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.

Centre for Longitudinal Studies Response

1. Do biomedical data have special significance?

1.1 It is not straightforward, or particularly useful, to define biomedical data as a distinct class of data. Biomedical data is so varied in its content and form that it is difficult to draw clear boundaries around what would constitute biomedical data (e.g. does it include self-reported subjective measures of health and well being?). Biomedical data is also very heterogeneous - arguably varying from an MRI scan image (which is extremely complex and detailed, and which an individual would have no access, to unless he or she engaged in a research study or had a medical reason for being scanned) to the answer to a question on self reported general health. The heterogeneity of biomedical data make its utility as a category of data questionable - a more specific and narrower definition might be more useful.

1.2 Arguably some forms of biomedical data fall into a specific subcategory because they reveal information about an individual that the individual is unlikely to know about himself or herself. In addition if this data is directly relevant to the capability of the individual to carry out tasks relevant to employment and daily living it could be particularly sensitive.

1.3 In recent years a great deal of use has been made of the linked genotype and phenotype data in the 1958 cohort, much of which has resulted in published research providing important insights into individual biological variation and health. In addition the 1958 British Birth Cohort Study provides control cases for research using the Wellcome Trust Case Control Consortium data. For example a few examples of papers published in 2013 include:

DI BERNARDO, M.C., BRODERICK, P., HARRIS, S., DYER, M.J.S., MATUTES, E., DEARDEN, C., CATOVSKY, D. and HOULSTON, R. (2013) Risk of developing chronic lymphocytic leukemia is influenced by HLA-A class I variation. ***Leukemia***, 27(1), 255-258.

HENRION, M, FRAMPTON, M, SCELO, G, PURDUE, M, YE, Y, BRODERICK, P, RITCHIE, A, KAPLAN, R, MEADE, A, McKAY, J, JOHANSSON, M, LATHROP, M, LARKIN, J, ROTHMAN, N, WANG, Z, CHOW, W-H, STEVENS, V.L, DIVER, R.W, GAPSTUR, S.M, ALBANES, D, VIRTAMO, J, WU, Z, BRENNAN, P, CHANOCK, S, EISEN, T and HOULSTON, R.S. (2013) Common variation at 2q22.3 (ZEB2) influences the risk of renal cancer. ***Human Molecular Genetics***, 22(4), 825-831.

LOPES, M.C, HYSI, P.G, VERHOEVEN, V.J.M, MACGREGOR, S, HEWITT, A.W, MONTGOMERY, G.W, CUMBERLAND, P, VINGERLING, J.R, YOUNG, T.L, VAN DUIJN, C.M, OOSTRA, B, UITTERLINDEN, A.G, RAHI, J.S, MACKEY, D.A, KLAVER, C.C.W, ANDREW, T and HAMMOND, C.J. (2013) Identification of a Candidate Gene for Astigmatism. ***Investigative Opthalmology & Visual Science***, 54(2), 1260-1267

SCOTT, I.C, SEEGOBIN, S.D, STEER, S, TAN, R, FORABOSCO, P, HINKS, A, EYRE, S, MORGAN, A.W, WILSON, A.G, HOCKING, L.J, WORDSWORTH, P, BARTON, A, WORTHINGTON, J, COPE, A.P and LEWIS, C.M. (2013) Predicting the Risk of Rheumatoid Arthritis and Its Age of Onset through Modelling Genetic Risk Variants with Smoking. ***PLoS Genetics***, 9(9), e1003808.

VIMALESWARAN, K.S, CAVADINO, A and HYPPÖNEN, E. (2013) APOA5 genotype influences the association between 25-hydroxyvitamin D and high density lipoprotein cholesterol. ***Atherosclerosis***, 228(1), 188-192.

The CLS searchable bibliography provides further examples - see www.cls.ioe.ac.uk

1.4 Genomic data sets present specific ethical challenges, particularly around the ethics of providing individual feedback to those who consented to provide their data for research purposes. Population-based, prospective longitudinal cohort studies increasingly face questions surrounding returning findings to individuals as a result of genomic and other medical research studies. Researchers are commonly conducting whole exome and whole genome sequencing and SNP genotyping on samples accessed from these collections and finding information that might or might not have been a part of their original investigation. As the number of genetic data resources increases, there is every expectation that the number of individual genomic research (IGR) findings discovered will also increase. While guidance is being developed for clinical settings, the process is less clear for those conducting longitudinal research. Beyond deciding which results to return, returning results in the cohort setting is complicated by questions of whether re-consent is needed and the possible impact on the study, whether there is a need to introduce third parties such as genetic services to assist in the feedback process and how that will be managed, and what resources are needed to support and manage the feedback process.

2. What are the new privacy issues?

2.1. Linking data across large numbers of databases can be shown to increase the chances that it is possible to identify a unique individual within the data, even if no explicitly identifying information (e.g. name, date of birth, address) has been recorded on the database.

2.2. There is a balance to be maintained between reaping the benefits of readily available large scale detailed datasets with research utility, which would include providing important insights into individual variability in susceptibility to disease, and respecting the rights and interests of individuals. More attention needs to be paid to how to obtain fully-informed broad consent from individuals when collecting their data so that the data can be used as widely as possible. The time and resources required to achieve this should not be underestimated (e.g. including the proper training and accreditation of survey interviewers, nurses, or other individuals, who are collecting the consent).

2.3 Gaining prospective, broad, informed consent is increasingly challenging given the rapid pace of change in what data are available (e.g. including data from social media etc) , and the increasing ability to link new forms of data. However this is arguably key to making most efficient use of data collected from specific samples of individuals for wider public benefit.

2.4 In the context of prospective longitudinal studies it is important to keep cohort members informed about how their data is being used (e.g. via targeted mailings and the use of a website aimed at participants). This will help ensure continued engagement with the study, increase response rates, and improve the representativeness and scientific value of the data collected. It is important to be aware that members of longitudinal studies are a very heterogeneous group and, due to the increase in open access, will increasingly have access to published scientific research. Scientists therefore need to be aware that the research they publish may well be read by the subjects of the research. This underlines the importance of gaining fully informed consent in the context of these studies.

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

3.1 Developments in information technology and the increased creation/existence of big data mean that we need new cadres of individuals with specific skills in data science and informatics. These individuals need to be fully integrated and valued within research teams and their skills fully recognised with clear career paths and appropriate reward structures. There needs to be very effective communication between researchers who can articulate interesting research questions and informaticians who fully understand both the potential and limitations of development in data science, and who can reliably estimate and plan for the resources needed to conduct research using these new data resources and technologies.

3.2 There needs to be a clearer articulation of what big data means in different contexts and what the challenges are of different types of big data i.e. to what extent is the volume vs. the complexity or quality of data an issue.

3.3 The rapid pace of data acquisition and turnaround of results, coupled with the multiple agencies involved in processing biological samples and generating for instance genomic data, had in some cases led to a lack of documentation on provenance, processing and quality assurance.

3.4 The lack of good metadata standards, enabling discovery of this data lags behind that for instance in other scientific areas such as social science. This has a consequent impact on the use and re-use of data.

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

4.1 The UK has excellent data resources that can be used for biomedical, public health and life sciences research. For example: UK Biobank is a major and significant resource; the portfolio of British Birth Cohort studies is unique in the world and these studies are multi-disciplinary including objective biomedical data; Understanding Society is a very large household panel study that also now has biomarkers attached. There are therefore major opportunities for use of linked biomedical data in research. These studies also almost all have consent to link to health records such as the Hospital Episode Statistics and Primary Care Data. However, my understanding is that additional work is needed to make linkage to primary health care data more practical at a national level.

4.2 It is difficult retrospectively to require researchers to allow others access to data that they have collected for further research. Journals are increasingly demanding that data on which research findings are based should be made available for replication purposes. It is important that when applications are made for funding for data collection and research that it is clear from the outset what the arrangements will be for making the data available more widely to other researchers. The consent obtained should also wherever possible allow the data to be accessed and analysed by those outside of the research team. It is also not just about data being available in theory, but the processes for obtaining the data being transparent and straightforward and also the data needs to be of high quality and well documented (e.g. in terms of what algorithms have been used to transform the raw data into research data). Recognition needs to be given to the intellectual property invested in the collection of high quality data for research so that there are not dis-

incentives for making data more widely available for research. The ability to submit data sets as outputs in the REF is an important step forward in this regard.

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

7.1 it is not clear that linked biomedical data requires distinct governance arrangements compared to the use of other personal data. There are issues around sensitivity and potential disclosure linked to many different types of data. For example data about financial circumstances and information about previous bankruptcy or criminal convictions would be viewed as sensitive by many individuals and could be potentially harmful if disclosed. In addition information about household/family composition can be uniquely identifying where families are relatively large and ages of family members are known.

7.2 It is not reasonable to expect individuals to have a high level of continuing involvement in how their data re used after they have been collected. Experience of working on longitudinal birth cohort studies, and evidence from qualitative research on a subset of the cohort, suggests that being part of even a very long-standing birth cohort study is not an important or significant aspect of most cohort members lives, and it would not be reasonable to expect a great deal of engagement with the study between sweeps of data collection. In addition, cohort members move, choose to cease participation, emigrate and die, and so it can be difficult to recontact the whole of a sample, who previously gave consent, in order to gain their reconsent for use of their data for different purposes. This is what makes it important to ensure that well-informed broad-based consent is obtained at the point of data collection.

7.3 It is difficult to state in principle whether an opt-in or opt out system should be used for people to decide whether to allow their personal medical data to be used for public benefit as it depends on the sensitivity of the data and the level of risk of disclosure. However, when possible an opt-out system is to be preferred as it maximizes the sample size and the representative nature of the sample while also giving individuals the opportunity to prevent their data from being used.