UNIVERSITY OF
EXETER

Nuffield Council on Bioethics Open consultation on "The linking and use of biological and health data"

Response by the Exeter Centre for the Study of the Life Sciences (Egenis), University of Exeter

Egenis respondents: Dr Nadine Levin nadine.sarah.levin@gmail.com and Dr Sabina Leonelli s.leonelli@exeter.ac.uk

We respond to this consultation as representatives of a team of researchers studying the increased intensity of biological and biomedical data collection, analysis, use, and retention in healthcare. We also situate our response in relation to response to growing ethical concerns over the development, use, and value of data-intensive research in the biomedical sciences, in addition to concerns over the management, use, and analysis of large data sets for the study of common diseases. Our comments attempt to move beyond issues of privacy and public interest: instead, they are rooted in questions about how new forms of biomedical research face key issues in, and enable normative frameworks for, engaging with approaches to health and disease.

Researchers at the Exeter Centre for the Study of the Life Sciences (Egenis) have been studying data practices in biomedicine for over a decade from the points of view of the history, philosophy and social studies of science. We also have long-term experience in collaborating with both laboratory-based and clinical scientists around strategies for effective data dissemination, and we have recently initiated a Data Studies group funded by the ESRC, Leverhulme Trust and the European Research Council (www.datastudies.eu). This group is led by Dr Sabina Leonelli, Associate Director of Egenis, who has ten years of research experience on data dissemination and the development of data infrastructure in biology and biomedicine. Dr Leonelli's work has focused on the ways in which genomic data are classified in order to be re-used in biological and clinical settings (e.g. Leonelli 2012a/b and 2013); and on the sustainability of the infrastructures used to that purpose particularly within plant science (e.g. Bastow and Leonelli 2010, Leonelli et al 2013). In 2013, Egenis was able to employ Dr Nadine Levin, whose recently concluded PhD research is based on cutting-edge ethnographic research on data practices within the Computational and Systems Medicine (CSM) laboratory at Imperial College London, as well as on interviews with members of the wider metabolomics community within the United Kingdom (e.g. Levin forthcoming, Levin under review).[1]

[1] The CSM represents one of the pioneering centres for the development of metabolomics methods and applications to the study of health and disease. Drawing on the network of research hospitals that exist in the Imperial College Healthcare NHS Trust, the CSM has established a variety of collaborations with health researchers and clinical practitioners, to advance the implementation of translational metabolomics research. Examples of CMS initiatives—which are being spearheaded by Professor Jeremy Nicholson (the Head of the Department of Surgery

General Questions

1. Do biomedical data have special significance

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?

We do not think it useful to define biomedical data as a distinct class of data, for two reasons: (1) most problems plaguing the dissemination and integration of biomedical data bear similarities to problems experience in other areas of research, and thus considering data practices across the sciences is a better way to identify problems and possible solutions (e.g. the Royal Society report from 2012); (2) within biomedicine, different types of data are created through different technologies and different cultures of professional training, and as such entail different notions of what constitutes biomedical research and its objects of investigation.

Indeed, despite the increasing importance of big data and related technologies within biomedical research, the definition of "data" remains elusive and poorly understood (see Borgman, 2012). In metabolomics research, data exist as extremely large datasets or matrices of biochemical information collected from nuclear magnetic resonance (NMR) or mass spectrometry (MS) experiments, which are subsequently analyzed with multivariate statistical and machine learning techniques. Many researchers have questioned the very idea that these data can be disseminated outside of its original production context, as metabolomic data are very sensitive to environmental conditions and thus potentially unreliable when transferred across experimental contexts (Leonelli 2012a, Leonelli et al 2013). This type of data is also very different from that which is collected as a routine part of clinical work, which often exists as reports, lists of the numerical results of laboratory tests, or collections of (sometimes digitized) images (Levin forthcoming).

As an example, metabolomics researchers collaborating with clinicians often spoke of the "gulf of understanding" that existed between them. This originated from struggles not only with different ways of carrying out research—of learning how to manipulate samples in laboratory environments versus interact with patients in hospital settings, or of learning how to balance the time demands of clinical work and laboratory research—but also with struggles on a fundamental level, on behalf of the clinicians, to understand the concepts, uses, and values underlying the data produced in metabolomics experiments. Because clinicians were not usually trained in the methods and ideas of biochemistry or data analysis, they struggled to understand how to interpret the graphs produced during metabolomics experiments, or to envision how metabolomics data could be used to assess patients in clinical practice. While metabolomics data was useful for showing the biochemical composition of tissues or for producing molecular signatures of disease, its use for diagnosing or treating disease was less clear. Thus, metabolomics and clinical data had fundamentally different forms/formats, roles, and meanings.

and Cancer)—include "surgical metabonomics," the development of the National Phenome Centre as one of the legacies of the London 2012 Olympics, and clinical trials involving the "intelligent knife" (see Kinross et al., 2011; Nicholson et al., 2012; Mirnezami et al., 2012). Many of these initiatives are so current that they still remain relatively undocumented in the publication record.

Several initiative under the heading of "translational research" or "personalized medicine" initiatives have attempted to integrate such different types of data, sometimes resulting in failure. A notable initiative has been the Cancer Biomedical Informatics Grid (caBIG), created in 2003 to function as a portal linking together datasets gathered by the research institutions and patient care centers under the purview of National Cancer Institute (NCI). caBIG was heavily critiqued for underestimating the complexities associated to integrating different types of datasets and disseminating them to a wide variety of stakeholders, and this multi-million-dollar infrastructure was eventually closed down in 2013 (Leonelli 2013). Examples such as this show that **the interlinking of different types of data requires** more than the development of databases, standards, or computer algorithms for analysis. It also requires

(1) **Interdisciplinary training for and interaction between various types of researchers and clinical practitioners** (to familiarize clinicians, for example, with multivariate statistical and machine learning techniques, and to familiarize researchers, vice versa, with the complex process of patient diagnosis and treatment)
(2) **New types of user-friendly data analysis interfaces and visualizations of complex data**, which enable researchers and clinicians to access and make sense of data produced by different types of experts
(3) **Consideration for ways to integrate and compare the intuitive, judgment-based, and often qualitative aspects of patient care with the quantitative, information-based measurements of research.** Here, it will be particularly important not to discount the value of more "subjective" forms of clinical knowledge and practice, and to remember—as scholars in social and historical studies of science have established—that biomedical research itself is never fully "objective" (see Latour and Woolgar, 1986; Daston and Galison, 2007).

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

The notion that data can be used to understand and treat individual biological variation and health has arisen in the context of recent attempts to develop "personalized medicine" or "stratified medicine," the notion that medical care can be tailored to individual or small groups of persons and instances of disease through data-intensive research. To some extent, medical care has always been personalized. Throughout the 20th century, physicians practiced patient-centered care and used the "art" of clinical judgment, avoiding a one-size-fits-all approach to medicine. What is new, therefore, about modern forms of personalized medicine is an emphasis on the use and value of large volumes and interlinked kinds of data for finding health patterns and maximizing biomedical knowledge.

Despite a discursive emphasis on the treatment of individuals (for example, with genetic tests for breast cancer), efforts to develop personalized or stratified medicine are fundamentally concerned with comparing groups of individuals. They are thus centrally concerned with populations, rather than with individuals. This is particularly evident when considering how personalized medicine is intricately tied to the establishment of biobanks and databases for the storage of samples and data: when taken in a wider perspective, efforts to study individuals are always tied to efforts to study populations (see Raman and Tutton, 2010; Foucault, 1990).

Though this is not necessarily problematic, it does prompt us to consider to what extent efforts to develop personalized or stratified medicine can truly provide insight into, or impact the diagnosis and treatment of, individual biological variation and health. Such efforts to study individuals are intensely immersed in and predicated upon the use of complex statistics: these are not concerned with individuals per se, but rather with defining or predicting the range of biological or health values that an individual *is statistically likely* to display. Though studies recognizing the highly individualized nature of disease are becoming increasingly prominent—through, for example, "n=1" clinical trials (van der Greef et al., 2006) or studies of the "patient journey" (Kinross et al., 2011)—the outcome of personalized medicine is still predominantly about statistical likelihood, chance, and variance.

Data-intensive methods for development personalized medicine can come into contrast and conflict with the inherently personalized practice of medical diagnosis and treatment by clinical practitioners. In the assessment of patients, practitioners rely on a combination of technological measurements (such as blood tests or imaging), but also subjective judgments building on years of training and experience in assessing patients. While such judgments are not foolproof or without problems, they are able at times to assess the course of a patient's illness or treatment in ways that algorithms or statistics-based approaches cannot. Because clinicians are able to draw upon information that is not easily quantified—for example, the pallor of a patient's skin, the broad picture of how a patient has changed over time—they have additional capacities that data-intensive approaches to personalized medicine do not. Thus, efforts to develop increasingly complex and personalized models of disease with the aid of engineering principles remain problematic, in that it is difficult to know whether the configuration of patients into a series of measurable and objective variables captures those elements of health that enable clinical practitioners to effectively carry out the diagnosis and treatment of disease.

Statistics-based approaches to personalized medicine do not necessarily make it easier for practicing clinicians to treat individual patients. If personalized algorithms can provide a percent chance or likelihood that, for example, a patient will require a liver transplant, such information is useful for understanding how to manage the burden of disease in populations, but not so useful for determining the care of patients who display unique combinations of symptoms or require unique types of treatment. While molecular and statistical approaches to personalized medicine are seen as more objective and accurate, they can prove unhelpful in the everyday practice of medical care, in which patients display individual cases of and trajectories for disease. Thus, as researchers allocate time and resources to the development of post-genomic and molecular approaches to personalized medicine, they must consider how their methods for understanding disease can place priority on the health of populations over individuals.

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

To what extent and in what ways has the availability of biomedical data and new techniques for analyzing them affected the way in which biomedical research is designed?

The advances in data science and technology, which have given rise to the increased availability and use of large volumes of data in biomedicine, have also change the types of questions and theoretical approaches used to ask questions about health and disease. With the rise of "big data" approaches to biology, which are paralleled in other aspects of society by big data approaches to business (Google and Facebook) and government, there has been an increased focus on the collection and storage of data. Efforts have been directed at the development of data mining, statistical and machine learning techniques, and visualizations to aggregate and collate the various types of data that exist for analysis. With such an emphasis on the collection of data, researchers have increasingly begun to conduct "hypothesis generating" rather than "hypothesis testing" type of work. They have, in other words, looked for the data to give rise to new areas and questions for investigation, rather than seeking to collect data in response to particular questions or lines of inquiry. Such hypothesis generating research is also supported by the re-use of data beyond the original scope of its collection, as researchers seek to understand what insights will "emerge" from data.

As evident from previous phases of the history of biology and medicine (e.g. Leonelli 2012b), these approaches are not new, and yet they have come to prominence over the last decade for their capacity to yield surprising findings by suggesting correlations between previously unconsidered or unlinked facets of life. **This way of conducting research can be extremely effective within medicine, where testing the efficacy of a treatment does not necessarily involve understanding the underlying biological processes.** At the same time, the development of personalized or precision medicine does increasingly require some understanding of the underlying causes or mechanisms for such correlations. **This is a major challenge for data-intensive research, which in itself is not enough to investigate the biological processes and structures responsible for a given pattern.**

For instance, metabolomics researchers frequently emphasize the recurring challenge not in generating, but rather in making sense of statistical and molecular data in relation to disease processes and outcomes, and particularly in relation to specific genes, metabolic pathways, or bodily systems. This is because the statistical patterns observed in metabolomics data often have no inherent or pre-existing connections to clinical outcomes. Researchers emphasize their difficulties in

(1) interpreting common metabolites (small metabolic molecules) that recurred across multiple experiments
(2) determining the physical or biological origins of metabolites
(3) assessing the range of metabolites that particular technologies or techniques were capable of detecting
(4) determining the variability of metabolites within particular samples.

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?

One of the main barriers to advances in data science and technology, and in particular those that are focused on hypothesis generating or data mining approaches, is linking the ways in which data are 'packaged' for dissemination (through databases and other infrastructures) with the ways in which they are interpreted and re-used across different research contexts. Currently, **large-scale efforts to disseminate data are relatively separate from efforts to generate and interpret data in biology and biomedicine:** different groups of researchers are involved, who bring different types of expertise that are not always easy to integrate (Leonelli 2013). Further, **database curators are struggling with the challenge of assessing and accurately portraying data production practices across biological and clinical research**, with important consequences for how efficiently data stored in databases are retrieved and interpreted (Leonelli 2012a). There are significant labor and time costs in organizing and assessing the meaning underling data-intensive research. Future research agendas and funding initiatives will need to consider **how researchers can be trained not only in data analysis techniques, but also with the skills required to make sense of the large volumes of data and the statistical patterns created with such techniques**. This will also require a serious considering for how automated forms of data analysis must be paired with the skills and abilities of trained professionals.

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

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

The ability to create and analyze large and complex digital data sets not only relies on the establishment of biobanks, but also on the establishment of various community databases and less formalized structures for the sharing of data. While a number of well-document ethical and social issues surround the storage, availability, sharing, and use of biological materials and data within biobanks, such dynamics are not well understood in relation to databases and informal networks of sharing. As reported by one of us, "Many scientists and science funders view databases as crucial tools to handle the vast amount of molecular data produced by technologies such as automated sequencing and microarray experiments (often referred to as 'big data'), and getting them to travel across the world quickly and easily" (Leonelli, 2013). However, **key issues arise from the variety of data used, the lack of standards available and the lack of clarity as to who should provide the structures and support for sharing** (Bastow and Leonelli, 2010; Leonelli et al., 2013).

The increasingly interlinked and collaborative nature of data-driven science also encounters issues with the tensions is places upon researchers, communities, and institutions, in encouraging them to share resources and data. Researchers must constantly evaluate how to advance their own careers while also advancing the overall gains in knowledge to the community and society. **This tension raises concerns about what types of incentives or disincentives for sharing data, resources, or knowledge might be built into the current systems of scientific research**. This is particular important when considering the current mechanisms by which people receive formal credit for the work they have carried out. While long-standing mechanisms exist for awarding and assessing citations for publications, such mechanisms have not yet been established with regards to the sharing of data and biological resources. Such issues are increasingly important in the context of the "Open Access" and "Open Data" movements, which encourage—and provide norms for—the sharing of publications and data, respectively.

For the interlinked and collaborative data-driven research to improve health outcomes, we must consider the incentives, structures, and modes of credit—for example, the establishment of data citation indices, community standards and databases—that might encourage the wider sharing of biological data and resources.

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

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?

Two key hopes for the implementation of increased volumes and linked types of data in medical practice, as espoused by efforts to develop "systems medicine" (Auffray et al., 2009) and "precision medicine" (Committee on a Framework for Development a New Taxonomy of Disease, 2011), are that

1) The "subjectivity" of current medical practice can be overcome with the use of data and molecular technologies. It is hoped that this will be accomplished by replacing a reliance on symptom reported with observations and measurements of molecular characteristics of biology, as well as replacing the self-reporting of patients with the direct measurement or surveillance of their bodies (for example, via biological tracking technologies or the Quantified Self Movement).
2) The improved allocation of resources and efficacy care can be achieved by stratifying patients into different disease phenotypes through multiple stages of their treatment or "patient journey" through hospital settings (Kinross et al., 2011)

Summarising our points above, key concerns and fears involve:

- **The difficulties in developing and implementing adequate databases and data infrastructure to support effective inferences from data**
- **The danger of excessive reliance on data-intensive methods, who are excellent tools to spot new correlations but can hardly be relied upon to unveil underlying causes of disease**
- **The difficulties in training clinical researchers to correctly implement statistical analyses of data, as well as understand data provenance so as not to misinterpret results**
- **The difficulties in widely disseminating certain kinds of data, such as metabolomic data, in the first place**
- **The danger of underestimating physicians' assessments in favor of population-level statistical evaluations, which may not be equally effective in the treatment of individual patients**

References

Auffray C, Chen Z and Hood L. (2009) Systems Medicine: The Future of Medical Genomics and Healthcare. *Genome Med* 1: 2.

Bastow, R. and Leonelli, S. (2010) Sustainable digital infrastructure. *EMBO Reports,* 11(10): 730-735.

Borgman CL. (2012) The Conundrum of Sharing Research Data. *Journal of the American Society for Information Science and Technology* 63: 1059-1078.

Committee on a Framework for Development a New Taxonomy of Disease. (2011) Toward Precision Medicine: Building a Knowledge Network for Biomedical Research and a New Taxonomy of Disease. Washington DC: National Academy of Sciences.

Daston L and Galison P. (2007) *Objectivity,* New York: Zone Books.

Foucault M. (1990) *The History of Sexuality,* New York Vintage.

Kinross JM, Holmes E, Darzi AW, et al. (2011) Metabolic Phenotyping for Monitoring Surgical Patients. *Lancet* 377: 1817-1819.

Latour B and Woolgar S. (1986) *Laboratory Life: The Construction of Scientific Facts,* Princeton: Princeton University Press.

Leonelli S. (2013) Global Data for Local Science: Assessing the Scale of Data Infrastructures in Biological and Biomedical Research. *BioSocieties* 8: 449-465.

Leonelli S, Smirnoff N, Moore J, et al. (2013) Making Open Data Work for Plant Scientists. *Journal of experimental botany* 64: 4109-4117.

Leonelli, S. (2012a) When Humans Are the Exception: Cross-Species Databases at the Interface of Clinical and Biological Research. *Social Studies of Science* 42(2): 214-236.

Leonelli, S. (2012b) Making Sense of Data-Driven Research in the Biological and the Biomedical Sciences. *Studies in the History and Philosophy of the Biological and Biomedical Sciences* 43(1): 1-3.

Levin, Nadine (forthcoming). Multivariate statistics and the enactment of complexity in metabolic science. *Social Studies of Science.*

Levin, Nadine (under review). What’s Being Translated in Translational Research? Making and Making Sense of Data between the Laboratory and the Clinic. *TECNOSCIENZA: Italian Journal of Science & Technology Studies.*

Mirnezami R, Nicholson J and Darzi A. (2012) Preparing for Precision Medicine. *New England Journal of Medicine* 366: 489-491.

Nicholson JK, Holmes E, Kinross JM, et al. (2012) Metabolic Phenotyping in Clinical and Surgical Environments. *Nature* 491: 384-392.

Raman S and Tutton R. (2010) Life, Science, and Biopower. *Science, Technology & Human Values* 35: 711-734.

van der Greef J, Hankemeier T and McBurney RN. (2006) Metabolomics-Based Systems Biology and Personalized Medicine: Moving Towards N= 1 Clinical Trials? *Pharmacogenomics* 7: 1087-1094.