ICR The Institute of Cancer Research

Big Data

Position Statement from

The Institute of Cancer Research, London

Summary

Technological developments in research and healthcare are generating unprecedented quantities of complex and varied data and information. Data collection is getting cheaper and faster, but this has not been matched with systems for analysing different types of data together and maximising their value for patient benefit. Big Data describes vast and complex datasets that challenge our capabilities to analyse and handle with traditional methods. To properly seize the opportunities afforded by Big Data, we need new approaches to data collection, curation and analysis. There will need to be Government support for Big Data infrastructure, and for education and training to ensure we are training people with the skills required in complex data analysis. In conducting research, we must work together with patients to ensure that they are receiving value through treatment advances emerging from research which uses data which they have generously donated, while safeguarding their privacy and security.

November 2017

Big Data

Background information

Technological developments in research and healthcare now mean vast quantities of genetic sequencing, imaging, clinical and other forms of data are being generated at an unprecedented rate. To give an idea of the scale, the ICR's cancer knowledge base, canSAR, contains 1.7 billion experimental results.

The term 'Big Data' indicates a volume, complexity and diversity of data which challenges our current capabilities to analyse and handle. It is not only the huge quantity of data which is important, also the diversity of data from different areas of science and medicine, which can include measurements across time and unstructured information such as free text. At the ICR, diverse datasets may include genomic data, structural data, images of cells, tissue and the whole body, chemical and pharmacological data, radiotherapy and clinical notes.

Using Big Data effectively will allow us to answer key questions in cancer research and treatment. The sharing and integration of diverse sets of data has the potential to uncover knowledge that cannot be observed through working on smaller, individual subsets. This approach to data has the potential to open up advances in clinical practice, especially in areas such as personalised adaptive therapy, where clinicians aim to predict and monitor the effectiveness of treatment, and modify it to respond to changes in the cancer.

A data-driven approach can help us meet two of the biggest challenges in improving cancer survival – ensuring early diagnosis and overcoming drug-resistant disease. The ICR's researchers have developed canSAR – the world's largest disease knowledge base – to bring together and integrate huge amounts of data on cancer biology and drug discovery, and use it to perform virtual experiments designed to identify potential leads for cancer treatment. Researchers used canSAR to identify 46 potentially 'druggable' cancer proteins that had previously been overlooked as drug targets. The ICR is also creating an internal Big Data platform, called the Knowledge Hub, which will integrate all our scientific and clinical research data.

But across science the rapid increase in the pace at which we generate data has not yet been matched by an ability to use this data effectively. Technologies for data storage, transfer, integration and analysis are emerging rapidly but are still insufficient. There is insufficient coordination between scientific organisations, in finding common ways of formatting and labelling data. Big Data is a growing field and research organisations are faced with a shortage of people with the necessary skills and experience.

Key ICR positions on Big Data

- We need new approaches to data collection, curation and analysis, along with the funding and infrastructure to allow fast and accessible handling of data. Failure to coordinate Big Data approaches across the sector risks developing solutions that are inconsistent and incompatible.

- We believe scientific organisations need to work together to develop compatible data management approaches that will enable effective data sharing between different organisations. There is a risk that huge volumes of data are currently being stored in ways that will not prove compatible between organisations, and this is likely to become an increasing problem as data sharing becomes more common. Common protocols on how to label, search and access data are required, together with compatible infrastructure to allow greater sharing between collaborators. We believe an organisation or group of organisations in each scientific area should take the lead in developing open and federated approaches for data management.

- We need changes in the practices for data collection during clinical trials and routine treatment to ensure data are collected and annotated in a way that enables Big Data analysis. Certain types of data have common agreed standards, whereas other forms – in particular clinical notes outside trials – do not. We need to establish clearer protocols for ownership, governance and sharing of data, and education programmes, in particular for clinicians, to drive changes in the culture surrounding data storage and exchange.

- We believe the current infrastructure for the storage and transfer of data is inadequate and technological development and Government investment is needed. Current software and hardware infrastructure will soon be unable to support the vast and ever increasing volume of data transfer required. A coordinated effort including Government, industry and academia is needed to design and implement the next generation of data transfer infrastructure and technologies in order to create Big Data transfer highways. We need innovative database technologies to keep pace with researchers' increasing need to access both old and new datasets. Advances using what are known as non-structured query language (NoSQL) technologies are needed to link multiple data points to create large interconnected networks of multidisciplinary data. We believe that Government funding is needed to support storage infrastructure requirements, in particular in NHS organisations.

- Effective use of Big Data requires research organisations to have access to people with the necessary skills. Because we are seeing such rapid technological change, higher education institutions need to train people with

fundamental skills that will allow them to adapt to future changes, rather than in specific technologies that risk becoming out of date. We need training and internship programmes to build the skills base on Big Data, along with a relaxation of visa restrictions for people with skills relevant to Big Data, who may not necessarily possess or require PhD qualifications.

- Access to patient data is vital for the research conducted at the ICR. To uncover new knowledge and advance our understanding of cancer we need to be able to learn from large patient datasets. We believe the regulatory system is too risk averse in its requirement for explicit patient consent where research has been ethically approved. Patient data must be stored safely and securely to ensure patient confidentiality – but such safeguards must not come at the cost of efficient access to patient data for research use. We are supportive of moves to clarify and streamline processes to access patient data for ethically approved clinical research.

- We need to get the most possible benefit out of the data that we collect from patients to maximise our impact. We need to develop and promote standards for gaining patient consent in a flexible way so that data can continue to be used in future – including in slightly different ways to what was originally envisioned. And we need to ensure we collect broad enough data to allow us to ask all the questions that we may later have, maximising what we can do with the data we collect from patients.

- It is critical that we build public trust in Big Data projects, particularly including the way organisations like the ICR use patient data. More information needs to be provided to both the public and clinicians to improve their understanding of the value of Big Data, as well as how patient data is stored and accessed. Researchers must work together with patients to ensure that research using patient data delivers value to patients.