House of Commons Science and Technology Committee: The Big Data Dilemma

Response by the Wellcome Trust

3 September 2015

Key Points

- Big data technologies hold huge potential for benefiting health, society, research and the UK economy.
- Ensuring the trustworthiness of systems that use big data is essential if the public are to trust how their data are gathered, processed and used. Appropriate safeguards, robust, transparent processes, and opportunities for people to object should all form part of a governance framework for data use.
- The risks and potential harms of misusing big data technologies, for example to re-identify individuals, must be recognised and understood, with strong sanctions imposed for those who deliberately misuse data about individuals.
- There is a pressing need to ensure the UK is equipped with the capacity, infrastructure and skills base to capitalise on big data and ensure we can maximise the value of these technologies as they develop.

INTRODUCTION

1. The Wellcome Trust fully supports the development of 'big data' as a priority area for Government investment. Enabling the availability and re-use of large and complex datasets has potential to deliver significant societal and economic benefits.
2. Many big data technologies do not use data from individuals, but where data use involves personal or potentially identifiable data, there are important ethical, legal and social issues that need to be addressed. In order to realise the potential opportunities of big data, it is vital that these risks are understood and addressed, and that we move forward with the trust and support of the public.
3. The Wellcome Trust is a passionate advocate of open access publishing and research data access, and we work actively with partners to build the policy frameworks, governance mechanisms and infrastructures to support these activities. A key example of this is through the work of the Expert Advisory Group on Data Access (EAGDA), a group convened by the Medical Research Council, Economic and Social Research Council, Cancer Research UK and the Wellcome Trust.[1]
4. In our response, we briefly describe some key areas in which the Government can build on investments to date to help ensure that the UK can realise the opportunities flowing

[1] www.wellcome.ac.uk/EAGDA

from big data technologies. We then discuss priorities for addressing the associated challenges and risks.

Realising the opportunities

1. Big data technologies have the potential to transform our understanding of health and illness, improve services and drive innovation. In the field of health and biomedical research, the Human Genome Project was an early manifestation of the power of utilising vast, complex datasets. Big data also has the potential to enable the aspirations of 'personalised medicine' to be realised, through combining large-scale datasets for the purposes of diagnostics, risk prediction and identifying optimum treatment pathways.

2. In the field of clinical trials, it is anticipated that substantial benefits will arise from pharmaceutical companies' joint efforts to improve transparency and make their clinical trial data available, through initiatives such as clinicalstudydatarequest.com. The potential to link across these vast and complex datasets holds great promise for medical research.[2]

Investments

3. We were pleased that the previous Government prioritised big data, including as one of the 'eight great technologies'. We also strongly supported its work to progress its open data strategy,[3] and enable greater access to data collected by Government. Investments in this area over the last five years have helped to establish some world-leading research centres and facilities, for example:
    - The Farr Institute – bringing together four UK centres of excellence in health informatics to progress cutting edge research in the use of electronic patient information.
    - The Open Data Institute – supporting innovative businesses utilising open data, and providing associated research, training and services.
    - The Administrative Data Research Network – enabling research access and secure linkage of datasets collected by Government departments, as part of ESRC's Big Data Network'.
    - Alan Turing Institute – a new centre of excellence for computational and data sciences.

4. Investments in key infrastructures of this type mean that the UK is well positioned to play a leading role in the big data arena. However, these facilities will require sustained long-term funding commitments to remain internationally competitive and ultimately deliver benefits to the UK economy and society.

5. The UK Government should consider its broader strategy for big data in an international context. Realising the full potential of big data requires that researchers and innovators can draw on data from all over the world. Forging international partnerships to allow data to be shared across borders in standardised formats will be crucially important. In the research space, the Global Alliance for Genomics and Health[4] is an excellent

[2] http://www.wellcome.ac.uk/About-us/Policy/Spotlight-issues/Data-sharing/Access-to-clinical-trial-data/

[3] Open Data White Paper (2012) https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/78946/CM8353_acc.pdf

[4] www.genomicsandhealth.org

example of an initiative which is building the IT and policy frameworks to allow genetic and associated clinical data to be shared safely and securely for the benefit of the global research community and ultimately to accelerate progress in improving human health..

Skills

6. It will be critical for the UK to develop the skills and capacity in data science that will be required by industry, academia and Government to capitalise on big data technologies, as recognised in the joint Government and Information Economy Council report.[5] We welcome moves to equip graduates with the skills required in the digital age and we hope that this makes a contribution towards providing a pipeline of individuals with specialist data science skills. However, the retention of data scientists will require career structures that recognise their contributions and provide opportunities for development. This is a particularly pressing issue in the academic sector, where career paths for data managers and data scientists have been lacking.

7. A further key barrier in the academic sector is that current incentive structures do not adequately recognise data re-use. At present, research papers remain the primary currency in research assessment processes and career advancement decisions, and little credit is given to researchers who make high quality datasets available for others to access and re-use. These issues were explored in a report of the Expert Advisory Group on Data Access (EAGDA) in 2014. As recommended by this report, a key priority should be to ensure that data outputs are explicitly recognised as valued research outputs in the next Research Excellence Framework.[6]

Regulation

8. Realising the potential benefits of big data will also require an enabling regulatory framework at UK and European level. We were delighted that the UK Parliament approved the introduction of an exception to Copyright to permit data and text mining for non-commercial research purposes. We call on the UK Government to participate actively in on-going discussions over copyright reform at EU level, advocating for an exception to allow data and text mining for both commercial and non-commercial purposes.

9. The UK Government has played an important role in promoting a positive outcome for research in negotiations on the European Data Protection Regulation and we urge them to continue this during the on-going trilogue process. In particular, it is vital to ensure that the Regulation includes exceptions for health and scientific research, including an alternative means to allow processing of personal data for research where consent is not practicable. We are deeply concerned that the current wording of the European Parliament's text would have a hugely detrimental impact on valuable biomedical research.[7]

---

[5] Seizing the data opportunity: a strategy for UK data capability (October 2013) https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/254136/bis-13-1250-strategy-for-uk-data-capability-v4.pdf

[6] Assessing the research potential for access to clinical trial data (March 2015) http://www.wellcome.ac.uk/About-us/Policy/Spotlight-issues/Data-sharing/EAGDA/WTP056496.htm

[7] Joint statement on Ensuring a healthy future for scientific research through the Data Protection Regulation 2012/0011 (COD) (July 2015) http://www.wellcome.ac.uk/stellent/groups/corporatesite/@policy_communications/documents/web_document/wtp059364.pdf

Public trust and understanding

10. The opportunities and benefits of 'big data' can only be realised if there is sufficient public understanding of what the use of these technologies means, what the risks are for individuals and what choices they have for objecting to data relating to them being used. The opportunity to object (or 'opt-out') is not the same as the provision of informed consent. Consent can only be meaningfully 'informed' if a person is able to make a non-coerced choice and the benefits, risks and safeguards around data use are clearly explained. If the systems in place for managing data are trustworthy, able to demonstrate they are honest, reliable, transparent and competent, and provide clear societal benefits, public trust in big data technologies may be earned.

Learning from care.data

11. The importance of getting the underlying governance, safeguards and communications right with regard to uses of large-scale datasets cannot be overstated: the cost of getting it wrong has been evident from the *care.data* programme, which was an example of proposed big data technology being embedded in a healthcare setting. As acknowledged in this Committee's previous report on *Responsible Use of Data*,[8] the programme failed to take seriously the need to put in place robust, transparent systems of governance and understand people's legitimate concerns about the privacy of their sensitive personal data. The options for opting out of the programme were not adequately explained and the consequent loss of trust had ripple effects throughout the health and social care data domain.

12. This resulted in many legitimate, valuable research projects that accessed data via the Health and Social Care Information Centre being substantially delayed or halted. The Wellcome Trust reported on these effects at an evidence session of the Health Select Committee in January 2015.[9] The potential benefits of linking large-scale primary and secondary health care data, for medical research and improvements in healthcare, have now been overshadowed by the concerns that emerged when *care.data* was initially announced.

Commercial access

13. A number of previous surveys and reports have indicated that while people are generally happy for their data to be used if there is a clear personal or public benefit, they are significantly less comfortable with the idea of companies accessing and using data about them.[10] Misuse of data, lack of control, making profit from people's data and privacy intrusions are frequently cited as major concerns when commercial interests are involved in data use.

14. The Wellcome Trust has commissioned Ipsos MORI to undertake a research project seeking to understand what factors influence people's attitudes towards commercial organisations accessing health, biomedical and genetic data, so that we can identify

[8] Responsible Use of Data, House of Commons Science and Technology Committee (November 2014) para. 27
[9] http://data.parliament.uk/writtenevidence/committeeevidence.svc/evidencedocument/health-committee/handling-of-nhs-patient-data/oral/17740.html
[10] See for example, Royal Statistical Society (July 2014) http://www.statslife.org.uk/news/1672-new-rss-research-finds-data-trust-deficit-with-lessons-for-policymakers; Digital Catapult Centre (2015) http://www.digitalcatapultcentre.org.uk/pdtreview/; Medical Research Council (2007) http://www.mrc.ac.uk/documents/pdf/the-use-of-personal-health-information-in-medical-research-june-2007/

what uses of data are deemed acceptable and unacceptable by the public, patients and health professionals, and why. The project will also explore what safeguards and governance mechanisms the research community and others can use to demonstrate their trustworthiness in handling people's data. The report is due to be published in early 2016 and we would be happy to share the findings with the Committee.

Re-identification risks and harms

15. Capitalising on the potential of big data requires the risks to be acknowledged and mitigated as far as possible. As datasets become more sophisticated and the capacities to link across different data sources and mine data develop, the technical possibility of undertaking 'jigsaw' re-identification of individuals increases, even from data that has been through a process of anonymisation. This means that even if identifying information has been removed from a dataset (e.g., name, date of birth, address details), linkage with other information may, under some circumstances, enable individuals to be re-identified. The degree to which a dataset is anonymised is a function of its environment, other sources of information that are available, and the motivations and expertise of an 'intruder', which the original data controller may not be able to anticipate. Simply removing identifiers does not necessarily mean that data is no longer sensitive.

16. At the same time, however, poorly-implemented efforts to de-identify data may create perturbations in the data that fail to preserve the useful statistical properties of datasets: this may render them either useless or, worse, create distortions and errors in analysis that can lead to poor decision-making. Efforts to protect data must be proportionate to the risks posed by its use and its potential benefits.

17. In 2013, EAGDA considered the risks of re-identifying individuals from the linking of large genomic datasets with other data.[11] It recognised that establishing systematic, scientifically sound risk evaluations is a complex science and set out a number of recommendations for funders about how to ensure the value of such data can be maximised while protecting the confidentiality of research participants. These recommendations included making honest appraisals of the risks involved when seeking consent from research participants; controlling access to data that may be inferentially disclosive; and enforcing data access agreements to prohibit attempts at re-identification.

18. These steps would be broadly applicable to the wider governance of big data technologies that utilise people's data. Both policy safeguards and technical de-identification processes are needed to ensure privacy can be protected, but it is unlikely that a 'one size fits all' approach could ever accommodate the vast range of uses, linkages and circumstances in which big data technologies may pose a potential privacy risk. It is worth noting that the risk of successful re-identification through linking de-identified big datasets may well be extremely low depending on the circumstances, and would likely require substantial resources and motivation to succeed.

19. Also in relation to the risks of large-scale data collection and use, EAGDA co-commissioned an evidence review that fed into the Nuffield Council on Bioethics' recent report on access to biological and health data,[12] focusing on the risks of harms arising

[11] Statement for EAGDA funders on re-identification (October 2013) http://www.wellcome.ac.uk/stellent/groups/corporatesite/@policy_communications/documents/web_document/wtp055972.pdf

[12] Nuffield Council on Bioethics report (February 2015) http://nuffieldbioethics.org/project/biological-health-data/

from the use of these kinds of data. The review found that legal definitions of harms were too narrow to fully account for the types of impacts that data misuses have on individuals: it is unclear whether current legislation adequately protects individuals' and their rights over their data in the era of big data collection, usage and linking. The review also did not find evidence of harms resulting from academic research uses of data. The academic community has a strong track-record of robust governance processes and security mechanisms for handling large-scale data.

20. Substantial harms will, however, arise if there is a loss of public trust in data use technologies, insofar as this will limit or prevent valuable research that would be of societal benefit. We urge the Committee to take the risks we have outlined seriously in considering how the Government should approach the governance, regulation and development of big data technologies in the UK.

Sanctions

21. We recognise that no system will ever be 100% secure and it is important to manage, rather than necessarily eliminate, the risks of data misuse. Criminal sanctions for unauthorised and unwarranted deliberate re-identification of individuals through big data technologies would be one of the most powerful available methods of managing this risk. In our view the potential risks of re-identification from big data strengthen the case for an appropriately stringent re-identification deterrent. At present, we do not believe that sanctions for misusing personal data, in the form of a civil monetary penalty imposed through the Information Commissioner's Office under the Data Protection Act, are strong enough to reassure the public that malicious re-identification would be treated as a crime.

22. We believe that criminal liability could encourage individuals to take their responsibilities towards protecting personal data seriously, and to be rigorous in assessing identity disclosure risks when determining whether and under what circumstances data could be shared with third parties or linked across different data sources.

23. Sanctions should be introduced across all uses of data about people, and focus on the intent behind any attempted re-identification. The majority of countries in the EU can impose custodial sentences on those who breach data protection laws. We consider such sanctions could be introduced in the UK without detriment to the legitimate use of data underpinning much academic research, so long as these were supported by a concerted drive for training and skills development for staff handling data to ensure that data is used appropriately to maximise its value. The legislative and regulatory framework governing uses of data is complex, and there is a danger that criminal sanctions could create a risk-averse culture if they are not accompanied by clear, pragmatic guidance clarifying the scope of the Data Protection Act and common law of confidentially in the context of big data.

24. The Information Commissioner's Office needs to be adequately resourced to develop and provide support for legitimate data use and act swiftly on breaches where they are identified. The Wellcome Trust has worked closely with the Information Commissioner's Office on anonymisation and re-identification issues. Although the ICO has not called for a specific criminal offence related to re-identification, it does share our concerns and recognises the need for strong and effective sanctions. We note that in its 'Anonymisation: managing data protection code of practice' (2012) the ICO states that *"where an organisation collects personal data through a re-identification process without an individual's knowledge or consent, it will be obtaining personal data unlawfully and could be subject to enforcement action… Where there is evidence of re-identification*

taking place, with a risk of harm to individuals, the Information Commissioner will be likely to take regulatory action, including the imposition of a civil monetary penalty of up to £500,000."

25. We believe that this is indicative of the potential seriousness of a re-identification breach but that the introduction of a criminal sanction would send a clearer message as to the serious unacceptability of deliberately re-identifying individuals from anonymised data sources.

The Wellcome Trust is a global charitable foundation dedicated to improving health. We support bright minds in science, the humanities and the social sciences, as well as education, public engagement and the application of research to medicine.

Our investment portfolio gives us the independence to support such transformative work as the sequencing and understanding of the human genome, research that established front-line drugs for malaria, and Wellcome Collection, our free venue for the incurably curious that explores medicine, life and art.