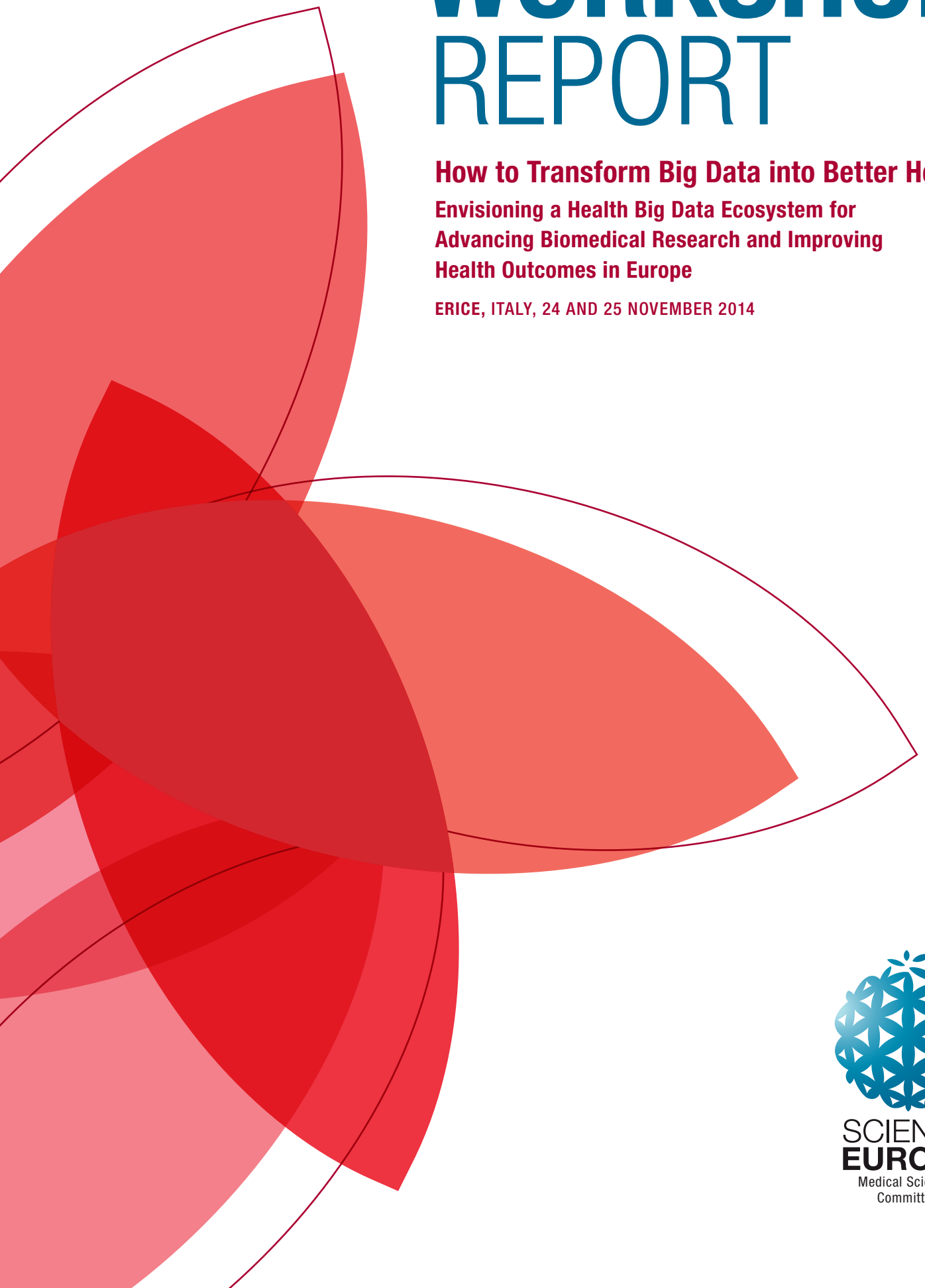
WORKSHOP REPORT

How to Transform Big Data into Better Health:

Envisioning a Health Big Data Ecosystem for Advancing Biomedical Research and Improving Health Outcomes in Europe

ERICE, ITALY, 24 AND 25 NOVEMBER 2014



SCIENCE EUROPE
Medical Sciences Committee

How to Transform Big Data into Better Health

Executive Summary

Overview

Europe faces rising health challenges due to demographic changes, an increase in communicable diseases, an inefficient R&D process in the biomedical research domain and an increase in disease complexity. Tackling big health challenges demands a 'Big Science' approach that requires integration of multi-layered health information and a highly elaborate ecosystem. Thanks to recent advances in biomedical (such as 'omics') and information and communication science technologies, we now potentially have access to a vast amount of complex health-related data, sometimes called 'Big Data' (biological, behavioural, clinical and environmental). Our ability to organise, integrate and transform health Big Data into better health outcomes will depend on an ability to create the conditions for transformation of Big Data into Big Science.

In this context, the Medical Sciences Committee of Science Europe, chaired by Prof. Richard Frackowiak (Centre Hospitalier Universitaire Vaudois/CHUV, Lausanne, Switzerland), in collaboration with the Italian National Institute for Nuclear Physics (INFN), organised a two-day workshop on 24 and 25 November 2014 at the Ettore Majorana Foundation and Centre for Scientific Culture in Erice, Italy. Entitled, 'How to Transform Big Data into Better Health: Envisioning a Health Big Data Ecosystem for Advancing Biomedical Research and Improving Health Outcomes in Europe', the workshop aimed to identify opportunities and challenges of the translation of 'Health Big Data' into 'Big Science'. The workshop was divided into two sessions. Session 1 explored potential opportunities, risks and challenges that Big Data hold for research across many fields of biomedical science. Session 2 focused on developing a long-term vision regarding the form that a European health Big Data ecosystem might take.

The workshop gathered 41 participants from 16 countries, with 14 speakers. Representatives of 11 Science Europe Member Organisations participated in the discussion, together with representatives of the European Commission, members of Science Europe Scientific Committees and other experts in this field. Ongoing initiatives aimed at crafting a health Big Data ecosystem were showcased to help participants actively engage in discussions on how to develop such a health Big Data ecosystem in Europe.

Participants recognised that tackling Europe's big health challenges needs a more systemic approach to Big Science. Such an approach requires combining multiple health-related dimensions represented by 'health Big Data' from the molecular level to the integration of information related to individual environments and lifestyles. It was also recognised that the main challenge for transformation of data into knowledge to improve health at the individual and population levels requires new analytical tools to discover novel relationships and patterns in a very heterogeneous data set. In turn, it was stressed that our ability to organise, integrate and transform health Big Data into better health outcomes depends on the ability to create a permissive and enabling ecosystem. Developing such an ecosystem in Europe relies on data sharing between multiple stakeholders, from public and private organisations involved in biomedical R&D to other disciplines (for example ICT, social sciences) that must put citizens and patients at its centre. Such an ambition requires development of an adequate framework to collect and interpret health-related data, as well as the right infrastructure, funding models, expertise and reward mechanisms to support data-driven science.

The many challenges to leveraging health Big Data into better health outcomes are not only scientific or technological in nature. Participants stressed on many occasions the need to develop policies that will facilitate the use and re-use of health-related data. The European Parliament's proposed amendments to the draft European Data Protection Regulation were widely regarded as a major hurdle, should they be enacted.

Key Highlights and Recommendations:

Facilitate the Integration of Multi-Dimensional Health-Related Data

Participants recognised that leveraging health Big Data represents an opportunity to advance different fields across the biomedical sciences and healthcare including personalised medicine, systems biology, clinical research, drug discovery, drug development and public health. This requires combining data from a range of levels (molecular, cellular, tissue, phenotype), from multiple sources and various disciplines (such as molecular and systems biology, medicinal chemistry, preclinical and clinical pharmacology, medical records, census records), many of which exist in different formats.

Many challenges have been identified, including:

- Data heterogeneity (accuracy, format);
- Data fragmentation (multiple databases, multiple owners/stakeholders);
- Data availability (protection for commercial or cultural reasons, or related to personal privacy);
- Data handling (data management, data access, data quality, data querying, data sharing);
- Data privacy and integrity (prevention of corruption and hacking); and
- Data conceptualisation (ontologies).

Recommendations:

- Currently, genomics is the fastest-moving area in the biomedical field. It aims to: integrate genomic data from individuals to investigate disease causation and develop biomarkers; predict disease phenomenology at the population level; help develop better diagnostic tools; and design new ways of managing diseases in the era of personalised medicine. However, genomics data are insufficient to capture biological processes that operate and are controlled as complex and modular systems both in the body and in its interactions with the environment. **To capture the complexity of the expression of disease, future efforts should be devoted to integration of -omics data with genomic characterisation and also with higher levels of complexity, including lifestyle and environmental data. Only in this way does it seem feasible to obtain a better understanding of clinical phenotypes. This effort requires development of a controlled and standardised vocabulary for describing entities and the semantic relationships between them.**

- Some countries such as Scotland, Brazil and Denmark have developed **national centralised databases** that contain medical records and some biological information from entire population cohorts. This approach is based on the attribution of a unique identifier to each citizen that allows linkage of various patient health data stored in different databases thus providing a unified approach to data handling. Such a clinically-led centralised model could serve as an information

commons from which, under defined conditions, various stakeholders could share and mine multi-dimensional data for research applications and optimisation of clinical care.

- Undergoing European pilot initiatives showcase other types of **models for data integration based on disease areas** such as brain disorders (the Human Brain project[1]), asthma (the U-BIOPRED project[2]), Alzheimer's disease and metabolic disorders (the European Medical Information Framework[3]). Each of these initiatives is developing its own procedures on how to handle data. Learning from these pilot projects should help to design appropriate strategies to federate and integrate disease-relevant information and develop a set of recommendations for data handling. It is clear that the ultimate goal is to unlock the immense added value in the wealth of clinical information and associated biological data stored in research, pharmaceutical company and hospital databases. These are currently largely unused or under-utilised for a variety of legitimate reasons, all of which should be scrutinised with the aim of solving them technologically and legally to permit their use in improving the health of the European population.

- The participants recognised that **best research practices should be developed** and implemented to increase the value and accuracy of Big Data. It was also recognised that **fostering a transparent reproducibility culture** was indispensable to avoid confusion and diffusion of irreproducible claims.

Facilitate the Transformation of Big Data into Knowledge

- Participants recognised that the value of data relies on finding their inter-connectedness. Integrating multi-dimensional data from molecular to behavioural levels is, however, complex and requires an integrative or systematic approach along with the development of supporting ICT tools and platforms. We are learning from systems biology that the relation between the multiple layers (for example from genes to phenotypes) is not linear. Rather, biological processes operate and are controlled as complex and modular systems. Yet, biology still relies on the central dogma from the 1950s that establishes a linear flow of information from genes to proteins to phenotypes. **Thus, an integrated approach to biology and the development of new mathematical modelling and simulation approaches in biology are of crucial importance to the transformation of Big Data into knowledge.**

- Analyses of large datasets require separation of signals from noise and multivariate statistical techniques that eliminate as far as possible the discovery of false or spurious correlations within them. Participants recognised that **Big Data analyses should be embedded in epidemiologically well-characterised and representative populations.**

- Analyses and integration of multi-dimensional health data will necessarily mean drawing on expertise from multiple disciplines. **Data-driven science will be multi-disciplinary, collaborative and less competitive than classical science and focussed on specific problems.**

Develop a Permissive and Conducive Health Big Data Ecosystem

Leveraging Big Data to advance biomedical research and healthcare is achieved by far more than the information technology needed to federate, integrate and analyse it. It requires an environment that supports and integrates a multi-disciplinary approach which, in turn, relies on a data sharing culture.

Many challenges have been identified, including:

- Health-related data are fragmented across multiple and unconnected data sources (patient registries, bio-banks, social networks, and others);

- There is no clear code of practice for data sharing. Data are stored in databases that belong to multiple institutions and stakeholders across the biomedical research and healthcare fields;

- The prevailing biomedical R&D model is segmented into basic, preclinical and clinical research silos. This 'compartmentalisation' of the biomedical R&D and healthcare data chain, with value expected for the citizen/patient as a passive end-user, is a major hurdle to a data-sharing culture;

- There is as yet no clear code of practice to ensure personal privacy while preserving openness in data sharing; and

- Current funding and career appraisal systems for biomedical researchers mainly recognise investigator-driven research. Mechanisms recognising collaborative inter-disciplinary networks are in their infancy.

Recommendations:

Development of a multi-disciplinary approach and a culture of data-sharing in Europe requires co-ordinated efforts at the legal, societal, organisational and investigator levels.

Legal level:

- Introduce appropriate legal and ethical frameworks to support **data-sharing** while developing appropriate security and oversight measures to reduce the risk of **personal data loss** (for example the European Data Protection Regulation). Big Data present other challenges with respect to **ownership** and **liability** that will need to be resolved.

Societal level:

- **Increase citizen and patient involvement in the management and processing of their own health data** and restore public trust in science (such as health data co-operatives)

Organisational level:

- Develop **codes of conduct and research practices** that define rigorous quality control mechanisms for all aspects of data handling, from collection and annotation through to storage and sharing between organisations;
- Develop **funding opportunities** for collaborative research networks; and
- Develop **recognition and reward mechanisms** for data sharing activities by individual researchers, especially in relation to career progression.

Investigator level:

- Develop **pilot experiments** to showcase evidence-based benefits of sharing data for researchers from the public and private sectors.

Future Actions

Monitoring and Mapping Big Health Data Initiatives

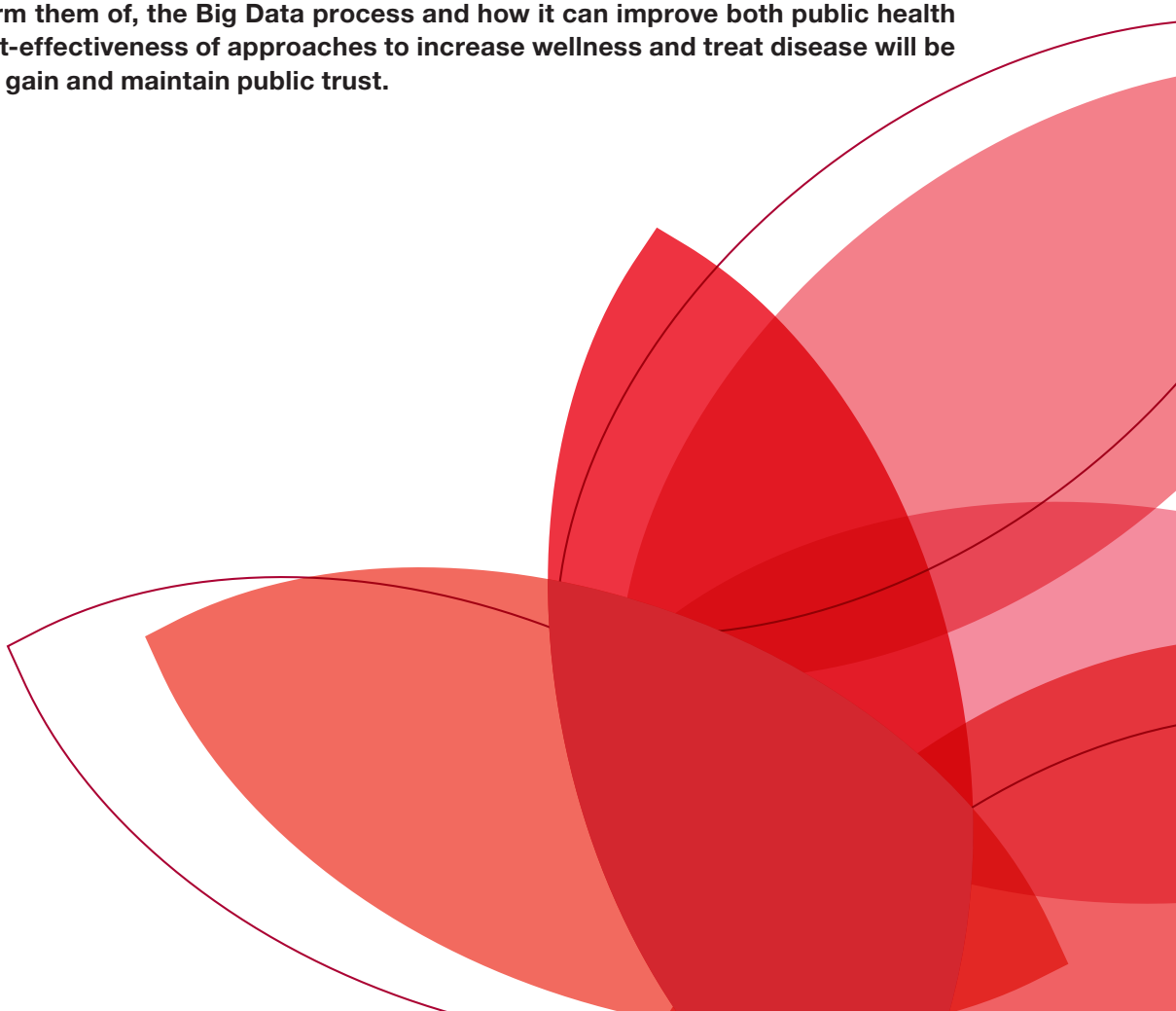
Many 'pilot experiments', some of which were presented during the workshop, show the potential that Big Data holds for advancing an understanding of human health and disease, and for transformation of healthcare to significantly impact on society. Monitoring and careful assessment of progress in, and obstacles to, these initiatives will be helpful, in order to identify obstacles at the cultural, scientific, technical, legal and institutional levels. Participants identified a need **to map undergoing Big Health Data initiatives in Europe before a set of best practices and recommendations for implementation of a Big Health Data ecosystem in Europe can emerge.**

Fostering Public-Private Partnerships

Big Data represents an economic value that is attracting large investment from the private sector. The creation of a European Public Private Partnership for Big Data involving the European Commission, industry and academia is currently underway. Fostering such a **partnership between stakeholders involved in biomedical and health research, including both public and private sectors, will be essential to leverage Big Data and implementation of a Big Science approach to health challenges.**

Gaining Public Trust

While public trust in the use of personal data mainly depends on the maintenance of data security, it also depends on perceptions about the utility and importance of the questions for which personal data based research is proposed. **Developing transparency practices that involve citizens in, and inform them of, the Big Data process and how it can improve both public health and the cost-effectiveness of approaches to increase wellness and treat disease will be essential to gain and maintain public trust.**



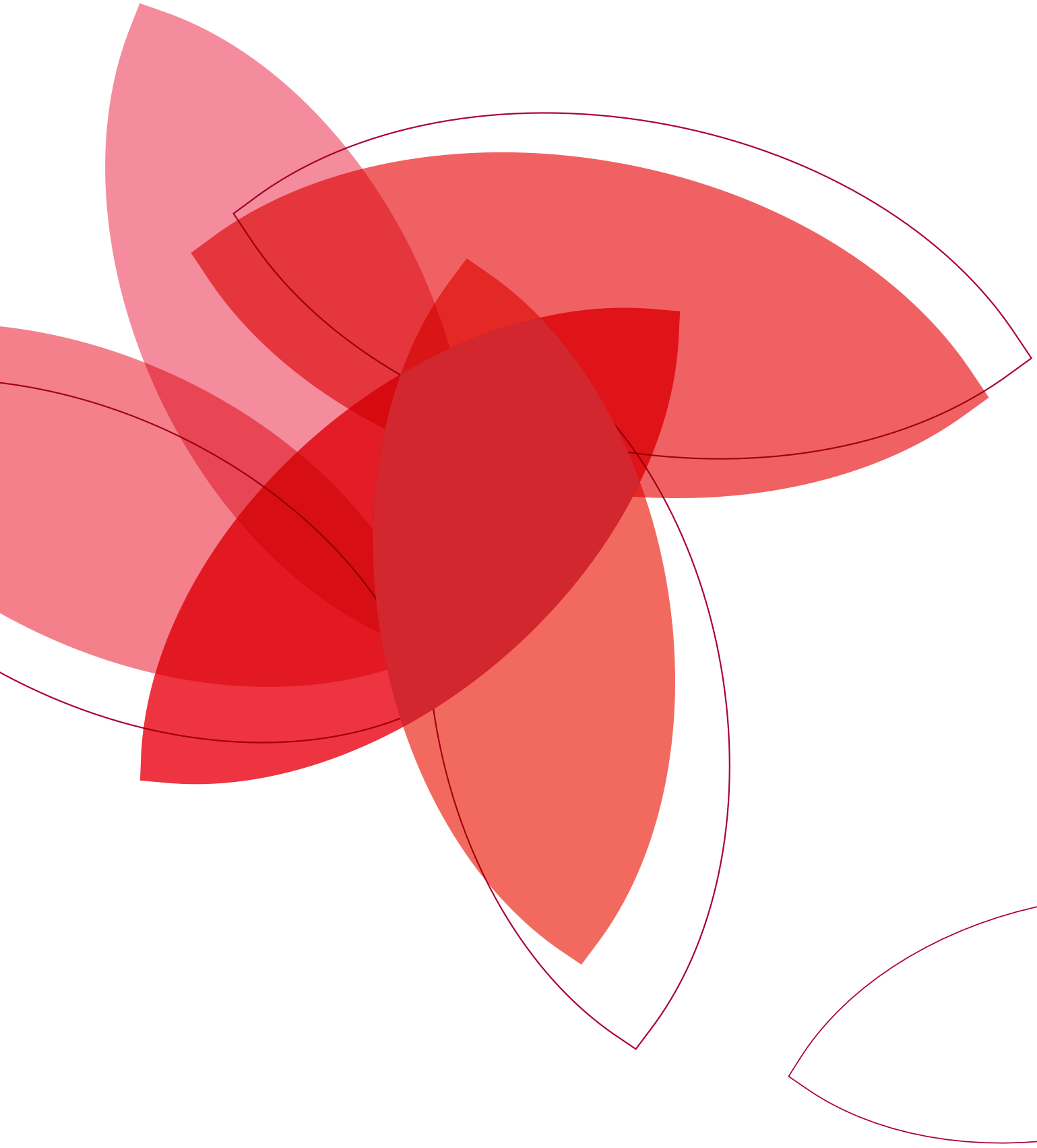


Table of Contents

1. Introduction: Big Science, Big Data and the Concept of a Big Data Ecosystem

1.1 'Big Science' will be needed to tackle Europe's future health needs

Europe is facing significant health challenges, including an ageing population, increased chronic diseases, a shift towards preventive and personalised medicine, and a lengthy and inefficient biomedical research and development (R&D) process, all portending an unsustainable economic burden. Common, complex diseases caused by a combination of genetic, environmental and lifestyle factors, many of which have not yet been identified, remain poorly understood yet they cause the great majority of morbidity in European populations. Tackling these big challenges demands a **'Big Science'** approach that requires integrated scientific knowledge with multiple partners all working in a highly co-ordinated fashion.

1.2 Leveraging health-related Big Data for Big Science

While healthcare challenges are undoubtedly growing, so too is our ability to generate, capture, store and analyse data in an unprecedented way. These data exist at a range of levels (molecular, cellular, tissue) and from various disciplines (such as molecular and systems biology, medicinal chemistry, preclinical and clinical pharmacology, the healthcare system) as well as from 'social media' and other 'non-traditional' sources. Data are accessible in real-time, in a continuous fashion, on new types of variables (for example geolocation, lifestyle, environment) and within structured and unstructured forms (for example email and social networks).

This multi-dimensional and heterogeneous health information has been termed health **'Big Data'**. If it were possible to combine these fine-grained, fragmented data in a dynamic fashion using analytic tools and interfaces this would offer biomedical researchers the opportunity to discover novel relationships and patterns among multiple variables. In turn, this knowledge could lead clinicians to personalise treatment by linking patients' care to their clinical, biological, lifestyle and environmental background. One concept that has been proposed is that of a 'human information system', analogous to a geographical information system, designed to capture, store, manipulate, analyse and integrate biomedical and health-related data holistically.

However, many challenges remain. While the integrated analysis of large datasets becomes more necessary and more common, new data models are needed to collect, share, integrate and interpret good quality and multi-dimensional data while protecting the citizen's privacy. Currently, health and biomedical data are stored and controlled by multiple health actors and organisations in innumerable, incompatible and unconnected databases. Data quality is uneven. Most citizens lack access to and control over their own health data. On the other hand, researchers encounter difficulties in accessing data which often depends on personal links they have established with those organisations that hold the data. Data integration and interpretation is also a key challenge. This fragmented data model and the lack of supporting structures to access, share and interpret increasingly larger and more complex datasets substantially reduces the effectiveness of data-driven innovation for healthcare and biomedical research.

1.3 The need for a 'health Big Data ecosystem'

To achieve this goal it will be necessary to build an enabling and supportive environment – essentially an 'ecosystem' – based on an open data-sharing model that enables efficient querying, retrieval and exchange of quality data by different stakeholders and institutions and which allows the interpretation and analysis of health Big Data while ensuring citizens' rights to privacy. Such a

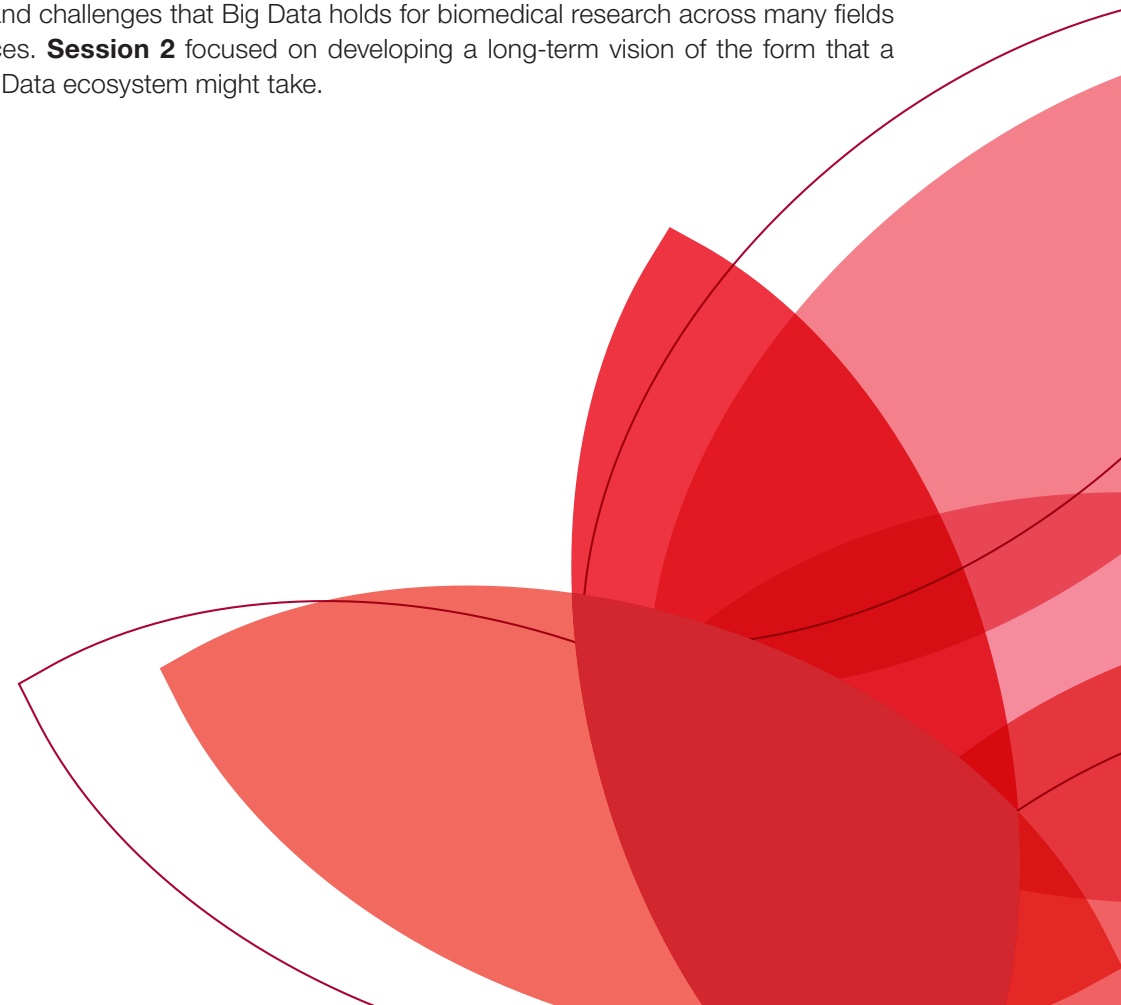
health **Big Data ecosystem** would have at its centre an 'information commons'[14] in which multi-dimensional health-related data will be collected across multiple sources from each individual and constitute, together with data generated from biomedical R&D, a health Big Data commons that will be broadly available for research use and clinical decision support, and a knowledge network that will create inter-connectedness between data and the interpretation of data into new information, leading to new knowledge.

The creation of a health Big Data ecosystem will depend on the ability to capture the opportunities as well as anticipate and overcome the technical, scientific, legal, regulatory, ethical, organisational and human resources challenges. New funding models will be needed, together with appropriate training of personnel and new ways to assess careers in what will necessarily be a highly interdisciplinary environment. Notwithstanding these challenges, to some extent this represents a cultural issue: the biomedical sciences have been slow to embrace the opportunities of Big Data in a way that, for example, particle physics, astronomy and meteorology have not.

1.4 The rationale for the Science Europe Workshop

In this context, the Science Europe Medical Sciences Committee organised a workshop entitled 'How to Transform Big Data into Better Health: Envisioning a Health Big Data Ecosystem for Advancing Biomedical Research and Improving Health Outcomes in Europe'. The meeting aimed to: first, promote better understanding and increased awareness of the promise and challenges that Big Data holds in transforming European biomedical research at all stages of the value chain, from understanding fundamental biological mechanisms to healthcare delivery; second, create a community of key decision-makers and performers across the biomedical, healthcare and ICT research fields; and third, leverage synergies between Science Europe Member Organisations and scientific experts to develop a long-term strategic vision for a health Big Data ecosystem for advancing health outcomes for citizens in Europe.

To reflect these aims, the workshop was divided into two sessions. **Session 1** explored the potential opportunities, risks and challenges that Big Data holds for biomedical research across many fields of biomedical sciences. **Session 2** focused on developing a long-term vision of the form that a European health Big Data ecosystem might take.



2. Big Data in Biomedical Research: Potential Opportunities, Risks and Challenges

Conventional models of drug discovery that have served the pharmaceutical industry for the past half century are no longer delivering results. It takes more than ten years to develop a single drug at an estimated cost of more than $1 billion and the attrition rate is high (only one out of 100 candidates at the preclinical stage will obtain Market Authorisation). This approach is wasteful and expensive and does not respond to the health challenges facing society.

As stated in the workshop by **R. Barker** (Centre for the Advancement of Sustainable Medical Innovation, Oxford University/UCL), Big Data represents an opportunity to take an alternative approach to this failing model. The interrogation of diverse sets of data, such as population-based health records, clinical trial data and omics, presents a new way of analysing and understanding diseases. This can lead to the identification of subsets of what were traditionally considered a single disease such as asthma, as shown by **S. Holgate** (School of Medicine, University of Southampton), and in turn to treatments or prevention appropriate to that subset. This is what is termed personal, precision or stratified medicine. In 2011, the National Research Council in the US published a report entitled 'Towards Precision Medicine: Building a Knowledge Network for Biomedical Research and a New Taxonomy of Disease'[4]. The report's authors, including **S. J. Galli** (Stanford University School of Medicine), concluded that advances in new information and concepts in biomedical research were difficult to incorporate into existing taxonomies of disease, and that the time was right to develop a new taxonomy of disease based on molecular biology. The creation of this new taxonomy would first require an 'Information Commons' in which data on large populations of patients are made widely available for research, and a 'Knowledge Network' that adds value to the data by highlighting their inter-connectedness and integrating them with evolving knowledge of fundamental biological processes.

Indeed, it is becoming increasingly clear that fundamental biological processes, for example metabolic pathways within living organisms, are far more complex than previously understood, involving multiple interacting networks with intricate and multifaceted regulatory systems. As nicely shown by **R. Aebersold** (Institute of Molecular Systems Biology/Eidgenössische Technische Hochschule Zürich/ETH Zurich), what have been considered key concepts of biology may need to be discarded or re-thought: the link between a gene locus and phenotype, the linear paradigm of gene to protein to function, and the central dogma of the flow of genetic information. Again, Big Data may help to represent a new way to understand these phenomena: as we generate more and more data at a variety of levels, from the genome, transcriptome, proteome and metabolome, and microbiomes, to behaviour and lifestyle, this offers the opportunity to develop new data-driven correlation approaches to provide information about biological systems.

Big Data is also being used to provide a standardised vocabulary of phenotypic abnormalities encountered in human disease. The Human Phenotype Ontology project[5],led by **P. N. Robinson** (Freie Universität Berlin), has developed a database that to date has more than 10,000 terms and around 115,000 annotations for some 7,000 diseases. The user enters individual symptoms into the system and receives a ranking of differential diagnoses. This approach has succeeded in making a diagnosis in previously undiagnosed subjects and has resulted in the identification of novel candidate genes for disease.

Within public health, Big Data is already making an impact in many parts of the world. For example, **A. Morris** (Farr Institute of Health Informatics Research) described how Scotland has developed a single clinical information system for diabetes patients that follows the journey of care of the patient from primary to secondary to tertiary. A real-time information system has been established for diagnostic tests such as retinal scanning, and every person registered with diabetes has had biological markers measured. This comprehensive, systematic approach to disease surveillance has resulted in a reduction in amputations. **M. L. Barreto,** (Universidade Federal da Bahia and the Oswaldo Cruz Foundation/FIOCRUZ), showed that in Brazil the main primary healthcare strategy is delivered through the Family Health Programme (FHP). Data analysis of the programme has demonstrated in areas of high coverage significant impact, on outcomes such as infant and child mortality, and mortality and hospitalisation due to cardiovascular and cerebrovascular events. When FHP and a cash transfer programme (Bolsa-Familia) were analysed together the impact on child health (measured by mortality and hospitalisation rates) was amplified in particular to poverty-related causes such as under-nutrition, diarrhoea and pneumonia. **S. Brunak** (University of Copenhagen), described a programme of population-wide genome sequencing which has been discussed in Denmark and which is an approach that may provide a basis for randomly-selected analyses rather than analyses that are confined to cohorts. The importance of a co-ordinated approach to data collection, and the need to be mindful of the wider health data environment across Europe, was illustrated by **G. Mihalas** (Romanian Academy of Medical Sciences, Timisoara) with the case of Romania, where a comprehensive database of citizens' health-related information was created but was subsequently found to be not interoperable with other health record systems in Europe.

An example of how Big Data is being used to help tackle one of science's greatest challenges is the EU's ten-year Flagship Programme the Human Brain Project[6], co-led by **R. Frackowiak** (CHUV, Lasuanne, Switzerland). A total of 70 laboratories and 120 principal investigators are involved in the project, together with industrial partners. The project will generate knowledge leading to a better understanding of brain function, develop new analytical tools, and federate and integrate data from scattered databases owned by different stakeholders – a key requirement for a Big Data ecosystem.

It is clear that the opportunities offered by a health Big Data ecosystem in Europe to drive a shift towards efficient data-driven biomedical research are significant. However, a central challenge is to ensure that the data are trustworthy. Extensive research over many years by **J. Ioannidis** and colleagues at the Stanford University School of Medicine has shown that a large proportion of research findings are false; they cannot be replicated or are subsequently refuted. This is due to a range of factors such as small sample sizes, poor experimental and statistical methodology and selective reporting, compounded by a lack of published raw data. To help overcome these problems a number of practices could be implemented, such as developing a culture of large scale collaborative research, adoption of a 'replication culture', and better training of the scientific workforce in methods and statistical literacy.

3. Laying the Foundations for a Europe Health Big Data Ecosystem

3.1 Some existing European initiatives

There are a number of existing initiatives within Europe that may play an important role in laying the foundations for a health Big Data ecosystem. As part of the EU's Innovative Medicines Initiative (IMI) the European Medical Information Framework[7] (EMIF) seeks to create an environment to allow the efficient re-use of existing health data. As highlighted by **B. Vannieuwenhuyse** (Janssen Pharmaceutica), the goal is to link clinical care with research by tracking the data generated when a person enters the healthcare system and feeding this into clinical research to create a 'virtuous circle' that leads on the one hand to new discoveries and on the other to better patient care. Key challenges involve recurring issues with the secondary use of data: where they reside, how good they are, and how to obtain access. In addition, ethical and privacy issues need urgently to be tackled.

The question of privacy and the use of a citizen's data remains a challenge. A new project from Switzerland, co-led by **E. Hafen** (ETH Zurich), aims to develop a system whereby people can regain their 'digital self-determination' through citizen-controlled data co-operatives through a project called MiData[8]. The idea is for members of the co-operative to be able to securely store, manage and control their secondary data as a new class of asset that has a financial value. It is envisaged that such a platform would develop in partnership with healthcare providers and with appropriate levels of transparency, and in an open source and open standards environment.

Y. Ioannidis (University of Athens) showed how existing European ICT infrastructures may play a role in a future Big Data ecosystem. The OpenAIRE project[9], for example, is the vehicle by which researchers can comply with the EU's Open Access policy for EU-funded research results and data. Around 8.5 million publications are currently indexed; the data remain in their own repositories distributed across the EU, with the OpenAIRE hub allowing text-mining and other end-user services. OpenAIRE is an example of information commons and a knowledge network, and could provide useful lessons for a healthcare Big Data ecosystem.

3.2 Issues identified by delegates arising from discussions during the workshop

3.2.1 Data sharing and quality assurance

3.2.1.1 Data sharing

Sharing data is central to the concept of a Big Data ecosystem. It will allow opportunities for replication, analysis and interpretation. However, the sharing of data can be problematic. Many institutions purport to support the concept of data sharing, but then fail to do so in practice, and many researchers are reluctant to share their data. The issue of privacy protection can be a significant barrier to the free sharing of data, but even so this is often put forward as a spurious or misunderstood reason to deny access to data. In addition there are moves by publishing houses to create their own data repositories, which could result in another commercial business model which has the effect of restricting free and open access to the data. It will be important to know who controls access to the data, and ways need to be found to accredit people who wish to work on the data. It will be necessary to develop a culture of data sharing and to put in place incentives that will encourage this culture to develop.

There is a degree of mistrust towards pharmaceutical companies who wish to access citizens' data. To ensure transparency, protocols could be put in place whereby anyone wishing to gain access to data should declare their intent and provide a comprehensive audit trail so that any studies undertaken with the data can be confirmed as being in line with the declared intention.

3.2.1.2 Quality assurance of data

Quality assurance systems should be developed that equip the scientific community with the information needed to undertake its own assessment of the quality of data and the quality of the techniques used to generate the data. Organisations such as the Global Alliance for Genomics and Health, which have an international perspective on these issues, could be invited to promote the development of suitable guidelines. The issue of what may be termed 'rotting and dying' will also need to be addressed. Data are accumulating rapidly and the ability to keep up with this inflow of information is lagging; an active mechanism is needed to assess data quality (data curation) and to let go of unused or erroneous data. How this can be done remains a challenging question.

3.2.2 Careers and training

Carrying out research within a Big Data ecosystem is necessarily inter-disciplinary, often involving multiple research teams distributed across many institutions. This can be problematic for a biomedical researcher given the way that careers are currently structured. Promotion generally depends upon a record of publication as a prominent author in journals that are fundamentally disciplinary. Sharing data can pose a risk to a researcher's career, as can performing bioinformatics research as part of a large interdisciplinary team. This contrasts with the case in other disciplines, notably particle physics. Here, a paper may have, literally, thousands of co-authors. Nevertheless, within this long list of authors the community is able to define the distinctive and crucial contribution to the collective effort by individuals and teams. Efforts are currently undergoing in the biomedical community to address the multi-authorship issue.

There is a strong need for a new breed of data scientists who integrate other scientific disciplines and a range of expertise. The need for such a service has been recognised in the U.S., where many departments in universities are setting themselves up as data science departments. Such

researchers will be desperately in demand to address issues of data curation and knowledge management, software development, text and data mining and analytics. The contribution of such researchers to the research effort will need to be recognised.

Current training of medical doctors in Europe has little or no emphasis on the development of a desire to conduct research, or to broaden an individual's set of skills to include expertise in disciplines that will be important for the medicine of the future, such as engineering and computer science. In the U.S., Medical Scientist Training Programs (MSTP) are in place at many top tier universities to permit students to combine training for the M.D. and Ph.D. degrees, with the Ph.D. programmes including traditional biomedical disciplines and others, such as bioengineering, chemical engineering, computer science and bioinformatics. However, only a minority of U.S. medical students pursue such training programmes.

Establishing in Europe new modes of training and improved incentives can go a long way towards overcoming these hurdles of training true physician-scientists and other investigators with interdisciplinary interests. Funding agencies and research performing organisations can provide financial incentives and recognition for data scientists who make a vital contribution towards large, multi-disciplinary projects with a strong bioinformatics element. There should be dedicated training for data scientists, and professional bodies that establish curricula in medicine should have a wider representation, including from those who will ultimately be treated by the profession.

3.2.3 Trust and the role of the citizen

There is consensus that the citizen should be at the heart of any effort to establish a Big Data ecosystem. Health data of citizens constitute a central part of Big Data and citizens should be empowered by being given control of their own data. In many countries citizens harbour deep suspicion about how their personal data will be used, and any attempt to develop a comprehensive data system containing such data – albeit anonymised or with any identifying details stripped out – will require the trust of the citizen. Many studies have shown that where there is demonstrable benefit to the public, there will be trust; conversely, if people think that their data, collected through public agencies, will be used primarily for the pursuit of profit, there will be distrust. The issue is heavily politicised in a way that, for example, the use of data from particle physics is not.

New, imaginative ways need to be developed to provide incentives to the citizen to have active control over his or her data – that is, digital self-determination. In other contexts, people appear to be willing to share many aspects of their personal lives, through social media platforms such as Facebook. A central facet of any strategy that works towards the creation of a Big Data ecosystem in health is to build trust and engage with the public.

Scotland provides an interesting case study of how this may be achieved. Here, a blueprint of guiding principles[10] was developed that highlighted best practice in issues such as public interest, privacy, consent, governance, access, data control processes, clinical trials, public involvement and benefit-sharing. This blueprint was published three years ago and has been incorporated into government policy. It is important that such initiatives find their way into the public sphere and do not remain within the academic domain. In Denmark, the default position is that a citizen's health data are made available for research purposes unless the citizen chooses to opt out.

3.2.4 Legal and ethical issues

Legal and ethical issues may well prove to be two of the biggest hurdles to be overcome in any effort to lay the foundations for a Big Data ecosystem across Europe. Procedures for ethical approval for the use of data vary widely between countries, and better advice and guidelines are required, particularly for the re-use of data whose collection was originally for a specific research purpose. The way in which data sharing interacts with various complex legal structures in different countries in Europe is confusing. In addition, it is important that European data protection laws are not

framed in a way that will hamper the use of data for legitimate research purposes whose aim is the improvement of the health and wellbeing of European citizens. The representatives of the research community, including Science Europe, should raise their voices to ensure that the opportunities for harnessing Big Data are not hampered by misguided new legislation.

The public must be involved in the formulation of ethical advice and guidelines, and harmonisation of these ethical considerations across Europe will be fundamental to good practice and governance.

3.2.5 Future actions

3.2.5.1 What is the best approach?

On one hand there is an argument that a 'Grand Challenge' approach would be an appropriate way to proceed as a single large programme involving multiple centres and with a clearly defined strategic goal. Such an approach can bring problems, such as disgruntlement among other researchers who perceive that funds are being siphoned away from smaller research groups carrying out fundamental research. There is also a question of whether, even under the umbrella of a single programme, the separate elements develop their own framework for data-related tools and operating procedures, or whether a global framework should be constructed.

There is a view that ecosystems will arise through a kind of Darwinian evolution, in that various schemes will develop, and those that work best will pick up speed and draw in money. At the same time programmes should be encouraged to share a communal interface. Here the analogy could be that of Apple Computers and apps, where diversity can flourish but with an underlying coherence.

An alternative to a 'Grand Challenge' approach is a series of small pilot studies, which require only relatively modest investments and which can be completed in a short time. If pilot studies are decided to be the best way forward, it may be sensible to select a relatively small number of studies, which have well-defined goals and which test specific technologies. Assuming these are successful, this approach would put people in a stronger position to argue for larger investment subsequently.

Whatever approach, or mixture of approaches, is adopted, it is important that health economists are included to demonstrate economic advantages, and that metrics to measure the outcome of the particular approach that is chosen are clearly defined at the outset.

A note of caution is also sounded. It can be counter-productive to promise the delivery of benefits that are then not realised in the timescales that were initially promised and anticipated.

3.2.5.2 The importance of public–private partnerships

It is important that everything possible be done at the European level to stimulate public-private partnerships. The private sector is moving rapidly into the arena of personal health, particularly in areas such as personal health monitoring devices. This market will continue to grow and there could be value in the public sector putting resources into this type of development as a joint enterprise. The ubiquity of portable, multi-media devices such as smartphones represents a major opportunity for data gathering and there is a strong argument that apps could provide a powerful route for garnering invaluable information to populate a Big Data ecosystem. This requires serious attention.

3.2.5.3 Moves to gain public trust

To help overcome public concerns about issues such as privacy and confidentiality there may be value in organising a high-profile partnership between patient organisations, clinicians, research bodies, public health agencies and the commercial sector to present a common and united front on the value of a Big Data ecosystem to the public, through the media and decision-makers such as the European Parliament. Surveys have shown that attitudes towards privacy and confidentiality are

not uniform across the countries of Europe, so it would be necessary to think carefully how best to engage the public in different countries.

In the UK, the recent Research Excellence Framework[11] evaluated research activity in all universities, and required the submission of case studies to demonstrate the impact of research. This has yielded many impressive examples of how research benefits society. It may be that a portfolio of case studies relating to the use of Big Data across Europe could be compiled as a way to illustrate best practice and the value of the concept to improving healthcare and economic outcomes.

The important issue of branding also needs to be addressed. Ultimately, politicians respond to public opinion, and so favourable public opinion towards endeavours in Big Data is essential. In this regard it would seem important to clearly spell out the potential benefits of using Big Data to advance public health, in terms of developing new ways to prevent or postpone dementia, for example, or for tackling more effectively diabetes or cancer. It is important that public and private bodies and patient support groups are fully behind these moves. If this is not achieved there is a strong risk that the research community in this area will simply come across as technocrats pursuing their own narrow interests, whom the politicians can afford to ignore.

3.2.5.4 Mapping the existing landscape

There is already much activity in Europe within the arena that is directly or indirectly related to Big Data. Before proceeding further it would be sensible to map this landscape to understand what is already being done, where there are gaps, and where there are opportunities to potentially add value. The present workshop took the first step into developing a roadmap for a European health Big Data ecosystem that will be required to overcome challenges.

4. Conclusions

Harnessing the potential of Big Data represents an opportunity to transform biomedical science, leading to new discoveries and better healthcare for the European citizen, with the attendant economic benefits that this brings.

Big Data alone is insufficient; it needs to be translated into Big Science. This can be done only through the development of an underlying information commons framework coupled with a knowledge network to create a Big Data ecosystem. The permissive environment to enable such an ecosystem to develop will also need to be established, with an appropriate regulatory framework, human resources, infrastructure, and a data-sharing culture.

Multiple issues remain to be resolved. One of the biggest hurdles relates to the issue of the privacy and confidentiality of the citizen and the re-use of data for purposes other than that for which it was gathered originally. However, EU legislation on privacy must not hamper biomedical research that is aimed at improving the quality of citizens' lives.

Data-sharing protocols and inter-operability of databases need to be developed to produce the environment that will enable the evolution of an information commons and knowledge networks.

New training and careers structures in biomedical research need to be devised that recognise and reward the contribution of individuals to large, interdisciplinary research efforts.

The creation of a European Public Private Partnership for Big Data involving the European Commission, industry and academia partners is currently underway. Fostering such a partnership between the stakeholders involved in biomedical and health research including the public and private sectors will be essential to leverage Big Data for developing a big science approach of health challenges.

As was stressed many times in the conference, beside industry and the regulators, nothing will be achieved without the trust of the public.

Annex

Monday 24 November 2014 // Lecture Hall: the BLACKETT INSTITUTE

9:00 - 9:30 Welcome and Introductions
9:00 – 9:05 **Ms A. Crowfoot**, Science Europe
9:05 – 9:10 **Prof. R. Frackowiak**, CHUV, Lausanne, Switzerland
9:10- 9:30 **Dr. N Kayadjanian**, Science Europe

Session 1 **Opportunities, Risks and Challenges for Big Data to transform Biomedical Research and Healthcare**

1.1 - Big Science needs Big Data
Chairs: Dr. N. Kayadjanian, Science Europe
Prof. H. Barros, University of Porto, Portugal

9:30 - 10:00 CASE STUDY 1: THE HUMAN BRAIN PROJECT
Prof. R. Frackowiak, CHUV, Lausanne, Switzerland

10:00 - 11:00 CASE STUDY 2: BIG DATA AND PERSONALISED MEDICINE
The European landscape
Prof. S. Holgate, School of Medicine, University of Southampton, UK

The US landscape: Precision Medicine: building a knowledge network for biomedical research and a new taxonomy of disease
Prof. S. J. Galli, Stanford University School of Medicine, Stanford, USA

11:15 - 11:45 Deciphering the biology of disease through Big Data system biology
Prof. R. Aebersold, ETH Zurich, Switzerland

11:45 - 12:35 Big Data and epidemiology: the Public Health perspective
Prof. A. Morris, Farr Institute, UK
Prof. M. L. Barreto, Universidade Federal da Bahia, Brazil

14:00 - 14:30 Big Data and the private pharmaceuticals/biotechnologies industry
Prof. R. Barker, Center for the Advancement of Sustainable Medical Innovation, Oxford, UK

1.2- Beyond the Hype of Big Data
Chairs: Prof. H. Billig, Gotheburg University, Sweden
Prof. R. Aebersold, ETH Zurich, Switzerland

14:30 - 15:00 How to make non-sense of Big Data
Prof. J. P. A. Ioannidis, Stanford University School of Medicine, Stanford, USA

Session 2 **Developing a Long-Term Vision for a Health Big Data Ecosystem**

2.1 - Opportunities and Challenges for creating a Health Big Data Ecosystem
Chairs: Prof. R. Frackowiak, CHUV, Lausanne, Switzerland,
Prof. A. Morris, Farr Institute, UK

15:00 - 15:30 Case-study: The European Medical Information Framework (EMIF) project
Dr B. Vannieuwenhuyse, Janssen Pharmaceutica, Beerse, Belgium

15:30 - 16:00 Mining health records for better research applications and clinical care
Prof. S. Brunak, University of Copenhagen, Denmark

16:00 - 16:30 State of the art of Information Communication Technologies:
Do they support a health Big Data ecosystem?
Prof. Y. Ioannidis, University of Athens, Greece

17:00 - 18:00 Round-table discussion
Panelists:
Prof. A. Morris, Farr Institute, UK,
Prof. R. Frackowiak, CHUV, Lausanne, Switzerland,
Mr. A. Alsop, Economic and Social Research Council, UK,
Prof. P. Elias, University of Warwick, UK,
Prof. E. Hafen, ETH, Switzerland

20:00 - 20:15 The INFN Big Data project **Dr. G. Maron**, INFN, Italy

Annex

Tuesday 25 November 2014 // Lecture Hall: the BLACKETT INSTITUTE

9:00 - 9:45 Private matter and Public concern: Adapting the structural framework for data protection and data sharing
Prof. E. Hafen, ETH, Zurich, Switzerland
Prof. G. Mihalas, Romanian Academy of Medical Sciences, Timisoara, Romania

2.2 - How to Adapt to the Big Data Revolution in Health Research: Developing new Organisational and Funding models and new Talents
Chairs: Prof. J. P. A. Ioannidis, Stanford University, USA,
Mr. A. Alsop, Economic and Social Research Council, UK

9:45 - 10:00 Phenotype Data for Diagnostics and Translational Research: Current Status and Future Challenges
Prof. P.N. Robinson, Freie Universität Berlin, Germany

10:00 - 11:00 Panel discussion
Prof. P. Elias, University of Warwick, UK,
Prof. P.N. Robinson, Freie Universität Berlin, Germany,
Ms G. Clement-Stoneham, Medical Research Council, UK,
Dr. G. Maron, INFN, Italy,
Prof. E. Hafen, ETH, Switzerland,
Dr. M. Radwanska, Science Europe

2.3 - Workplan: Opportunities, Challenges, and Recommendations
Chairs: Prof. S. J. Galli, Stanford University, USA,
Prof. S. Holgate, School of Medicine, University of Southampton, UK

11:30 - 13:00 Panel discussion
Prof. R. Frackowiak, CHUV, Lausanne, Switzerland,
Prof. J. P. A. Ioannidis, Stanford University, USA,
Mr. A. Alsop, Economic and Social Research Council, UK,
Prof. A. Morris, Farr institute, UK

2.4 - Closing remarks

13:00 - 13:15
13:00 – 13:10 **Dr. B. Wolff-Boenisch**, Science Europe
13:10 – 13:15 **Prof. R. Frackowiak**, CHUV, Lausanne, Switzerland

14:15 Meeting closes

Organising Committee

Nathalie Kayadjanian, Henrique Barros, Hakan Billig, Monique Capron, Adam Cohen, Gilles Dubochet, Richard Frackowiak, Annette Grüters-Kieslich, Stephen Holgate, Stephan Kuster, Tullio Pozzan, Martin Vingron, Bonnie Wolff-Boenisch, Magdalena Radwanska.

Acknowledgements

The Organising Committee would like to thank the Italian Institute of Nuclear Physics (INFN) and the Ettore Majorana Foundation, in particular Dr. Valerio Vercesi and Ms Fiorella Ruggiu, for their help in organising and hosting the workshop.

For further information contact:
Dr Nathalie Kayadjanian, Senior Scientific Officer, Medical Sciences
Email: office@scienceeurope.org

References

[1] https://www.humanbrainproject.eu/

[2] http://www.imi.europa.eu/content/u-biopred

[3] http://www.imi.europa.eu/content/emif

[4] National Research Council. Toward Precision Medicine: Building a Knowledge Network for Biomedical Research and a New Taxonomy of Disease. Washington, DC: The National Academies Press, 2011

[5] http://www.human-phenotype-ontology.org/

[6] https://www.humanbrainproject.eu/

[7] http://www.emif.eu/

[8] Hafen et al. (2014). Health data cooperatives - citizen empowerment. Methods Inf Med.,53, 82-86.

[9] https://www.openaire.eu/

[10] http://www.scot-ship.ac.uk/sites/default/files/Reports/SHIP_BLUEPRINT_DOCUMENT_final_100712.pdf

[11] http://www.ref.ac.uk/

Science Europe is a non-profit organisation based in Brussels representing major Research Funding and Research Performing Organisations across Europe.

More information on its mission and activities is provided at: www.scienceeurope.org.

To contact Science Europe, email office@scienceeurope.org.

SCIENCE EUROPE
Shaping the future of research

Science Europe
Rue de la Science 14
1040 Brussels
Belgium

Tel +32 (0)2 226 03 00
Fax +32 (0)2 226 03 01
office@scienceeurope.org
www.scienceeurope.org