NUFFIELD
COUNCIL ON
BIOETHICS

The linking and use of biological and health data

Open consultation

17 October, 2013

(Closing date: 10 January, 2014)

Contents

Part One

Background to the consultation

The Nuffield Council on Bioethics has convened a Working Party to examine ethical issues that arise in relation to the linking, analysis and use of biological and health data. Information about the project and the Working Party can be found on the Council's website. A glossary is provided at the end of the document; terms explained in the glossary appear in (hyperlinked) blue text when they are first used in each section.

Some relevant developments

We are undertaking this work now because technological advances are leading to new global opportunities and developments in biomedical data collection and linkage in the areas of health-related research, clinical practice and governmental activities.

Developments in which we are particularly interested include:

Developing resources:

- Large-scale biomedical research resources ('biobanks') collecting data from many participants that combine the comprehensive description of observable characteristics of people, their health records, analyses of their genomes (for example, whole genome sequencing) and/or other large data sets, such as proteomics, metabolomics and other 'omics', or imaging results (MRI, CT, etc.).
- The increasing intensity of biomedical data collection, analysis, use and retention in health care.
- The development of telemonitoring and assessment devices to provide biomedical data for research, clinical care or personal use.
- The availability of increasing volumes of open access data, both from research studies and from other sources that are potentially useful in combination (e.g. user generated/ self-published data/ social networking data).

Developing methods:

- Advances in data science and technology that support the management and analysis of the large data sets available to research, such as cloud computing, statistical imputation, machine learning, visualisation, and 'big data' approaches generally.
- The increased emphasis on the use of very large data sets and computationally intensive methods to attempt to understand underlying models and to generate hypotheses in areas such as the genetic epidemiology of common diseases.
- Developments in the use of data and predictive analytics to inform health interventions (e.g. identifying individual risk profiles for hospital readmission to devise personalised, preventative primary care plans) and use of algorithms that use diverse types of data in health care.

- Development of global interoperability standards that allow the combination of data from distinct sources to increase the statistical power of research studies, and the increasing globalisation of data linking and use (e.g. population biobanks and international genome research collaborations).

Developing contexts:

- An impetus from research funders and others for researchers to share their data, analyses and results with other researchers, and an accompanying movement towards open access publication of research results.
- The movement of biomedical data, medical records and other personal data internationally for research use under different regulatory frameworks.
- The movement of data use outside traditional, regulated research and health care contexts (e.g. Medicine 2.0, 'citizen science', apomediation).
- The potential for increased use of biomedical data and data linking for non-health purposes (e.g. government administration, insurance, marketing).
- Policy and legal developments that favour and facilitate the extraction of value through the reuse of data outside the scope of the purpose for which they were originally collected (e.g. using health records in biomedical research).

What ethical issues do these developments raise?

We are interested in the ethical challenges raised by these developments and their social and legal implications. Some of these ethical challenges will be familiar from other contexts but need to be reassessed in the light of the new opportunities, risks and uncertainties created by these new technological advances. At the same time, issues may arise that are specific to the new contexts created by these developments.

The main ethical questions in which we are interested relate to how, and how far, the different interests that are engaged by the use of biological and health data should be satisfied, and the ethically appropriate basis for doing so. In particular, they relate to determining ethically appropriate ways of avoiding or resolving conflicts between these different interests, where they occur.

We are interested in mapping the full range of ethical issues relating to the linking and use of data. However, it is striking that, in discussions on this topic, the issues are often articulated in relation to the concepts of privacy and the public interest. These include:

- How we understand and recognise people's relation to their biological and health data (e.g. in terms of ownership, personal rights, identity, interest, etc.)
- How the pervasive use of information technologies and their impact on wider culture affect behavioural norms (including moral norms of privacy) and expectations (e.g. about public and private spaces).
- The extent to which people should be able to exercise control over how data relating to them are used and accessed by third parties and the mechanisms through which this

may be achieved (such as consent procedures) or made unnecessary (e.g. through anonymisation).

- The extent to which people should be enabled to opt out of socially beneficial systems and relationships (e.g. a right to anonymity/to be forgotten) and the consequences of this (e.g. if joining in is linked to effective access to public services).
- The circumstances in which individual privacy interests can be qualified by the interests of others (e.g. the privacy of family members, of groups or the interests of commercial firms) or the wider public interest.
- The role of public authorities and democratic processes in setting norms, defining options, and developing appropriate governance mechanisms for the use of biomedical data.
- The proper role and rewards for commercial firms in delivering public benefits through the use of biological and health data (e.g. developing commercial products, such as medicines, and delivering public goods).
- The cultivation of solidarity and the degree to which "free-riding" can be tolerated (i.e. when people benefit from goods without bearing their fair share of the cost of making them available).
- The ways in which public trust in biomedical research and health care data processing institutions may be created, maintained and lost, and the role of governance mechanisms in this.
- The implications of government or commercial control over what sets of options are available to individuals (e.g. making services available only on condition that an individual's personal data can be re-used for other purposes).
- Obligations (e.g. of public authorities and service providers) to those who are excluded or who exclude themselves from utilised data sets, for example groups that may be hard to reach, be wary of stigmatisation, etc.

Comment [IH1]: The wording of seems unduly perjorative to me, and could discredit the consultation.

Comment [IH2]: This is linked to bullet 4 above

Our aim is to develop an ethical analysis that will frame policy and governance approaches to linking and use of linked biomedical data. Where appropriate, we will make specific, practical recommendations for action. Our report will be published in 2014.

Below we set out a number of questions on which we would welcome responses to help to shape and inform our deliberations. In each case there is a general question, and some more specific questions that you may wish to tackle in your response. Please feel free to respond to as many or as few questions as you wish, and to suggest other questions that, in your view, we ought to tackle in our deliberations.

1 Introduction to my response

The Nuffield is undertaking '*this work now because technological advances are leading to new global opportunities and developments in biomedical data collection and linkage in the areas of health-related research, clinical practice and governmental activities.*' My response focuses on the sharing of data recorded in patient records that clinicians and care workers create and use during the provision of personal care. In what follows, biomedical data would be treated much as any other data in the patient record.

1.1 I wish to see the maximum exploitation of patient data for patient care and secondary purposes such as managing the NHS, research and development. But for me the major driver for this consultation is not the '*new opportunities and global developments*' but the '*ethical issues that arise in relation to the linking, analysis and use of biological and health data*'. Without urgent dialogue and compromise an opportunity to help create a high quality and sustainable NHS and boost the UK life sciences industry will become the subject of an increasingly bitter battle between privacy advocates and many patients and clinicians caring for them, and a powerful faction of those who see it as their right to use identifiable patient data for secondary purposes with few if any constraints.

1.2 Patient privacy in the NHS today

2.1 Lobbying by Big Data and others over a number of years (see 'Fair Shares for All', chapter 6) resulted in the Health & Social Care Act 2012 (HaSC Act 2012) greatly reducing the legal protection for identifiable patient data. The Secretary of State (SoS), NHS England, Monitor, NICE and the Care Quality Commission may now direct the Health & Social Care Information Centre (HSCIC) to collect identifiable patient data from providers of NHS-funded care without patient consent, without constituting a breach of the confidentiality that clinicians owe their patients. The Act also frees directions from the checks and balances previously applied to all requests to use unconsented identifiable patient data by the independent Ethics and Confidentiality Committee (ECC), now the Health Research Authority Confidentiality Advisory Committee (CAG). The HSCIC may provide identifiable patient data it collects to end users without patient consent unless approved by the CAG. The Act provides five other new statutory gateways that various bodies can use to disclose or request identifiable data, but these are unlikely to involve patient data and using them does not override any obligation of confidentiality involved.

2.2 The first major direction, NHS England's collection – care.data -of identifiable patient data from general practitioners (GPs) to form the first element in the care episode database database of linked data to be built by the HSCIC has been notable for the gradual and belated way that it has 'informed' clinicians, the public and patients of its content, character and intentions. It has also apparently exhibited major scope creep in both the range of intended users, the purposes for which it may be used, and in the near future in the data it will collect. The HSCIC may provide the data to NHS, academic, government and commercial organisations for a very broad range of purposes.

2.3 Collective top level NHS governance arrangements have been rendered ineffective by the Act's abolition of the National Information Governance Board for Health & Social Care (NIGB), the independent body responsible for developing information governance policy in the NHS and monitoring its implementation.

2.4 Other than actions of GPs and care providers as controllers of patient data, the greatly weakened Data Protection Act 1998 (DPA 1998) and the Human Rights Act 1998 (HRA 1998), the other mechanisms that intended to balance the public benefit against patient confidentiality - the General Extraction Service Advisory Group (GPES IAG) and the Confidentiality Advisory Group - are advisory only. GPES IAG cannot even recommend that a proposal is rejected: the most it can do is ask the proposer to reconsider its application and then resubmit it, hopefully in revised form.

2.5 The only counterbalance to this erosion of protection is the patient opt-out to the sharing of their identifying data held by GPs and/or the HSCIC for secondary purposes announced by the SoS in October 2013. This is a pledge in the NHS Constitution rather a legal right. The fact that it has been explicitly stated that the opt-out rate for care.data is being monitored is somewhat disturbing, and the way in which NHS England has informed clinicians and patients about its proposals is likely to result in a much larger opt-out rate than it need have been.

2.6 With the rapid increasing availability of identifiable business and social data on the web and the increasing availability of tools to analyse and match unstructured or poorly structured data, including images and sounds, the chance of someone being able to re-identify non-identifying data is likely to be rising rapidly. This will, if permitted, increase the pressure to process data currently regarded as non-identifying in well-controlled contexts under formal data sharing agreements, and ideally in situ at the organisation that is holding it.

2.6 While the Data Protection Act 1998 may not be the best vehicle for dealing with the new opportunities and developments in data creation and handling – see 6 below for more on this – it plus the common law obligation of confidentiality and application s251 of the NHS Act 2006 should still be key mechanisms for determining the balance between personal, professional and societal interests when it comes to examining applications to use identifiable data for secondary purposes other than collections mandated by directions made under the HaSC Act 2012.

3 Public and patient views

3.1 A more comprehensive review of these is given in chapters 4 & 5 of 'Fair Shares for All' a review of patient information governance issues published by BCS Health and the BCS Primary Healthcare Specialist Group in early 2012. What follows is a summary of that material. Sharing relevant elements of that data with others involved in their care as and when needed is generally accepted by patients and clinicians as covered by an implicit consent that starts when care is sought, although the patient may ask that particularly sensitive items are not shared without his or her consent: their clinician will usually agree and most clinical systems for general practitioners provide facilities to flag items as such.

3.2 The majority of patients know little of the secondary uses their data is put to, and whether they use identifying or non-identifying data. A majority are content for their effectively de-identified data to be used for secondary purposes without consent, as the law permits, although a substantial minority believe that patient data cannot be effectively de-identified. Clinical trials must be law only involve patients with their explicit consent. Hospital departments often recruit patients with explicit consent as and when they are seen for treatment. There are a set of circumstances, largely to do with the health of third parties, public health and crime, where clinicians are obliged by statute to share identifiable information about their patients without patient consent.

3.3 Although patients wish to be asked for consent before their identifiable data is used for secondary purposes (including by researchers to identify and contact suitable subjects for research although they are happy for this to be carried out by practice staff on the researcher's behalf), they also understand that seeking consent to use identifiable data is not always possible. Where they are aware that a independent and open mechanism with effective patient and public representation is available to act as a proxy for obtaining consent and understand how it operates, 30% of those asked were happy to see it make recommendations about who may use their identifying data without their consent, vide 'Fair Shares for All' Chapter 5.

3.4 Patients have preferences about who should use their data for secondary purposes. The NHS comes top, academe second and commercial firms a poor third. A significant percentage of patients doubt that the profit motive and respect for confidentiality make comfortable bed fellows. This attitude may be bolstered by the view that patient data used in the course of making a profit should be paid for.

4 The ethical considerations underlying patient data sharing

4.1 The ethical foundation of health care is the trust between patients and those caring for them. Patients expect their professional carers to act in the patients' best interests. This includes respecting the confidentiality of the data generated during the provision of care and storing it securely. It is on this basis that patients provide clinicians with the (sometimes very confidential) information that is regarded as necessary for care. Some spectacular data losses of NHS patient data have dented patient willingness to share confidential data for their own care, in some cases leading them to seek care in different institutions or more rarely outside the NHS.

4.2 When data is stored and shared in a way that respects patient privacy in line with patient expectations it creates a virtuous circle and encourages patients to share their data for care and secondary purposes. If it is not it undermines public confidence in the NHS, a major public interest, and discourages patients from providing data to their clinicians and others, clinicians from recording what they are told, and in extreme cases patients from seeking treatment. It will lead some patients to hold records to which they alone control access, and to use whatever means they have to prevent their data being used for secondary purposes.

4.3 Most patients are very supportive of healthcare-related research, and are keen to support it – in other words they are aware of their moral obligation to provide data to improve the lot of patients in general as well as their own. This is recognised in the NHS Constitution and to some degree in the DPA 1998, s251 of the NHS Act 2006 and the various codes of practice on how patient data may and should be used. However patients and their clinicians would like secondary users to respect this goodwill by treating them as partners in a joint enterprise rather than merely as inert data sources. This means, among other things, providing as much accurate information about proposed secondary uses and patient options as possible as soon as possible, and ensuring that any consent / opt outs are well documented, easy to use and place the minimum burden on busy clinicians. Asking for consent to before using identifiable data for secondary purposes is seen as a mark of respect. . The introduction of the English Summary Care Record and the care.data collection are vivid reminders of what not to do. Unsurprisingly the arrogance occasionally shown by secondary users goes down very badly with patients and the public, see 'Fair Shares for All'.6.1 bullet 3.

4.4 For some their attitude to particular kinds of secondary use are coloured by strongly held beliefs. For example Jehovah Witnesses would be horrified if their data was used in any form for work related to the production or use of blood products for systemic use. Similarly many practising Catholics would be very disturbed if their data was involved in research into biochemical contraception or terminations of pregnancy.

4.5 The DPA 1998 makes it clear, in line with public opinion, that identifying data about health and healthcare should only be used for secondary purposes outside the list of exemptions listed in Schedules 2 and 3 of the DPA 1998 with explicit consent, but does not bypass the need to respect the obligation of confidentiality clinicians owe their patients.

4.6 The ethics of opt outs are curious and open to different interpretations. Those in favour of using them to avoid seeking explicit consent:

(a) consider it reasonable to use a default that reflects the majority view of data subjects on the topic, and

(b) that the data subjects involved are aware of what is being proposed, that an opt out is available and how to use it.

Opponents point out that without some explicit signal from the data subjects it is not possible to know whether (b) is true or not. It has been compared to knocking on someone's front door, and if there is no reply assuming that it is OK to enter the house and take what is required. The onus is clearly on the body wanting the data to take all reasonable steps to ensure that (b) is true. If that is not possible, the alternative is to approach an independent body that has substantial representation of the data subjects, the general public and the data controllers, such as the Confidentiality Advisory Group, for permission to use the subject's data without consent.

5 New techniques that impact the need to use identifiable data

5.1 Very few secondary purposes can only be achieved using identifying data, and linking data from different data sources is no exception. The need to extract and use identifiable data for secondary purposes is dramatically reduced when the data to be shared is de-identified and allocated a pseudonym at source where it is to be linked with data extracted from other source(s) using the same pseudonym generation technique. Pseudonyms can be generated by enciphering a known unique identifier, NHS number being the obvious one in the English NHS, or one or more commonplace but partially unique identifiers, such as date of birth, name and postcode of residence. NHS number will give very high linking rates where NHS numbers are accurately and comprehensively present in records. Fuzzy matching using commonplace identifiers produces lower matches, 87 – 93% being typical.

5.2 Proven open source and commercial software is available for the purpose, and is in use in at least two research data repositories, Q Research and THIN. The Scottish Personal Information Research Environment (SPIRE) has also opted to use it. However it must be emphasized that pseudonymisation at source will have to be accompanied by other measures to prevent re-identification where the data in the records concerned is very rich, e.g. has been linked with data from other sources, and/or and could be used in combination with other data its holder has or can have.

5.3 Attachment of location-specific data, such as local deprivation indices, should also be done at source if practicable, to avoid having to export detailed location data, such as full postcode.

5.4 Where complete partial identifiers such as date of birth and dates of clinical consultations and / or postcode, are essential to the intended purpose, the data should be treated as though it is identifiable.

6 Limitations of the DPA 1998

6.1 The DPA 1998 works best when data is collected for a well-defined and fairly stable set of purposes and users known beforehand. Its application to the creation and use of comprehensive databases of linked data collected/created prospectively for largely unknown purposes and users is much more problematic. These make it very difficult to provide fair processing information for consent or opt out purposes. Prospective collections are also likely to be very rich in order to cater for a wide variety of largely unknown users and purposes, although many users may only use a fraction of the data available.

6.2 Who should provide fair processing information to data subjects where a data controller is directed to share identifiable data by another body? It is required by law where data subjects are being asked for consent / offered an opt out, and good practice when any major collection of individual data is planned that is not identifiable. The DPA states that the Data Controller should do this, e.g. the GP for care.data. GPs consider this unreasonable where the extraction is mandated by another body (NHS England) using a statutory power. NHS England and the HSCIC have finally agreed to provide the information to patients via a maildrop to each household after discussions with various bodies, including the Information Commissioner. This raises the broader question of whether this should be the norm where a data controller is being mandated to provide data by an organisation using a statutory power.

6.3 Is data identifiable or not?
Linked data is may well be so rich, e.g. contain dates of patient attendances at named care facilities, the clinician(s) seen, rare conditions, etc, that it should be treated (as the DPA states) as identifying data in conjunction with other data that its data controller has, or has access to. One result of this is that data that leaves a data controller in an effectively de-identified form may offer such a significant risk of re-identification in the hands of its new controller that it should be treated as identifiable. This can make it difficult for data controllers to decide whether to agree to share data and whether data subject consent is required before doing so, and for patients to assess the breach risk before deciding whether to consent or not. If consent or an opt out is sought for the future sharing of warehoused data before the data collection starts and when the individual users and uses are not known, as is the case for care.data, the situation becomes extremely difficult to resolve.

6.4 Is informed consent is only possible when an application is received from a particular body to use specific data for a specific purpose? And to add to the difficulties, obtaining patient consent to use data collected some time ago becomes progressively more difficult as time goes on, and in the case of people who have died, impossible (as well as possibly being unnecessary). One solution is an independent mechanism such as the CAG or an equivalent that enables representatives of patients, clinicians and the public to make the necessary decisions on their behalf after receiving expert advice on the likely data requirements and potential risks, harms and benefits, etc. Without such a mechanism that is stable, known and generally respected by the public, most patients asked for consent to or who are aware of an opt out to the collection of data to be warehoused are likely to play safe and refuse consent / opt out.

7 What do the big data enthusiasts want and how can it be done?

7.1 NHS England & HSCIC's goal understandably appears to be a convenient warehouse:

- that contains a large and growing pool of cleansed anonymised / pseudonymised linked and unlinked comprehensive patient-related data, incrementally kept up to date where cohorts are being followed (for care.data the cohort is all English GP patients who do not opt out). It has to be comprehensive as the data requirements of all end users are not known at the outset.
- that end-users – including NHS England - can draw on as and when they wish without further reference to the original data controllers.
- where much data would be routinely collected, e.g. GP, hospital and mental health patient data, while other types would only be collected on demand.
- where the body managing the data pool would link some data routinely (e.g. the data from all care providers that it routinely collects) and generate the rest on demand. This might involve linkage with data outside the warehouse, including Cancer Registry information and non-health data, e.g. from ONS, social services, education, etc.
- where collected and linked data would be retained indefinitely if it was thought to be necessary.

7.2 The ultimate end users and their purposes would not be known at the time the relevant data collection starts except in the most generic terms, and the same would be true of the purposes for which it would be used. This information would be of little help in determining the risks to patient confidentiality that its end-uses will create. The HSCIC was set up under the HSC Act 2012 to carry out the functions above, although other bodies, academic and otherwise, are planning to provide similar services, generally on a smaller scale.

7.3 In many ways, doing these activities once for many clients makes good sense. Collecting, cleansing and linking patient data and keeping it secure takes considerable expertise, time and resources. The necessary quid pro quo for patients and their clinicians is that they feel confident that the centre is open and honest in its dealings with them, robust in its dealings with people requesting data collections and would-be end users and goes the extra mile to ensure patient data confidentiality – no saying "we don't have to do this so we won't".

7.4 If we accept this idea, the next question is: how should such a body operate in order to keep their side of the bargain and generate and maintain clinician and patient trust? The following would encourage this:

- Data collections respect patient wishes not to share items in their records flagged as especially sensitive
- Data collected should be anonymized at source, and allocated pseudonyms if it is to be linked after collection. Only the source can use the pseudonyms to re-identify the patient

- Bearing in mind the unknown nature of future uses and end users, and the difficulty of knowing whether the data may pose a significant risk of re-identification after linking and/or in a future end user context, the norm should be that patients become contributors by opting in rather than by default.
- All requests for collections of identifiable data where it is not thought possible to obtain consent, and those involving data that is aggregated or de-identified at source, would be determined by an independent committee that contains all major stakeholders (including a majority of patients and patient data controllers) whose conclusions and the reasoning behind them would be published. Where possible, the committee would ensure that the data collected did not exceed that needed to satisfy the proposed end uses, where these are known. The same would apply to all requests from end users for access to the warehoused data.
- All data held at the warehouse is encrypted
- End users either use data at the data warehouse (the preferred solution as it minimises the risk to data subject confidentiality) or (second best) may extract effectively de-identified data
- End users would have a public data sharing agreement with the centre. Data would only be allowed to use the data for the agreed specific purpose(s). Data abuse, such as patient re-identification. further sharing and use for purposes other than those agreed with the data provider would be explicitly forbidden (the former would only be permitted at the original data source) and ideally criminalised, and significant sanctions would be applied to the user if it is found to have occurred.
- Any copy of the data held by end users must be destroyed after use.
- Catalogues would be maintained giving a full description of :
    - The data collected, including whether incrementally or not.
    - Each datasets created by linking
    - Each use of data by an end user.
- The organisation's information governance arrangements and performance would be audited by an independent body at regular intervals, and described in a public report intended for a wide audience, including clinicians, patients and the public.

8 Glossary

The following definitions of data about a person are used in the Consultation response.

Identifying data - data about a dead or living person that identifies him or her:

1 because it includes one or more unencrypted commonplace identifier (e.g. name, date of birth and postcode) and/or other unique identifier which is in general use, (e.g. NHS number, Hospital Number or National Insurance Number), and/or

2 when combined with other data that the data controller has or could have.

Data about a person that contains no identifiers is known as **identifier-free data***.* The opposite of identifying data is **non-identifying data**.

The problem with definition 2 is that the risk of re-identification is a continuum that can vary from minute to almost certain depending on the richness of subject data and what identifiers the other data contains. What level of risk makes the data identifiable?

Personal data - identifying data about a living person. The Data Protection Act 1998 (DPA) only applies to personal data.

Anonymised data –identifier-free data that contains a meaningless unique identifier. For example it may be a number generated randomly as each record is anonymised.

Pseudonymised data - identifier-free data that contains an intrinsically meaningless unique identifier or pseudonym that can under certain conditions be used by specified people to establish the person's identity. The pseudonym may be an encrypted variant of one or more commonplace / unique identifier, in which case its creator must ensure that it is extremely unlikely to be decrypted by unauthorised personnel.

Potentially identifying data – identifier-free data that offers a significant risk of re-identification in conjunction with other information that the person holding the data and/or others with access to it know or have access to. The DPA states that if re-identification is possible in this way, the data should be treated as identifying data. Potentially identifying data may be anonymised or pseudonymised.

Effectively de-identified data – identifier-free or aggregate data that is considered to pose an insignificant intrinsic risk of re-identification. Identifier-free data may be anonymised or pseudonymised. However if the cohort used is not a random sample, repeated tracker querying may isolate individuals or small groups with a unique combination of properties and enable someone with knowledge of the population to identify them.

Aggregate data is derived from records about more than one person, and expressed in summary form, such as a statistical table comprising a number of cells, each of which records the number of people with one or more properties / property value in a particular range. While intrinsically much more resistant to re–identification, it may be possible where the source population is not a random sample and some cells contain small numbers (typically <5).

Part Two

Consultation questions

1 Do biomedical data have special significance?

We are interested in the linking and use of human biomedical data. This includes measurements and test results (e.g. blood test results, DNA results, X-ray and MRI scan images), reported information (e.g. descriptions of symptoms and responses to medicines), and experimental findings (e.g. clinical trial data) among many other things. Data of this kind might be collected as part of a diagnostic or treatment procedure, health assessment, research project, drug trial, etc.; they may be stored in databases, medical records or other types of record. We are also interested in the use of data about behaviour, lifestyle, social relationships, sexual history, occupational and environmental exposures, etc. where these are relevant to understanding human biology and health, the development of medical science or the delivery of health care. Our focus, however, is principally on the ways in which these data may be linked and analysed together in order to generate insights that can be applied in the treatment of individuals and populations.

Comment [IH3]: I am not clear whether you are assuming that these are human biomedical data or not.

Comment [IH4]: This is also done in order to develop new clinical techniques, approaches and products to use for personal and population healthcare

Consultation question 1:

Do biomedical data have special significance?

Possible aspects to consider:

- Is it useful (or even possible) to define biomedical data as a distinct class of data? If it is, what are the practical and ethical implications of different ways of defining this class?

Background to the Consultation

- What factors contribute to the belief that personal biomedical data deserve special protection? Does the sensitivity of biomedical data depend entirely on context or do biomedical data have special attributes that make them intrinsically more sensitive than other kinds of data?

Some biomedical data, e.g. about a person's mental health and sexual behavior, is undoubtedly regarded by patients as among the most sensitive there is, and the Data Protection Act 1998 recognises all data about a person's health as one of the most sensitive types of personal data. I doubt whether genomic data, including DNA sequences, are any more sensitive than healthcare data in general.

- How are changes in the scope of the data in use providing meaningful insights into individual biological variation and health?

Having genomic data enables us to see what specific collections of genetic and epigenetic features affect:-

- Susceptibility to particular forms of morbidity

- functional performance, longevity and other aspects of health,

- The efficacy of, and tolerance to, systemic treatments

- Do some sub-sets of biomedical data (such as genomic data sets) present particular ethical challenges or offer ethically important benefits?

I don't believe that they do in a large way, but see the succeeding answer.

- To what extent should genomic data sets be regarded as belonging to one individual and to what extent should other interests (e.g. of family members sharing genomic sequences) be recognised? What implications might this have for consent to collection of such data, for feedback concerning the data and for its broader use?

Comparatively little genomic data is unique to an individual, other than his or her complete genome sequence. But this shouldn't mean that the data should be handled any differently to any other data. It should only be collected in identifiable form with consent for specific purposes, and shared according to that consent, or a specific additional consent or as permitted by the law. Sharing it with others, such as related carers or family, may have to be done with care where it has implications for them, and/or they may realise the implications of its contents for the person it belongs to. I cannot see that there would be special problems when it was shared with other clinicians caring for the patient, but, like other patient data, it should not be shared more widely in identifiable form without patient consent.

DNA matching is generally accepted as the most reliable way of uniquely identifying people, and a complete DNA sequence is the ultimate unique identifier in its own right even if it is not commonplace one, i.e. intelligible to any casual competent observer:

2 What are the new privacy issues?

The protection of privacy is considered as a basic tenet of a civil society and is protected through established ethical and legal frameworks. Biomedical data relating to living individuals is usually given special protections because of the potentially sensitive nature of the data. However, the value and the protections attached to a given type of data may vary according to the circumstances and the social context. Furthermore, people's individual privacy preferences may be complex, and are sometimes contrary to the privacy interests of others or to the perceived public interest in using personal data for the benefit of the wider society.

Comment [IH5]: Which is why we have a Data Protection Act.

Willingness to disclose personal data may depend on many things, including how sensitive people consider the data to be, the benefits they believe they or others may receive from disclosure, the perceived risks to their own interests or those of others, the level of confidence they have in how subsequent use will be governed, the level of control they may retain over the data, and norms of social behavior in similar situations, among many other things. Given the public and commercial interests in biomedical data, some decisions must be made collectively about what measures are most appropriate and desirable to ensure adequate protection of privacy, and to facilitate voluntary participation and legitimate activity. We are therefore interested in understanding the values that underlie the concepts of privacy and public interest in the use of biomedical data; we are interested, too, in understanding the nature and significance of actual harms and benefits involved.

Comment [IH6]: Including above all the purpose for which it was provided (which unfortunately may be with implicit consent)

Comment [IH7]: There is an existing set of these decisions, as represented by statute and the common law obligation of confidentiality owed by clinicians to their patients.

Consultation question 2:
What are the new privacy issues?

Possible aspects to consider:

- Do new information technologies and 'big data' science raise privacy issues that are new in kind or in scale?

I don't believe they do unless the Big Data enthusiasts seek to remove data protection mechanisms, as they have successfully done in the English HASC Act 2012..

- What are the implications for individual anonymity of linking data across large numbers of databases?

Minimal if linking is done with encrypted identifiers (i.e. pseudonyms) and it is only used for specific purposes stated at the time it was originally collected into the contributing data bases. But the linked output is likely to be too rich to publish, unless major additional steps are taken to make re-identification more difficult, and may have in DPA 1998 terms to be treated as identifying.

- What is the 'public interest' in biomedical data? What benefits do we want to obtain? In what circumstances might the public interest take precedence over individual and minority group interests?

I believe that new techniques, such as pseudonymisation at source will mean that we will be increasingly be able to have the benefits without risking patient data confidentiality by using identifying data. In the diminishing % of cases where identifying data is required, *mechanisms such as the CAG should be used to assess the necessity to use it without patient consent, and where it is considered necessary, the net harm/ benefit of doing so.*

- What are the actual harms we should seek to avoid in using biomedical data (e.g. discrimination, stigmatisation)? What evidence is there of these harms having occurred?

We must ensure that the data is not use to disadvantage patients, e.g. by unreasonably denying them access to employment, services, products or benefits.

- In what ways does it matter if people's data are used in ways of which they are unaware but that will never affect them?

Please explain what is meant by 'will never affect them'. Just becoming aware that this is generally happening is would seriously erode the trust that is the basis of the patient-clinician relationship, which is definitely not in the public interest (vide the emphasis on "no surprises" in the 2003 NHS of Code of Practice Patient Data Confidentiality). I find it disturbing that it should be considered necessary to ask this.

- How are applications of computer-based technology (e.g. social networking, image sharing, etc.) affecting concepts of privacy, identity and social relatedness? How are related behavioural norms influenced (e.g. willingness to share and publish data)?

Not as much as one might expect, see 'Fair Shares for All', in particular Chapter 4 and references 47-63.

- Would it be helpful to treat biomedical data as 'property'?

I don't believe so; as most biomedical data forms part of patients records which are about the interaction between the patient, clinicians and possibly carers, not just about the patient. Clinicians need patient data for a variety of purposes besides patient care, and so co-ownership would be necessary. This would complicate matters for no obvious gain over the present arrangement, which recognises separate data subject and data controller roles, and that a record may contain information about multiple data subjects, such as carers, children & siblings in healthcare records. The notion of 'data ownership' is favoured by those who wish to make personal data a marketable good, and who see this as supplementing, perhaps even replacing, the need for data protection legislation. I am not aware of evidence that suggests that such a move would make it easier for third parties to access patient data for secondary purposes, or patients more willing to share their data. It is more realistic and in accordance with the particular nature of data (e.g. it's infinitely replicability) to speak of people's rights over, and responsibilities concerning it.

On the other its use in a commercial environment should be for a fee which is used by the data provider, e.g. the HSCIC, for the benefit of all and not just the making a profit at the HSCIC.

3 What is the impact of developments in data science and information technology?

Holding biomedical data electronically, in digital form, allows their efficient storage, retrieval, transmission, replication, and manipulation. The main developments that make our inquiry timely are those associated with the unprecedented availability of digital data, and the increasing power of computing technologies and analytical techniques available to manipulate it. Combined with technologies that generate large data sets (such as genome sequencing), advances in fields such as database design and management, artificial intelligence, bioinformatics, statistical methods, machine learning, and visualisation all offer powerful and rapid ways to 'mine' large, varied and complex data sets for significant patterns. As a consequence of the growth of digital biomedical data, new opportunities are emerging to correlate data from diverse sources, including research collections, disease registries, and even social networking, internet browsing and geolocation data. So-called 'big data' technologies offer new approaches to enquiry that promise insights that could not be derived previously. Extracting value from rich data resources has become a priority for the knowledge economy. We are interested in the significance of developments in areas of knowledge through use of biomedical data, especially data that were initially collected for a limited purpose (e.g. a medical diagnosis) but that could now be reused for a further purpose (e.g. biomedical research), or many other purposes.

Consultation question 3:

What is the impact of developments in data science and information technology?

Possible aspects to consider:

- To what extent and in what ways has the availability of biomedical data and new techniques for analysing them affected the way in which biomedical research is designed and funded? Is there any evidence that these factors have affected (or are likely to affect) research priorities?

I doubt whether health related research priorities would be affected, but the new techniques available and being developed to analyze structured and unstructured data (including images, sounds, social application data and scanned documents, viz the work of Autonomy, using Hadoop technology, etc) will create new opportunities for research.

Realization that data is now available for linking, thus offering an opportunity to obtain more complete episode of care data, which in turn will give a better basis for establishing which care processes, pathways and products work best. It will also enable earlier detection of major health issues, and better forecasting of care service use for individual patients, offering opportunities to improve outcomes and reduce care consumption.

- What are the main interests and incentives driving advances in data science and technology that can be applied to biomedical data? What are the main barriers to development and innovation?

The vast increase in the amount of personal data in digital form in healthcare records and social applications is exciting people, as is the increasingly 'always connected' nature of society (via tablets, phones, wearable devices, etc). Data quality (e.g. data incompleteness, errors and the multiple ways in which the same data can be represented in digital storage) *are a barrier to melding data from different sources, but see answer above, as are the non- universal use of unique identifiers, such as NHS number (although this is steadily improving).*

- Does ‘big data’ need a more precise definition or is it a useful concept in the life sciences even if loosely defined? Has enthusiasm for ‘big data’ led to over-inflated expectations on the part of governments, researchers and/or the general public?

Big data is a useful concept in the life sciences.

It is in the early part of its life cycle, and its reputation is currently more hype- than evidence-based. Its true significance has yet to be established but I believe it will be very significant. ‘Big data’ is unlikely to remove the need for projects collecting specific patient data.

- What are the significant developments in the linking or use of biomedical data, including any we have not mentioned, to which we should pay attention in our deliberations?

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

4 What are the opportunities for, and the impacts of, the use of linked biomedical data in research?

Life sciences research in general, and epidemiological and biomedical research in particular, are beneficiaries of the capacity to create and analyse large and complex digital data sets. Significant developments include the generation of data sets of unprecedented size in areas such as genomics and medical imaging. These large data sets are being created in many countries and used in different ways. Biobanks around the world hold, for example, DNA markers, information on lifestyle and environmental factors, and other health-related data (e.g., imaging data, laboratory test results and other quantitative data) from millions of individuals, and are a very valuable resource for medical research. The data collected when people are recruited to biobanks can be linked to pre-existing data, for example from health records, administrative databases or disease registries. These data sets are made available for research into the determinants of multifactorial diseases – such as cancer, diabetes and coronary artery disease. Researchers using these data may link them with other pre-existing data sets to answer new research questions (e.g. linking fertility treatment and childhood cancer registers to identify childhood cancer risk following assisted conception). Research groups and biobanks are also combining their data sets through national or international consortia to tackle research questions not answerable by one study alone. These collaborative efforts take place in both the private and public sector, with academic groups working alone, or with commercial or non-profit collaborators.

Access to data is dependent partly on the types of data held; anonymous data may be openly available on the internet while potentially identifiable data will require approval by an access committee and acceptance of rules guiding its use (including requirements for consent). Some data may not be available outside the original study. This may be true of commercial companies that have responsibilities to shareholders. It can also be true in academic studies where researchers do not wish to share data or consent has not been gained to allow use beyond those who gathered the data. There is increasing pressure from patients, participants and funders to ensure data is being made available. Commercial companies are responding to patient calls to release clinical trial data and many funding bodies now require researchers to specify plans for data sharing.

Comment [IH8]: Only 'effectively anonymised' data could be made openly available on the Internet, and with the increasing availability of other data (including identifiable data, such as the UK electoral register) only the simplest of data about individual, especially where the dataset is not about a random sample, could be said to pose a sufficiently low risk of re-identification to be considered 'effectively anonymised's

Comment [IH9]: What evidence do you have for this? I see much pressure from the research lobby (from Wellcome downwards), Big Data aficionados, some from commerce, and a lot from Government (who see it as a major cure for the UK's economic ills) but patients, where they are interested at all in the issue, seem as concerned with ensuring the confidentiality of their data as sharing it more widely for secondary purposes.

Comment [IH10]: This is because clinicians, patients and the public have become aware of the selective suppression by companies of trial data that does not show products in a favourable light: it has little to do with new research.

Comment [IH11]: Again I think you'll find this is due to that a few significant and well publicized studies that have turned out to be fraudulent. It

Different governance frameworks, consent arrangements, national jurisdictions, and political priorities condition what research is carried out, what data are collected, in what ways data can be linked and by what means, and for what further purposes data may be accessed. Our interest here is identifying how these systems align with underlying ethical and social values.

Consultation question 4:

What are the opportunities for, and the impacts of, use of linked biomedical data in research?

Possible aspects to consider:

- What are the hopes and expectations associated with data use for biomedical, public health and life sciences research? What are the main concerns or fears?

Apart from using genome sequencing data, the hopes and fears of patients and the public are as they are with other particularly sensitive data, and probably less than the fears they have about the confidentiality of data about their sexual and mental health and care.

- To what extent do the kinds of collaborations required for data-driven research (e.g. international or multi-centre collaborations) generate new ethical and social issues and questions to those in other forms of research?

Only where they involve the movement of identifiable individual data across boundaries that correspond to significant changes in data protection legislation and/or practice. Where patients consent to the transfer of data across such boundaries after being adequately informed of the risks to their confidentiality, or only research results are collated across them (as happens in many studies), no new issues are raised.

- Should researchers be required to allow others to access data they have collected for further research?

If they are, and identifiable data is used, patient consent will be required for all the purposes involved. If the 'further research' is solely to verify the results obtained by the original users, consent will not be needed for additional purposes. But if consent is not obtained at the outset for more than one research team to use it, then additional consent must be sought for their use.

- What sorts of concerns are raised when research is carried out by a commercial firm?

People in general believe that commercial firms are less likely to behave ethically than NHS or academic institutions, and are generally considerably less willing to provide their individual data to commercial firms for research purposes (see 'Fair Shares for All', 4.7.3).

Also obtaining information about how the data is used may be difficult if the firm concerned claims that it is commercial in confidence.

All these aspects need to be sorted out by the data controllers providing the data and the firm(s) and the results recorded in a data sharing agreement before use by a commercial firm is permitted.

5 What are the opportunities for, and the impacts of, data linking in medical practice?

As in research, the use of digital data is also expected to produce significant further transformations in health care. Increasing quantities of data are generated and recorded in the course of routine medical practice, including a variety of imaging and pathology test results (increasingly including the results of DNA tests), and metadata. These data support increasingly stratified or 'personalised' clinical decisions (adjusted to a greater number of the specific characteristics of the individual patient). Such data might be used to profile patients for particular interventions (for example, to tailor drug dosages to their individual metabolic response) or to predict future disease. The collection and retention of health data raises questions about how we should approach the possibilities for making use of it, not only to optimise individual medical treatment but also for wider public benefit. It is also important to consider how individuals should be able to control access to records about themselves to protect their privacy or to conceal sensitive information that they do not wish others to see.

Consultation question 5:

What are the opportunities for, and the impacts of, data linking in medical practice?

Possible aspects to consider:

- What are the main hopes and expectations for medical practice associated with increased use of linked electronic data? What are the main concerns or fears?

It is expected that linking patient data from many sources will provide a much better basis for forecasting individual patient needs for care, and planning for their satisfaction in the most economical way possible (e.g. by using interventions in the community to avoid expensive hospital admissions / attendances at A&E departments). It will also provide a basis for comparing the efficacy and true cost of various kinds of interventions and care pathways in different clinical circumstances. This is a key weapon in the fight to ensure that the NHS remains sustainable in the face of increasing demand for a wider variety of care from an aging and generally less fit population. The fears of some clinicians and patients is that this will done in a way that compromises the confidentiality of the patient- clinician relationship, and that it may damage the personal quality of care. The truth is that it can be done with data that has been pseudonymised at source rather than identifiable data, and re-identification of any patients found to require a revised care plan can also be done at source.

- What can be said about public expectations about the use of health care data, in terms of appropriate use, information and control? To what extent would members of the public expect health care data to be shared with other agencies or bodies?

Patient attitudes, expectations and concerns are discussed is depth in chapters 4 and 5 of 'Fair Shares for All' q.v. and the references on which it is based. In summary patients and the public are content for – indeed expect - clinicians caring for them to share relevant data with other clinicians caring for them as required: Most would like a facility to flag very sensitive items in their record as not be shared without their explicit consent. As the recommendation is that this should only be done with the clinician's agreement, it is unlikely to be used hide data about abuse or that would affect treatment options. While *most patients would be pleased for their data to be used for research purposes, they wish to be asked for consent to use their identifiable data, although a minority would be happy for this decision to be taken on their behalf by an independent advisory group such as the Confidentiality Advisory group. A majority are happy for their de-identified data to be used for secondary purposes without consent, but a substantial minority doubt that data can be effectively de-identified. Most patients have little or no awareness of who uses their data for what secondary purposes, and whether the uses involve identifiable data or not.*

- Is there potential for privacy controls to hide secrets, such as abuse, or to disadvantage people in unintended ways (by preventing best treatment, perhaps)?

See preceding answer.

- Are there particular issues raised by 'risk-profiling' where individuals at high-risk (e.g. of type 2 diabetes) are identified and approached for specific interventions? What might make the difference between this being intrusive and it being supportive?

Risk profiling should be done in-situ, or by others acting as DPA 'data processors' for the patients' clinicians. But it raises complications, as the process may involve linking the data held be two or more care providers (tho' this doesn't require identifiable data). Who is the data controller of any linked data? The HSCIC? Who may see it? If the data is identifiable, is patient consent required so that a clinician in one care provider can see information collected by someone in another, or does the implicit consent to share assumed when the patient seeks care cover this situation? Ideally patients should be informed beforehand that they are likely to be profiled, and will offered specific interventions if they are likely to be of benefit to them.

- What are the implications of episodes of treatment across different care providers being used routinely as research data? How might this affect the ethical basis of the doctor-patient relationship?

There is no reason why data for all providers involved in an episode of care cannot be linked (providing there is consensus about what constitutes the start and end of an episode), and identifiable data is not needed to do this. If non-identifiable data is used and the linked data does not to offer a significant implicit risk of re-identification, then it may be used for any secondary purpose where the context of use does not create a significant risk of re-identification. As above patient consent should be sought before identifiable data for a complete care episode is used for secondary purposes

- To what extent does the possibility that biomedical data can contribute to a research base to advance the effective treatment of others create a moral obligation to allow them to be used in this way? What might limit this obligation? How should we regard (and provide for) those who refuse to allow their data to be used?

It all depend on what you mean by '..can contribute to a research base'. Is this a decision about ongoing contribution, authorized once at the start. Are you only talking about identifiable patient data?

There is no 'contract' in place which states that NHS treatment is contingent on the patient accepting that his data will be put into a research database, but many patients would accept a moral obligation to provide it. In general patients are in favour of their data being *used for research, but are keen to be asked for consent where identifiable data is required (which is may not be easy to do at scale). Where data is effectively anonymised, the majority of patients consider that patient consent is not required and but would like to know who is using it for what (which is not easy to do effectively: some patients will be very unhappy when even effectively anonymised data is used for particular purposes or by a particular body). Non-consent is only likely to be significant where patients and/or their clinicians do not trust the body(s) proposing and/or carrying out the collection, such as NHS England and the HSCIC. Causes of this include, but are not limited to:*

Patients are not given full & accurate fair processing information at the outset

Scope creep occurs, e.g. of the data collected, the purposes it's for and the user set

The stated purpose(s) for which the data is being collected are very generic

Identifiable data is required by the end users

There appears to be little control of how end-users may use the data

Only a very generic description of the end users is provided

The data is to be, or may be used, by one or more commercial organization

I do not believe that any action should be taken against patients who do not contribute data collected about them for research, other than periodically reminding them that they can contribute their data if they wish to and how to do so.

Some clinicians caring for patients, especially GPs, have similar views to their patients. For example GPs are generally unhappy with the 'opt out' method used to control the collection of identifiable data (care.data) for a wide range of secondary purposes by a wide range of bodies requested by NHS England. At least two GPs plan to opt all their patients out of care.data, only opting each in if the patient agrees at their next contact.

6 What are the opportunities for, and the impacts of, using biomedical data outside biomedical research and health care?

Most detailed human biomedical data are collected and stored for purposes of either health care provision or biomedical research. The circumstances in which they are collected often entail clear restrictions on their further disclosure, deriving from requirements of medical confidentiality, consent and data protection. Aside from medical practice and research there is also a range of other purposes for which biomedical data may be used, including uses as diverse as criminal investigations, genealogy, and marketing. In many cases, access to the data will be governed by specific measures (as in the case of forensic purposes). However, not all biomedical data are collected in a setting of medical confidentiality: the information may be offered voluntarily, for example, where it is required in order to receive a service such as insurance. Individuals may also provide biomedical data to commercial companies, for example, when purchasing direct-to-consumer health tests, in some cases (unwittingly or deliberately) in circumstances allowing it to be used for other purposes, such as marketing by the company or its affiliates. Data may be stored in private online health records, shared through patient support websites or mobile apps, or self-published on blogs, message boards, or social networking websites (such as quantifiedself.com or even Facebook). Finally, predictive analytic techniques can, in some cases, impute biomedical data (for example, pregnancy inferred from supermarket purchasing patterns), to inform targeted marketing of products and services. Here we are interested in the implications of the use of data not being definitively constrained by the purpose for which they were initially collected, and the possibility of unforeseeable subsequent uses of the data.

Consultation question 6:

What are the opportunities for, and the impacts of, using biomedical data outside biomedical research and health care?

Possible aspects to consider:

- What are the main hopes and expectations associated with the wider use of biomedical data (outside biomedical research and clinical practice)? What are the main concerns or fears?

Companies seek better targeted marketing of care related products & functional aids, including healthcare, medicines, devices of all sorts, insurance. More accurate risk assessment of personal suitability for credit provision, personal insurance of various sorts, employment. Some generic work can be done with de-identified data that is related to anonymous purchase data, but better targetted marketing depends upon knowing at least some of the properties of the possible purchasers, and ideally their identity.

The main concerns / fears are that people who are less fit / at higher risk of disease and/or who have functional impairment will be discriminated against if identifiable biomedical data about them is widely available outside clinical practice & academic research establishments.

- What factors are relevant to determining the legitimate scope of further uses of biomedical data? For example, should it be restricted to a ‘compatible purpose’ (and, if so, how might this be defined)? To uses that are in the 'public interest'? To use only by public authorities (and those providing public services under contract)? To non-commercial or non-profit uses/users?

As you point out, how do you decide whether a purpose is compatible with the purpose for which the data was originally collected / shared? In general, data that is identifying should not be distributed beyond its origin without patient consent, unless an independent expert body such as the Confidentiality Advisory Group considers that the public interest is on balance best served by doing so and further sharing by the new data controller is banned.

- What are the ethical implications of using predictive analytic tools with biomedical data outside health care and research (e.g. in recruitment or workforce management)?

Nothing new - health assessment is already mandatory prior to and during employment in the armed forces, public service driving, and for airline pilots, divers, astronauts, etc. The key fact is that these form part of the accepted recruitment process for these occupations, and should be conducted in a way that is clearly visible to the would be recruit, who will have to give his consent for access to his GP record, possibly taking a physical examination, etc, etc. It is also relevant to insurance taken out by employers on their staff for the benefit of the former, a practice now frowned upon but commoner in the USA than here.

- Would the ability of individuals to maintain direct control over the use of data about them be likely to affect the range of further uses to which they would allow the data to be put?

In an ideal world, yes, In some cases it may not be possible to seek further consent – the patient may be dead, have moved, etc. And the effectiveness of any de-identification proposed cannot be readily assessed without knowing the intended context(s) of use (notably by whom, for what and under what constraints), which even a data controller may find difficult to do. This is one of the tasks best handled by the Confidentiality Advisory Group or a more local equivalent, and the major reason for setting up its predecessor, the NIGB Ethics and Confidentiality Committee. The need for further data subject control will be greatly reduced when people realise that data that has been pseudonymised at source can be used for almost all data driven studies that only require existing linked or unlinked patient data which removes it from the purview of the DPA 1998 and the common law obligation of confidentiality..

- Should individuals be able to profit from the use of their biomedical data (e.g. by selling access to the data to commercial companies)?

This can only be done if the individuals are considered to 'own' clinical data about them, which for reasons explained in the answer to q2 bullet 7 is not really a good or practical idea.

7 What legal and governance mechanisms might support the ethical linking of biomedical data?

Knowledge of the identity of individuals is not always necessary for data to be used beneficially: many kinds of research can be undertaken using anonymised data and public health measures can be applied to populations or subgroups generally, rather than to individuals. The same is true of other public and commercial services. However, we are interested in cases in which strict anonymisation is either impossible (because of the possibility of identifying individuals within data sets) or the use of anonymised data is not feasible (for example, where clinical characteristics and biological data are being linked, as in the case of linking different disease registers). We are interested in how the risks of identifying individuals can be avoided or mitigated, in the context of incentives to collect ever more data, extract secondary value out of existing data collections, and to increase the efficiency and power of research.

In health systems, confidentiality usually creates a barrier to the broader disclosure of data while allowing it to be used to support the provision of treatment within the system. The stipulation that data use must be subject to consent, which is a legal requirement in most jurisdictions, allows individuals to exercise control over the use of data about them. However, consent operates at different levels of specificity and requires different levels of commitment. When research data is collected it is increasingly common to ask participants for 'broad' consent to a wide range of types of health-related research; this model is usually used because the full scope of future research studies is not known at the time of recruitment. Instead participants can give consent knowing that governance mechanisms (e.g. oversight committees, data access rules) will be in place to ensure their data is accessed by only approved researchers who agree to use it for scientifically valid, ethical and appropriate research activities. In fact, consent may be neither necessary nor sufficient to protect individual privacy.

Some linking is possible using 'pseudonymised' or key-coded data, where a code may be assigned to data in a recognised safe haven (or by a trusted third party) before data are more widely released. Alternatively, the data can be interrogated using algorithms supplied to trusted third parties and applied by them, with the commissioning party (a biomedical researcher, for example) receiving only the results and without them seeing the raw data at any stage. These safeguards are not perfect; as more data become available through a variety of sources, protected and open, it has been shown that identification of individuals from 'anonymous' data is possible. This has led some to suggest that a guarantee of confidentiality is increasingly implausible and that genetic data, for example, should be openly shared, with regulations and sanctions put in place for misuse; others are exploring additional means by which data can be protected, such as ever more secure encryption. It is unlikely that a single approach will be appropriate to all contexts, although there is value in approaches being consistent (in terms of underlying values) and interoperable (given advantages of linking data between contexts).

Consultation question 7:

What legal and governance mechanisms might support the ethical linking and use of biomedical data?

Possible aspects to consider:

- What ethical principles should inform the governance of biomedical data? For example, should the principle of 'respect for persons' be given primacy here? How might this relate to principles such as solidarity and tolerance?

See section 4 of the Introduction to my response.

- Does the use of linked biomedical data require distinctive governance arrangements compared to the use of other personal data?

The linking of data can be done quite satisfactorily with pseudonymised data (which would not be regarded as identifiable and therefore subject to the DPA 1998 and other controls). The resulting linked dataset may be so rich that it offered a significant enough risk of re-identification to become identifiable in the eyes of the DPA 1998. This is especially true if very rich data is collected prospectively for largely unknown secondary uses and end users, i.e. the data centre is running a large data warehouse. Tighter controls are therefore required over the linked output and its subsequent consumption by the end users to render the risk of re-identification acceptable. These are made more necessary by the fact that cleansing and linking data are non-trivial tasks requiring significant expertise, with the result that relatively few bodies will be equipped to do it.

But to ensure that the risk of re-identification stays low, further constraints are required on the activity of the data centres and the end users including an independent information governance body that involves all stakeholders with a majority of members representing the source data controllers and data subjects, i.e. clinicians and their patients. The IG body should be responsible for assessing the need for, and IG risks of, all proposed data collections and end uses of the collected/linked data, before deciding whether they should be permitted, and if so what constraints should be placed on them.. It would also ensure that identifiable data was only collected (or released) in the increasingly rare cases where the proposed use(s) could not be achieved without it, obtaining consent was not feasible and the public benefit outweighed the potential public damage and risks of breaching patient confidentiality.

- Are the current principles of consent – including the principle that consent can be withdrawn – still 'fit for purpose' in relation to the linking of biomedical data?

It is possible to implement the withdrawal of consent in a data centre that collects and links patient data, although it might be cumbersome and the data subject would have to accept that analyses using his or her data before consent was withdrawn would not be affected by it. The more difficult task would be ensuring that end users to whom it had already been distributed make no further use of it and destroy their copies of it. Making end users use the data at the data centre itself would make the problem significantly more tractable. I do not envisage that much -if any - linked data would be published for anyone to use as they wish: if it is, withdrawal becomes meaningless where would-be users are allowed to copy the data at will without registering with the data centre providing it..

- What level of continuing involvement is it reasonable to expect individuals to have in how their data are used after they have been collected?

Other than via the data centre governance mechanism outlined above, it is difficult to see that individual data subjects can be offered any control over how their data may be used. This is the reasoning behind the need for an effective and trusted independent governance mechanism for the data centre, representing all stakeholders but with the data subjects & clinician data controllers are in the majority.

- Should there be an opt-in or an opt-out system for people to decide whether to allow their personal medical data to be used for public benefit?

Where Identifiable data is being sought, the best solution is an opt in, as this ensures that patients are given the information needed for properly informed consent / dissent. This is particularly important where the prospective users and uses of the data are only defined in the most generic of terms. Having said, it may well be that consent in such ill-defined cases would not be regarded as sufficiently informed to be valid. Using an opt out is much less satisfactory still, as the data controller, as in the care.data example, is being forced to assume that the patient goes along with the answer that NHS England and the HSCIC want and there is no guarantee that the patient has digested the information necessary to make a decision. It is also uncertain whether the practice would survive a challenge under the UK Human Rights Act 1998.

I would however emphasise that pseudonymisation at source should remove the need to collect most identifiable data.

- Under what conditions ought individuals to be content to delegate authorisation of the use of health and biological data about them?

This should be when and if patients believe that there is an effective independent body on which they are copiously represented and which has the necessary expertise to determine the (rare) applications by someone to use unconsented identifiable or potentially identifiable patient data for secondary purposes. NIGB's Ethics and Confidentiality Committee (ECC) was such a body, and it is to be hoped that its successor, the Confidentiality Advisory Group (CAG), will prove itself fit for the task. However the CAG, like the ECC, can only recommended a course of action to the SoS, not determine it.. Better still, maximum use should be made of the evolving techniques that enable the use of pseudonymised instead of identifiable data for almost all purposes. This should greatly reduce the need to use for the CAG's services.

- What role should public engagement and democratic processes play in the determination of governance measures? In what circumstances, if any, might the outcome of democratic procedures mandate overriding individual interests?

This has already happened with the implementation of the DPA 1998 and the implementation of section 251 of the NHS Act 2006 and the creation of the ECC/CAG. The DPA gives a list of purposes where personal data can be used without consent where the public interest is considered to outweigh the duty of confidentiality the clinician owes his patients, see DPA 1998 Schedule 2 and 3 and section 35.. The Secretary of State (SoS) can decide, and the Health Research Authority Confidentiality Advisory Group can recommend *to the SoS, after considering the need to use identifiable data, the practicability of obtaining consent, the overall effect on the public interest and the impact on the patients concerned, whether or not he should permit a given set of unconsented identifiable patient data to be used by a specific body for a given secondary purposes.*

This arrangement, which worked well despite the protestation to the contrary from some secondary users lobbyists, has been damaged by the provision of extra and unnecessary gateways in the HaSC Act 2012 which enable the Secretary of State, NHS England, Monitor, CQC and NICE to direct the Heath and Social Care Information Centre to collect identifiable data from any provider of NHS-funded care without the need to seek patient consent or review by the CAG. These powers invoke section 35 of the DPA 1998, and have been used (unnecessarily) to request the identifiable GP patient data that will form the basis of the care.data dataset.

The damage has been greatly enhanced by the abolition of the National Information Governance Body for Health and Social Care (NIGB), the independent body responsible for the formulation of IG policy in the NHS and the monitoring of its implementation. Taking its advice during the drafting of the HaSC Act 2012 would have minimized the damage done, and taken into account the needs of new bodies such as Clinical Commissioning Groups and Commissioning Support Units .

- What inconsistencies exist in current ethical guidance and governance structures relating to biomedical data?

See the preceding answer

- What examples are there of innovative initiatives that promote privacy while encouraging

Pseudonymisation and identifier encryption at source are now being used routinely on an increasing scale to extract linkable de-identified data from multiple sources, and both open source and commercial sotware are available to do the job. For more about these techniques see 'Fair Shares for All', particularly Appendices H & I. The resultant linked data may well be so rich as that it poses a significant risk of re-identification, and so further constraints, e.g. extracting a random sample rather than the whole population and only permitting subsequent processing in a controlled environment are necessary to reduce the likelihood of re-identification. With these techniques, it is very difficult to re-identify individual records other than at the original data sources. With more elaborate key management, the same techniques can be used to incrementally generate linked data for the same cohort over long periods of time.

Glossary

Algorithm: a method for calculation or problem-solving in a finite set of steps, which may be automated.

Anonymisation: the removal of personally identifying information from a record in order to prevent the identification of an individual subject of the record.

Apomediation: the involvement of agents (individuals, groups, tools, or processes) to guide consumers in the acquisition of products or services that fulfill specific criteria (e.g. quality) without acting as a gatekeeper (i.e. the use or agreement of such agents is not a pre-requisite to access).

Big data: a relatively informal term used to describe data sets large or complex enough to pose problems for traditional methods of storage, management and analysis, thus necessitating the development of new techniques to replace and supplement those methods. The term is increasingly used to characterise approaches employing these new techniques and methods to extract value out of 'big data' resources.

Biobank: a repository of biological material samples and associated data for research use. The content of biobanks varies, but may include tissue, blood, and urine samples, and also information associated with the sample donors (such as phenotypic or lifestyle information).

Bioinformatics: the scientific field of processing, storing, distributing, analysing, and interpreting biological data. Encompasses many disciplines, including biology, computer science, and engineering.

Biomedical data: Data relating to the state or functioning of human beings as biological systems, e.g. those derived from biological research or clinical investigations.

Citizen science: the inclusion of non-scientists in the scientific research process.

Cloud computing: provision of computational capacity as a virtual 'resource pool', available over a network connection.

Confidentiality: the restriction of access to certain data. This may derive from an agreement or implicit understanding between two or more parties, or from the nature of the relationship between them. For example: a contractual agreement to restrict access to commercially sensitive data, or the doctor-patient relationship (the confidential nature of which is assumed and obliges the doctor to keep secret some information imparted while providing care.)

Data set: a collection of data in a structured format, such as a database.

Epidemiology: the study (or the science of the study) of the patterns, causes and effects of health and disease conditions in a defined population.

Ethics: the study of values and moral reasoning, and their application to human conduct; formally a branch of philosophy.

Genome: the full complement of genetic material (or hereditary information) in the cells of an individual organism or species; the totality of the DNA sequences of an organism or organelle.

Genome sequencing: the biochemical procedure by which the complete DNA sequence of an organism's genome is determined.

Health record: a transcript of information regarding a patient's medical data, often from multiple sources and over time (as distinct from a 'medical record' which normally concerns only a single episode of care at one institution).

Machine learning: a branch of computer science concerned with developing techniques to enable a computer program to improve its own performance.

Medicine 2.0: a term used to describe the relationships between medicine and 'Web 2.0' (second generation internet services, characterised by greater web-based participation, interactivity and collaboration, and the move towards these behaviours and away from the use of 'static' webpages; exemplified by wikis and social networking). Medicine 2.0 therefore refers to the possibilities those methods and behaviours have for the delivery of health care.

Metabolomics: the systematic study of metabolic responses to external stimuli.

Metadata: data about data. For example, data attached to a digital picture describing the time of its creation, or data describing the structure of a database.

Omics: an informal term used to describe biological fields of investigation ending in '-omics'. It refers particularly to fields concerned with complex, integrated biological systems, the study of which is intended to lead to an understanding of the organism as a whole. Examples include: genomics, metabolomics, microbiomics, proteomics, and transcriptomics.

Predictive analytics: the use of data science techniques to predict future events in a variety of fields (business, health care, insurance, sports etc.). Draws on many disciplines, including statistics, machine learning, and data mining.

Privacy: the right (or ability) to acquire and maintain freedom from intrusion or public attention. It can relate to a number of different conditions, such as physical seclusion/concealment, freedom from external interference, or positive control over access to one's personal information.

Proteomics: the systematic study of all the proteins encoded by the genome of an organism.

Pseudonymisation: the process of removing identifying associations between data and the subject of that data, possibly including the replacement of those associations with artificial identifiers.

Statistical imputation: the process of identifying missing data and replacing them with substituted values.

Telemonitoring: the monitoring of patients at a distance (i.e. not in the same location as the health care provider). Data from the monitoring devices are transmitted for storage, analysis and recording elsewhere. For example, values such as blood pressure, heart rate, and blood glucose levels can be monitored regularly or continuously using home-based, wearable or even implantable devices.

Visualisation: visual representations of data, primarily for the purpose of rapidly conveying possible meanings of complex data (i.e. the animated representation of a sifting algorithm.