A Roadmap for Restoring Trust in Big Data

Reference: THELANCETONCOLOGY-D-18-00670R1

Lawler M[1-3], Morris AD[4], Sullivan R[5], Birney E[3, 6], Middleton A[7], Makaroff L[8], Knoppers BM[3, 9], Horgan D[2], Eggermont A[2, 10]

[1]Centre for Cancer Research and Cell Biology, Queen's University Belfast, UK
[2]European Alliance for Personalised Medicine, Brussels, Belgium
[3]Global Alliance for Genomics and Health, Boston, USA
[4]Health Data Research UK, London, UK
[5]Institute for Cancer Policy, Kings College London, UK
[6]European Bioinformatics Institute, Cambridge, UK
[7]Welcome Genome Campus, Society and Ethics Research, Cambridge, UK
[8]European Cancer Patient Coalition, Brussels, Belgium
[9] Centre for Genomics and Policy, McGill University, Montreal, Canada;
[10]Gustave Roussy Cancer Campus Grand Paris, Villejuif, France

Fallout from the Cambridge Analytica/Facebook scandal marks a significant inflection point in the public's trust concerning Big Data. The health-science community must use this "crisis-in-confidence" to redouble its commitment to talk openly and transparently about benefits and risks, and act decisively to deliver robust effective governance frameworks, under which personal health data can be responsibly used. Activities such as the Innovative Medicines Initiative's "Big Data for Better Outcomes" [1] emphasise how a more granular data-driven understanding of human diseases including cancer could underpin innovative therapeutic intervention. Health Data Research-UK is developing national research expertise/infrastructure to maximise the value of health data-science for the National Health Service and ultimately UK citizens. [2] Comprehensive data analytics are crucial to the US Cancer Moonshot, the UK's 100,000 Genomes project

and other national genomics programmes. Cancer Core Europe, a research partnership between 7 leading European oncology centres,[3] has personal data-sharing at its core. The Global Alliance for Genomics and Health recently highlighted the need for a global cancer knowledge network[4] to drive evidence-based solutions for a disease that kills >8.7 million global citizens annually. These activities risk being fatally undermined by the recent controversy. We need to restore the public's trust in data science and emphasise its positive contribution in addressing global health and societal challenges. An opportunity to affirm the value of data science in Europe was afforded by DigitalDay 2018 In Brussels, where European Health Ministers signed a declaration of support to link existing/future genomic databanks across the EU, through the Million Genomes European Alliance.

So how do we address evolving challenges in analysis, sharing and storage of information, ensure transparency and confidentiality and restore public trust? We must articulate a clear 'Social Contract', where citizens (as data donors) are at the heart of decision-making. We need to demonstrate integrity, honesty and transparency as to what happens to data and what level of control people can, or cannot, expect. We must embed ethical rigour in all our data-driven processes. The *Framework for Responsible Sharing of Genomic and Health Related Data*(https://www.ga4gh.org/ga4ghtoolkit) represents a practical global approach, promoting effective and ethical sharing/use of research/patient data, while safeguarding individual privacy through secure and accountable data transfer.

We must guarantee that data analysis/storage are not compromised by data breaches that reveal personal information. Sanctions for such breaches must be clear with meaningful effect. Developing solutions whereby security is achieved in concurrent layers is required: reducing data travel, separating personal identifiable data from payload data, using effective anonymisation/encryption methods, thus maximising secure global sharing of data representing modern heterogeneous populations. Use of accredited data safe-havens, provision of honest broker services and, where necessary, maintenance of personal data at host locations for *in situ* analysis with easy-to-use freely available Application-Programme-Interfaces, can democratise data analysis for maximum scientific/clinical value. In Scotland, the electronic Data Research Innovation Service(eDRIS) established a "multi-tenant" research environment, whereby researchers access dedicated secure computing environment workspaces.[5] Each researcher must be approved/pass a certified data use course; each project is independently reviewed, with a privacy impact assessment performed before any data are released. This safe-haven capability "safe people, working on safe data, in safe places, with safe outputs" establishes robust research/governance infrastructure necessary for Big Data analyses.

Restoring public trust is the most daunting challenge. Resolving it will underpin a citizen-focussed data sharing cooperative that can help realise the promise of Big Data, without

compromising personal privacy. Patients,[6] and in particular cancer patients,[4] want to share their data but unless we assuage privacy concerns, confidence will be eroded. We must continually engage with the public through initiatives such as *Understanding Patient Data*[7] and *Your DNA Your Say*.[6] Clear, balanced and understandable information empowers citizens to make informed decisions about how their data is used. Each data safe-haven should have a "shop-window" to the public, indicating data projects being pursued, by whom and for what purpose. But we must do even more, elevating discourse to an international level,[4] making citizens more comfortable with their personal health data and removing the "fear factor" early, by engaging with schools/local communities.

It is incumbent on us all to advocate for robust, trustworthy and effective data collaborative approaches to enhance human health and eradicate diseases such as cancer globally. We must ensure a reputation for reliability, honesty and competency in how and why we use data, Transparency, with public involvement at every step is vital, if we are to maintain the social license for data-driven research. Otherwise, rather than shooting for the moon, we risk flying too close to the sun.

References
1. http://bd4bo.eu/ Last accessed 29th March 2018
2. https://hdruk.ac.uk/ Last accessed 29th March 2018
3. Eggermont AM, Caldas C, Ringborg U, Medema R, Tabernero J, Wiestler O. Cancer Core Europe: a consortium to address the cancer care-cancer research continuum challenge. Eur J Cancer. 2014; 50:2745-6.
4. Lawler M, Haussler D, Siu LL, Haendel MA, McMurry JA, Knoppers BM, Chanock SJ, Calvo F, The BT, Walia G, Banks I, Yu PP, Staudt LM, Sawyers CL. Clinical Cancer Genome Task Team of the Global Alliance for Genomics and Health. Sharing Clinical and Genomic Data on Cancer - The Need for Global Solutions. N Engl J Med. 2017; 376:2006-2009.
5. http://www.isdscotland.org/Products-and-Services/eDRIS/ Last accessed 11th April 2018
6. Middleton A. Society and personal genome data. Hum Mol Genet. 2018 Mar 7. doi: 10.1093/hmg/ddy084. [Epub ahead of print]
7. https://understandingpatientdata.org.uk/ Last accessed 29th March 2018