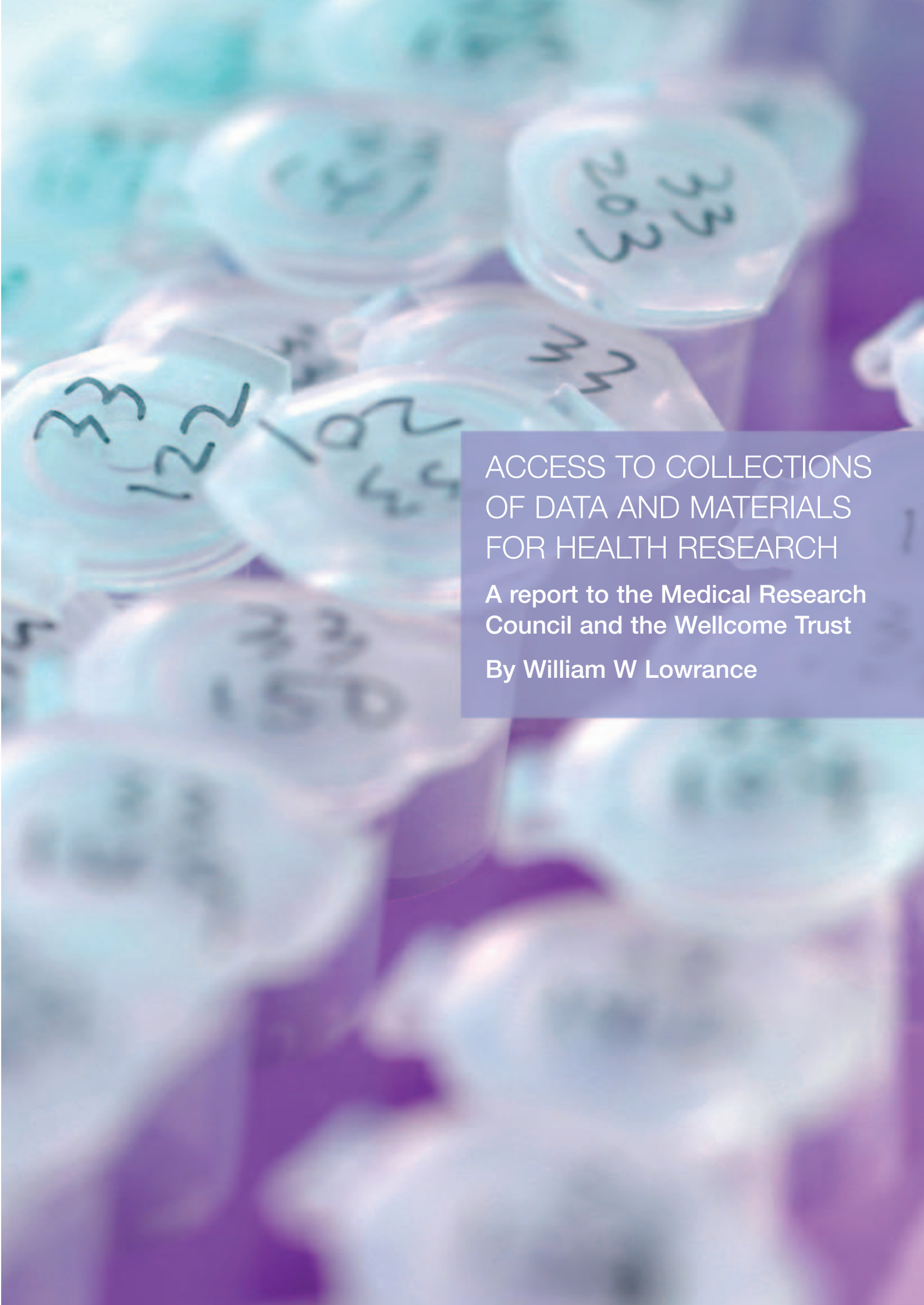

ACCESS TO COLLECTIONS OF DATA AND MATERIALS FOR HEALTH RESEARCH

A report to the Medical Research Council and the Wellcome Trust

By William W Lowrance

This report was commissioned by the Medical Research Council (MRC) and the Wellcome Trust to review various issues surrounding research access to population-based collections of data and materials in the UK. The views and opinions expressed are the author's own and do not necessarily reflect those of the MRC or the Wellcome Trust. However, as the report contains valuable material that can inform discussions and policy making in this area, it is being made widely available to all interested parties.

Available online at www.wellcome.ac.uk/accessreport and www.mrc.ac.uk/research_collection_access

The Medical Research Council (MRC) is the UK's leading publicly funded medical research organisation. Its aim is to improve human health by supporting high-quality biomedical research. MRC-funded research has led to many of the most significant discoveries in medical science, both in the UK and worldwide.

The Wellcome Trust is an independent charity whose mission is to foster and promote research with the aim of improving human and animal health.

Contents

Preface

This is a report from an independent consultant to the Medical Research Council (MRC) and the Wellcome Trust. The charge was to review the issues surrounding research access to population-based collections of data and materials in the UK, principally collections that the two organisations fund or have some responsibility for.

The funding organisations asked the consultant to review the access arrangements across the range of collections supported in the UK by the MRC and the Trust, separately or in partnership, to aid them in deciding:

1. The extent to which current access arrangements are **standardised**
2. Whether there is scope for **greater standardisation**, given the ethical, legal, and practical considerations
3. Whether there is scope for a **model governance structure** and if so, whether this might usef ully include elements centralised to cover several collections
4. Whether there is scope to develop **guidelines** of general applicability to the range of collections supported.

Early along it became clear that answers to these questions depend on consent and confidentiality concerns, scientific concerns, proprietary concerns, issues of decision authority and oversight, and some practical considerations, and the funders accordingly broadened the remit.

The study proceeded via discussions with many principal investigators, data custodians, archivists, representatives of the funding organisations, and chairs of oversight or advisory committees, complemented by review of working documents and technical literature. The author is very grateful to all the busy people who took the time to share their experiences and insights. (Those consulted are listed in Appendix 1.)

The review was helpfully informed by, and builds upon, a report prepared for the MRC in 2002 by Dr Louise Corti and Ms Melanie Wright of the UK Data Archive.[1]

1 Corti L, Wright M. 'MRC Population Data Archiving And Access Project Consultants' Report: Developing an MRC policy for population data archiving and access' (draft). Medical Research Council; September 2002. www.mrc.ac.uk/prn/index/strategy-strategy/strategy-science_strategy/strategy-strategy_implementation/strategy-other_initiatives/strategy-data_sharing/pdf-ukda_draft_report.pdf-link [accessed 21 December 2005].

Executive summary

Access to collections can be improved, and most scientists hope it will be. But if access is to be optimised, not only will barriers have to be reduced but the provision of access will have to be actively facilitated, guided, funded and rewarded.

Arguments for increasing access include:

- deriving greater informational value and higher return on invested funding and effort, and thereby increased health benefit for society
- serving the reason above and the tradition of scientific openness, not denying researchers access to resources that are special or that would be difficult or expensive to duplicate
- allowing verification or variant replication of studies
- reducing unproductive duplication of effort
- minimising the need for patients or members of the public to participate, donate samples or give permissions
- facilitating the linking, pooling, or comparing of data sets or materials with other data sets or materials
- improving the quality of collections and their richness as more, and more diverse, researchers analyse them, gain competence, improve methods and procedures, publish and return results, and identify new avenues of investigation.

If proper, efficient access is to be increased, the following must be attended to:

- making sure that as access is provided, the original promises to the participants, such as those relating to consent, confidentiality, and rights of withdrawal, are strictly kept
- not allowing the public reputation of the project or its relationship with participants to be jeopardised
- upholding the scientific reputation of the project
- protecting the interests and rewarding the hard work and goodwill of the developers and custodians of the resource
- fairly compensating the effort and costs incurred in enabling access, such as in anonymising, documenting, and archiving data, reviewing applications for access and negotiating access agreements, preparing data and/or materials, and assisting with scientific interpretation
- judiciously managing intellectual property.

The following are precis of responses to the funders' initial questions.

Q: To what extent are current access arrangements standardised?

They are not very standardised, although there are many commonalities.

Q: Is there scope for greater standardisation, given the ethical, legal and practical considerations?

Yes. The core terms of access and material transfer agreements are begging for standardisation, for instance. Standard default criteria for access to materials could be useful. And there are other opportunities.

Q: Is there scope for a model governance structure and if so, might this usefully include elements centralised to cover several collections?

The most common mode of governance is that of supervision of collections and data sharing by oversight committees. With some variation, most of these have similar remits and structural features. A model structure could be derived, incorporating the best features of the current oversight bodies.

The report identifies examples of situations in which multiple oversight or advisory committees might be consolidated.

Q: Is there scope to develop guidelines of general applicability to the range of collections supported?

Clarification and revised guidance are urgently needed on aspects of consent, confidentiality, and anonymisation. Guidance on the characteristics of thoroughly anonymised 'limited data sets' would be helpful.

Whether guidelines would be the best instrument is not evident, but it is important now to sort out the rights and obligations of data providers (such as regarding the screening of the bona fides of data requesters, the merits of proposals, and the quality of manuscripts) and the rights and obligations of data requesters (such as regarding whether they can resist working in collaborative mode if they prefer to, and whether they can be required to publish negative-association results).

At least informal guidance is needed on such matters as publication of collections' access policies and procedures. This might be coupled with guidance on the elements of access agreements.

Criteria or guidelines deserve to be drafted on the documentation required to support scientifically sound independent use of data sets.

So, much can be done.

1. The impetus for data sharing

Recent years have brought many calls for the optimisation of data sharing for research, with the intention of deriving maximal societal benefit from the contributions made by research participants, the investment of public and charitable funds, and the support provided by universities and medical centres.

Data sharing involves a variety of issues, central among them the ones that this report will examine as 'access' issues (as compared with, for example, archiving issues).

1.1 MRC and Wellcome Trust policies

1.1.1

The MRC has adopted a 'Statement on Data Sharing and Preservation Policy', the opening of which declares:

> MRC expects that the valuable data arising from the research it supports will be made available to the scientific community to enable new research with as few restrictions as possible. Such data must be shared in a timely and responsible manner.
>
> New studies enabled through data sharing should meet the high scientific quality, ethical and value for money standards of all MRC research; and should add distinctive value to that of the original dataset.[2]

1.1.2

Then the Statement so clearly articulates the challenge that it must be quoted extensively here:

> MRC supports the view that those enabling sharing should receive full and appropriate recognition by funders, their academic institutions and new users for promoting secondary research.
>
> Such research is often most fruitful as a collaboration between the new user and the original data creators or curators, with the responsibilities and rights of all parties agreed at the outset.
>
> A limited, defined period of exclusive use of data for primary research is reasonable: different disciplines require different approaches, reflecting the nature and value of the data and the way they are generated and used. Ongoing research contributing to the completion of datasets must not be compromised by premature or opportunistic sharing and analysis. Sharing should always take account of enhancing the long-term value of the data.
>
> MRC policy is not intended to discourage filing of patent applications in advance of publication, and recognises that it may be necessary on occasion to delay publication for a short period to allow time for filing to be considered and drafted.
>
> Medical research involving personal data has special responsibilities associated with it, particularly in relation to consent and confidentiality. It is essential that the appropriate regulatory permissions – ethical, legal and institutional – are in place prior to sharing data of this type. ... Researchers, research participants and research regulators must ensure that, within the regulatory requirements of the law, opportunities for new uses are maximized without unnecessary restriction. Potential research benefits to patients and public should outweigh identified risks. Risks such as inappropriate disclosure of personal information must be managed in a proportionate yet robust manner.
>
> To enable effective sharing, data must be properly curated over its life-cycle and released with the appropriate high-quality metadata. This is the responsibility of the data owners who are usually those individuals or institutes that have received funding to create or collect the primary data. ...
>
> From 1 January 2006, all applicants submitting funding proposals to MRC must include a statement explaining their strategy for data preservation and sharing. ... Applicants who consider data arising from their MRC-funded proposals not amenable to sharing must provide explicit reasons for not making the data available.[3]

2 'MRC Statement on Data Sharing and Preservation Policy'. Medical Research Council; September 2005. www.mrc.ac.uk/index/strategy-strategy/strategy-science_strategy/strategy-strategy_implementation/strategy-other_initiatives/strategy-data_sharing/strategy-data_sharing_policy.htm [accessed 21 December 2005].

3 Ibid.

1.1.3
Although the Wellcome Trust does not have an omnibus policy on access to collections, its leaders and advisory bodies are increasingly, programme by programme, encouraging the provision of as wide access as makes sense. Many of its programmes have specific data-sharing requirements. The Trust insists that human genome sequence information should be placed in the public domain and made freely available, and it requires that the projects that it funds submit genotype data to public databases as rapidly as possible.

1.2 NIH policy

The US National Institutes of Health (NIH) believes that "data should be made as widely and freely available as possible while safeguarding the privacy of participants and protecting confidential and proprietary data".[4] It requires that applicants for grants exceeding $500 000 include in their application a plan for sharing final research data, or a clear justification of not sharing. The NIH's rationale is that:

> Sharing data reinforces open scientific inquiry, encourages diversity of analysis and opinion, promotes new research, makes possible the testing of new or alternative hypotheses and methods of analysis, supports studies on data collection methods and measurement, facilitates the education of new researchers, enables the exploration of topics not envisioned by the initial investigators, and permits the creation of new datasets when data from multiple sources are combined.[5]

1.3 Other sources of impetus

1.3.1
In 2003 the Wellcome Trust convened a meeting in Florida to discuss issues of pre-publication release of data, especially genomic sequence data, which led to a statement known as the Fort Lauderdale Agreement.[6] This agreement encourages rapid release of sequence data to public databases, and it outlines roles for funding organisations, resource producers, and resource users to facilitate this. Rapid posting of raw sequence data on public databases – sometimes as quickly as at the end of a SNP-mapping day – has become a cultural habit of genomic research.

1.3.2
The Fort Lauderdale group "recommended that the principle of rapid pre-publication release should apply to other types of data from other large-scale production centers specifically established as community resource projects". It did not develop this point, though, probably because it was not the best group to do so.

1.3.3
Many countries are encouraging data sharing, for reasons similar to those described above. Broader access resonates with the development of web-based databases, DNA banks, public molecular libraries, registration and publication of clinical trial data, and the movement toward more open scientific publication generally.

1.3.4
One other pressure for data sharing should be mentioned: the concerns of members of the public who have participated in cohorts or clinical trials, donated specimens, consented to analysis of their medical records, or otherwise contributed to data generation. Most such volunteers naturally hope that constructive, rapid progress will be made using the data or materials, which usually implies wide access.

4 'Final NIH Statement on Sharing Research Data'. US National Institutes of Health; 2003. http://grants.nih.gov/grants/guide/notice-files/NOT-OD-03-032.html [accessed 21 December 2005].

5 Press release. US National Institutes of Health; 1 March 2002. http://grants.nih.gov/grants/guide/notice-files/NOT-OD-02-035.html [accessed 21 December 2005].

6 'Sharing Data from Large-scale Biological Research Projects: A system of tripartite responsibility'. London: Wellcome Trust; 2003. www.wellcome.ac.uk/assets/wtd003207.pdf [accessed 21 December 2005].

2. The collections, and terminology

2.1 Sketches of some collections

The reason access is so important becomes clear when one appraises the diversity and richness of the resources that potentially can be tapped. The following are some of the major collections supported in the UK by the MRC and/or the Wellcome Trust, most of which are referred to in this report, with a few others added to fill out the picture.

1946 Birth Cohort (National Survey of Health and Development, NSHD).[7] An ongoing study of the health and development of people born during one week in March 1946. Supported by the MRC. Conducted by the Department of Epidemiology and Public Health of University College London, with Professor Michael Wadsworth leading. One of the longest-running cohorts, it has been studied in great depth and through many cycles of data collection. Access has mainly been provided via direct collaborations.

Aberdeen 'Children of the 1950s' Cohort.[8] A historical cohort revitalised. The cohort comprises 12 000 people born in Aberdeen during 1950–56 who took part in a school learning disability study in 1962, and about whom birthweight, gestational age, height, weight, tests of cognition, socioeconomic indicators and other data were collected. Starting in 1998 the majority of the participants have been traced, new data have been collected, and linkages have been made to data in the Scottish Morbidity Records system. Now the resource is being used for studies of lifecourse and trans-generational influences on health.

1958 Birth Cohort (National Child Development Study, NCDS).[9] An ongoing study of the physical, educational, social, and economic development of people living in the UK who were born during one week in March 1958. Developed by several social-science organisations, it is now run by the Centre for Longitudinal Studies of the Institute of Education at the University of London. Funded mainly by the Economic and Social Research Council. A biomedical survey of the cohort, funded by the MRC, was completed recently. The Wellcome Trust has funded development of immortalised cell lines from the samples. Many leading researchers are involved. One of the biomedical access uses of the data set is case-controlling some external studies.

1970 British Cohort Study.[10] An ongoing study of people born during one week in April 1970. Although it addresses such health issues as childbearing, principally it is a social science study of such matters as literacy and employment. Since 1998 it has been housed at the Institute of Education at the University of London. Dr Jane Elliott is the Principal Investigator. Neither the MRC nor the Wellcome Trust seems to have been involved, and the study does not have a biomedical component like that of the 1958 Birth Cohort. It is mentioned here because it can be seen as part of the (virtual) series of UK birth cohorts.

ALSPAC (Avon Longitudinal Study of Parents and Children).[11] Also known as the 'Avon Children of the 90s Study'. An ongoing study of the factors that affect child health and development, based on 14 500 pregnancies in the Bristol area enrolled during 1991–92. Hosted by the University of Bristol. Funded mainly by the MRC and the Wellcome Trust but also underwritten by the University. Upon the retirement in summer 2005 of Professor Jean Golding, who served as the Executive Director from the project's inception, the directorship was assumed by Professor George Davey Smith. Having generated hundreds of millions of data points and holding half a million samples, from placentas to milk teeth, the ALSPAC trove has been the subject of many collaborative studies.

Millennium Cohort Study.[12] A cohort of more than 18 000 children born in the four countries of the UK during 2000–01 whose starts in life have been thoroughly documented. Funded by the Economic and Social Research Council and a consortium of Government departments. Hosted by the Centre for Longitudinal Studies (CLS) of the Institute of Education

7 www.nshd.mrc.ac.uk.

8 See Batty GD et al. The Aberdeen Children of the 1950s cohort study: background, methods and follow-up information on a new resource for the study of life course and intergenerational influences on health. Paediatr Perinat Epidemiol 2004;18:221–39. Article posted at www.epi.bris.ac.uk/staff/gdaveysmith/pdf/P427%20The%20Aberdeen%20Children.pdf [accessed 21 December 2005].

9 See www.clseros.ioe.ac.uk.

10 See www.cls.ioe.ac.uk.

11 www.alspac.bris.ac.uk.

12 See www.cls.ioe.ac.uk.

at the University of London. Professor Heather Joshi is the Project Director. Many collaborators are involved, studying many questions.

Southampton Women's Survey.[13] An ongoing study of several thousand pregnancies, examining effects of a variety of preconceptual and prenatal factors on the later health of offspring. In some ways complementary to ALSPAC. Conducted by the MRC Epidemiology Resource Centre in Southampton, with Dr Hazel Inskip leading.

Hertfordshire Cohort Study.[14] An ongoing study of 3000 people born during 1931–39 who are still living in Hertfordshire, examining associations between genomic and early environmental factors and later development of chronic diseases. Conducted by the MRC Epidemiology Resource Centre in Southampton, collaborating with many external researchers.

Arthritis Research Campaign (ARC) Epidemiology Unit collections.[15] A part of Manchester University, headed by Professor Alan Silman, the Unit is principally funded by the ARC but it is also supported by the MRC, the Wellcome Trust and others. The Unit has amassed rich collections of data and materials relating to rheumatic and musculoskeletal disorders, which have been the focus of many cooperative studies.

Whitehall and Whitehall II.[16] Starting in 1967, the Whitehall project examined inequalities in health in relation to the 'social gradient' of men in the civil service, resulting in the landmark suggestion that psychosocial factors might be as important as dietary and other deprivation factors in explaining the inequalities. Since 1985, Whitehall II, also known as the 'Work, Stress and Health' study, has extended and broadened these investigations. Conducted with very diverse funding. Led by Professor Sir Michael Marmot of the Department of Epidemiology and Public Health of University College London.

Caerphilly Prospective Study.[17] During 1979–99 the MRC Epidemiology Unit South Wales conducted a study of cardiovascular and other aspects of several thousand Welshmen's health. After the closing of that Unit in 1999 the data and samples have been managed as an active legacy collection under the custodianship of the Department of Social Medicine at the University of Bristol. A steering group makes decisions on access.

EPIC (European Prospective Investigation of Cancer) – Norfolk.[18] An arm of a project across ten European countries analysing associations between lifestyle (especially diet), genetic factors, and cancer and other chronic diseases. The UK cohort now comprises 24 000 people living around Norwich. Affiliated with Addenbrooke's Hospital and the University of Cambridge.

MRC DNA collections.[19] These 13 collections, located at various universities, are disease-specific, focusing on: coronary artery disease, late-onset Alzheimer's disease, asthma and eczema, breast cancer, colorectal cancer, type 2 diabetes, glomerulonephritis, hypertension, acute leukaemia, age-related macular degeneration, multiple sclerosis, Parkinson's disease and unipolar depression.

MRC DNA Bank.[20] Stores samples from the above and other collections. Comprises the Centre for Integrated Genomic Medical Research in Manchester and the European Collection of Cell Cultures in Porton Down.

UK Biobank.[21] Being developed as a large multipurpose resource with the provision of wide access as its very mode of operation. Starting in 2006 it will recruit 500 000 volunteers in the age range 40–69 and administer questionnaires, collect blood and other materials, and link to the participants' NHS records and other databases. Data collection will involve many universities and healthcare institutions.

13 www.swsurvey.soton.ac.uk.

14 See www.mrc.soton.ac.uk.

15 See www.medicine.manchester.ac.uk/arc/.

16 www.ucl.ac.uk/whitehallII.

17 www.epi.bris.ac.uk/caerphilly/caerphilly.htm.

18 www.srl.cam.ac.uk/epic.

19 www.postgenomeconsortium.com/cigmr.

20 See www.medicine.manchester.ac.uk/cigmr/ and www.ecacc.org.uk.

21 www.ukbiobank.ac.uk.

Funded by the MRC, the Wellcome Trust, the Department of Health, and the Scottish Executive. The coordinating centre is hosted by the University of Manchester. UK Biobank hopes to follow the cohort for several decades.

Generation Scotland.[22] An initiative to investigate the factors and treatments of major diseases by studying thousands of family members (15 000 in the first phase and a further 35 000 in the second), including a selection of families suffering from target chronic illnesses. Supported by the Scottish Higher Education Funding Council, the Scottish Executive, and other funders. Involves all four Scottish medical schools (Aberdeen, Dundee, Edinburgh, Glasgow) and many other institutions. Pharmacogenetics is among its interests. Genetic extraction sample management will be handled by the Genetics Core Laboratory at the Wellcome Trust Clinical Research Facility in Edinburgh. Meant to be highly collaborative.

Other collections. There are thousands of other research collections in the UK alone, organised with respect to particular regions, age ranges, gender, exposures, diseases, social status or experiences, health services, or clinical trials or other medical interventions.[23,24] There are record-linkage resources that serve as virtual collections for research. And there are countless pathology and other tissue banks that can be considered for research. The MRC and the Wellcome Trust are, or have been, involved with many of these, one way or another.

2.2 Terminology in this report

'Access' means being allowed to use data or materials. Depending on circumstances such use may include seeing, duplicating, manipulating, analysing, translating, encrypting, linking, forwarding, storing, archiving or destroying. Access may be to a data set or some portion of it, to materials, or to research participants or subjects.

'Secondary access' will be used in a few places to make clear that what is meant is access by researchers who are not the primary custodians. But 'access' in this report always means secondary access.

'Data subjects' are the people to whom personal data pertain, as is defined in the UK 1998 Data Protection Act.

'Anonymisation' means conversion of personal data to a form not readily associable with a person. Because identifiability runs a spectrum, anonymisation is relative.

'Materials' mainly refers to human tissues or to DNA or other molecules derived from tissues, but it may also refer to urine or other collected substances.

'Resources' is occasionally used to refer to collections, more often to those meant to be widely accessible. Obviously the distinction is not strict, as all collections are resources.

'Documentation' in this arena is whatever detailed description of selection, measurements, validation, coding, programmes, sample handling and so on – including the changes over time – that a stranger to the data set would have to know in order to conduct a respectable independent analysis.

'Principal investigators' are senior scientists who are responsible for the actions of a research team.

'Custodians' are the persons – normally one or more principal investigators – who hold formal responsibility for the protection and use of collections.

'Stewards' occasionally is used in place of 'custodians'. The author prefers its more active tone, implying the protecting, nurturing, and growing of holdings, and the judicious providing of access to them. (Parallels include forest stewards, seed bank stewards, stewards of churches or synagogues, and stewards of rare manuscripts or other cultural treasures.)

22 www.generationscotland.org.

23 Directory of Clinical Databases, www.lshtm.ac.uk/docdat.

24 UK Data Archive, www.data-archive.ac.uk/findingdata/majorstudies.asp.

3. Basics of data and access

3.1 Access

3.1.1
As was said above, 'access' means being allowed to use data or materials, which depending on circumstances may include seeing, duplicating, manipulating, analysing, translating, encrypting, linking, forwarding, storing, archiving or destroying.

3.1.2
In general access is achieved by:

- being sent copies of data or samples of materials
- downloading data online
- visiting a centre in person and using data on site.

3.1.3
Access-in-effect, although perhaps stretching the term 'access', sometimes is provided not by data custodians' allowing access to data or materials directly (usually because of ethical or legal reservations) but by their performing particular analyses upon request and providing the results.

3.1.4
Sometimes access involves contact or communication with study participants, material donors or other data subjects. For ethical reasons the stewards of collections usually must resist allowing external researchers such access, but they may offer to communicate as go-betweens, such as by asking the participants whether they agree to be contacted by the external researchers or by the custodians for collection of additional data or possible recruitment to a study. (Incidentally, some cohort programmes report that they are experiencing higher demand for access to participants than to data or materials.)

3.1.5
Access can be sought to:

- study the collection itself
- use the collection as a platform upon which to base other data collection and analysis
- link the data with other data to establish a larger resource
- scan the collection to identify potential recruits for other studies
- select data from the collection as case controls for other studies
- pool the data with other data for meta-analysis.

3.1.6
A remark on the terms, 'access' and 'transfer'. 'Data access' amounts to 'data transfer' under the UK Data Protection Act, and 'material transfer' may in practice refer simply to 'access to material' involving no transport, as when samples are stored nearby, perhaps even in the same building. But this causes little confusion; at issue is whether a research activity is provided with some data or materials.

3.2 The life stages of data

3.2.1
Data sets go through several stages, and this has implications as regards access. Data begin as **raw data,** data as initially measured and recorded. These are transformed into **cleaned data** by being quality-controlled and having redundancies removed, evident errors of transcription or coding corrected, and so on, and their filing and indexing improved. As they are studied they become **augmented data** by incorporating derivative or 'built' data, i.e. inferences drawn from multiple initial data (such as the date of onset of illness, established by reviewing clinical measurements along with interview data), or by receiving new data from studies based on the resource, or by having data from the analysis of materials added to them. As **mature data,** which means different things for different data, data sets are held in databases, stored, or archived. Anonymised versions may be prepared.

3.2.2
Collections evolve in other ways over time. Paper files may be translated into electronic format, and earlier electronic data systems may be upgraded to more modern ones. Newer coding systems may be adopted. Phenotypic health or social data collections may have genotypic data spliced onto or linked with

them. Data may be linked with materials. Once-dynamic collections may reach the fullness of age, or lose support and become legacy collections, for which decisions have to be made about caretaking or destruction. At any time custodianship or financial sponsorship may change.

3.2.3
Whether in any instance secondary access is appropriate depends on, among other things, how mature the data set is, how thoroughly it is documented, and how searchable it is. Raw data are rarely of use to people outside the unit collecting them. Although the original collectors may know how the subjects were selected and why, what questionnaires were used, what measurement techniques were employed, how samples were treated, what codes were used, and so on, they may not have taken the (considerable) trouble to write all this up in a way that would allow a competent but 'cold' accessor to analyse the data properly. Moreover, because the current custodians may not have been involved in earlier stages, even they may be unaware of all the soft areas and pitfalls. Data sets cannot be used effectively by secondary researchers – or by primary researchers either, for that matter – unless the collection and its variables are thoroughly documented. Access decisions must take all of this into account.

3.2.4
Upon any transferring of data to third-party archives or depositing of samples in storage centres beyond the direct custodianship reach of the resource builders, a variety of safeguarding, documentation, and access issues must be attended to.

3.3 Degrees of accessibility

3.3.1
Collections range from ones that do not lend themselves to wider access, to ones that were designed from the start to be resources for multiple uses and users. In between are many collections that are oriented to the research interests of their custodians but that can also be used by others under controlled conditions.

3.3.2
Applicants for access may be close colleagues of the custodians, perhaps even members of the same institution. Or they may be external researchers known to and respected by the custodians. Or they may be researchers far away who have no relationship with the custodians whatsoever. The relationships may change over time, of course, as the parties interact.

3.3.3
Access may amount to a truly cooperative effort – symbiotic collaboration – in which researchers who are not members of the custodial team bring complementary resources or capabilities to a joint project, and the custodians make interpretive or other intellectual contributions beyond merely supplying data or materials. At the other extreme, access may be provided for research completely independent of the data custodians. In between are a variety of cooperative situations. Depending on the circumstances, collaboration can be viewed, by access providers or access requesters, either as an advantage or as a burden.

3.3.4
So in addition to the data life-stage considerations, whether secondary access is appropriate also depends on the preparedness of the custodians to provide access and the extent to which access would involve collaboration.

3.3.5
Industrial or other commercial applications for access tend to be treated specially because of intellectual property or public image concerns.

3.4 ‘Public data’ and ‘open access’

A caution about these terms. Sometimes data are referred to as ‘public data’ when what is meant is data collected by government agencies or by programmes sponsored by governments or public charities; the trouble is that the data may not actually reside in the public domain (in the sense that the catalogue of the British Library is in the public domain) but are sequestered and accessible only conditionally. And sometimes true ‘open access’ obtains, as it does with genome sequences posted on the world wide web, but more often what is meant is that the collection is open to applications for access. Care should be taken with both expressions, as too-casual use can mislead the public.

4. Conditions of access

This section succinctly discusses access agreements and their terms, which allows the telegraphing of many issues. The more pivotal of these are then discussed in later sections.

4.1 Agreements

4.1.1
Access to data that are not in the public domain is negotiated via agreements, contractual undertakings specifying terms and referring to policies and laws.

4.1.2
On the providing side, agreements may be executed by funders, principal investigators or other custodians, supervisory bodies, universities or other institutions that host collections, or a combination of these parties.

4.1.3
On the receiving side, agreements are executed by principal investigators or other leaders, or by universities or other institutions, who assume responsibility for ensuring that the conditions are met and are complied with, including by staff and students.

4.1.4
Conformance with laws, regulations, and governance is not negotiable, but agreements may note some specifics or timing with regard to NHS research ethics review or compliance with NHS Research Governance, MRC guidance, the Human Tissue Act, the Data Protection Act or other laws, or the ethics guidance of international organisations such as the Council for Organizations of Medical Sciences (CIOMS) or the Human Genome Organisation (HUGO).

4.1.5
Agreements may refer to the decisional authority of oversight bodies or other governance mechanisms.

4.2 Terms of agreements

4.2.1
Access agreements differ with circumstances, and surprisingly little standardisation is evident. Some arrangements are more formal than others, but all amount to contracts. Probably no agreement covers all of the following points, but set out here as a menu are the more common elements of data access and material transfer agreements, which might be considered when policies or agreements are being drafted or revised.

4.2.2
Confirmation of professional competence. Applicants may be asked to provide evidence of experience with database or genomic research, or of having published articles on the health topic of concern. Such a provision is meant to protect the reputation of the resource from incompetent analysis, avoid wasting the efforts of resource stewards, and serve some of the other concerns that follow below.

An example of a set of qualifying criteria is this one, used by the Juvenile Diabetes Research Foundation/ Wellcome Trust Diabetes and Inflammation Laboratory (JDRF/WT DIL) in Cambridge:

> 'Bona fide researcher' may be defined one of two ways. For access to data held solely under the control of the JDRF/WT DIL, it means a person who has authored a relevant peer-reviewed article that we can locate on PubMed, and who is still working in the field. Where data on JDRF/WT DIL data subjects is held as part of a wider consortium, the decision as to who is or is not a bona fide researcher is the primary responsibility of the consortium data access committee.[25]

The obligation of custodians to evaluate and confirm the competence of data or sample requesters is a matter of debate.

(See discussion of this and other scientific concerns in section 6.)

4.2.3
Screening of scientific merit and relevance. Some collections, either in the access agreement or by their tradition, reserve the right to screen the scientific compellingness of the research proposal (such as its creativity, analytic power or relevance), and some insist on the right to vet resulting publications in draft. Others believe that such screening is simply not within the purview of the custodians.

25 'Human Genetic Data Access Agreement (version 2)'. Juvenile Diabetes Research Foundation/Wellcome Trust Diabetes and Inflammation Laboratory; 2005. www.gene.cimr.cam.ac.uk/todd/access-agreement.html [accessed 21 December 2005].

4.2.4
Specification of what is to be provided. Data access agreements usually specify how and when the data – and possibly also associated data codes, computer programmes, and general documentation of the data set – will be provided. Reference may be made to ancillary technical and IT specifications.

Material transfer agreements usually address how and when the materials will be delivered and how they will be preserved in transit. Reference may be made to protocols on handling and storage.

4.2.5
Consent. This is a core consideration whenever personally identifiable data or materials are involved and always must be addressed in the agreement. It may refer to the original consent, or it may refer to some subsequent or special consent or to a waiver of consent requirements.

(See discussion of consent in section 5.)

4.2.6
Purpose limitation. This relates to consent, and when included it most often specifies that only certain health conditions or diseases can be studied. Purpose limitation is a traditional constraint, embodied in data protection and other laws and guidance, and it can be a concern of data subjects and specimen donors.

A purpose question that occasionally must be addressed is whether data or materials can be used for case controls vis-à-vis other data. Another is whether a collection can be used to identify and contact potential subjects for clinical trials or other projects.

In many agreements the pursuit of commercial purposes, or research that is recognised as having potential to lead to profit, is disallowed unless explicitly granted and an intellectual property agreement is signed.

Occasionally agreements say something about the use of data or materials for local professional educational purposes, as a sideline of research use.

(One can question whether narrow purpose-limitation should be encouraged in this era of research on multifactorial conditions. Also, it should be remarked that purpose limitation can be difficult to audit or enforce once data or materials have been transferred.)

4.2.7
Confidentiality. Solemn reminders may be made as to common law obligations of medical confidentiality or requirements of the Data Protection Act, the Human Tissue Act, MRC guidelines, or other laws or guidance.

Aspects of anonymisation or disclosure protection may be discussed; sometimes these are technically complex. If anonymisation is reversible – 'linked anonymised' in MRC terminology – issues must be addressed as to how the identifiers will be held, who will have authority to use the key to re-identify, and the criteria and procedures governing re-identification.

Usually agreements specify that data recipients will make no attempt to re-identify, trace or contact the subjects or their relations, or to link to databases containing identifiable data, or to use the received data or materials in any way that could infringe the rights of the data subjects or otherwise affect them adversely.

Confidentiality restrictions may also apply to data about healthcare providers or institutions, other researchers, or relatives of the data subjects.

(See discussion of confidentiality in section 5.)

4.2.8
Research ethics approval. Usually, NHS Research Ethics Committee or equivalent ethics approval is a condition of access. The access agreement may specify the stage in the application process by which ethics approval must be secured.

4.2.9
Limiting onward transfer. Almost always the recipients must promise not to pass the data or materials on to unauthorised parties. An authorised party might be, for instance, a researcher in another institution who has signed a similar agreement.

4.2.10
Linking. Conditions may be imposed on the linking of the provided data with other data or with materials. Agreements increasingly are having to address the linking of health or other data to human materials that the applicants already hold, or vice versa.

4.2.11
Recontacting the data subjects or material donors. Usually recontacting without intermediate steps of consent or representation is forbidden. Some agreements mention that contact might be facilitated via the custodians.

4.2.12
Maintaining the quality of the resource. Usually it is required that any errors or degradation of data, materials, coding, programs, methods or documentation that become apparent must be notified to the original custodians. And it may be required that upon notice from the data providers, erroneous or outdated data must be destroyed.

4.2.13
Publication. Publication of results in the open scientific literature, posting them on a website or depositing them in an archive, is usually required. Coding formats and documentation may be specified, as may limits on delay before submitting manuscripts for publication.

To reduce publication bias of association studies and avoid wasteful duplication by other researchers, some agreements require that if the results of such studies are not published in the literature they must be offered for posting on the collection's or another website.

4.2.14
Acknowledgements. Always the contributions of the resource and its curators and funders must be acknowledged in publications. Usually this is to be expressed in standard acknowledgement notes at the beginning or end of articles. Wording may be specified by the agreement. In some instances database copyright also must be acknowledged.

4.2.15
Co-authoring with the stewards of the resource. The criteria for determining co-authorship of individual scientists or the custodian group vary with collections. A group example is that articles based on research using the ALSPAC resource must list as the last author 'The ALSPAC Team'. A perennial issue everywhere is what co-authorship implies and whether it amounts to more than just pro forma credit.

4.2.16
Enriching the resource. Agreements may require returning findings to the resource, posting them on a public database, or depositing them in an archive. Methods and timing may be specified, as may who bears what costs. This kind of feeding-back enrichment is becoming customary.

4.2.17
Archiving. If serious, accessible archiving is expected from the start, agreements usually refer to separate technical specifications for the submission of data and documentation, and address the attendant issues of cost-bearing and intellectual property.

4.2.18
Assigning or waiving of intellectual property (IP) rights. IP rights may be asserted over access to data or materials in the first place, or to subsequent sale of data or samples, or to publications, patents, copyrights or royalties resulting from the access.

Basic access agreements usually refer to separate detailed legal IP agreements, to which on the data-providing side the funders, the Department of Health or a hosting university, for example, may be parties. Or the agreement may simply declare that the data providers retain no IP rights.

In some instances exclusivity of access to some data or materials is granted, such as for periods while articles or patent applications are prepared. Or a principle of non-exclusive access may be stated.

(See discussion of IP in section 7.)

4.2.19
Responding if consent is withdrawn. It may be required that if a research participant or data subject withdraws consent, or if consent is for any other reason invalidated, holders of data or materials, including secondary users, must destroy the data or materials and/or sever links and certify the actions to the custodians.

4.2.20
Prioritisation of access to limited resources. The agreement may state that access to depletable samples, analytic services or other limited resources is subject to prioritisation as determined, for example, by precedence of application, comparative scientific merit against other applications, or lottery. Mention may be made of a committee that makes or oversees priority decisions.

4.2.21
Access fees or royalties. Fees may be assessed to cover the costs of preparing and transferring data, materials and so on, and possibly to help offset infrastructure costs as well. For commercial users the rate may reflect the prospects of profit.

4.2.22
Returning or destroying materials. This may be required at the end of the project, or in the event of noncompliance with the terms of the access agreement.

4.2.23
Transborder enforcement. If data or materials are being provided to recipients outside of local national legal jurisdiction, then special ethics review, transborder data protection and contract enforcement undertakings may be included.

4.2.24
Termination. A clause may state what will happen if the primary custodian responsibility has to be ceded, such as if the unit closes. Often this stipulates that curatorship will pass to a similar charitable research unit or the funders.

4.2.25
Disclaimers. Standard legal disclaimers of responsibility for errors or inaccuracies, or for consequences of use of the provided data or materials, invariably are included.

4.3 Opportunities for guidance

4.3.1
Even a cursory survey of various programmes' access and material transfer agreement documents reveals that there is little consistency, even within programmes. Many projects seem not even to have drafted access policies or agreements, or at least to have developed final template versions – final in the sense of being approved by their university's lawyers, for instance. And the collections use different material transfer agreements.

4.3.2
Boring though it can be, the exercise of drafting access policies and agreements does focus attention on the terms that must be negotiated, and it induces discussion of the issues within the custodian team and its institution. Agreements make positions clear and enforceable. And research governance requires that this be attended to.

4.3.3
Almost certainly the funders and their advisory groups would find it useful to develop guidance on the core terms of access and material transfer agreements.

5. Consent and confidentiality concerns

Almost all biomedical data and materials to which researchers want access are originally collected in confidence, perhaps medical confidence, and with the person's consent to the collecting and possibly to current or future use in research. Secondary access to collections for research must therefore deal carefully with consent and confidentiality, and with the instrument of confidentiality, anonymisation.[26]

Aspects of consent and confidentiality are currently being examined by the MRC, the Academy of Medical Sciences,[27] the Department of Health and other bodies, so this report will not deal with them in detail. But because the issues are central to access decisions it must describe some considerations.

5.1 Consent

5.1.1
Most collections like those that are the subject of this report cover the provision of secondary access in the original consent. For access, either the original terms of consent must be respected, or special consent to the access and use must be sought by the custodians.

5.1.2
Even if the terms of consent are not rehearsed in detail in access agreements, the implications, such as those regarding confidentiality, purpose limitation or intellectual property, invariably are, and a chain of consent is maintained by such restrictions as limiting onward transfer of the data or materials. The record of consent must accompany data sets or materials when they are archived.

5.1.3
The general view seems to be that when providing access, most custodians and oversight committees of the kinds of collection that are the subject of this report extend the chain of consent responsibly.

5.1.4
A contingency that some collections, and their accessors, have to prepare for is responding down the way if data subjects or participants annul some aspect of consent, which may require severing linkages or destroying data or materials.

5.2 Confidentiality and anonymisation

5.2.1
It is standard practice to protect confidential data by such safeguards as maintaining physical security, training personnel, protecting computer passwords, recording use-logs and holding sensitive data in non-networked computers.

5.2.2
Data providers must assure themselves, at least via the access agreements, perhaps making reference to NHS or ISO (International Standards Organisation) security standards, that accessors are committed to protecting the data or materials. Rarely, though, it must be remarked, is any auditing conducted to verify that such commitments are kept.

5.2.3
The most common safeguard when providing access is anonymising the data or materials, i.e. reducing the risk of deductive re-identification by: stripping off overt and indirect identifiers; encrypting, coarsening or aggregating the data; suppressing extreme, easily re-identifiable cases; and possibly injecting statistical noise.[28]

5.2.4
For linked anonymised data, a practical question always is: who should hold and be designated to use the key for re-linking? In many instances the custodian principal investigator holds the key, as is, for example, required by MRC guidance on human materials.[29] The related question is: what are the conditions on use of the key?

26 The present author addressed many aspects of these issues in a 2002 report to the Nuffield Trust, 'Learning from Experience: Privacy and the Secondary Use of Data in Health Research'. www.nuffieldtrust.org.uk/publications/detail.asp?id=0&PRid=45 [accessed 21 December 2005].

27 See 'Personal Data for Public Good: Using health information in medical research'. Academy of Medical Sciences; January 2006. www.acmedsci.ac.uk/p47.html [accessed 30 January 2006].

28 A technical survey is Domingo-Ferrer J (ed.). Inference Control in Statistical Databases. Berlin: Springer-Verlag; 2002.

29 'Human Tissue and Biological Samples for Use in Research: Operational and ethical guidelines'. Medical Research Council; 2001. Section 9.4. www.mrc.ac.uk/pdf-tissue_guide_fin.pdf [accessed 21 December 2005].
Also 'Clarification Following Passage of the Human Tissue Act 2004'. Medical Research Council; 2005. Section 3.6. www.mrc.ac.uk/pdf-ethics_guide_human_tissue_clarification_april_2005.pdf [accessed 21 December 2005].

5.2.5
Several cohorts carry intrinsically elevated identifiability risks. For example, the people in the Hertfordshire Cohort Study and in EPIC-Norfolk reside in those regions, and those in the 1946 and 1958 birth cohorts are publicly known to have been born during a particular week in March. Such factors have to be taken into account when providing access.

5.2.6
To mention one example of precautions taken by data providers, the confidentiality agreement of the NSHD (1946 Birth Cohort) says:

> 2. We must check the draft of any publication or presentation to ensure that members cannot be identified. ...
>
> 5. When the research is complete the data in all forms (including computer files), copies and extracts must be destroyed. We keep a copy of the data provided to you (and documentation provided). We can provide it again should you wish to resume research on the data.[30]

5.2.7
In the extreme, custodians can perform analyses upon request, as the Office of National Statistics (ONS) does with some highly detailed personal data known as 'microdata'. UK Biobank is considering providing analyses of materials, or contracting with specialist laboratories to provide them, so as not to have to release the materials from its direct custody.

5.3 Safe settings and data enclaves

5.3.1
A strategy that is gaining appeal is the providing of access within safe settings or data enclaves (which may be distributed but secure networks), situations tightly supervised by the custodians. For example, ALSPAC invites researchers to come to Bristol and analyse data in a stand-alone computer. The NSHD allows examination of unusually sensitive data only on the Survey's premises. The Centre for Longitudinal Studies operates a similar safe setting. The ONS provides access to some fine-grain individual-level data via a 'virtual microdata laboratory', in which researchers work from a secure server in London to access data held at ONS sites around the UK.

5.3.2
A question all such approaches must face is whether they are striking the right balance between access and exclusivity. Other questions are how the visitor's analytic manipulations are recorded, whether a take-away summary is provided to the researcher, and how other researchers later can verify a locked-up analysis.

5.4 Limited data sets

5.4.1
Another useful approach is the preparation of data sets in which identifiability has been reduced sufficiently that they can be more widely released. Considerable experience with limited data sets has been accrued by the ONS.

5.4.2
Anyone thinking of deriving a limited data set may find it useful to review the experience being gained in the US under a recently implemented Privacy Rule governing data collected in the provision or payment of healthcare.[31] The Rule specifies the kinds of potentially identifying data that must be stripped from data by their custodians to qualify them under Federal law as limited data sets for research.

Under the Rule a limited data set must exclude – for the patients and for their relatives, employers and household members – specified identifiers (addresses, telephone and telefax numbers, medical record numbers, dates, vehicle licence plate numbers, fingerprints and so on, down through 16 categories). The data may be used for research only under an access agreement in which, among other things, the recipient specifies the uses, names the individuals who will be using the data, commits to enforcing safeguards, and states that he will not identify the data subjects or attempt to contact them. The problem of course is that some of the data removed to reduce identifiability, such as detailed geo-locators, may be the very data needed for some kinds of research.

30 Personal communication, Professor Michael Wadsworth, NSHD Study Director.

31 An overview is 'Protecting Personal Health Information in Research: Understanding the HIPAA Privacy Rule'. US National Institutes of Health; 2003. http://privacyruleandresearch.nih.gov/pr_02.asp [accessed 21 December 2005].

5.4.3
Some researchers believe it would be helpful if the UK had guidance (from whom?) on the characteristics and access conditions of limited data sets.

5.5 Need for resolution and guidance

5.5.1
Few issue clusters are identified by researchers as so urgently needing resolution as those surrounding confidentiality and anonymisation. They are not unique to data sharing activities, but they are central to them.

5.5.2
The difficulty starts with the interpretation of several ethico-legal fundamentals. For instance, one line of interpretation, for short the 'risk-management view', argues that if data are anonymised they are not 'personal' and so should be widely (though still carefully) usable, especially for such a public-good purpose as health research. Another and conflicting line of interpretation, for short the 'trailing-rights view', holds that even if data are thoroughly anonymised, the wishes of people regarding use of data about themselves must be respected. In the UK at present both views have some support, guidelines are somewhat ambiguous, and the implications for data sharing practice are far from clear.

5.5.3
Whatever other bodies may advise, biomedical researchers look to the MRC for guidance on these matters. Access will continue to be hampered by uncertainty until more definitive guidance is provided. Among the issues most strongly affecting access are:

- Given that identifiability runs a spectrum, what constitutes full, or at least an acceptable degree of, anonymisation?
- If data are anonymised, are they still considered 'personal data' under the Data Protection Act?
- Can linked anonymised data be considered to be *functionally anonymised to a researcher* if the researcher has no access to the key for re-identifying, safeguards the data, and promises not to attempt to re-identify? What dispensation does this gain the researcher? And who can hold the key, and how?

6. Scientific concerns

A number of access issues having to do with scientific openness, integrity and quality are controversial. Mostly they devolve to roles and obligations.

6.1 Stance and preliminaries

6.1.1
Requests for access surely have the best chance of being approved if they are based on a proposal that has passed scientific peer review and research ethics review and been funded. But before committing to doing all that work, researchers need at least tentative assurance that the custodians will grant access.

6.1.2
Many researchers express frustration at the obscurity that they perceive surrounds some collections, and urge that collections publish clear access policies and their access- and material-transfer agreements (or templates), and offer to hold preliminary discussions with interested researchers. Some collections do these things; some do not. Probably some are simply misunderstood as being more stand-offish than they really are.

6.1.3
Early negotiations between access requesters and custodians must take account of how collaborative the project will be and what this will imply. Forced data sharing can put pressure on project teams to collect data or materials, assist with data analysis, or do other work that they have not intended to do. Conversely, an insistence by custodians on keeping research 'inside' can in effect coerce applicants to collaborate.

6.2 Screening of applicants and proposals

6.2.1
What is the obligation, or indeed the right, of a custodian principal investigator or oversight committee to judge the competence of requesters or the scientific robustness or relevance of applications? Must requesters commit at the outset to investigating a specific hypothesis, or can they trawl through the data informally or do opportunistic genotyping, looking for associations and leads? Whose research agenda is it? Relevance with respect to what?

6.2.2
Regarding competence, many people think that custodians need not go further than confirming that applicants appear to be legitimate – as might be indicated by their curricula vitae, publication records and current appointments, for instance – and make the requisite legal undertakings on such core matters as confidentiality, consent, limitation of use and further transfer of data or materials, and intellectual property. Others believe that custodians must take a hard look at indicators of competence, indeed that this is required by Research Governance.

6.2.3
Regarding research protocols, one can ask: do principal investigators or oversight committees have the competence to review the scientific details of disparate proposals? If a proposal has passed peer review for funding, does that not suffice? Can the custodians manage the subtle or not-so-subtle conflicting interests that may be involved? (Among the responses are that reviewer expertise can be tapped ad hoc, that funding review may not address all of the complexities, and that oversight and other mechanisms buffer conflicting interests.)

6.2.4
More broadly one can ask, as from time to time custodians themselves do: is it really the duty of custodian principal investigators to head off or bat down inadequate science? Why can the greater intellectual marketplace not be relied on to sort out scientific quality and usefulness? As a general matter, in the light of experience, are apprehensions about projects 'being brought into disrepute' justified? Must resource stewards feel an obligation to be result stewards?

6.2.5
There can be no question about the difficulty of analysing the kinds of collection this report is discussing. Many cover thousands of variables –

the 1946 Birth Cohort covers some 13 000 – each with its peculiarities. Countless matters of sampling, validity of measurements and questions, confounding, and so on have to be dealt with in statistical detail.

6.2.6
The rationales for careful scientific screening and control by collection stewards include:

- making sure that consent is respected and disclosure risks are minimised
- protecting the resource and its subjects/participants from uses that might reflect negatively on the project or discourage participation
- carrying out the 'guarantor' role as regards the quality of the resource (including new data returned to the resource by accessors?)
- generally enforcing scientific quality, a community duty of all scientists
- avoiding having to undertake countering analyses, write opposing articles or even meet with the press in order to deal with spurious conclusions
- avoiding the wasting of effort, expense and materials.

6.2.7
A variety of filtering precautions can be taken. Before providing genotype data several custodians ask for evidence that an experienced statistician will be involved. At least one collection requires that laboratories that request DNA demonstrate their capability by successfully sequencing some pilot samples. Material collections demand justification for consuming limited samples or requiring freeze-thaw cycling to take aliquots (which can degrade delicate protein components).

6.2.8
To avoid unintentional duplication, several cohort projects have a policy of telling new requesters if another project is already pursuing a similar line of investigation using the resource, and offering to put the two groups in contact.

6.2.9
Note should be taken of the Research Governance Framework's admonishment (in its section 2.3.1): "Research which duplicates other work unnecessarily, or which is not of sufficient quality to contribute something useful to existing knowledge, is unethical."[32] By their actions it is clear that the MRC and the Wellcome Trust agree.

6.3 Control after access is granted

6.3.1
Some resources require, at least as the default policy, that every study using the collection have at least one principal investigator who is a member of the custodian group. Some require that articles based on the resource be submitted to the custodians for vetting before being submitted for publication.

6.3.2
They are not unique, but ALSPAC's policies illustrate this set of issues clearly. As stewards of all cohorts do, the ALSPAC team view their database as being complicated to use and vulnerable to misuse, and so they control access and use carefully. ALSPAC predicates access on its approving the scientific purpose and quality of proposals. As it explains: "Information is obtained in confidence from the study parents and linked to results of biological assays. The study team have the responsibility to ensure that the data are only accessed by scientists with bona fide projects of high scientific probity who have promised to abide by the study rules."[33] Applications for external funding to use the database must be approved by the Director before being submitted. Projects that then become 'ALSPAC projects' must include a member of the ALSPAC Directorate as a co-principal investigator. And ALSPAC vets publications in draft.

32 'Research Governance Framework for Health and Social Care: Second edition'. Department of Health; 2005. www.dh.gov.uk/PublicationsAndStatistics/Publications/PublicationsPolicyAndGuidance/PublicationsPolicyAndGuidanceArticle/fs/en?CONTENT_ID=4108962&chk=Wde1Tv [accessed 21 December 2005].

33 'Collaboration with ALSPAC'. Avon Longitudinal Study of Parents and Children. www.alspac.bris.ac.uk/protocol/collaboration_and_funding.shtml [accessed 21 December 2005].

6.3.3
So: screening of applicants' bona fides, the scientific merit of protocols, and applications for study funding; co-principal-investigatorship; vetting of drafts; and co-authorship. These policies, which relate to notions of scientific and ethical responsibility as well as local territorial interests, may be entirely defensible. And perhaps they are especially appropriate for cohort studies.

6.3.4
A very different view is promoted by many who are pushing for more open access. An example is this clause of the NIH policy requiring data sharing by all major projects:

> When making data available, researchers cannot place limits on questions or methods nor require coauthorship as a condition for receiving data. Proper documentation is needed to ensure that others can use the dataset and to prevent misuse, misinterpretation, or confusion.[34]

No doubt underlying this is a conviction that question-asking should not be constrained, and a recognition that epidemiological methods and assumption-making come in marvellous variety.

6.3.5
Importantly, coupled with the NIH policy's reducing of restrictions is its requiring proper documentation of the data set. 'Documentation' is whatever detailed description of selection, measurements, validation, coding, programmes, sample handling, and so on – including the changes over time – that a stranger to the data set would have to know in order to conduct a respectable independent analysis. Value resides not in data, but in data situated within their full investigatory matrix.

6.4 Tending the scientific ethos

6.4.1
The above issues are fundamental and deserve continuing discussion by the scientific community and the funders. Some may be addressed as part of guidance in future, and some should be reviewed when programmes revise their access policies.

6.4.2
The new policies requiring wider data sharing, such as the NIH policy, are a test of whether independent (i.e. non-collaborative) research on documented data sets can generate as high-quality results as collaborative research does. Concomitantly they are a test of data-set documentation. All of this deserves to be followed, although it will not be easy to evaluate.

34 'Final NIH Statement on Sharing Research Data'. See footnote 4.

7. Territorial and proprietary concerns

Issues of 'ownership' – with the word set in quotation marks here because its connotations can be ambiguous – are among the most awkward and deeply felt factors of data sharing, and until they are dealt with more openly and effectively, goodwill and access will suffer.

7.1 Custodians' vested interests

7.1.1
After devoting years of hard work to planning a project, securing funding and ethics approval, negotiating all the institutional arrangements, recruiting the participants, collecting the data and materials, performing analyses, managing the collection and infrastructure, controlling technical quality, and generally nourishing the project through waves of funding and maturation, teams inevitably feel proprietary toward 'their' collection and 'their' participants and have a sense that they have earned a priority right to exploit the resource they have built.

7.1.2
At the same time custodians are very sensitive about protecting the identities and rights of people who have allowed data about themselves, or samples from their bodies, to be collected and shared in research. And beyond the ethical and legal obligations, custodians have the practical need to maintain the fidelity and long-term engagement of the participants, especially in longitudinal studies that involve repeated contact and data gathering with volunteer subjects.

7.1.3
All of these concerns together result in a strong sense of professional ownership. Naturally they also lead to an expectation of proper credit and compensation for providing (careful, effective, high-quality) access to others.

7.2 Intellectual and academic credit

The MRC data sharing policy emphasises that "data creators should receive full and appropriate recognition by funders, their academic institutions and new users for enabling the secondary research".[35] In addition to the respect accorded by fellow scientists, credit is given in the form of acknowledgement in publications and sometimes co-authorship. But principal investigators complain that the work of building collections, maintaining infrastructure and facilitating access must somehow be accorded more concrete recognition than it now receives, such as in academic promotions and Research Assessment Exercises. How this is to be accomplished is not obvious. The academic community should be invited to propose specific remedies.

7.3 Cost coverage

7.3.1
Whether one-off or continual, the provision of access entails work – skilled manipulation of data and materials, careful attention to information technology and linkages, expert advisory and review work, and administration and management. Many collections now charge access fees to recover at least the operational and associated administrative costs, and funders are willing to subsume access charges in project grants.

7.3.2
Less fully covered in some cases, apparently, are the costs of preparing data for storage or archiving, such as anonymising the data and documenting the variables, or posting on an accessible database. Some centres remark that they even have difficulty recouping the costs of securely storing older paper or digital files.

7.3.3
As collections mature and the access request load increases, the need increases for financial support to hire junior academic or technical staff to carry out some of the demanding but routine data-sharing work, relieving senior investigators to do other things.

35 'MRC Statement on Data Sharing and Preservation Policy'. See footnote 2.

7.3.4
Basic financial questions for any collection are: are the costs of providing access covered by the resource's core budget, or must they be recovered via access charges? And in either event, is the compensation adequate?

7.3.5
No doubt the funders are aware of all of these support issues. But many custodians believe that access will be fostered if the funders take fuller account of them.

7.4 Collectors' privileges

7.4.1
Can and should those who build a resource exert preferential rights to its use? The MRC data-sharing policy affirms that "a limited, defined period of exclusive use for primary research is reasonable".[36] Such a privilege is exercised by many collections de facto as they simply resist, or at least do not encourage, data sharing.

The issue is widely known to have been a central one in the development of UK Biobank, which is asking academic Regional Coordinating Centres to lead the recruitment of 500 000 participants and the collection of data and samples, in return for which many of those who will do the work have asked for inside-track access, say for the first several years or for research in their speciality areas.

7.4.2
There is no single right answer. What is crucial is that funders and those they support in building collections reach explicit agreement on this when core support is negotiated.

7.5 Interim exclusivity

7.5.1
Separately, then: is it legitimate to grant external researchers any kind of interim exclusivity, which they often ask for? It is not uncommon for collections to concede such protection, allowing specified use of specified data or materials until the users have completed their project and had reasonable time to submit the results for publication or apply for a patent, and during this time not providing the data to other researchers, or perhaps not for similar research purposes.

7.5.2
In one view, interim exclusivity is seen as an incentive to wider access and use. But in the view that sees no objection to researchers' using the same or similar data to investigate the same or similar questions, it is seen not only as a disincentive but as a violation of scientific openness. This issue deserves continuing attention, and it should be revisited in future guidance on data sharing.

7.6 Intellectual property

7.6.1
The possessive right that can be claimed for data and materials is intellectual property (IP), the right to exploit the informational potential and exclude others from exploiting it without the permission of the IP holder.

7.6.2
Generally in the UK, hosting institutions such as universities act as legal entities and serve as formal custodians of the collections they host, unless the funders retain custodianship (as they do of some large collections). In this role these legal entities are assigned and manage IP. In recent years universities have become adept at negotiating IP rights.

7.6.3
Universities delegate operational custodianship to senior academic staff who are their employees. Principal investigators themselves do not own data or materials as property; for instance, they have no sellers' rights.

36 'MRC Statement on Data Sharing and Preservation Policy'. See footnote 2.

7.6.4
Access agreements now invariably address IP rights, reflecting policies that are stated or referred to in grant awards.

7.6.5
Incidentally, data protection law never focuses on 'ownership' or on data as property, but solely on how the use of data affects persons. Thus the UK Data Protection Act has no relevance to IP.

7.6.6
Special concern attends access to collections by commercial organisations or their agents, but the funders have no fundamental objection and have policies on this. If an industrial firm applies for access to the MRC DNA collections, for example, a principal investigator may negotiate an agreement, but then must submit it to the MRC for legal review and approval.

7.6.7
Many academic scientists express confusion, if not consternation, over the meaning and practical implications of possession, custodianship, ownership, database rights and IP generally. Some do not understand their roles in between the university and the funders, and some complain that they hear differing explanations from the two sides. Some are unsure as to whether materials have different IP status from data. All of this is pertinent to the sharing of data and materials. A publication from the funders explaining these matters and the alternative legal arrangements, with examples, no doubt would be enlightening.

8. Decisions on access

8.1 Who should make what decisions?

8.1.1
The tradition in which custodian principal investigators themselves made access decisions has generally been giving way to more consultative decision-making with independent input. This has been a response to the increase in the number of interested parties, the complexity of projects as more specialities and organisations interact, clinical and research governance requirements, and the interest in broadening access.

8.1.2
Although there is no standard arrangement, probably the most common model now is decision-making by, or presided over by, fairly independent oversight bodies. Some of these groups are formally constituted, have terms of reference and hold regular meetings. Others are casual, rarely meeting but existing to be consulted from time to time by the custodian and in a position to address serious problems should any arise. Some committees are charged with strategic, planning or management roles as well as roles regarding access decisions.

8.1.3
The prevailing view seems to be that custodian principal investigators should not now be making important access decisions alone. This is for the reasons mentioned above, and to protect them and their institutions from accusations of excessive territoriality as well. For many collections, however, in effect the principal investigators still make most access decisions.

8.2 Oversight and advisory structures

8.2.1
Usually the funding organisations appoint and charter the oversight groups, for example by approving their terms of reference, in consultation with interested parties including the custodians. Usually oversight bodies have 'strong advisory' relationships with the funders.

8.2.2
The funders tend to retain ultimate control. One clear statement of authority is this MRC policy on tissue collections:

> The MRC reserves the right to specify the arrangements for management and access of collections as a condition for awarding funding. In the case of jointly funded collections, these arrangements would be negotiated with the other funders. This will allow us to ensure that collections are managed appropriately to maintain their usefulness, and to ensure that optimum use is made of them.[37]

8.2.3
Most committees include experts who are not otherwise involved with the particular resource or host institution, and an independent chair.

8.2.4
An issue is whether committees should include as members the custodian principal investigators of the collections they oversee, and whether these custodians should have voting power. Many, perhaps most, such committees do include one or more of the custodians. The custodians undoubtedly have influence, whether or not they vote. Probably formal voting power is an issue only in the rare event of an extreme controversy, and in such situations the funders would have the ultimate say.

8.2.5
The general view seems to be that the funders should remain involved with the oversight bodies, although probably not serving as voting members. The funding organisations at least maintain a watching brief. Simply having their representatives at the table, even if only with the status of participating observers, should get the funders' views heard. Probably they should retain ultimate prerogative. But while it is recognised that the funders have the right, and even the duty, to exert general direction, there is concern that they not interfere too much with the judgement of the experts and leaders they appoint to such committees. The funders are familiar with this kind of balancing.

37 'Human Tissue and Biological Samples for Use in Research: Operational and ethical guidelines'. See footnote 28.

8.2.6
For each collection, the powers of the several bodies – oversight and advisory committees, custodian principal investigators, hosting universities, funders – should be spelled out. In some situations they seem not to be.

8.3 The plethora of committees

8.3.1
Given the relatively small number of gurus (for short) in such fields as epidemiology and genomics, it is no surprise that there is a lot of overlap among data custodians, data accessors, and membership on boards of the funding organisations, programme review committees, and advisory and oversight bodies. Potentials for conflicting interests have to be managed. Also, there are just an awful lot of committees. (Given all the committee duties, one can wonder how senior scientists get around to doing their own research.)

8.3.2
The 13 MRC disease-specific DNA collections illustrate the situation. Each aims to have its own Guardianship Committee, the membership of which is supposed to be approved by the MRC. Apparently the multiplicity of committees exists because the projects developed rather independently, with several of them in operation before the current MRC programme was established, and because each collection has its own local context and character. In an effort to bridge the diversity, the MRC holds Collectors Network Meetings at which practices and problems are discussed.

In addition to the Guardianship Committees of the individual collections, a DNA Banking Steering Committee advises the MRC on the central banking of samples. One of the Steering Committee's remits is to make sure that access is well managed. As all banking involves making sure that deposits and withdrawals are transacted properly, the Committee focuses on such issues as access criteria, material-transfer agreements and risk management.

Whether this clutch of DNA committees is the optimal set-up is a matter for the MRC and the scientists involved to decide.

8.3.3
Similar multiple-committee situations exist at other programmes. ALSPAC, for example, has an Independent Steering Committee, an Executive Committee, an Ethics and Law Advisory Committee, a Genetics Advisory Committee, and several expert scientific committees.

8.4 Oversight needs

8.4.1
In future, as access broadens and the funders take a more strategic approach to large long-term programmes, as must be inevitable, surely there will be some appeal in consolidating some of the oversight duties under stronger central oversight bodies.

8.4.2
Consolidation could foster consistency, facilitate and elevate the development of policies and good practices, and enhance efficiency by reducing the advisory workload of busy scientists. Central committees could set access criteria, help establish access agreements, set priorities for use of depletable samples, and provide guidance and assistance to both custodians and accessors.

8.4.3
One hesitates to make suggestions on high-level oversight and coordination, given that several distinguished MRC and Wellcome Trust boards are responsible for overall leadership. But for collection-centred research there may be advantages in tasking a group with watching over clusters of programmes, advising on strategic development and generally working to optimise scientific return on investment. It could make sure that guidance evolves to keep up with developments in science, law, IT and good practice, and it could address important issues having to do with management of materials, linking of databases or the fate of legacy collections. Possibly, too, it could advise on applications having questionable purposes, or hear appeals of access denials.

8.4.4
All oversight systems must respect the specific circumstances of collections and their relationships with participants/data subjects, helpful advocacy groups, other funders and hosting institutions.

9. Storage for data sharing

9.1 Refining and documenting data

9.1.1

The 2002 MRC Population Data Archiving and Access consultants' review reported that:

> The complexity, quirks, or lack of adequate documentation of data were seen, in particular by researchers in the older and larger established studies, to be major barriers to re-using data properly, particularly without the input of the [custodian] team. A high level of investigator support was considered to be essential for supporting new users. ... For some studies, only parts of the dataset (e.g. variables of particular interest) were even cleaned, making the dataset as a whole unusable, even by the in-house team.[38]

9.1.2

Understandably, project teams refine and document subsets of their data holdings according to interest and resources. Documentation can be in good enough shape for internal use, but not written up into the integrated, explanatory, free-standing form that would enable solid analysis by a 'cold' accessor working at a distance.

9.1.3

Not every data set merits full refinement, and it is a matter of return on investment. The MRC consultants' report said:

> Investigators saw the need for a policy to determine relative levels of investment in data preparation and documentation according to different types of datasets and for different end uses. For example, for cohort studies, or studies that have potential for follow-up, the not insignificant costs of good project housekeeping, high-quality data documentation, anonymisation, sample maintenance and sufficient document retention facilities (including paper storage, or digitisation costs) were valued as good investments.[39]

Data-sharing propositions, though, raise difficult portfolio questions of investment of what by whom, for return of what to whom.

9.2 The ESRC/UKDA example

9.2.1

It is useful to note how the Economic and Social Research Council (ESRC) provides for archiving and access. The ESRC requires that all studies that it funds offer their data, in anonymised form, for deposit in the UK Data Archive (UKDA), a centre it supports at the University of Essex in Colchester. The Archive holds data from thousands of studies, and qualitative data as well as quantitative.[40,41]

9.2.2

The UKDA only accepts data that meet quality and format standards, are judged to be sufficiently anonymised, and are accompanied by full documentation of the variables (which it calls 'metadata', data about the data) and a description of the consent. The ESRC stipulates that unless a waiver has been agreed in advance, qualitative data and machine-readable quantitative data must be offered to the UKDA within three months of the end of the funding award, else the final payment of the award is withheld. Whenever necessary, the UKDA offers to assist researchers in preparing data sets for archiving.

The ESRC requirement is that:

> The dataset must be deposited to a standard which would enable the data to be used by a third party, including the provision of adequate documentation.[42]

Once a depositor agreement is executed and a data set is accepted, the ESRC pays the ongoing costs of archiving.

9.2.3

The UKDA maintains a richly detailed online catalogue of its holdings and actively fosters access. Researchers may register and have their bona fides authenticated, whereupon they are allowed access to most of the data in the Archive except for some highly sensitive data that are specially protected.

38 'MRC Population Data Archiving And Access Project Consultants' Report: Developing an MRC policy for population data archiving and access', section 4.5.3. See footnote 1.

39 Ibid, section 4.5.8.

40 'Datasets policy'. Annex C of 'Research Funding Guide'. Economic and Social Research Council; 2005. www.esrcsocietytoday.ac.uk/ESRCInfoCentre/opportunities/research_funding [accessed 21 December 2005].

41 www.data-archive.ac.uk.

42 'Datasets policy'. See footnote 40.

9.2.4
Among many other functions, the UKDA manages the Economic and Social Data Service (ESDS), a unit of which, ESDS Longitudinal, holds the anonymised data from the 1958, 1970, Millennium, and other cohorts.[43]

9.3 Archiving biomedical data

9.3.1
Archiving for access, and preservation for the long term, is a much more robust arrangement than just basic storage. Communal research archives that provide a variety of services are different from local archives. The increasing experience now of biomedical researchers in lodging and using birth-cohort and other data in the UKDA may suggest either that the UKDA should be considered as a home for more biomedical data, or that some kind of national biomedical data archive would be advantageous.

9.3.2
Incidentally, when thinking about the different kinds of data, it is important to recognise that the distinction between 'social' and 'biomedical' data is soft, and has more to do with the auspices of data collection than with the intrinsic nature of the information. Many social science data (such as those collected under ESRC support) have to do with diet, reproduction, ageing, pain, mental health, alcoholism, sensory capacity and other matters of central concern in biomedical research; and many biomedical data (such as those collected with MRC or Wellcome Trust support) have to do with diet, reproduction, ageing, behaviour, mobility, caretaker burdens and other matters of central concern in social research.

9.4 Registers and portals

9.4.1
Given the mountains of data in the various population-based projects, the temporal overlapping of the birth cohorts and the push for more efficient access, there may be some advantages in establishing registers, indexes or portals. Such structures could take forms such as:

- An **online register** of the auspices, data holdings, current status of various data tranches, linkages to other databases or materials, documentation, administration, oversight, and access conditions and procedures of major collections. If topics were searchable, it would serve as an index. Possibly it could list current studies using the collections. Entries could be made either by the collections themselves or by some central unit. It would require serious commitment to keep up-to-date. (One partial analogue might be the Directory of Clinical Databases.)[44]
- An **online portal** to anonymised versions of major collections. This would be more dynamic than a register. It would have indexing functions as above, but additionally it would provide distributed access, perhaps through registration, to a federation of databases. (Thus it might have some of the features of the UK Data Archive or the genome databases.)

9.4.2
Repositories of these kinds could also disseminate good practice publications and provide access to questionnaires or other research instruments. Some lessons may be evident from the experience with Current Controlled Trials and other clinical trial registries.[45] The funders and research community may wish to explore setting up such structures.

43 www.esds.ac.uk/longitudinal.

44 www.lshtm.ac.uk/docdat/.

45 www.controlled-trials.com.

10. Observations and conclusions

10.1 Observations

10.1.1

Access can be improved, and most scientists hope it will be. But if access is to be optimised, not only will barriers have to be reduced but the provision of access will have to be actively facilitated, guided, funded and rewarded.

10.1.2

Potential for access. The collections differ greatly in their potential for sharing data, depending on their auspices, purposes, funding, capacities and attitudes. Two of the most important factors are:

- whether a collection is a longitudinal programme requiring repeated interaction with participants, with implications for protecting the interests of the participants and maintaining good relations with them
- whether a collection's mandate and funding cast it as a resource for wide use or as a locally delimited project.

10.1.3

Access demand. For most collections, the demand for access seems not to be well known. Nor is there an estimate of how the demand might increase if access were more actively invited and facilitated. Demand is not easy to evaluate, and the differences in the nature of the collections makes generalising difficult. But some indication could be gained by polling or by reviewing records of inquiries (if such are kept), access logs, material transfer records, and publications based on the resource.

10.1.4

Collaborative access. Collaborative modes of research find much favour. Indeed, they are encouraged by the MRC access policy, which remarks that "such research is often most fruitful as a collaboration between the new user and the original data creators or curators, with the responsibilities and rights of all parties agreed at the outset".[46] Of course collaboration tends to be selective and self-fulfilling, in that once they have committed to co-labouring, the parties have a mutual interest in making it succeed.

10.1.5

Independent access. What is less clear is how fruitful *non*-collaborative studies can be, under favourable conditions. If access broadens, custodians will not be able to be involved in depth with the scientific substance of all studies. Independent use needs to be examined. To this end the consequences of the NIH and ESRC policies (sections 1.2 and 9.2 above) that enforce the sharing of anonymised, well-documented data, and the MRC's slightly less directive policy, deserve evaluation. How well is consent tracked and respected? How effective is the disclosure limitation? How thorough does documentation have to be to support good science? (And can criteria be developed?)

10.1.6

Overseeing use of materials. Access to materials is treated differently from access to data, and in some situations it is subject to separate oversight. In part this is determined by the auspices of the sample collecting and the terms of consent. In part it follows from the need to allocate depletable materials carefully. But in part it is a response to the public's concerns – not necessarily felt by researchers themselves – about risk of re-identifiability, or of potential for abuse by the police, insurers, banks or other external parties. Whether separate oversight of materials is necessary, effective or efficient should be monitored.

10.1.7

Broad suggestions. Among large-scale initiatives that might be considered are:

- establishing registers or portals to a range of collections
- developing a large national case-control collection (is this UK Biobank?), in part to relieve some other access demands
- establishing a robust national biomedical data archive.

46 'MRC Statement on Data Sharing and Preservation Policy'. See footnote 2.

10.2 Conclusions

Brief responses to the funders' original questions are as follows.

10.2.1
Q: To what extent are current access arrangements standardised?

They are not very standardised, although there are many commonalities. It is difficult to generalise, however, because many projects' policies are informal or difficult to obtain. Even some projects that provide access do not publicise that fact or post their policies or conditions on their websites. A surprising variety of access and material transfer agreements are used.

10.2.2
Q: Is there scope for greater standardisation, given the ethical, legal and practical considerations?

Yes. The core terms of access and material transfer agreements are begging for standardisation, for instance (section 4 above). Standard default criteria for access to materials could be useful. And there are other opportunities.

10.2.3
Q: Is there scope for a model governance structure and if so, might this usefully include elements centralised to cover several collections?

The most common mode of governance – which implements and complements NHS Research Governance and the authority of the funders – is that of supervision of collections and data sharing by oversight committees. With some variation, most of these have similar remits and structural features.

Surely a model structure could be derived, incorporating the best features of the current oversight bodies. Among other things it could address composition, chairing, voting status of members, relationship with the activities supervised, relationship with the funders, reporting and powers.

The report identifies examples of situations in which multiple oversight or advisory committees might be consolidated, and others seem obvious, but the present review was not in any position to evaluate this. Centralisation could elevate attention to the issues (many of which are common among projects), foster consistency and make more efficient use of committee members' time.

As was remarked in section 8.4.3, repeated here for convenience:

For collection-centred research there may be advantages in tasking a group with watching over clusters of programmes, advising on strategic development and generally working to optimise scientific return on investment. It could make sure that guidance evolves to keep up with developments in science, law, IT and good practice, and it could address important issues having to do with management of materials, linking of databases or the fate of legacy collections. Possibly, too, it could advise on applications having questionable purposes, or hear appeals of access denials.

10.2.4
Q: Is there scope to develop guidelines of general applicability to the range of collections supported?

Clarification and revised guidance are urgently needed on aspects of consent, confidentiality, and anonymisation. Guidance on the characteristics of 'limited data sets' would be helpful.

When the Human Tissue Authority issues its Codes of Practice, naturally guidance on the implications will be essential.

Whether guidelines would be the best instrument is not evident, but it is important now to sort out the rights and obligations of data providers (such as regarding the screening of the bona fides of data requesters, the merits of protocols, and the quality of manuscripts), and the rights and obligations of data requesters (such as regarding whether they can resist working in collaborative mode if they prefer to, and whether they can be required to publish negative-association results).

At least informal guidance is needed on such matters as publication of collections' access policies and procedures. This might be coupled with guidance on the elements of access agreements.

Criteria or guidelines deserve to be drafted on the documentation required to support scientifically sound independent use of data sets.

10.2.5
So, much can be done.

Appendix 1
People consulted during the project

Dr Yoav Ben-Schlomo	Department of Social Medicine, University of Bristol
Professor Paul Burton	Department of Health Sciences, University of Leicester
Ms Tara Camm	Legal Department, the Wellcome Trust
Dr Sheila Casserly	UK Biobank, Manchester
Dr Brian Clark	National Cancer Tissue Resource, London
Professor David Clayton	Diabetes and Inflammation Laboratory, Cambridge Institute of Medical Research, University of Cambridge
Professor Cyrus Cooper	MRC Environmental Epidemiology Unit, Southampton
Dr Louise Corti	UK Data Archive, University of Essex, Colchester
Professor George Davey Smith	ALSPAC, and Department of Social Medicine, University of Bristol
Professor Ian Day	Division of Human Genetics, School of Medicine, University of Southampton
Professor Carol Dezateux	Paediatric Epidemiology and Biostatistics Unit, Institute of Child Health, University College London
Professor Richard Durbin	Wellcome Trust Sanger Institute, Cambridge
Professor Shah Ebrahim	Department of Social Medicine, University of Bristol
Professor Peter Elias	Institute for Employment Research, University of Warwick, Coventry, and Special Advisor (Data Resources) to the ESRC
Professor Paul Elliott	Department of Epidemiology and Public Health, School of Medicine, Imperial College London
Professor Jean Golding	ALSPAC, and Unit of Paediatric and Perinatal Epidemiology, University of Bristol
Dr Hazel Inskip	MRC Environmental Epidemiology Unit, Southampton
Professor Heather Joshi	Centre for Longitudinal Studies, Institute of Education, University of London
Professor Kay-Tee Khaw	EPIC-Norfolk, and Addenbrooke's Hospital, Cambridge
Professor Diana Kuh	Department of Epidemiology and Public Health, University College London Medical School
Professor David Leon	Department of Epidemiology and Population Health, London School of Hygiene and Tropical Medicine
Professor Sally Macintyre	MRC Social and Public Health Sciences Unit, University of Glasgow
Dr Kevin Moreton	Programme for Basic Genetics, MRC
Dr Andrew Ness	ALSPAC, and Unit of Paediatric and Perinatal Epidemiology, University of Bristol

Professor William Ollier	Centre for Integrated Genomic Medical Research, University of Manchester
Professor Catherine Peckham	Paediatric Epidemiology and Biostatistics Unit, Institute of Child Health, University College London
Professor David Porteous	Medical Genetics Section, Centre for Molecular Medicine, University of Edinburgh
Professor Christine Power	Paediatric Epidemiology and Biostatistics Unit, Institute of Child Health, University College London
Professor Alan Silman	Arthritis Research Campaign Research Unit, University of Manchester
Professor David Strachan	Department of Public Health Sciences, St George's Hospital Medical School
Dr Richard Sullivan	Clinical Programmes, Cancer Research UK
Mrs Joanne Sumner	Biomedical Ethics Programme, the Wellcome Trust
Professor John Todd	Diabetes and Inflammation Laboratory, Cambridge Institute of Medical Research, University of Cambridge
Professor Sir James Underwood	University of Sheffield, and President of the Royal College of Pathologists
Professor Michael Wadsworth	Department of Epidemiology and Public Health, University College London Medical School
Dr Heike Weber	Stem Cell Programme, MRC
Ms Melanie Wright	UK Data Archive, University of Essex, Colchester

Appendix 2
The author

William W Lowrance, PhD, is a consultant in health policy and ethics, based in Geneva, working on issues surrounding database research, genetic and genomic research, and pharmaceutical research.

After earning a PhD in organic and biological chemistry at The Rockefeller University, Dr Lowrance has taught and conducted research on science and technology policy, environmental policy, health policy, and risk decision-making, at Harvard, Stanford, and Rockefeller Universities.

He has served as the Executive Director of the International Medical Benefit/Risk Foundation and as a member of many government, industry, and public-interest advisory committees.

Among his many publications are the books *Of Acceptable Risk: Science and the Determination of Safety* and *Modern Science and Human Values.*

In 2002 he prepared an extensive report, *Learning from Experience: Privacy and the Secondary Use of Data in Health Research*, for the Nuffield Trust. During the startup of the UK Biobank project he chaired the Interim Advisory Group on Ethics and Governance.

E lowrance@iprolink.ch

The MRC is the UK's leading publicly funded medical research organisation. Its mission is to

- Encourage and support high-quality research with the aim of improving human health.
- Produce skilled researchers, and to advance and disseminate knowledge and technology to improve the quality of life and economic competitiveness in the UK.
- Promote dialogue with the public about medical research.

The Wellcome Trust's mission is to foster and promote research with the aim of improving human and animal health. During 2005–2010, our principal aims are:

Advancing knowledge: To support research to increase understanding of health and disease, and its societal context

Using knowledge: To support the development and use of knowledge to create health benefit

Engaging society: To engage with society to foster an informed climate within which biomedical research can flourish.

In support of these aims, we also recognise the importance of promoting the development of individuals we fund, enhancing the environment for research and its application, and constantly improving the way we operate.

The Wellcome Trust is a registered charity, no. 210183.
Its sole Trustee is The Wellcome Trust Limited, a company registered in England, no. 2711000, whose registered office is 215 Euston Road, London NW1 2BE.

DP-3564p/1k/03-2006/TU