This response was submitted to the consultation held by the Nuffield Council on Bioethics on *The linking and use of biological and health data* between 17 October 2013 and 10 January 2014. The views expressed are solely those of the respondent(s) and not those of the Council.

1. Do biomedical data have special significance?

Biomedical data that reference individuals are, and have always been, a special class of data subject, in general, to restricted access. When biomedical data consisted primarily of patient records in GP's offices and hospitals, and were mostly paper based, managing such data was as easy as controlling physical access to the files. With digitisation making biomedical data accessible (in principle) to anyone with an internet connection, and with the advent of large molecular datasets the ethical concerns have changed in that there is now a much higher likelihood that such data may be accessed by someone without the proper permissions. Furthermore, it is clear that 'omics data, are more useful when aggregated so health care providers are keen to collect large data sets for population-level analyses.

Such analyses, used for instance to identify disease-causing genetic differences, genetic differences that affect drug metabolism, and changes in the genomes of cancer cells, have led to improved prevention, treatment, and diagnosis for many patients. Such benefits are likely to increase steadily as more data from more individuals is aggregated and studied, and this is a real revolution in medical treatment.

Genomic information is often classified as being special because it is unique – i.e., the data contain enough information to specifically identify a person. This does not mean it is a straightforward task to take this information and then identify a specific person in the population from it. This distinguishes it from information that is potentially less unique but easier to use for identification (e.g. post codes). The uniqueness of genomic data is a feature shared with other (potentially as widespread) datasets, such as RNA or protein levels; these are often grouped together as ‘omics data.

The real feature of genomes (as well as transcriptomes, proteomes, etc.) is that their information content is unique to a single individual. This characteristic is shared by traditional single dimensional biometric data types, such as fingerprints and iris scans, but is also shared by other non-biometric multidimensional, and indeed mundane, aspects of human behaviour such as typing and driving speed patterns.

The combination of such datasets associated with each individual is likely to increase, and to become unique, in particular if the datasets are rich enough to be interesting also for health care research. We believe that over time the “special position” of genomic information will progressively merge with the more general problem of providing access to all types of information that uniquely identify an individual. The ethical framework is likely to require a switch from trying to guarantee that the information does not uniquely identify a particular individual towards preventing inappropriate use of information by researchers, or to inadvertent use.

With respect to genomics it is clear that family members may share an interest in the findings for a particular individual, but other less unique types of information such as postcodes and religion, albeit with less certainty of transmission, also share this characteristic. The clinical genetics community has already had to navigate complex between-family-member scenarios for the release of information, and also the discovery and informing of unexpected findings, e.g., of misattributed parentage. It is worth using the long experience of this community to help structure how the release and use of information is propagated amongst family members.

2. What are the new privacy issues?

Biological “big data” like traditional paper-based medical records, may contain private information that patients do not wish made public, and that medical systems by custom keep private. “Big data” are different in that, by definition, they are held in a computational infrastructure and are therefore more accessible, and in the scale of information – including both constant lifestyle measurements as well as unique information such as the genome.

Big biological data can be separated in two types: molecular data such as genomes, transcriptomes, etc. which are unique to an individual, but which are difficult to analyse and therefore unlikely to be used for identification even if inadvertently released, and lifestyle and personal phenotype information that might be less unique but carry more risks if inadvertently released. We would advocate an approach that balances the risks of harm on identification with the benefits to society for the aggregated research.

The risk of harm has three components; the first is the protections (both technological and practice based) to prevent malicious use (eg, by signing agreements on research use) and minimising the likelihood of inadvertent release. The second is the ability to change a unique piece of information (genomic, transcriptomic, lifestyle, or another measurement) to identify an individual. The third is the harm which an individual would suffer if information can be assigned back to that specific individual. The benefits are society-wide in terms of better health care practice, better discovery for biomedicine and potentially lower costs in the health care system. Currently most focus has been on the first of the risks, and not the second two, and the case for benefits is often implicitly and not well argued. We believe there should be a more systematic and broader outreach about the benefits of more accessible data, while acknowledging and minimizing, but not eliminating, the risks of harm.

We note that the value of aggregate data is not changed if the data (e.g. a genome) from one or a few people are not included. Thus, provided that most individuals do consent to release of their data, which seems likely, the reluctance of a small percentage of the population to release their data will not affect research outcomes. Given the often widespread uptake of cohort studies in the UK (eg, BioBank, GoSHARE, NIHR BioResource) this means that an active consent process is likely to accrue most of the benefits.

Since most individuals do not have backgrounds in biological science it seems likely that most individual’s data will be used in ways that they are unaware of. This should be noted when consenting the data. In some ways this is already true of traditional medical records, where reporting of (de-identified) infectious disease cases is required by law in some jurisdictions, and where overall counts for numerous diseases are routinely collected for large areas and even entire countries.

Social networking and the sharing of information have become the norm for most people, and certainly for younger people. It seems likely that individuals who actively participate in social networks will be accustomed to the notion sharing data and will be less concerned by the aggregation of data for society-wide health care benefits (as they are comfortable for aggregation of data for commercial gain by companies providing “free” services to enable this aggregation, e.g., Facebook).

As we note above, biomedical data are functionally equivalent to the data in traditional medical records, which are treated as private information but not as property. It would seem sensible to continue treating newer data types the same way. If these data are deemed to be “property” we suspect that distribution for research and public health uses will be curtailed, or made more difficult, while having little additional value for the individual.

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

New technologies for DNA sequencing, transcriptomics, proteomics, metabolomics, and other ‘omics" methods have fundamentally changed the boundaries of what is possible for biomedical researchers. These technologies generate vast quantities of data that have enabled large-scale big picture experiments that look at, for instance, the genomic variation in whole populations, or the genetic and metabolomic structure of the gut microbiome over time. These technologies also allow very fine-scale experiments to examine ever-finer aspects of cellular and organismal biology, for example how gene expression or metabolic processes change over time or in response to specific environmental stimuli.

In response to these new technologies scientists are now designing experiments with broader reaching goals aimed at understanding more complex questions and funding agencies have responded with programmes that implicitly assume the use of "big data" projects. For example, the 2011-2015 BBSRC delivery plan explicitly lists data intensive science as a priority, and the other grand challenges: food security, industrial biotechnology, and fundamental bioscience to improve wellbeing; all assume the use of data intensive technologies to address these challenges.

Should "big data" be defined? This is difficult. For a single scientist the output of a single run on a next generation sequencing machine is "big data" since storage and analysis of this single dataset might exceed the physical capacity of the his or her lab's computational infrastructure and the lab may not have anyone with the expertise to deal with the data. At the other end of the spectrum, for large organisations such "big data" is much bigger and is measured in petabytes of disk, or in the output of thousands of next generational sequencing runs. For the single researcher or a large organisation data become "big" when they approach or exceed storage and analysis capacity.

The experience shared by single researchers, large organisations, and everyone in between is that at every scale our ability to generate data is growing faster than our ability to manage, store, and analyse those data. This, perhaps, is how "big data" should be defined: not as a quantity, but as a rate.

Among biomedical scientists this growth problem is a well known and well discussed problem. There are numerous large-scale initiatives to address the problem. Within Europe the ELIXIR research infrastructure is developing a large distributed infrastructure to store and analyse big biomedical datasets and, importantly, also includes numerous initiatives for training researchers (and clinicians) to understand and analyse big data. Other European initiatives include EUDAT, specifically aimed at dealing with the "long tail" of data that are not addressed by the larger infrastructures, as well as the Research Data Alliance, which is a worldwide effort to coordinate storage and analysis of large data.

141

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

The opportunities for linked data are enormous. From a public health and medical perspective big data presents opportunities in personalised medicine (genotype-based drug dosing); cancer diagnosis and treatment; and epidemiology (identifying and tracing infectious disease, data mining health records). There are a number of major challenges to real advances in this area. It is important to regularly restate the fundamental benefits of analysing large scale cohorts in discovery in biomedicine; without this analysis much of the health care that we have today would not be in place, and if we do not continue to enable and grow our ability to gather data we seriously jeopardise the chance of effective and efficient healthcare in the 21st century.

Commercial firms have played an important role in many aspects of biomedicine from the development of drugs to devices. Much of this development requires appropriate safety measures which often require large patient populations (clinical trials). Therefore many commercial firms by definition do large scale, “big data” cohort research now as part of their operations. This is likely to increase in the future.

When use of and access to cohorts is driven by a biomedical endpoint, there is little to distinguish commercial and academic research. However, there should be appropriate controls that commercial companies do not reposition the data outside of the biomedical context (for example, allowing the data to be combined with other information about individuals for targeted marketing).

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

Again, the fundamental benefits of cohort based research need to be regularly restated. Patient cohort based research is a key part of many of the medical innovations currently used (not lease the clinical trials of successful drugs) which has both extended life and improved quality of life for the entire population. In the coming century we will need to deepen our understanding of disease aetiology and treatment, with proactive screening, better diagnosis, better treatment and better on-going care. To do this effectively and efficiently we need to have large scale, information rich cohorts with both lifestyle, physiological and molecular readouts. A key concern we have is that the risk of harm to individuals is overstated whilst the benefits to individuals as part of society is understated.

In a number of health care scenarios there is already appropriate sharing of information between the health care system and other services; for example in the integration of health, social services and educational information for children. Already a number of cohorts have the ability to link beyond just healthcare data, such as the SHIP system in Scotland. We note that this further linkage includes less unique information that may nevertheless be easier to use for identification and therefore there may incur a higher risk of harm. However, biomedical studies in a number of areas (for example, dyslexia or extreme behavioural disorders) by definition will cross these boundaries. With appropriate controls to minimize risk, this linkage seems appropriate.

The use of this information for identifying “high risk” individuals is effectively a screening procedure. There is extensive research and experience on how to assess and model effective screening for public health, which requires a careful analysis of both the benefits and harms (for example, even a low false positive rate can mean a well meaning screening program does not have population level benefits). Any use of information as a screening mechanism should be rigorously assessed, and only put into place after such an assessment.

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

There are two levels of wider use of biomedical data. The first is the use of biomedical data with respect to associated components of the public services, such as social services and education. Where it is in the public benefit there should not be a strong boundary between “biomedical data” and “other social data”, though noting again that other social data might be less unique but carry more risks on inadvertent release.

A special case is the use in forensics. As more biomolecular (in particular genomic) information is determined for health care usage it needs to be clear to all parties – patients, health care providers and forensic officers that there are limits to the use of health care data. We are unclear on the current status of health care data with respect to criminal or civil investigations; the presence of larger amounts of DNA will allow new potential forensic routes if desired. This needs to be examined and debated.

The second level is the use of this information outside of the public benefit, such as in commercial companies to help improve the profiling of individuals. This is inappropriate and not desired.

The question of predictive analytic tools for aspects such as recruitment is presumably something which interacts with discrimination of people by other means. This data should be seen in the light of discrimination policies and laws.

Once an individual’s data has been used in aggregate in analysis, it is effectively impossible to remove that information inherently from aggregate analysis.

As people routinely exchange personal information for access to other goods (in particular in social networking and many commercial offers) it seems impossible to ban individuals from profiting from their own data. However, we do not think that direct monetary profit is the best motivation for individuals to provide information for the public good, and the broad uptake of many cohorts stress this.