the complete set of DNA within a single cell of an organism.

Genomics: a discipline in genetics that applies technologies such as genome sequencing methods and bioinformatics to the study of function and structure of genomes.

Genome sequencing: techniques which allow researchers to read the genetic information in DNA.

Genome-wide association studies (GWAS): studies involving large numbers of people with and without a particular disease, each of whom is genotyped at several hundred thousand markers throughout the genome. Comparisons are then made between these groups to identify genetic markers associated with the disease or its absence.

Genotype: the genetic makeup of a cell, an organism, or an individual usually with reference to a specific characteristic under consideration.

Guthrie cards: a paper card conserving blood from a baby's heel prick, named after Robert Guthrie, who introduced these in Scotland in 1965. The cards are used for screening for a variety of metabolic disorders such as phenylketonuria, congenital hypothyroidism, blood disorders such as sickle cell disease and HIV infection in more than 20 countries.

Health-related findings: Findings discovered during research with human participants that relate specifically to an individual participant's health.

Hippocratic oath: an oath historically taken by physicians to show their commitment to upholding ethical standards, including patient benefit, avoidance of harm and *Confidentiality*.

Identifying information: information that relates to an individual (or individuals) and from which their identity can be determined either directly (as in the case of a proper name) or, deductively, in combination with other available information. See *Personal data***.**

Information: in this report, data that have gained informational value and meaning in a particular context.

Informational privacy: an interest or right in the disclosure and withholding of information, founded in respect for persons; an aspect of privacy provided for in legal instruments such as the European Convention on Human Rights.

Knowledge discovery in databases (KDD): the process of extracting patterns and knowledge from data in large databases*. See also Machine learning, Big data.*

Knowledge economy: an economy whose focus is information rather than physical products or processes (such as mining or manufacturing).

Life logging: the process of tracking personal data generated by someone's activities such as exercising, sleeping, and eating, in particular with the help of wearable technology and the supporting digital services and applications. See for example the "Quantified Self" Movement.

List cleaning: Validating demographic data to ensure it is accurate and improve linkage outcomes.

Lloyd George record: the traditional paper (GP) medical records introduced in 1911 by then health Minister David Lloyd George.

Machine learning: a branch of computer science that deals with the construction and study of algorithms that can learn from data.

Material Transfer Agreement (MTA): a contract that governs the transfer of tangible research materials between two organisations, when the recipient intends to use it for their own research purposes, but the provider typically retains right in the materials and any derivatives.

Medical confidentiality: see *Confidentiality* and *Hippocratic oath*.

Medical record: a record, whether paper or electronic, of medical history, such as medication, allergies, results of health tests, lifestyle and personal information, which is created in the context of clinical care and whose purpose is recording that care and facilitating the care of the patient in the future.

Meta-analysis: a quantitative statistical analysis of several separate but similar clinical trials or biomedical studies in order to gain more precise estimates of treatment effects or to investigate factors which may explain heterogeneity of outcomes.

Metadata: data that describe the contents of substantive records and the circumstances of their creation and processing, for example the size of data files, the time or location at which they were created, identity of the author, and technical characteristics of the data.

Open data: data anyone is free to access, use, modify, and share, but which may have sharing conditions so that it is correctly attributed or that further use is not constrained (see 2.32).

Participant-driven research: see *Citizen science*.

Patient-reported outcome measures (PROMs): in the NHS, “health gain in patients undergoing hip replacement, knee replacement, varicose vein and groin hernia surgery in England, based on responses to questionnaires before and after surgery”. (HSCIC)

Personal confidential data (PCD): a term (used in Caldicott2) to describe personal information about identifiable individuals who are owed a duty of confidentiality, i.e. the information was given ‘in confidence’ and should be kept private unless there is a legal basis or overriding public interest for disclosure. PCD includes information about deceased as well as living individuals and is therefore different in scope from ‘personal data’ under DPA 1998.

Personal data: identifiable or identifying information relating to living natural persons (data subjects). Data protection law applies to such data. See also *Sensitive personal data*.

Personalised medicine: a concept in medicine and health care policy according to which diagnostic testing is employed for selecting appropriate and optimal therapies based on the context of a patient’s genetic content.

Phenotype: the composite of an organism's observable characteristics or traits, such as its morphology, development, biochemical or physiological properties, and behaviour. A phenotype results from the expression of an organism's genes as well as the influence of environmental factors and the interactions between the two.

Precision medicine: see *Personalised medicine*.

Privacy: the interest and right people have in controlling access to themselves, their homes, or to information about them. What counts as private can change depending on social norms, the specific context, and the relationship between the person concerned and those who might enjoy access. Privacy is exercised by selectively withholding or allowing access by

others or through limits on acceptable behaviour in others. In the UK and the rest of Europe, privacy is guaranteed by the European Convention on Human Rights. *See Informational privacy*.

Pseudonymisation: processes of de-identification which are reversible, for example by replacing names with patient numbers, used for example to enable further data about individuals to be added over time.

Public-private partnerships: government services provided and financed by one or more private sector entities.

Re-identification: the act of identifying data that have been de-identified to protect the data subject's *Privacy*.

Safe havens: initially medical record libraries in hospitals or health authorities, more recently used for data centres providing a pseudonymisation and linkage service, so that medical records whose names have been removed can be linked up with other records from other providers that refer to the same individual patients. See also *Trusted third party.*

Sample bias: see *Ascertainment bias*.

Secondary use: 'reuse', 'secondary use' or 'repurposing' of data: any data use that goes beyond the use intended at the time of data collection and for which the patient gave consent. It typically means the use of GP and hospital medical records for research and administrative purposes. In this report we are also concerned with unpredictable future uses, and uses for incompatible purposes.

Sensitive personal data: in data protection law, personal data of specified kinds, including, for example, data on physical or mental health or condition, racial and ethnic origin, sexual life, political opinions, membership of a trade union, or lawbreaking. Additional data protection measures apply to sensitive personal data.

Summary Care Record (SCR): nationwide system containing a GP record summary (initially, current prescriptions and allergies) that would facilitate out-of-hours care and could also enable patients to view their own records via a mechanism called Healthspace promoted by Connecting for Health.

Stratified medicine: see *Personalised Medicine*.

Tabulation: a statistical table of aggregate data (HSCIC).

Trusted third party: a trusted organisation with secure facilities for linking data which is normally independent of institutions which hold data**.** See also *Safe havens.*

Whole genome sequencing: see *Genome sequencing.*

List of abbreviations

ALSPAC	Avon Longitudinal Study of Parents and Children
ARDC	Administrative Data Research Centre
BMA	British Medical Association
CAG	Confidentiality Advisory Group (of the HRA)
CMO	Chief Medical Officer
CPRD	Clinical Practice Research Datalink
DAAG	Data Access Advisory Group (of the HSCIC)
DH	Department of Health
DPA	Data Protection Act
EAGDA	Expert Advisory Group on Data Access
EBM	evidence-based medicine
ECC	Ethics and Confidentiality Committee (esp. of the NIGB)
ECHR	European Convention for the Protection of Human Rights and Fundamental Freedoms
ECtHR	European Court of Human Rights
EGC	Ethics and Governance Council (esp. of UK Biobank)
EGF	Ethics and Governance Framework (esp. for biobanks)
EHR	electronic health record
EPR	electronic patient record
ESRC	Economic and Social Research Council
FDA	Food and Drug Administration (USA)
FOI	Freedom of Information
GCHQ	Government Communications Headquarters
GDPR	General Data Protection Regulation (UK)
GeL	Genomics England Limited

GMC	General Medical Council
GP	general practitioner
GPES	General Practice Extraction Service
GWAS	Genome-Wide Association Studies
HES	Hospital Episodes Statistics
HRA	Health Research Authority
HSCIC	Health and Social Care Information Centre
ICO	Information Commissioner's Office
IIGOP	Independent Information Governance Oversight Panel
KDD	knowledge discovery in databases
MHRA	Medicines and Healthcare products Regulatory Agency
MRC	Medical Research Council
NHS	National Health Service
NHSE	National Health Service England
NIGB	National Information Governance Board for Health and Social Care
NIHR	National Institute for Health Research
NPfIT	National Programme for IT
NSA	National Security Agency (USA)
OECD	Organisation for Economic Co-operation and Development
OLS	Office for Life Sciences
ONS	Office for National Statistics
PDS	Personal Demographics Services
PGP	Personal Genome Project
PIAG	Patient Information Advisory Group
PLM	PatientsLikeMe
PLR	Patient-led research
QOF	Quality and Outcomes Framework
RCGP	Royal College of General Practitioners

RCT	randomised controlled trial
REC	Research Ethics Committee
SCR	Summary Care Record
SHIP	Scottish Informatics Programme
SPIRE	Scottish Primary Care Information Resource
SUS	Secondary Uses Service
UKBB	UK Biobank

Index