

OPINION

Open Access



# Making sense of big data in health research: Towards an EU action plan

Charles Auffray<sup>1,2\*</sup>, Rudi Balling<sup>3\*</sup>, Inês Barroso<sup>4</sup>, László Bencze<sup>5</sup>, Mikael Benson<sup>6</sup>, Jay Bergeron<sup>7</sup>, Enrique Bernal-Delgado<sup>8</sup>, Niklas Blomberg<sup>9</sup>, Christoph Bock<sup>10,11,12</sup>, Ana Conesa<sup>13,14</sup>, Susanna Del Signore<sup>15</sup>, Christophe Delogne<sup>16</sup>, Peter Devilee<sup>17</sup>, Alberto Di Meglio<sup>18</sup>, Marinus Eijkemans<sup>19</sup>, Paul Flicek<sup>20</sup>, Norbert Graf<sup>21</sup>, Vera Grimm<sup>22</sup>, Henk-Jan Guchelaar<sup>23</sup>, Yi-Ke Guo<sup>24</sup>, Ivo Glynne Gut<sup>25</sup>, Allan Hanbury<sup>26</sup>, Shahid Hanif<sup>27</sup>, Ralf-Dieter Hilgers<sup>28</sup>, Ángel Honrado<sup>29</sup>, D. Rod Hose<sup>30</sup>, Jeanine Houwing-Duistermaat<sup>31</sup>, Tim Hubbard<sup>32,33</sup>, Sophie Helen Janacek<sup>20</sup>, Haralampos Karanikas<sup>34</sup>, Tim Kievits<sup>35</sup>, Manfred Kohler<sup>36</sup>, Andreas Kremer<sup>37</sup>, Jerry Lanfear<sup>38</sup>, Thomas Lengauer<sup>12</sup>, Edith Maes<sup>39</sup>, Theo Meert<sup>40</sup>, Werner Müller<sup>41</sup>, Dörthe Nickel<sup>42</sup>, Peter Oledzki<sup>43</sup>, Bertrand Pedersen<sup>44</sup>, Milan Petkovic<sup>45</sup>, Konstantinos Pliakos<sup>46</sup>, Magnus Rattray<sup>41</sup>, Josep Redón i Màs<sup>47</sup>, Reinhard Schneider<sup>3</sup>, Thierry Sengstag<sup>48</sup>, Xavier Serra-Picamal<sup>49</sup>, Wouter Spek<sup>50</sup>, Lea A. I. Vaas<sup>36</sup>, Okker van Batenburg<sup>50</sup>, Marc Vandelaer<sup>51</sup>, Peter Varnai<sup>52</sup>, Pablo Villoslada<sup>53</sup>, Juan Antonio Vizcaíno<sup>20</sup>, John Peter Mary Wubbe<sup>54</sup> and Gianluigi Zanetti<sup>55,56</sup>

## Abstract

Medicine and healthcare are undergoing profound changes. Whole-genome sequencing and high-resolution imaging technologies are key drivers of this rapid and crucial transformation. Technological innovation combined with automation and miniaturization has triggered an explosion in data production that will soon reach exabyte proportions. How are we going to deal with this exponential increase in data production? The potential of “big data” for improving health is enormous but, at the same time, we face a wide range of challenges to overcome urgently. Europe is very proud of its cultural diversity; however, exploitation of the data made available through advances in genomic medicine, imaging, and a wide range of mobile health applications or connected devices is hampered by numerous historical, technical, legal, and political barriers. European health systems and databases are diverse and fragmented. There is a lack of harmonization of data formats, processing, analysis, and data transfer, which leads to incompatibilities and lost opportunities. Legal frameworks for data sharing are evolving. Clinicians, researchers, and citizens need improved methods, tools, and training to generate, analyze, and query data effectively. Addressing these barriers will contribute to creating the European Single Market for health, which will improve health and healthcare for all Europeans.

## European healthcare systems and the potential for big data

Medicine has traditionally been a science of observation and experience. For thousands of years, clinicians have integrated the knowledge of preceding generations with their own life-long experiences to treat patients according to the oath of Hippocrates; mostly based on trial and

error. Knowledge generation is changing dramatically. The digitalization of medicine allows the comparison of disease progression or treatment responses from patients worldwide. Whole-genome sequencing allows searching and comparing one's own genome to millions and soon billions of other human genomes. Eventually, the entire world population could be used as a reference population in order to link genome information with many other types of physiological, clinical, environmental, and lifestyle data. For many, this is a vision full of opportunities, whereas for others it provides a wealth of technical

\* Correspondence: [cauffray@eisbm.org](mailto:cauffray@eisbm.org); [rudi.balling@uni.lu](mailto:rudi.balling@uni.lu)

<sup>1</sup>European Institute for Systems Biology and Medicine, 1 avenue Claude Vellefaux, 75010 Paris, France

<sup>3</sup>Luxembourg Centre for Systems Biomedicine, University of Luxembourg, 7 Avenue des Hauts Fourneaux, 4362 Esch-sur-Alzette, Luxembourg

Full list of author information is available at the end of the article

challenges, unanticipated consequences, and loss of privacy and autonomy.

The quality of conclusions on the etiology of diseases follows a law of large numbers. Cross-sectional cohort studies of 30,000 to 50,000 or more cases are required to separate the signal from noise and to detect genomic regions associated with a given trait in which disease-related genes or susceptibility factors are located [1, 2]. Whole-genome sequencing studies often identify only a few genomic regions that contain elements with large effects on the penetrance or expressivity of gene products but hundreds of genomic regions that have small effects and are highly dependent on genetic background, environmental factors, or social and lifestyle determinants [3]. There is also a need to study disease pathogenesis on genome, epigenome, transcriptome, proteome, and metabolome levels and combine these dimensions through multi-omics research. Furthermore, individual variation responsible for normal and disease phenotypes is high as a result of somatic mutations or variation in transcription, splicing, or allele-specific gene expression between individuals [4–6].

Vast amounts of temporal and spatial parameter data are now available. But what are we going to do with the data? It takes hard work to condense useful information from big data and turn this information into knowledge and action. The challenge will be to make a smart choice between situations when less is more versus less is less but also when more is more versus more is less.

Here, we briefly describe the key challenges that result from making sense out of big data in health and using these data for the benefit of the patient and the healthcare system. We also highlight key technical, legal, and ethical issues that we face to develop evidence-based personalized medicine. Finally, we put forward five recommendations for the European Union (EU) and member states' policy makers to serve as a framework for an EU action plan that could help to reach this ambitious goal.

### **Making sense of big data in health research**

On 30 October 2015, the Health Directorate of the Directorate-General for Research and Innovation at the European Commission (EC), the executive body of the EU, organized in Luxembourg a workshop entitled “Big data in health research: an EU action plan” [7]. The aim of the workshop was to ask stakeholders in the “big data revolution” for their input on how European funding for health research should take into account the opportunities, limitations, and concerns of the anticipated developments in health and healthcare. Participants included bioinformaticians, computational biologists, genome scientists, drug developers, biobanking experts, experimental biologists, biostatisticians, information and communication technology (ICT) experts, public health researchers,

clinicians, public policy experts, representatives of health services, patient advocacy groups, the pharmaceutical industry, and ICT companies.

### **What do we mean by “big data”?**

“Big data” has a wide range of definitions in health research [8, 9] and to create a single definition for all uses (“one size fits all” approach) may be too abstract to be useful. However, a workable definition of what big data means for health research or at least a consensus of what this term means was proposed during the workshop in Luxembourg. “Big data in health” encompasses high volume, high diversity biological, clinical, environmental, and lifestyle information collected from single individuals to large cohorts, in relation to their health and wellness status, at one or several time points. Big data can only be dealt with by adopting a strong governance model and best practices of new technologies, e.g., in large-scale data production compliant with community-based quality standards, coupled with interoperable data storage, data integration, and advanced analytics solutions [10]. Another goal of the workshop was to develop an EU action plan for research funders towards the integration of big data into policy development, biomedical research, and clinical practice in health and wellness management. Big data comes from a variety of sources, such as clinical trials, electronic health records (EHRs), patient registries and databases, multidimensional data from genomic, epigenomic, transcriptomic, proteomic, metabolomic, and microbiomic measurements, and medical imaging. More recently, data are being integrated from social media, socioeconomic or behavioral indicators, occupational information, mobile applications, or environmental monitoring [11]. Big data comes in a wide range of formats. Data streams have to be assessed and interpreted in a timely manner to benefit patients affected by diseases and to help citizens remain in good health [8, 12].

### **Importance of patient registries**

Patient registries have for decades served as a key tool for assessing clinical outcome and clinical and health technology performance [13–15]. Rare disease registries pool data to achieve a sufficient sample size for epidemiological and/or clinical research [16, 17]. The European Organization for Research and Treatment of Cancer (EORTC) [18] opened a prospective registry for patients with melanoma in June 2015 [19]. The European Network of Cancer Registries (ENCR) [20], established within the framework of the Europe Against Cancer Programme of the EC, promotes collaboration between cancer registries, defines data collection standards, provides training for cancer registry personnel, and regularly disseminates information on incidence and mortality from cancer in the European Union and other European countries.

Patient registries provide significant potential for research and public health improvements in the EU, owing to the large volume of patients in each registry and the variety of quality medical information related to each patient. Patient registries are increasingly important to monitor patients' treatments and for safety assessment and the identification of trends in translational medicine (e.g., registry-based clinical trials, personalized medicine) [21].

Patient registries allow informed policy decisions at the local, regional, national, and, in some cases, the international level. As a result, hundreds of registries have been set up that range from national to international rare disease initiatives, coupling clinical and genetic data and biobanks. However, for various reasons, including data protection and the fragmentation of regulatory frameworks, the combination of these disparate information sources to guide health research and decision-making in the clinic has so far lagged behind the use of large-scale, big data collections in other sectors. Other disciplines, such as electronic and mechanical engineering, and whole industries, such as building airplanes, weather forecasting, or robotics, have demonstrated computational modeling and simulation as an essential component that is based on data sharing and their experience could help overcome the barriers experienced in health research [22–24].

### **The potential benefits of big data for healthcare**

Big data in health can be used to improve the efficiency and effectiveness of prediction and prevention strategies or of medical interventions, health services, and health policies [25–27]. Access to well-curated and high quality health-related data will likely have a number of benefits in a diverse range of situations. In clinical practice, these data will improve outcomes for individual patients through personalization of predictions, earlier diagnosis, better treatments, and improved decision support for clinicians in cyclic processes. (Cyclic processes are usually composed of the definition of policy/decision options, the selection of the best alternative, and the subsequent implementation and validation of this option. Integrating feedback from a continuous evaluation on the process completes this cycle [28].) These improvements should eventually lead to lowered costs for the healthcare system.

Likewise, the integration of fragmented information systems into the clinical life cycle will allow the discovery of medically relevant associations, early signals, or changed disease trajectories and should, therefore, enable better patient management strategies and improved quality and safety of care. For clinical trials, more expansive, interoperable health records should make it much easier to find suitable participants and to design and assess the feasibility of new studies [29, 30]. Moreover, better

management of big data would enable a more systematic identification of drug safety signals, such as earlier detection of adverse drug reactions [31], while allowing personalized medicine analyses via appropriate patient and/or population stratification methodologies. This in turn should lead to improved treatment responses for biologically or clinically defined patient subgroups, which will also avoid unnecessary rejection of potent drugs and devices. As a result, patient communities will benefit and the unsustainable trend of escalating costs in hospital and community care management as well as diagnostic and drug development costs by the biopharma industries will stop or slow down. Health economy specialists need to provide suitable metrics to monitor key performance indicators of success in big data pilot projects. Such metrics might include the change in response rates in stratified patient subpopulations or the number of adverse drug reactions after systems medicine-based companion diagnostics.

Big data also has many potential benefits for translational research into health and well-being. Integrated data sets should improve models of common disease to better understand the progression of rare diseases [32]. They may also enable the detection of population-level effects, such as the off-target and adverse effects of drugs or the occurrence of co-morbidity [33].

Biomarkers constitute a key building block of precision medicine, yet the development and clinical validation of new biomarkers is a lengthy process and relatively few such markers have yet reached routine clinical practice [34, 35]. However, a sizeable number of biomarkers are now widely used in routine clinical diagnostics, which include—but are not limited to—targeted cancer therapy [36–39]. Multidimensional signatures that take into account a wealth of prior information, both from the patient's previous life history and state-of-the-art information from the literature and relevant databases, will hopefully deliver a much higher predictive power than the single biomarkers used today. There is also potential for research on the impact of healthcare interventions and monitoring trends in infectious diseases to inform public health policies [40].

Finally, there is an opportunity to engage with the individual patient more closely and import data from mobile health applications or connected devices. This interaction with the patient will result in the collection of more detailed clinical, environmental, and lifestyle information, such as heart frequency and body temperature, physical activity and nutrition habits, and sleep and stress management, which will prevent risk exposure and disease onset [41]. Personal monitoring over time should aid the early detection of deviations from a healthy state and trajectories should lead to actionable recommendations, making it possible for individuals to maintain themselves in good health [12].

### The challenges ahead for the effective use of big data in healthcare

To refine the recommendations for an EU action plan, we identified the main challenges that exist for the use of big data in healthcare research and in the clinic. The challenges have been reported elsewhere [42, 43] and include clinical, technical, legal, and cultural hurdles. These challenges vary depending on whether the data are preclinical based on cellular or animal models or from patients in clinical settings, on the intended type of analysis and interpretation, on cross-cultural aspects of privacy, and on ethical and legal considerations. We are on the cusp of having access to vast personal data—for example, on physiological, behavioral, molecular, clinical, environmental exposure, medical imaging, disease management, medication prescription history, nutrition, or exercise parameters—that could potentially be used to track the health of individuals and populations in considerably more detail than ever before. The integration of structured and unstructured data, using natural language processing and other sophisticated machine learning tools, is being tested and it is hoped this will lead to a new level of integration of prior information with up-to-date clinical information [44].

Over a thousand Mendelian disorders are linked to genetic defects and, for many of these, genetic testing is performed to inform clinical practice. The most successful integration of basic and clinical data can be observed in oncology [45, 46] and in research on rare diseases [47–49]. However, the medical relevance of the large amount of genetic variation revealed by genomic sequencing is still unknown in most cases.

Data acquisition is undergoing rapid change. Wearable devices, integrated sensors, and continuous monitoring capabilities are available for all scales of measurements [50]. Several legal issues will have to be tackled, for example, when a consumer device becomes a diagnostic device and the quality assurance and regulatory approval are more stringent [51].

Data storage issues include security, accessibility, and sustainability. Should data be stored centrally or in a federated manner? There are concerns about entrusting health-related data to public clouds. As a result, there is a strong need to come up with alternatives. The decades of experience in big data management for the particle physics community at The European Organization for Nuclear Research (CERN) that led to the development of the World Wide Web [52] will be valuable. However, many aspects that are specific to big data in health research need to be taken into account, such as data heterogeneity, institutional and legal fragmentation, and strong data protection standards. There will be a massive increase in big data production in all areas of biomedical research, which includes studies at the preclinical

levels, such as animal or cellular models and translational studies, but also clinical research that involves patients or public health research. To make the most use of the information produced, several technical challenges should be addressed, such as the combination of structured data, such as genotype, phenotype, and genomics data, with semi-structured and unstructured data, e.g., medical imaging, EHRs, lifestyle, environmental, and health economics data [53–56]. Recent successful examples show the feasibility of combining such data for translational and clinical research [57–59].

### Technical challenges related to the management of electronic health records

Adoption of EHRs across Europe varies greatly. Estonia [60] and the Valencia Community in Spain (Josep Redón i Màs, personal communication) have moved entirely to EHRs. Integration is supported with auxiliary systems, for instance drug–drug interaction alert systems that warn physicians and pharmacists about potential prescription clashes, clinical risk groups calculation and costs (e.g., Valencia region, Spain), and drug–gene interaction alert systems that guide physicians to adjust the dose of a prescribed drug in aberrant drug metabolizers (e.g., The Netherlands). The USA have taken steps towards a “patient-driven economy” [61]. In such a scenario, the patient owns his/her data. This ownership requires the development of an appropriate health-record infrastructure but provides a wide range of new health service business opportunities with major economic potential. Empowering patients to take control of their data could be of particular importance for cross-border healthcare and health research activities in Europe where healthcare is highly fragmented and multinational. To transfer medical data from one country to another in the EU is very difficult. Ownership of data by patients could overcome these obstacles and unleash new ways to stimulate a competitive health-driven economy.

Furthermore, patient records can be computationally opaque, for example, in the form of free text, recorded speech, or medical images; translation into a format compatible with computational analyses will be necessary. Data in different languages and time-consuming searches and identification are other important barriers.

There are some best practices for the management of EHRs. For example, the International Rare Diseases Research Consortium (IRDiRC) [62] develops and implements standards and harmonized methodology across diseases and medical cases [63]. Several European collaborative projects, such as the European project p-medicine, have created IT infrastructures that will facilitate translational research and the development of personalized medicine [64]. ELIXIR, one of the European infrastructures for life sciences [65], has facilitated the collection, quality control, and



archiving of large amounts of life science data such as translational medicine data [66].

### **Technical challenges related to data analysis and computing infrastructures**

Basic as well as clinical researchers need new computational tools to improve data access and aid user-friendly data analysis for efficient decision making in the clinic. Clinicians need new tools that track, trace, and provide fast feedback for individual patient care. Researchers need tools that can be adapted for different data sets and analyses such as those used in a wide range of EU-funded projects through the Innovative Medicines Initiative (IMI)-funded eTRIKS consortium platform [67]. Accessing tool repositories to search for the best tool to answer specific research or clinical questions will be a prerequisite. Equally important are traceable computational environments that maintain data provenance information from patient to sample and from sample to clinically actionable results. In December 2012, the UK announced the 100,000 Genomes Project [68], which aims to sequence 100,000 genomes, from around 70,000 people, with the focus on patients with rare diseases or cancer. The US and China have recently announced plans for similar studies on one million individuals. The goal of these projects is to yield further insights into human health and disease and to build a framework with which to integrate genomics into standard public healthcare programs in the near future. Data continue to increase at an exponential rate and the need for cross-border exchange of biomedical and healthcare data, cloud-storage, and cloud-computing is inevitable [69, 70]. Until many issues of data safety and security are solved, however, local solutions will be favored [71].

### **Data quality, acquisition, curation, and visualization**

The quality and structure of health data available is inconsistent. A major challenge for preclinical and clinical research is to obtain and achieve access to sufficient high quality, informative data. Owing to a lack of harmonized methods, in most cases health data cannot be directly used for secondary purposes, such as quality of care, pharmacovigilance, safety and efficacy of treatments, health technology assessment, and public health policy. Efforts are underway, in both Europe and the US, to develop and implement standardized data collection, storage, and analysis [10, 72, 73]. The European Open Science Cloud, created by the EC, will offer Europe's 1.7 million researchers and 70 million science and technology professionals a virtual environment to store, share, and re-use their data across disciplines and borders [74].

Data curation is often neglected but vitally important to warrant high-quality, informative data [75]. Research

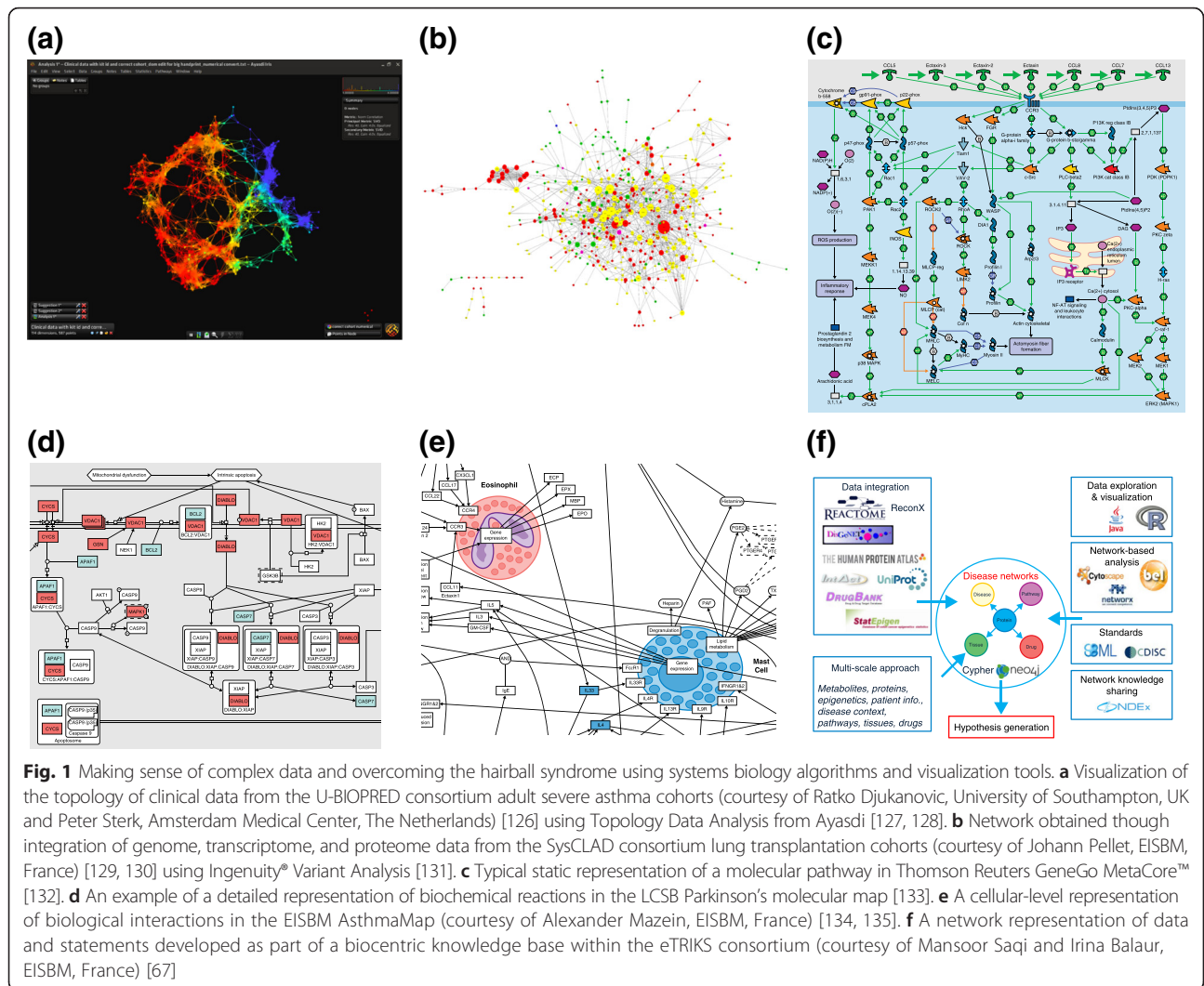
funders need to make sure that sufficient attention is paid to data quality at the experimental and study design stages, for example, by ensuring data management plans and appropriately reviewed data sharing procedures are in place for all funded research.

*"Seeing is believing"*. This phrase is relevant not only for high-resolution microscopy and imaging technologies but also for the presentation and visualization of health-related data. We need to progress from the current display of "hairballs", incomprehensible comprehensive networks, or ranking tables that nobody has the time or motivation to look at. If we want to provide clinicians with updated, relevant information and clinical decision support systems, the devices have to be user-friendly and intuitive with an interoperable format. The concept of disease-specific maps, with a common computational framework, might be one way to make progress, as demonstrated in several EU-funded projects (Fig. 1).

### **Computational modeling and simulation**

One of the pathways for exploitation of big data is its combination with predictive, mechanistic models [76] such as those provided by the European Molecular Biology Laboratory–European Bioinformatics Institute (EMBL-EBI) [77]. The Virtual Physiological Human (VPH) community has also endeavored to develop a descriptive, integrative, and predictive computational framework of human anatomy, physiology, and pathology with support from the EC Directorate General for Communications Networks, Content & Technology (DG CONNECT) [78, 79], following the path opened by the IUPS Physiome Project [80]. Predictive computational approaches are associated with infrastructural challenges, particularly for the integration of data with analytical tools and workflows. Online environments such as VPH-Share and projects such as p-medicine have appropriate infrastructures for these applications [81].

Another approach to make sense of big data is based on a systems-level understanding of health and disease [82]. Systems medicine integrative approaches are gradually gaining visibility and enable translation of the human biology complex and voluminous data into a toolbox to demonstrate clinical impact [83]. However, a full appreciation of the power of systems biology and computational modeling for the upcoming changes in health research and healthcare is still missing. Currently, with the exception of oncology, there are still few highly convincing use cases where systems biology approaches have found applications in routine clinical care [45, 46, 84, 85]. Mathematical, computational disease models are unlikely to be routine in health research anytime soon. Achieving necessary changes will need strong support from funders to foster this paradigm shift in methodology.



**Fig. 1** Making sense of complex data and overcoming the hairball syndrome using systems biology algorithms and visualization tools. **a** Visualization of the topology of clinical data from the U-BIOPRED consortium adult severe asthma cohorts (courtesy of Ratko Djukanovic, University of Southampton, UK and Peter Sterk, Amsterdam Medical Center, The Netherlands) [126] using Topology Data Analysis from Ayasdi [127, 128]. **b** Network obtained through integration of genome, transcriptome, and proteome data from the SysCLAD consortium lung transplantation cohorts (courtesy of Johann Pellet, EISBM, France) [129, 130] using Ingenuity® Variant Analysis [131]. **c** Typical static representation of a molecular pathway in Thomson Reuters GeneGo MetaCore™ [132]. **d** An example of a detailed representation of biochemical reactions in the LCSB Parkinson's molecular map [133]. **e** A cellular-level representation of biological interactions in the EISBM AsthmaMap (courtesy of Alexander Mazein, EISBM, France) [134, 135]. **f** A network representation of data and statements developed as part of a biocentric knowledge base within the eTRIKS consortium (courtesy of Mansoor Saqi and Irina Balaur, EISBM, France) [67]

### Legal and regulatory aspects

A crucial aspect to be addressed concerns the regulatory acceptance of big data for the evaluation of novel pharmacological or biological therapies to complement large randomized clinical trials [86]. Collaborative pilot projects that test the use of big data in observational and/or interventional large clinical trials with the contribution of regulatory agencies can bridge different methodological approaches and determine adapted quality standards. Universities and hospitals do not have the procedures in place to effectively capture and share data with other organizations and countries. We need to develop and adopt high quality standards for data generation and processing to ensure that meaningful and valid data with well-defined semantics are processed and shared. The quality of data generation as well as the processing and regulatory acceptance of big data are addressed at the international level. Research initiatives such as

the International Cancer Genome Consortium (ICGC, 2016) [87], the International Human Epigenome Consortium (IHEC, 2016) [88], the Genomic Standards Consortium (GSC, 2016) [89] and the Clinical Data Interchange Standards Consortium (CDISC) [90] and by ISO standards committees (e.g., ISO TC276 WG5, 2016) provide some examples [91]. The recently published FAIR Data Principles of Findability, Accessibility, Interoperability and Reusability for scientific data management should help stakeholders from academia, industry, funding agencies, and non-commercial publishers support the reuse of scholarly data [92]. Given the complexity and high number of stakeholders involved in the implementation of data standards within hospital and university settings, the biggest chance for success comes with highly focused pilot projects. Key factors include flexibility, expansion through modular strategies, and the identification and involvement of key healthcare actors providing them with immediate benefits.

Linking existing initiatives and building new initiatives on clinical data interchange are also important. The Global Alliance for Genomics and Health (GA4GH, 2016) [93] initiative is working towards technical, ethical, and legal frameworks to address and resolve some of these issues. The Coordinated Research Infrastructures Building Enduring Life-science (CORBEL) Services, a recently launched European consortium, will also contribute to the above data-sharing challenges [94]. CORBEL is an initiative of 11 new Biological and Medical Science Research Infrastructures (BMS RIs), who together will create a platform for harmonized user access to biological and medical technologies, biological samples, and data services (e.g., BRIDGEHEALTH consortium [95]). The Genomics England policy [68] of storing all data within the National Health Service (NHS) with highly regulated restricted access to prevent abuse of private information and user protocols might be a way to go forward. This policy will need to be complemented by that of the UK Personal Genome Project allowing volunteers to donate their personal genome from Genomics England and other sources to the public domain [96].

The processes and legal agreements for data sharing across registries and European Member States are seldom established. The harmonization of regulatory frameworks is crucial while also ensuring personal data protection and compliance with current legal frameworks, which includes provisions on how to prevent, handle, and prosecute potential abuse of the system. For example, there is no consensus within international law on whether specific requirements should be applicable to genetic information. Several documents exist at the regional and international levels that include useful guidelines, such as the UNESCO International Declaration on Human Genetic Data (2003) [97] and the Organisation for Economic Co-operation and Development (OECD) Guidelines on Human Biobanks and Genetic Research Databases (2009) [98]. The GA4GH has developed the Framework for Responsible Sharing of Genomic and Health-Related Data [99].

### **Privacy protection and data sharing policies**

There are broad differences within and across Europe with regards to privacy protection and data sharing policies [100]. The workshop in Luxembourg emphasized that the “one size fits all” approach will not be applicable in Europe. The EC proposal for the General Data Protection Regulation (2012/0011COD) [101] attempts to harmonize the fragmented situation that exists under the current Data Protection Directive (95/46/EC, European Parliament and Council, 1995). In the compromise text concluded in the trilogue negotiations between Parliament and the Council, a paragraph is included in the preamble of the new act which defines DNA and RNA as personal data [102]. A Code of Practice on Secondary Use of Medical Data in

European Scientific Research Projects has been developed [103] and is being deployed in the IMI-funded project eTRIKS [67]. There is also a need to have a much higher level of security than is possible today. One suggestion was to explore block-chain technology, which makes use of a digital, distributed transaction record, digital events, with identical copies maintained on multiple computer systems, shared between many different parties. Once entered, the block-chain contains a certain and verifiable record of every single transaction [104]. Originally used as the technology underlying “Bitcoin”, the potential to make secure transactions of biomedical and healthcare data is being explored [105]. Another possibility would be to make use of differentiated privacy approaches as practiced in health information exchanges [106].

### **Research infrastructures**

Similarly, research infrastructures are instrumental to support the harmonization of legal and ethical frameworks in European countries, as demonstrated by the Common Service on Ethical, Legal and Social Implications (CS ELSI) of Biobanking and BioMolecular resources Research Infrastructure Consortium (CS ELSI BBMRI-ERIC, 2016) [107]. The goal of ELSI BBMRI-ERIC is to facilitate and support cross-border exchanges of human biological resources and data attached for research uses, collaborations, and sharing of knowledge, experiences, and best practices.

Existing computational infrastructures are coping with storage of big data, but the challenge within the EU is the lack of a large-scale European infrastructure and methods of secure data distribution in a cross-border setting [108]. It is crucial to ensure that the infrastructures that exist and evolve are coordinated and sustainable. Initiatives such as ELIXIR [65] and the CS IT BBMRI-ERIC [109] have begun to address these issues but there is a need for coordination and significant strategic investments to ensure that organizations such as these are equipped to support the rapid growth and evolution of healthcare informatics over the next decade. Distribution of expertise and facilities, consistent operation, and federation throughout Europe are essential for scalability and long-term sustainability. This has become one of the key challenges of distributed infrastructures such as BBMRI-ERIC [110], ECRIN [111], and ELIXIR [65], which could benefit from the long experience of CERN in particle physics as discussed earlier [52].

### **Training and education: many health data, insufficient health data scientists**

One of the biggest bottlenecks and challenges is the availability of healthcare professionals and clinical researchers that are able to use the latest information technologies developed in the big data analytics era [112, 113]. Data

managers with good insight into the specificities of the health application domain are rare. An equally important bottleneck will be the lack of trained clinical scientists to deal with big data. The majority of university hospitals face a daily struggle to balance their budget. Clinical research rarely brings in money to pay the costs for clinical care. As a result, many university hospitals cease to maintain their culture of research as an essential basis for top-level healthcare. Once the chain of training the next generation of clinical scientists is broken by the retirement of the current trainers, the situation will change dramatically and result in a catastrophic shift. Therefore, there is a pressing need for programs that support the careers of clinical scientists with state-of-the-art training in data analysis and management.

There is a clear lack of cross-disciplinary education and training, which means that employees in the clinical environment often do not have the expertise to deal with big data in clinical research and healthcare. Coordinating Action Systems Medicine (CASyM) [83] has developed modules of multidisciplinary training for the next generation of researchers and medical doctors. Furthermore, despite compulsory requirements of data transparency applicable to clinical trials data, researchers and clinicians often have little incentive to make data fully available. Another challenge may be public skepticism about the security of an integrated healthcare system. However, several global initiatives have shown that individuals are ready to share their medical data for advancing science (Personal Genome Project) [96], which highlights the potential contribution of citizen science to big data in health research. Data donor cards would provide an incentive for people to make their data publicly available and would work in the same way as organ donor cards, thereby reusing a system already understood by many people. Legislative approaches should include opt-in and or opt-out solutions. For a successful transformation of healthcare, we need to push the boundaries of interdisciplinarity, which comprises the natural sciences such as biology and medicine, engineering, the social sciences, and the humanities. Projects fail more often because of the underappreciation of the complexities of ethical, legal, and social factors than for technological reasons.

The workshop in Luxembourg brought together a wide range of experts and stakeholders to discuss the key developments, challenges, and potential solutions that we face with using big data for the benefit of the patients, the health care industry, and Europe as a whole. The workshop resulted in specific recommendations for European policy-makers. There was no doubt among the participants that big data and the revolution in ICT will transform healthcare. There was also a sense of urgency to implement rapidly the possible and to tackle the yet impossible.

## **Recommendations for an EU action plan**

### **Launch pilot projects on the application of big data to inform health**

The primary recommendation is for the launch of pilot projects on the application of big data that involve healthcare providers, health technology developers, policy-makers, and advisory bodies. Pilot translational research projects that involve healthcare workers and patients could bring big data closer to the clinic and prove the value of collecting and analyzing such information using the latest mathematical and computational tools. The design principles for achieving integrated healthcare information systems [114] might serve as guidance on how small pilot projects can be used for future expansion.

### **Leverage the potential of open and citizen science for the exploitation of big data in health**

The concept of “open science” includes open access to publications and raw data, transparency of tools and methodologies, and networking of researchers across fields and countries [115]. Open science provides significant added value in pilot studies and its broad implementation in the scientific community and society is under discussion. For example, a high-profile effort to switch all peer-reviewed publishing to open access within the next years is envisaged [116–119].

The second recommendation is to encourage leveraging the complementarity between open and citizen science in the context of big data in health. It will be important to inform and involve the public not only about data collection but also about all aspects of health research [120]. Consumer genomics companies are already successful at gathering metadata through engagement with their customers. The field of rare diseases has also benefitted greatly from the involvement of parents of children with such diseases, using non-traditional techniques such as social media to build a network of related cases of a particular syndrome. “Citizen science” is also becoming increasingly important because of the increased uptake of mobile health devices, consumer electronics, and household appliances and is well-aligned with the EC focus on “responsible research and innovation” that includes elements of open science in its ongoing Horizon 2020 Framework Programme policy [121].

### **Catalyze the involvement of all relevant stakeholders in projects**

The third recommendation is to involve in projects all relevant stakeholders, which includes clinicians, patient organizations, researchers, software providers, healthcare managers, ethical and legal experts, regulatory authorities, policy-makers, pharmaceutical companies, and funding



bodies. Multidisciplinary involvement is required to secure an effective translation from basic research to applied healthcare and to bridge the organizational and cultural differences in data sharing practices across Europe and within the different health sectors in a worldwide context. Clark and colleagues have laid out “*a core set of lessons that should become part of a basic training for researchers interested in crafting usable knowledge for sustainable development*” [122]. One of the most important lessons is to understand that research is a social and political process, not just a process of discovery, and that stakeholders are diverse and need to be involved in the team building process at an early stage.

It is likely that bioinformaticians, biostatisticians, and computational scientists will more often be included in the near future as natural members of research and clinical teams and healthcare administration, as already carried out by global pharmaceutical organizations. Important towards this direction is cross-disciplinary training and to improve the dialogue between the information technology experts, biologists, and clinicians, especially as these groups have the potential to affect greatly the practical outcome of research.

#### **Support a rapid transition to new computational, statistical, and other mathematical methods of analysis**

The fourth recommendation is to foster the transition to new computational, statistical, and other mathematical methods of analysis that enable the integration of data across the multiple scales of time and space typical of complex biological systems in their healthy and diseased states [123]: traditional methods of analysis are no longer scalable for such big data diversity. The roadmap developed by the Avicenna Coordination Support Action provides a vision on how computer simulation will transform the biomedical industry by developing “*in silico* clinical trials” [124].

The need for new methods spans a wide range of topics. We need effective methods for data integration, collection, and data provenance management, for example, the integration of genomics information and patient registries with EHRs and the integration of model organism data into disease models. We also need improved methodologies and tools to support data entry by those recording data, such as visual and physiological information. Innovative statistical methods, such as models for predictive analytics and computational models tailored to big data, are required to enable hypothesis generation, estimation of risk models, and study design. The Infrastructure, Design, Engineering, Architecture, and Integration project (IDeAI) [125] is taking steps in this direction by developing new methods for gene selection to tailor the design for small population group trials. There may even be a requirement for new

types of data and data formats. The development and use of interoperable data, technology standards, and harmonized operating procedures for data collection and analysis are paramount to enable data integration and to support data flow and federated access between public and private partners. Furthermore, applicable data protection standards and maintaining public trust are important to realize the full potential of big data in health research for European citizens and, by extension, worldwide. In this regard, we need a definition of core data sets that could serve as a common standard for any individual health state.

Using the big data revolution to drive the transformation of healthcare requires resources for state-of-the-art ICT infrastructure, training programs, and pilot projects that can serve as a role model. These costs, however, will be overcompensated by the gains that will come with the implementation of functioning digital workflows and sophisticated health data analytics and the creation of a new health and wellness industry.

#### **Accelerate the harmonization of regulatory frameworks in Europe for health-related research and data sharing**

The final recommendation is to agree on the necessity for, and the high priority of, accelerating the harmonization of the European policy and regulatory frameworks that affect health-related research and data sharing and the distribution of biological material used for the generation of data necessary for research. There should be a balance between the protection of an individual’s privacy, while acknowledging that many patients are much more open about data sharing than current policies seem to assume, and the ability to proceed with research to ensure that Europe remains competitive in health research. EU and national funding bodies should take stock of the existing best practices and catalyze their adoption in transnational health research.

#### **Conclusions and future perspectives**

The digital revolution is underway. A number of industries have already transformed their activities or have now become inoperative. The driving forces are miniaturization, automation, and now increasingly the convergence of artificial intelligence, deep learning, and robotics. Healthcare will not escape these developments. In fact, big data as a driving force will play an even more important role than in most industries. In Europe, working across borders is the only way to master the challenges of this scientific, technological, and industrial revolution. The single most important factor is the workforce. Countries that are ahead in ICT competence and have an understanding of cultural differences and an ability and willingness to work together have the best chance to succeed.

## Abbreviations

BBMRI, Biobanking and Biomolecular Resources Research Infrastructure; BMS RI, Biological and Medical Sciences Research Infrastructure; CASyM, Coordinating Action Systems Medicine; CDISC, Clinical Data Interchange Standards Consortium; CORBEL, Coordinated Research Infrastructures Building Enduring Life-science Services; EBI, European Bioinformatics Institute; EC, European Commission; EHR, electronic health record; EISBM, European Institute for Systems Biology and Medicine; ELSI, ethical, legal, and social implications; EMBL, European Molecular Biology Laboratory; ENCR, European Network of Cancer Registries; EORTC, European Organisation for Research and Treatment of Cancer; ERIC, European Research Infrastructure Consortium; EU, European Union; EURORDIS, Rare Diseases Europe; GA4GH, Global Alliance for Genomics and Health; ICGC, International Cancer Genome Consortium; IHEC, International Human Epigenome Consortium; IML, Innovative Medicines Initiative; ISO, International Organization for Standardization; LCSB, Luxembourg Centre for Systems Biomedicine; NHS, National Health Service; OECD, Organization for Economic Co-operation and Development; UNESCO, United Nations Educational, Scientific and Cultural Organization; VPH, virtual physiological human

## Acknowledgements

We would like to thank the organizers of the EC workshop Anders Colver, Tomasz Dylag, Christina Kyriakopoulou, and Sasa Jenko, who serve as scientific officers at the EC Health Directorate.

Big data in health research: an EU action plan workshop was organized by the Health Directorate of the Directorate-General for Research and Innovation at the European Commission with the contribution of the Innovative Medicines Initiative office, Digital Society, Trust & Security Directorate of the Directorate General for Communications Networks, Content & Technology, Health systems and products Directorate of Directorate-General for Health and Food Safety, Joint Research Centre and EUROSTAT <http://bigdata2015.uni.lu/eng/European-Commission-satellite-workshop>.

<http://ec.europa.eu/research/health/index.cfm#>

We thank Alvar Agusti, Jacques Beckmann, Laurent Nicod, Andres Metspalu, Damjana Rozman, Philippe Sabatier, Ferran Sanz, Peter Sterk, Giulio Superti-Furga, Jesper Tegnér, Olaf Wolkenhauer, and two anonymous reviewers, who provided insightful comments that helped to improve the manuscript. Figure 1 was prepared by Bertrand De Meulder, Alexander Mazein, Johann Pellet, Mansoor Saqi, and Irina Balaur.

Workshop participants represent the following projects supported by the European Union's Horizon 2020 and the Seventh Framework Programme: AETIONOMY (Organising mechanistic knowledge about neurodegenerative diseases for the improvement of drug development and therapy, IML-n° 115568), ASTERIX (New methodologies for clinical trials for small population groups, FP7-n°603160), BBMRI ERIC, BLUEPRINT (A Blueprint of Haematopoietic Epigenomes, FP7-n°282510), BRIDGEHealth (BRIdging Information and Data Generation for Evidence-based Health policy and research, H2020-n°664691), CANCER-ID (Cancer treatment and monitoring through identification of circulating tumour cells and tumour related nucleic acids in blood FP7-n°115749), CASyM (Coordinating Action Systems Medicine—Implementation of Systems Medicine across Europe, FP7-n°305033), COMBIMS (A novel drug discovery method based on systems biology: combination therapy and biomarkers for Multiple Sclerosis, FP7-n°305397), DECIPHER PCP (Distributed European Community Individual Patient Healthcare Electronic Record, FP7-n°288028), ECHO (European Collaboration for Healthcare Optimization, FP7-n°242189), ELIXIR (European Life-science Infrastructure for Biological Information, FP7-n°211601), EMIF (European Medical Information Framework, IML-n°115372), EpiGeneSys (Epigenetics towards systems biology, FP7-n°257082), ERA-IB (ERA-Net for Industrial Biotechnology 2, FP7-n°291814), ERASynBio (Development and Coordination of Synthetic Biology in the European Research Area, FP7-n° 291728), ERASysAPP (Systems Biology Applications, FP7-n°321567), ESGI (European Sequencing and Genotyping Infrastructure, FP7-n°262055), eTRIKS (Delivering European Translational Information & Knowledge Management Services, IML-n°115446), EU-MASCARA, EUROBIOFORUM, European Lung Foundation, IDeAI (Integrated Design and Analysis of small population trials, FP7-n°602552), KConnect (H2020-n°644753), MedBioinformatics (Creating medically-driven integrative bioinformatics applications focused on oncology, CNS disorders and their comorbidities, H2020-n°634143), MeDALL (Mechanisms of the Development of ALLergy, FP7-n°261357), MIMOMics (Methods for Integrated analysis of Multiple Omics datasets, FP7-n°305280), MULTIMOD (Multi-layer network modules to identify markers for personalized medication in complex diseases, FP7-n°223367), IML ND4BB TRANSLOCATION

(New Drugs 4 Bad Bugs, IML-n°115525), PARENT (PAtient REgistries iNitiative, CHAFAEA Project Grant n°2011 23 02), p-medicine (From data sharing and integration via VPH models to personalized medicine, FP7-n°270089), PREDEMICS (Preparedness, Prediction and Prevention of Emerging Zoonotic Viruses with Pandemic Potential using Multidisciplinary Approaches, FP7-n°278433), PREPARE (Platform for European Preparedness Against (Re-)emerging Epidemics, FP7-n° 602525), READNA (Revolutionary Approaches and Devices for Nucleic Acid analysis, FP7-n°201418), CHAARM (Combined Highly Active Anti-retroviral Microbicides, FP7-n°242135), ProteomeXchange (International Data Exchange and Data Representation Standards for Proteomics, FP7-n°260558), PSIMEx (Proteomics Standards International Molecular Exchange—Systematic Capture of Published Molecular Interaction Data, FP7-n°223411), RADIANT (Rapid Development and Distribution of Statistical Tools for High-Throughput Sequencing Data FP7-n°305626), SEMCARE (Semantic Data Platform for Healthcare, FP7-n°611388), SPRINTT (Sarcopenia and Physical fRailty IN older people: multi-component Treatment strategies, IML-n° 115621), STATEGRA (User-driven Development of Statistical Methods for Experimental Planning, Data Gathering, and Integrative Analysis of Next Generation Sequencing, proteomics and Metabolomics data FP7-n°306000), SysCLAD (Systems prediction of Chronic Lung Allograft Dysfunction, FP7-n° 354457), SYSMEDIBD (Systems medicine of chronic inflammatory bowel disease, FP7-n°305564), U-BIOPRED (Unbiased BIOMarkers for the PREDiction of respiratory disease outcomes, IML-n°115010), VPH-share (Virtual Physiological Human: Sharing for Healthcare - A Research Environment, FP7-n°269978), Genom Austria (member of the Global Network of Personal Genome Projects).

SJ, PF, and JAV acknowledge support from the European Molecular Biology Laboratory.

IB acknowledges support from the Wellcome Trust (WT098051).

## Authors' contributions

We thank chairs and panelists Ana Conesa, Haralampos Karanikas, Inês Barroso, Ivo Gut, Jerry Lanfear, Niklas Blomberg, Norbert Graf, Pablo Villoslada, Paul Flicek, Rod Huse, Rudi Balling, Tim Hubbard, Yike Guo, Charles Auffray, Mikael Benson, Gianluigi Zanetti and Jeanine Houwing-Duistermaat for their contributions to the manuscript preparation. All authors contributed to the content of the manuscript. Sophie Janacek drafted the initial version of the manuscript, which was subsequently thoroughly edited by Rudi Balling, Charles Auffray, and Christoph Bock with the support of Maria Manuela Nogueira. Diane Lefaudeux helped with the bibliography. All authors read and approved the final manuscript.

## Competing interests

CA, RB, RDH, CD, KP, EBD, WM, MK, JR, LV, JAV, NG and IB declare that they have no competing interests. AK is employed by ITTM S.A. and is an expert at the ISO/TC 76 WG 5; he does not have any competing interests. TK is employed by Vitromics Healthcare Holding, which is a member of EuropaBio. Pablo Villoslada has received consultancy fees from Roche, Novartis, Araclon, and Health Engineering, is founder and hold stocks in Bionure Inc. and Spire Bioventures, and works as an academic editor for *Neurology & Therapy*, *Current Treatment Options in Neurology*, *Multiple Sclerosis & Demyelinating diseases*, and *PLoS One*. PF is a member of the Scientific Advisory Board for Omicia, Inc.

## Author details

<sup>1</sup>European Institute for Systems Biology and Medicine, 1 avenue Claude Vellefaux, 75010 Paris, France. <sup>2</sup>CIRI-UMR5308, CNRS-ENS-INSERM-UCBL, Université de Lyon, 50 avenue Tony Garnier, 69007 Lyon, France. <sup>3</sup>Luxembourg Centre for Systems Biomedicine, University of Luxembourg, 7 Avenue des Hauts Fourneaux, 4362 Esch-sur-Alzette, Luxembourg. <sup>4</sup>Wellcome Trust Sanger Institute, Wellcome Genome Campus, Hinxton, Cambridge CB10 1SA, UK. <sup>5</sup>Health Services Management Training Centre, Faculty of Health and Public Services, Sannelweis University, Kútvolgyi út 2, 1125 Budapest, Hungary. <sup>6</sup>Centre for Personalised Medicine, Linköping University, 581 85 Linköping, Sweden. <sup>7</sup>Translational & Bioinformatics, Pfizer Inc., 300 Technology Square, Cambridge, MA 02139, USA. <sup>8</sup>Institute for Health Sciences, IACS - IIS Aragon, San Juan Bosco 13, 50009 Zaragoza, Spain. <sup>9</sup>ELIXIR, Wellcome Genome Campus, Hinxton, Cambridge CB10 1SD, UK. <sup>10</sup>CeMM Research Center for Molecular Medicine of the Austrian Academy of Sciences, Lazarettgasse 14, AKH BT25.2, 1090 Vienna, Austria. <sup>11</sup>Department of Laboratory Medicine, Medical University of Vienna, Lazarettgasse 14, AKH BT25.2, 1090 Vienna, Austria. <sup>12</sup>Max Planck Institute for Informatics, Campus E1 4, 66123 Saarbrücken, Germany. <sup>13</sup>Príncipe Felipe Research Center, C/

Eduardo Primo Yúfera 3, 46012 Valencia, Spain. <sup>14</sup>University of Florida, Institute of Food and Agricultural Sciences (IFAS), 2033 Mowry Road, Gainesville, FL 32610, USA. <sup>15</sup>Bluecompanion Ltd, 6 London Street (second floor), London W2 1HR, UK. <sup>16</sup>Technology, Data & Analytics, KPMG Luxembourg, Société Coopérative, 39 Avenue John F. Kennedy, 1855 Luxembourg, Luxembourg. <sup>17</sup>Department of Human Genetics, Department of Pathology, Leiden University Medical Centre, Einthovenweg 20, 2333 ZC Leiden, The Netherlands. <sup>18</sup>Information Technology Department, European Organization for Nuclear Research (CERN), 385 Route de Meyrin, 1211 Geneva 23, Switzerland. <sup>19</sup>Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Heidelberglaan 100, 3508 GA Utrecht, The Netherlands. <sup>20</sup>European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Wellcome Genome Campus, Hinxton, Cambridge CB10 1SD, UK. <sup>21</sup>Department of Pediatric Oncology/Hematology, Saarland University, Campus Homburg, Building 9, 66421 Homburg, Germany. <sup>22</sup>Project Management Jülich, Forschungszentrum Jülich GmbH, Wilhelm-Johnen-Straße, 52428 Jülich, Germany. <sup>23</sup>Department of Clinical Pharmacy & Toxicology, Leiden University Medical Center, Albinusdreef 2, 2333 ZA Leiden, The Netherlands. <sup>24</sup>Data Science Institute, Imperial College London, South Kensington, London SW7 2AZ, UK. <sup>25</sup>CNAG-CRG, Center for Genomic Regulation, Barcelona Institute for Science and Technology (BIST), C/Baldiri Reixac 4, 08029 Barcelona, Spain. <sup>26</sup>Institute of Software Technology and Interactive Systems, TU Wien, Favoritenstrasse 9-11/188, 1040 Vienna, Austria. <sup>27</sup>The Association of the British Pharmaceutical Industry, 7th Floor, Southside, 105 Victoria Street, London SW1E 6QT, UK. <sup>28</sup>Department of Medical Statistics, RWTH-Aachen University, Universitätsklinikum Aachen, Pauwelsstraße 30, 52074 Aachen, Germany. <sup>29</sup>SYNAPSE Research Management Partners, Diputació 237, Àtic 3<sup>a</sup>, 08007 Barcelona, Spain. <sup>30</sup>Department of Infection, Immunity and Cardiovascular Disease and Insigneo Institute for In-Silico Medicine, Medical School, University of Sheffield, Beech Hill Road, Sheffield S10 2RX, UK. <sup>31</sup>Department of Statistics, School of Mathematics, University of Leeds, Leeds LS2 9JT, UK. <sup>32</sup>Department of Medical & Molecular Genetics, King's College London, London SE1 9RT, UK. <sup>33</sup>Genomics England, London EC1M 6BQ, UK. <sup>34</sup>National and Kapodistrian University of Athens, Medical School, Kristou Lada 6, 10561 Athens, Greece. <sup>35</sup>Vitromics Healthcare Holding B.V., Onderwijsboulevard 225, 5223 DE 's-Hertogenbosch, The Netherlands. <sup>36</sup>Fraunhofer Institute for Molecular Biology and Applied Ecology ScreeningPort, Schnackenburgallee 114, 22525 Hamburg, Germany. <sup>37</sup>ITM S.A., 9 avenue des Hauts Fourneaux, 4362 Esch-sur-Alzette, Luxembourg. <sup>38</sup>Research Business Technology, Pfizer Ltd, GP4 Building, Granta Park, Cambridge CB21 6GP, UK. <sup>39</sup>Health Economics & Outcomes Research, Deloitte Belgium, Berkenlaan 8A, 1831 Diegem, Belgium. <sup>40</sup>Janssen Pharmaceutica N.V., R&D G30, Turnhoutseweg 30, 2340 Beerse, Belgium. <sup>41</sup>Faculty of Life Sciences, University of Manchester, AV Hill Building, Oxford Road, Manchester M13 9PT, UK. <sup>42</sup>UMR3664 IC/CNRS, Institut Curie, Section Recherche, Pavillon Pasteur, 26 rue d'Ulm, 75248 Paris cedex 05, France. <sup>43</sup>Linguamatics Ltd, 324 Cambridge Science Park Milton Rd, Cambridge CB4 0WG, UK. <sup>44</sup>PwC Luxembourg, 2 rue Gerhard Mercator, 2182 Luxembourg, Luxembourg. <sup>45</sup>Philips, HighTechCampus 36, 5656AE Eindhoven, The Netherlands. <sup>46</sup>Department of Public Health and Primary Care, KU Leuven Kulak, Etienne Sabbelaan 53, 8500 Kortrijk, Belgium. <sup>47</sup>INCLIVA Health Research Institute, University of Valencia, CIBERobn ISCIII, Avenida Menéndez Pelayo 4 accesorio, 46010 Valencia, Spain. <sup>48</sup>Swiss Institute of Bioinformatics (SIB) and University of Basel, Klingelbergstrasse 50/70, 4056 Basel, Switzerland. <sup>49</sup>Agency for Health Quality and Assessment of Catalonia (AQuAS), Carrer de Roc Boronat 81-95, 08005 Barcelona, Spain. <sup>50</sup>EuroBioForum Foundation, Chryssantstraat 10, 3135 HG Vlaardingen, The Netherlands. <sup>51</sup>Integrated BioBank of Luxembourg, 6 rue Nicolas-Ernest Barblé, 1210 Luxembourg, Luxembourg. <sup>52</sup>Technopolis Group, 3 Pavilion Buildings, Brighton BN1 1EE, UK. <sup>53</sup>Hospital Clinic of Barcelona, Institute d'Investigacions Biomèdiques August Pi Sunyer (IDIBAPS), Rosello 149, 08036 Barcelona, Spain. <sup>54</sup>European Platform for Patients' Organisations, Science and Industry (Epossi), De Meeûs Square 38-40, 1000 Brussels, Belgium. <sup>55</sup>CRS4, Ed.1 POLARIS, 09129 Pula, Italy. <sup>56</sup>BBMRI-ERIC, Neue Stiftingtalstrasse 2/B/6, 8010 Graz, Austria.

Published online: 23 June 2016

## References

- Ideker T, Dutkowsky J, Hood L. Boosting signal-to-noise in complex biology: prior knowledge is power. *Cell*. 2011;144:860–3.
- Wood AR, Esko T, Yang J, Vedantam S, Pers TH, Gustafsson S, et al. Defining the role of common variation in the genomic and biological architecture of adult human height. *Nat Genet*. 2014;46:1173–86.
- Cooper DN, Krawczak M, Polychronakos C, Tyler-Smith C, Kehrer-Sawatzki H. Where genotype is not predictive of phenotype: towards an understanding of the molecular basis of reduced penetrance in human inherited disease. *Hum Genet*. 2013;132:1077–130.
- Pickrell JK, Marioni JC, Pai AA, Degner JF, Engelhardt BE, Nkadori E, et al. Understanding mechanisms underlying human gene expression variation with RNA sequencing. *Nature*. 2010;464:768–72.
- Vernot B, Stergachis AB, Maurano MT, Vierstra J, Neph S, Thurman RE, et al. Personal and population genomics of human regulatory variation. *Genome Res*. 2012;22:1689–97.
- Piraino SW, Furney SJ. Beyond the exome: the role of non-coding somatic mutations in cancer. *Ann Oncol Off J Eur Soc Med Oncol ESMO*. 2016;27:240–8.
- European Commission satellite workshop 'Big data in health research: an EU action plan'. <http://bigdata2015.uni.lu/eng/European-Commission-satellite-workshop>. Accessed 20 May 2016.
- Raghupathi W, Raghupathi V. Big data analytics in healthcare: promise and potential. *Health Inf Sci Syst*. 2014;2:3.
- Baro E, Degoul S, Beuscart R, Chazard E. Toward a literature-driven definition of big data in healthcare. *BioMed Res Int*. 2015;2015:639021.
- Meldolesi E, van Soest J, Damiani A, Dekker A, Alitto AR, Campitelli M, et al. Standardized data collection to build prediction models in oncology: a prototype for rectal cancer. *Future Oncol Lond Engl*. 2016;12:119–36.
- Fernández-Luque L, Bau T. Health and social media: perfect storm of information. *Healthcare Inform Res*. 2015;21:67–73.
- Hood L, Price ND. Demystifying disease, democratizing health care. *Sci Transl Med*. 2014;6:225ed5.
- Wade TD. Traits and types of health data repositories. *Health Inf Sci Syst*. 2014;2:4.
- Ludvigsson JF, Andersson E, Ekblom A, Feychting M, Kim J-L, Reuterwall C, et al. External review and validation of the Swedish national inpatient register. *BMC Public Health*. 2011;11:450.
- DiMarco G, Hill D, Feldman SR. Review of patient registries in dermatology. *J Am Acad Dermatol*. 2016. doi:10.1016/j.jaad.2016.03.020.
- Orphanet. Rare Disease Registries in Europe. <http://www.orpha.net/orphacom/cahiers/docs/GB/Registries.pdf>. Accessed 6 May 2016.
- 2013 EURORDIS policy fact sheet - Rare Disease Patient Registries. [http://www.eurordis.org/sites/default/files/publications/Factsheet\\_registries.pdf](http://www.eurordis.org/sites/default/files/publications/Factsheet_registries.pdf). Accessed 8 May 2016.
- EORTC: European Organisation for Research and Treatment of Cancer. <http://www.eortc.org>. Accessed 6 May 2016.
- EORTC opens prospective registry for patients with Melanoma. <http://www.eortc.org/news/eortc-opens-prospective-registry-for-patients-with-melanoma>. Accessed 8 May 2016.
- ENCR: European Network of Cancer Registries. <http://www.encreu.eu>. Accessed 6 May 2016.
- PARENT: Patient Registries Initiative. <http://patientregistries.eu/deliverables>. Accessed 6 May 2016.
- Kaplan G, Virginia Mason, Bo-Linn G, Gordon and Betty Moore Foundation, Carayon P, University of Wisconsin, et al. Bringing a systems approach to health. *National Academy of Engineering of the National Academies and Institute of Medicine of the National Academies*; Jul 2013. <https://www.nae.edu/File.aspx?id=86344>. Accessed 6 May 2016.
- Bulger M, Taylor G, Schroeder R. Data-driven business models: challenges and opportunities of big data. *Oxford Internet Institute. Research Councils UK: NEMODE, New Economic Models in the Digital Economy*; 2014. [http://www.nemode.ac.uk/wp-content/uploads/2014/09/nemode\\_business\\_models\\_for\\_bigdata\\_2014\\_oxford.pdf](http://www.nemode.ac.uk/wp-content/uploads/2014/09/nemode_business_models_for_bigdata_2014_oxford.pdf). Accessed 20 May 2016.
- Delfino A, Faure Ragani A, Telpis V, Tilley J, McKinsey & Company. Mature quality systems: what pharma can learn from other industries. *Pharm Manuf*. 26 Feb 2015; <http://www.pharmamanufacturing.com/articles/2015/mature-quality-systems-what-pharma-can-learn-from-other-industries/>. Accessed 20 May 2016.
- Rumsfeld JS, Joynt KE, Maddox TM. Big data analytics to improve cardiovascular care: promise and challenges. *Nat Rev Cardiol*. 2016;13(6):350–9.
- Monteith S, Glenn T, Geddes J, Whybrow PC, Bauer M. Big data for bipolar disorder. *Int J Bipolar Disord*. 2016;4:10.
- Janke AT, Overbeek DL, Kocher KE, Levy PD. Exploring the potential of predictive analytics and big data in emergency care. *Ann Emerg Med*. 2016; 67:227–36.



28. Khandani S. Engineering design process: education transfer plan. 2005. <http://www.saylor.org/site/wp-content/uploads/2012/09/ME101-4.1-Engineering-Design-Process.pdf>. Accessed 8 May 2016.
29. Abugessaisa I, Saevarsdottir S, Tspiras G, Lindblad S, Sandin C, Nikamo P, et al. Accelerating translational research by clinically driven development of an informatics platform—a case study. *PLoS One*. 2014;9, e104382.
30. Cano I, Lluch-Ariet M, Gomez-Cabrero D, Maier D, Kalko S, Cascante M, et al. Biomedical research in a digital health framework. *J Transl Med*. 2014;12 Suppl 2:S10.
31. Koutkias VG, Jaulent M-C. Computational approaches for pharmacovigilance signal detection: toward integrated and semantically-enriched frameworks. *Drug Saf*. 2015;38:219–32.
32. Espay AJ, Bonato P, Nahab FB, Maetzler W, Dean JM, Klucken J, et al. Technology in Parkinson's disease: challenges and opportunities. *Mov Disord Off J Mov Disord Soc*. 2016. doi:10.1002/mds.26642.
33. Austin C, Kusumoto F. The application of Big Data in medicine: current implications and future directions. *J Interv Card Electrophysiol Int J Arrhythm Pacing*. 2016. doi:10.1007/s10840-016-0104-y.
34. Poste G. Bring on the biomarkers. *Nature*. 2011;469:156–7.
35. Sawyers CL. The cancer biomarker problem. *Nature*. 2008;452:548–52.
36. Barlesi F, Mazieres J, Merlio J-P, Debieuvre D, Mosser J, Lena H, et al. Routine molecular profiling of patients with advanced non-small-cell lung cancer: results of a 1-year nationwide programme of the French Cooperative Thoracic Intergroup (IFCT). *Lancet Lond Engl*. 2016;387:1415–26.
37. Holderfield M, Deuker MM, McCormick F, McMahon M. Targeting RAF kinases for cancer therapy: BRAF-mutated melanoma and beyond. *Nat Rev Cancer*. 2014;14:455–67.
38. Kalia M. Biomarkers for personalized oncology: recent advances and future challenges. *Metabolism*. 2015;64:S16–21.
39. Semrad TJ, Kim EJ. Molecular testing to optimize therapeutic decision making in advanced colorectal cancer. *J Gastrointest Oncol*. 2016;7:S11–20.
40. Hay SI, George DB, Moyes CL, Brownstein JS. Big data opportunities for global infectious disease surveillance. *PLoS Med*. 2013;10, e1001413.
41. Zheng Y-L, Ding X-R, Poon CCY, Lo BPL, Zhang H, Zhou X-L, et al. Unobtrusive sensing and wearable devices for health informatics. *IEEE Trans Biomed Eng*. 2014;61:1538–54.
42. OECD Publishing. Health data governance: privacy, monitoring and research - policy brief. OECD; Oct 2015. <https://www.oecd.org/health/health-systems/Health-Data-Governance-Policy-Brief.pdf>. Accessed 6 May 2016.
43. Eisenstein M. Big data: the power of petabytes. *Nature*. 2015;527:S2–4.
44. Doyle-Lindrud S. Watson will see you now: a supercomputer to help clinicians make informed treatment decisions. *Clin J Oncol Nurs*. 2015;19:31–2.
45. Cesario A, Marcus F. Cancer systems biology, bioinformatics and medicine: research and clinical applications. 1st ed. Netherlands: Springer Science & Business Media; 2011.
46. Cancer Genome Atlas Research Network, Weinstein JN, Collisson EA, Mills GB, Shaw KRM, Ozenberger BA, et al. The Cancer Genome Atlas Pan-Cancer analysis project. *Nat Genet*. 2013;45:1113–20.
47. Gahl WA, Wise AL, Ashley EA. The undiagnosed diseases network of the national institutes of health: a national extension. *JAMA*. 2015;314:1797–8.
48. Taruscio D, Groft SC, Cederroth H, Melegh B, Lasko P, Kosaki K, et al. Undiagnosed Diseases Network International (UDNI): White paper for global actions to meet patient needs. *Mol Genet Metab*. 2015;116:223–5.
49. Thompson R, Johnston L, Taruscio D, Monaco L, Bérout C, Gut IG, et al. RD-Connect: an integrated platform connecting databases, registries, biobanks and clinical bioinformatics for rare disease research. *J Gen Intern Med*. 2014;29:780–7.
50. Yaman H, Yavuz E, Er A, Vural R, Albayrak Y, Yardimci A, et al. The use of mobile smart devices and medical apps in the family practice setting. *J Eval Clin Pract*. 2016;22:290–6.
51. American Bar Association, Health Law Section, ABA Section of Science & Technology Law and Center for Professional Development. Medical device law: compliance issues, best practices and trends. 2015. [http://www.americanbar.org/content/dam/aba/events/cle/2015/10/ce1510mdm/ce1510mdm\\_interactive.authcheckdam.pdf](http://www.americanbar.org/content/dam/aba/events/cle/2015/10/ce1510mdm/ce1510mdm_interactive.authcheckdam.pdf). Accessed 6 May 2016.
52. Di Meglio A. Big data management—from CERN/LHC to personalised medicine. *Ajaccio, France: MEDAMI*; 2016. doi:10.5281/zenodo.50739.
53. Murphy SN, Weber G, Mendis M, Gainer V, Chueh HC, Churchill S, et al. Serving the enterprise and beyond with informatics for integrating biology and the bedside (i2b2). *J Am Med Inform Assoc*. 2010;17:124–30.
54. Chen J, Qian F, Yan W, Shen B. Translational biomedical informatics in the cloud: present and future. *BioMed Res Int*. 2013;2013:658925.
55. Hofmann-Apitius M, Ball G, Gebel S, Bagewadi S, de Bono B, Schneider R, et al. Bioinformatics mining and modeling methods for the identification of disease mechanisms in neurodegenerative disorders. *Int J Mol Sci*. 2015;16:29179–206.
56. Tenenbaum JD. Translational bioinformatics: past, present, and future. *Genomics Proteomics Bioinformatics*. 2016;14:31–41.
57. Denny JC, Bastarache L, Ritchie MD, Carroll RJ, Zink R, Mosley JD, et al. Systematic comparison of genome-wide association study of electronic medical record data and genome-wide association study data. *Nat Biotechnol*. 2013;31:1102–10.
58. Gustafsson M, Gawel DR, Alfredsson L, Baranzini S, Björkander J, Blomgran R, et al. A validated gene regulatory network and GWAS identifies early regulators of T cell-associated diseases. *Sci Transl Med*. 2015;7:313ra178.
59. Landau DA, Carter SL, Stojanov P, McKenna A, Stevenson K, Lawrence MS, et al. Evolution and impact of subclonal mutations in chronic lymphocytic leukemia. *Cell*. 2013;152:714–26.
60. Leitsalu L, Alavere H, Tammesoo M-L, Leego E, Metspalu A. Linking a population biobank with national health registries—the estonian experience. *J Pers Med*. 2015;5:96–106.
61. Mandl KD, Kohane IS. Time for a patient-driven health information economy? *N Engl J Med*. 2016;374:205–8.
62. IRDiRC: International Rare Diseases Research Consortium. <http://www.irdirc.org>. Accessed 8 May 2016.
63. RARE-Bestpractices. <http://www.rarebestpractices.eu/home>. Accessed 8 May 2016.
64. p-medicine - from data sharing and integration via VPH models to personalized medicine. <http://www.p-medicine.eu>. Accessed 8 May 2016.
65. ELIXIR: A distributed infrastructure for life-science information. <https://www.elixir-europe.org>. Accessed 6 May 2016.
66. Ison J, Rapacki K, Ménager H, Kalaš M, Rydza E, Chmura P, et al. Tools and data services registry: a community effort to document bioinformatics resources. *Nucleic Acids Res*. 2016;44:D38–47.
67. eTRIKS: European Translational Research Information and Knowledge Management Services. <https://www.etriks.org>. Accessed 6 May 2016.
68. Genomics England 100,000 Genomes Project. <http://www.genomicsengland.co.uk>. Accessed 6 May 2016.
69. Rosenthal A, Mork P, Li MH, Stanford J, Koester D, Reynolds P. Cloud computing: a new business paradigm for biomedical information sharing. *J Biomed Inform*. 2010;43:342–53.
70. Chen Y-C, Horng G, Lin Y-J, Chen K-C. Privacy preserving index for encrypted electronic medical records. *J Med Syst*. 2013;37:9992.
71. Griebel L, Prokosch H-U, Köpcke F, Toddenroth D, Christoph J, Leb I, et al. A scoping review of cloud computing in healthcare. *BMC Med Inform Decis Mak*. 2015;15:17.
72. IMI: Innovative Medicines Initiative - Ongoing projects. <http://www.imi.europa.eu/content/ongoing-projects>. Accessed 8 May 2016.
73. Hughes R, Beene M, Dykes. The significance of data harmonization for credentialing research. Washington, DC: Institute of Medicine of the National Academies; 2014. <http://nam.edu/wp-content/uploads/2015/06/CredentialingDataHarmonization.pdf>. Accessed 8 May 2016.
74. European Open Science Cloud. <http://ec.europa.eu/research/openscience/index.cfm?pg=open-science-cloud>. Accessed 9 May 2016.
75. Howe D, Costanzo M, Fey P, Gojorbori T, Hannick L, Hide W, et al. Big data: the future of biocuration. *Nature*. 2008;455:47–50.
76. Liberles DA, Teufel AJ, Liu L, Stadler T. On the need for mechanistic models in computational genomics and metagenomics. *Genome Biol Evol*. 2013;5:2008–18.
77. EMBL-EBI: European Molecular Biology Laboratory – European Bioinformatics Institute. <http://www.ebi.ac.uk/biomodels-main>. Accessed 8 May 2016.
78. Viceconti M, Hunter P, Hose R. Big data, big knowledge: big data for personalized healthcare. *IEEE J Biomed Health Inform*. 2015;19:1209–15.
79. Virtual Physiological Human (VPH) Institute. <http://www.vph-institute.org>. Accessed 6 May 2016.
80. IUPS Physiome Project. <http://physiomeproject.org/software/fieldml>. Accessed 6 May 2016.
81. Marés J, Shamardin L, Weiler G, Anguita A, Sfakianakis S, Neri E, et al. p-medicine: a medical informatics platform for integrated large scale heterogeneous patient data. *AMIA Annu Symp Proc*. 2014;2014:872–81.
82. Schmitz U, Wolkenhauer O. Systems medicine. 1st ed. New York: Humana Press; 2016.
83. CASyM: Coordinating Action Systems Medicine Europe. <https://www.casym.eu>. Accessed 6 May 2016.



84. Pemovska T, Kontro M, Yadav B, Edgren H, Eldfors S, Sz wajda A, et al. Individualized systems medicine strategy to tailor treatments for patients with chemorefractory acute myeloid leukemia. *Cancer Discov*. 2013;3:1416–29.
85. Roca J, Cano I, Gomez-Cabrero D, Tegnér J. From systems understanding to personalized medicine: lessons and recommendations based on a multidisciplinary and translational analysis of COPD. *Methods Mol Biol Clifton NJ*. 2016;1386:283–303.
86. Kemp R. Legal aspects of managing big data white paper. 2014. Kemp IT Law, <http://www.kempitlaw.com/wp-content/uploads/2014/10/Legal-Aspects-of-Big-Data-White-Paper-v2-1-October-2014.pdf>. Accessed 6 May 2016.
87. ICGC: International Cancer Genome Consortium. <https://icgc.org/>. Accessed 6 May 2016.
88. IHEC: International Human Epigenome Consortium. <http://ihec-epigenomes.org>. Accessed 6 May 2016.
89. GSC: Genomic Standards Consortium. <http://gensc.org>. Accessed 6 May 2016.
90. CDISC: Clinical Data Interchange Standards Consortium. <http://www.cdisc.org>. Accessed 8 May 2016.
91. ISO TC276 WG5: Technical Committee 276 on Biotechnology, Working Group 5 on Data Processing and Integration. [http://www.iso.org/iso/home/standards\\_development/list\\_of\\_iso\\_technical\\_committees/iso\\_technical\\_committee.htm?commid=4514241](http://www.iso.org/iso/home/standards_development/list_of_iso_technical_committees/iso_technical_committee.htm?commid=4514241). Accessed 6 May 2016.
92. Wilkinson MD, Dumontier M, Aalbersberg IJJ, Appleton G, Axton M, Baak A, et al. The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data*. 2016;3:160018.
93. GA4GH: Global Alliance for Genomics and Health. <http://genomicsandhealth.org>. Accessed 6 May 2016.
94. CORBEL: Coordinated Research Infrastructures Building Enduring Life-science Services. <https://www.elixir-europe.org/about/eu-projects/corbel>. Accessed 6 May 2016.
95. BRIDGEHEALTH. <http://www.bridge-health.eu/content/integrate-information-injuries>. Accessed 8 May 2016.
96. Personal Genome Project. <http://www.personalgenomes.org>. Accessed 8 May 2016.
97. UNESCO. International Declaration on Human Genetic Data. Oct 2003. [http://portal.unesco.org/en/ev.php-URL\\_ID=17720&URL\\_DO=DO\\_TOPIC&URL\\_SECTION=201.html](http://portal.unesco.org/en/ev.php-URL_ID=17720&URL_DO=DO_TOPIC&URL_SECTION=201.html). Accessed 6 May 2016.
98. Publishing OECD. Guidelines for Human Biobanks and Genetic Research Databases (HBGRDs). 2009. <http://www.oecd.org/sti/biotechnology/hbgrd>. Accessed 6 May 2016.
99. Knoppers BM. Framework for responsible sharing of genomic and health-related data. *HUGO J*. 2014;8:3.
100. DLA Piper, Data protection laws of the world. <https://www.dlapiperdataprotection.com/index.html#handbook/world-map-section>. Accessed 6 May 2016.
101. Proposal for a Regulation of the European parliament and of the Council on the protection of individuals with regard to the processing of personal data and on the free movement of such data (General Data Protection Directive) 2012/0011 (COD). [http://www.europarl.europa.eu/RegData/docs\\_autres\\_institutions/commission\\_europeenne/com/2012/0011/COM\\_COM\(2012\)0011\\_EN.pdf](http://www.europarl.europa.eu/RegData/docs_autres_institutions/commission_europeenne/com/2012/0011/COM_COM(2012)0011_EN.pdf). Accessed 6 May 2016.
102. General data protection regulation, compromise text concluded in the trilogue negotiations between the Parliament and the Council (17 December 2015). [http://www.emeeeting.europarl.europa.eu/committees/agenda/201512/LIBE/LIBE%282015%291217\\_1/sitt-1739884](http://www.emeeeting.europarl.europa.eu/committees/agenda/201512/LIBE/LIBE%282015%291217_1/sitt-1739884). Accessed 6 May 2016.
103. Bahr A, Schlünder I. Code of practice on secondary use of medical data in European scientific research projects. *Int Priv Law*. 2015;5:279–91.
104. Why you should care about blockchains: the non-financial uses of blockchain technology. Nesta. <http://www.nesta.org.uk/blog/why-you-should-care-about-blockchains-non-financial-uses-blockchain-technology>. Accessed 8 May 2016.
105. Barnes R. Blockchain and digital health—first impressions. *DNA Dig*. <http://dnadigest.org/?s=block+chain+digital+health>. Accessed 8 May 2016.
106. Tang Y, Liu L. Searching HIE with differentiated privacy preservation. San Diego, USA: 2014 USENIX Summit on Health Information Technologies HealthTech '14; 2014.
107. CS ELSI BBMRI-ERIC: Common Service on Ethical, Legal, and Social Issues of Biobanking and BioMolecular resources Research Infrastructure. <http://bbmri-eric.eu/common-services>. Accessed 6 May 2016.
108. Georgatos F, Ballereau S, Pellet J, Ghanem M, Price N, Hood L, et al. Computational infrastructures for data and knowledge management in systems biology. In: Prokop A, Csukás B, editors. *Systems Biology*. Netherlands: Springer; 2013. p. 377–97.
109. CS IT BBMRI-ERIC: Common Service on Information Technology of Biobanking and BioMolecular resources Research Infrastructure. <http://bbmri-eric.eu/common-service-it>. Accessed 6 May 2016.
110. BBMRI-ERIC: Biobanking and BioMolecular resources Research Infrastructures. <http://bbmri-eric.eu>. Accessed 8 May 2016.
111. ECRIN: European Clinical Research Infrastructure Network. <http://www.ecrin.org>. Accessed 6 May 2016.
112. Cascante M, de Atauri P, Gomez-Cabrero D, Wagner P, Centelles JJ, Marin S, et al. Workforce preparation: the Biohealth computing model for Master and PhD students. *J Transl Med*. 2014;12 Suppl 2:S11.
113. Rozman D, Acimovic J, Schmeck B. Training in systems approaches for the next generation of life scientists and medical doctors. *Systems Medicine*. 1st ed. New York: Humana Press (Springer Protocols). Schmitz U and Wolkenhauer O; 2016. p.73–86.
114. Jensen TB. Design principles for achieving integrated healthcare information systems. *Health Informatics J*. 2013;19:29–45.
115. Open science definition. [https://en.wikipedia.org/wiki/Open\\_science](https://en.wikipedia.org/wiki/Open_science). Accessed 8 May 2016.
116. Butler D. Dutch lead European push to flip journals to open access. *Nature*. 2016;529:13–3.
117. Swedish Research Council. Proposal for National Guidelines for Open Access to Scientific Information. Swedish Research Council; Feb 2015. <https://publikationer.vr.se/en/product/proposal-for-national-guidelines-for-open-access-to-scientific-information/>. Accessed 8 May 2016.
118. Bauer B, Blechl B, Bock C, Danowski P, Ferus A, Graschopf A, et al. Recommendations for the transition to open access in Austria. Nov 2015. <http://zenodo.org/record/34079#.Vy-nijY03q0>. Accessed 8 May 2016.
119. Berlin declaration on open access to knowledge in the sciences and humanities. 22 Oct 2003. <https://openaccess.mpg.de/Berlin-Declaration>. Accessed 8 May 2016.
120. Follett R, Strezov V. An analysis of citizen science based research: usage and publication patterns. *PLoS One*. 2015;10, e0143687.
121. Horizon 2020 Framework Programme policy on open science (open access). <http://ec.europa.eu/programmes/horizon2020/en/h2020-section/open-science-open-access>. Accessed 8 May 2016.
122. Clark WC, van Kerkhoff L, Lebel L, Gallopin GC. Crafting usable knowledge for sustainable development. *Proc Natl Acad Sci U S A*. 2016;113:4570–8.
123. Wolkenhauer O, Auffray C, Brass O, Clairambault J, Deutsch A, Drasdo D, et al. Enabling multiscale modeling in systems medicine. *Genome Med*. 2014;6:21.
124. Viceconti M, Henney A, Morley-Fletcher E. In silico clinical trials: how computer simulation will transform the biomedical industry. Brussels, Belgium: Avicenna Coordination Support Action; 2016. <http://avicenna-iscit.org/wp-content/uploads/2016/01/AvicennaRoadmapPDF-27-01-16.pdf>. Accessed 20 May 2016.
125. IDEAI: Infrastructure, Design, Engineering, Architecture, and Integration. [http://www.uspto.gov/about/vendor\\_info/current\\_acquisitions/ideaihom.jsp](http://www.uspto.gov/about/vendor_info/current_acquisitions/ideaihom.jsp). Accessed 8 May 2016.
126. Shaw DE, Sousa AR, Fowler SJ, Fleming LJ, Roberts G, Corfield J, et al. Clinical and inflammatory characteristics of the European U-BIOPRED adult severe asthma cohort. *Eur Respir J*. 2015;46:1308–21.
127. Ayasdi. <http://www.ayasdi.com>. Accessed 6 May 2016.
128. Lum PY, Singh G, Lehman A, Ishkanov T, Vajdemo-Johansson M, Alagappan M, et al. Extracting insights from the shape of complex data using topology. *Sci Rep*. 2013;3:1236.
129. Pellet J, Lefaudeux D, Royer P-J, Koutsokera A, Bourgoïn-Voillard S, Schmitt M, et al. A multi-omics data integration approach to identify a predictive molecular signature of CLAD. *Eur Respir J*. 2015;46, OA3271.
130. Pison C, Magnan A, Botturi K, Sève M, Brouard S, Marsland BJ, et al. Prediction of chronic lung allograft dysfunction: a systems medicine challenge. *Eur Respir J*. 2014;43:689–93.
131. Ingenuity®. <http://www.ingenuity.com>. Accessed 6 May 2016.
132. Thomson Reuters GeneGo MetaCore™. <https://portal.genego.com>. Accessed 8 May 2016.
133. Fujita KA, Ostaszewski M, Matsuoka Y, Ghosh S, Glaab E, Trefois C, et al. Integrating pathways of Parkinson's disease in a molecular interaction map. *Mol Neurobiol*. 2014;49:88–102.
134. Mazein A, Auffray C. EISBM AsthmaMap. <http://www.eisbm.org/projects/disease-maps>. Accessed 6 May 2016.
135. Mazein A, De Meulder B, Lefaudeux D, Knowles R, Wheelock C, Dahlen S, et al. The AsthmaMap: towards a community-driven reconstruction of asthma-relevant pathways and networks. Estoril, Portugal: The 14th ERS Lung Science Conference; 2016.