Open Data in Health Care

Tom Pollard[*] and Leo Anthony Celi[†]

[*]University College London (UCL), London, UK

[†]Massachusetts Institute of Technology (MIT), Cambridge, USA

Signs of Life

As we pass through life in the digital era we leave a health trajectory in our wake. Phones, shopping habits, and visits to the doctor create a trace of data that can be used to not only assess our past and present wellbeing, but also forecast the future. To some, this is an unparalleled opportunity to improve health care, whereas to others it is an emerging threat to civil liberty. Most of us camp somewhere between the two poles: we see the rewards and we acknowledge the concerns. The question is how we move past this point, when business models and legal frameworks, built for

How to cite this book chapter:
Pollard, T. and Celi, L. A. 2014. Open Data in Health Care. In: Moore, S. A. (ed.) *Issues in Open Research Data*. Pp. 129–140. London: Ubiquity Press. DOI: http://dx.doi.org/10.5334/ban.h

a pre-internet world, struggle to keep up with the pace of change (Park & VanRoekel 2013).

The movement to give us open access to research articles began roughly fifteen years ago[1]. Before the dust has settled, there is now a strong push from researchers, funders, and publishers to open the data that underpins those articles. The suggestion to share research data is hardly new—Sir Francis Galton entertained this thought in 1901 (Hrynaszkiewicz & Altman 2009)—but technology now exists to enable sharing with relative ease. Culture is largely the barrier that restricts flow of research data, and for data sharing to be adopted there are challenges to overcome around privacy, competition, and incentives to share (Wellcome Trust 2014).

Improving Care

The medical and biomedical research professions have come under heavy criticism in recent years (Celi 2014). The Institute of Medicine's 1999 report 'To err is human', for example, estimated that between 44,000 people and 98,000 people die in US hospitals each year as a result of preventable medical errors, with even the lower estimate exceeding mortality of threats such as AIDS and breast cancer (eds. Kohn, Corrigan & Donaldson 2000). Further high profile blows were delivered in 2013, with the US National Research Council's report on 'Shorter Lives, Poorer Health' and The Economist's 'Unreliable research: Trouble at the lab' (National Research Council 2013; The Economist, 2013). 'Half of what we know might be wrong, and the other half useless,' is perhaps the

[1] Two key reference points are Steven Harnad's 'subversive proposal' in 1994 and the founding of the Budapest Open Access Initiative in 2001.

most damning appraisal of the state of medical knowledge, coming from Professor John Ioannidis in his editorial 'How Many Contemporary Medical Practices are Worse than Doing Nothing or Doing Less?' (Ioannidis 2013).

It is widely acknowledged that better handling of information could address many of the criticisms, potentially helping to transform the quality of research and care (Institute of Medicine 2000; Wellcome Trust 2014). When data is not shared, quality of care suffers through inefficiencies, proliferation of errors, and wasted opportunities for learning. An open approach enables refinement of knowledge and collaborative growth towards united goals (Ioannidis et al. 2014).

When efforts are collaborative, progress can be rapid. One such example was the global research effort in 2011 to sequence and analyse the genome of a toxic strain of *Escherichia coli*, quickly helping to control the outbreak and prevent further deaths (The Royal Society 2012; Rohde et al. 2011). Transparency, through open data, can also highlight potential cost savings in our health systems. A recent study in England suggested potential savings of over £300 million pounds per year through switching to generic equivalents of two branded drugs (Allen 2012).

Qualifying 'Open'

The definition of open data is unequivocal: 'A piece of data is open if anyone is free to use, reuse, and redistribute it—subject only, at most, to the requirement to attribute and/or share-alike' (Open Knowledge, 2013). This is a copyright-centric model of sharing, facilitated by the adoption of 'copyleft' licences that allow reproduction and reuse (Hrynaszkiewicz & Cockerill 2012; Korn & Oppenheim 2011). This approach to sharing means there are few downstream restrictions, allowing, for example, reuse in classrooms, industry, research, and 'citizen science', maximising the potential of the data.

Where we are dealing with sensitive information, as we often are in health, it is fair to accept that there is a limit to what can be shared openly. Unless explicit consent for sharing has been obtained, details may have to be abstracted or removed to protect the individuals. Finding the appropriate balance between anonymisation and retaining useful detail is not straightforward, often involving a trade-off between risk and value.

As a result of this trade-off, John Wilbanks, who worked for years at Creative Commons[2], suggests that the copyright-centric approach may be unsuitable for health data. Wilbanks champions an alternative model built on trust (Howard 2012). Projects that have adopted this privacy-centric approach include his Portable Legal Consent study and Sage Bionetworks' clinical research studies, which seek to match participants willing to share their data with networks of researchers under contract to 'play fair' by returning research insights and not attempting to re-identify individuals.

It is likely and desirable for data sharing to progress on both privacy- and copyright-centric branches: we will get better at sharing 'true' open data with few restrictions on downstream reuse, and we will also develop platforms for sharing within trusted networks. Complementing both approaches are practical measures of openness, which assess whether data can be found, accessed, and reused. Open Knowledge has assembled a list of examples of 'bad data', which emphasise, lightheartedly, that there is more to sharing data than dropping files onto

[2] Creative Commons: http://creativecommons.org/

a public website (Open Knowledge, 2014a). Another initiative by Open Knowledge, the Open Data Index, provides a series of questions to assess the availability and openness of data, asking, for example, whether data is machine readable (e.g. text instead of image), available in a non-proprietary format (e.g. CSV instead of XLS for tabulated data), and openly licensed (e.g. with a Creative Commons licence) (Open Knowledge, 2014b).

Doing No Harm

Radicals may be prepared to bare all on the web, but the majority of us have expectations that certain information will remain within trusted networks. Our desire for privacy goes beyond avoidance of embarrassment. Revealing identifiable information that relates to our physical, mental, and social wellbeing has risks, for example by enabling discrimination by insurers or employers. While the level of risk can be debated and varies from case to case, it is clear that damage is possible. In a well-referenced case in 2008, for example, a nurse's career was compromised when confidential health information was leaked to her employer (European Court of Human Rights 2008).

All health data is sensitive and should be treated with respect, but the specific legal provisions that regulate data processing and sharing vary by location (UK Parliament 1998; United States Congress 1996). Regardless of the legal framework, regulation is implemented to achieve a similar effect—protection of the data subjects—and so rather than discuss detail in specific locations we give an introduction to the general concepts here. Our aim is to protect the individual, and so whether or not an item is defined as 'personal information' we should err on the side of caution when sharing data to mitigate the risks of harm. Open data must either not identify the individual or there must be explicit consent to share.

Anonymisation is a method that can be employed to open up health data, by separating information from the individual. The EU Data Protection Directive, for example, states that ‘the principles of protection shall not apply to data rendered anonymous in such a way that the data subject is no longer identifiable. In addition, the UK Information Commissioner’s Office’s Anonymisation: Managing Data Protection Risk Code of Practice document notes:

> There is clear legal authority for the view that where an organisation converts personal data into an anonymised form and discloses it, this will not amount to a disclosure of personal data. This is the case even though the organisation disclosing the data still holds the other data that would allow re-identification to take place.
>
> (Information Commissioner’s Office 2012)

Successful anonymisation is not straightforward, however, and there are examples of both failure and success (El Emam et al. 2012; Neamatullah et al. 2008; Ohm 2010; Parry 2011).

In cases of breaches of privacy, regardless of the cause, proportionality is important and failures need to be considered in context. Treating breaches with ‘witch-hunts’ and exorbitant fines may not have the desired effect. Rather than creating a positive environment for safe data sharing, we create a culture of fear and lockdown, with inadequate systems and individuals unwilling to take responsibility. Researchers have argued that this has resulted in a tense environment, in which it becomes:

> … easy for the public, and regulators, to lose sight of how easily the increasingly broad body of restrictions limiting access to medical and public health data can

> undermine efforts to better understand and improve public health.
>
> (Wartenberg & Thompson 2010)

In medicine, we are learning that ‘naming, shaming, and blaming’ does not contribute to a safety culture, and this is a lesson that also needs to be learnt when it comes to data (Leslie 2014).

Our Future Selves

As the saying goes, our future self is the first recipient of shared data. Imagine trying to work with your data a year or two down the line – perhaps while writing up a thesis or perhaps while finally getting round to sorting out the revisions on a paper. You don’t want to be dealing with a smattering of unlabelled disks, containing a bunch of old files in unrecognised formats, on a desk that belongs to a previous employer. If data is well described, organised, and in non-proprietary formats, it will be easier to sort through, share and reuse.

Often we are required to register with a project or ethics committee prior to collecting or accessing health data, so it makes sense to sketch out a data management plan at this point. There are resources on the web to help create the plan, such as the Digital Curation Centre’s DMPonline[3] and the UK Data Archive’s Data Management Checklist (UK Data Archive 2014). Best practice is developing rapidly, so a specialist such as an academic librarian or local information governance manager should be involved where possible.

If a project requires consent from participants, it is important to clearly set out any intentions for data sharing. Good

[3] https://dmponline.dcc.ac.uk/

communication is crucial and keeping people informed from the outset will help to establish trust. Where it is not possible to obtain consent, or where consent has not been obtained for retrospective data, a local ethics committee should be approached for advice (Hrynaszkiewicz et al. 2010). Approval by the committee may be given where data is anonymised, but care is needed to maintain privacy. The British Medical Association offers a toolkit outlining key factors to take into account when sharing data, and the UK Information Commissioner's Office offers an overview of approaches to anonymisation (British Medical Association 2014; Information Commissioner's Office 2012).

Data that is not properly described is unlikely to be reused, so good metadata is vital. At the simplest level, metadata can be a description of the important details of the data. Reviewing data papers, such as those published in *Open Health Data* and *Scientific Data*,[4] may help to identify useful information to include. More formal metadata standards are established according to discipline and should be adopted where appropriate. A directory of standards is maintained by the Digital Curation Centre (Digital Curation Centre 2014). In cases where data cannot be shared due to privacy issues, it should almost always be possible to share the descriptive metadata, making the data discoverable and potentially reusable.

In terms of data publication, there are an increasing number of options, such as creating new instances of web-accessible databases (for example, via DataVerse), depositing in an institutional repository, or sharing via data publishers such as Dryad and figshare (King & Crosas 2014).[5] Most importantly, the service

[4] *Open Health Data*: http://openhealthdata.metajnl.com/; *Scientific Data*: http://www.nature.com/sdata/

[5] Dryad: http://datadryad.org/; figshare: http://figshare.com/

should offer some reassurance that data will be sustained for the foreseeable future, and a unique identifier such as a digital object identifier (DOI) should be provided to enable accurate citation and tracking of reuse.

Anyone sharing data along these lines is leading the way, at the front of a community that is working towards better, collaborative science. With mechanisms for researchers to cite data, and funders increasingly recognising the importance of data sharing, a system that gives proper recognition to those who share data must now be on the horizon.

Acknowledgements

Thanks to Ross Mounce, Cameron Neylon, and Paul Wicks for helpful comments via social media, and to Samuel Moore and John Wilbanks for giving feedback on the content before publication.

References

Allen, K. (2012). Un-needed branded drugs "cost NHS millions." *Financial Times Data*. Retrieved August 06, 2014, from http://blogs.ft.com/ftdata/2012/12/06/un-needed-branded-drugs-cost-nhs-millions/.

British Medical Association. (2014). Confidentiality and disclosure of health information tool kit. Retrieved August 05, 2014, from http://bma.org.uk/practical-support-at-work/ethics/confidentiality-tool-kit.

Celi, L. (2014). The Outing of the Medical Profession: Data marathons to open clinical Research Gates to Frontline Service Providers. *London School of Economics Impact Blog*. Retrieved August 06, 2014, from http://blogs.lse.ac.uk/.

Digital Curation Centre. (2014a). DMPonline. Retrieved August 05, 2014, from https://dmponline.dcc.ac.uk/.

Digital Curation Centre. (2014b). List of Metadata Standards. Retrieved August 06, 2014, from http://www.dcc.ac.uk/resources/metadata-standards/list.

Dryad Data. (2014). Retrieved August 06, 2014, from http://datadryad.org/.

El Emam, K., Arbuckle, L., Koru, G., Eze, B., Gaudette, L., Neri, E., … Gluck, J. (2012). De-identification methods for open health data: the case of the Heritage Health Prize claims dataset. *Journal of Medical Internet Research*, *14*(1), e33. doi:10.2196/jmir.2001.

European Court of Human Rights. I v. Finland (2008).

figshare. (2014). Retrieved August 06, 2014, from http://figshare.com/.

Howard, A. (2012). The risks and rewards of a health data commons. *O'Reilly Radar*. Retrieved August 05, 2014, from http://radar.oreilly.com/2012/08/health-data-commons.html.

Hrynaszkiewicz, I., & Altman, D. G. (2009). Towards agreement on best practice for publishing raw clinical trial data. *Trials*, *10*, 17. doi:10.1186/1745-6215-10-17.

Hrynaszkiewicz, I., & Cockerill, M. J. (2012). Open by default: a proposed copyright license and waiver agreement for open access research and data in peer-reviewed journals. *BMC Research Notes*, *5*(1), 494. doi:10.1186/1756-0500-5-494.

Hrynaszkiewicz, I., Norton, M. L., Vickers, A. J., & Altman, D. G. (2010). Preparing raw clinical data for publication: guidance for journal editors, authors, and peer reviewers. *Trials*, *11*, 9. doi:10.1186/1745-6215-11-9.

Information Commissioner's Office. (2012a). *Anonymisation: Managing Data Protection Risk Code of Practice* (p. 108).

Information Commissioner's Office. (2012b). *Anonymisation: managing data protection risk code of practice*. ICO. Retrieved from http://ico.org.uk/for_organisations/data_protection/topic_guides/anonymisation.

Institute of Medicine. (2000). *To Err Is Human*. (Committee on Quality of Health Care in America, Ed.) (p. 312). Washington D.C.: National Academy Press.

Ioannidis, J. P. A. (2013). How many contemporary medical practices are worse than doing nothing or doing less? *Mayo Clinic Proceedings*, *88*(8), 779–81. doi:10.1016/j.mayocp.2013.05.010.

Ioannidis, J. P., Greenland, S., Hlatky, M. A., Khoury, M. J., Macleod, M. R., Moher, D., … Tibshirani, R. (2014). Increasing value and reducing waste in research design, conduct, and analysis. *Lancet*, *383*(9912), 166–75. doi:10.1016/S0140-6736(13)62227-8.

King, G., & Crosas, M. (2014). The Dataverse Project. Retrieved August 06, 2014, from http://datascience.iq.harvard.edu/dataverse.

Korn, N., & Oppenheim, C. (2011). *Licensing Open Data: A Practical Guide (Version 2.0)*. Retrieved from http://discovery.ac.uk/files/pdf/Licensing_Open_Data_A_Practical_Guide.pdf.

Leslie, I. (2014, June). How mistakes can save lives: one man's mission to revolutionise the NHS. *New Statesman*. Retrieved from http://www.newstatesman.com/2014/05/how-mistakes-can-save-lives.

National Research Council. (2013). *U.S. Health in International Perspective: Shorter Lives, Poorer Health*. The National Academies Press. Retrieved from http://www.nap.edu/openbook.php?record_id=13497.

Neamatullah, I., Douglass, M. M., Lehman, L. H., Reisner, A., Villarroel, M., Long, W. J., … Clifford, G. D. (2008). Automated de-identification of free-text medical records. *BMC Medical Informatics and Decision Making*, *8*, 32. doi:10.1186/1472-6947-8-32.

Ohm, P. (2010). Broken Promises of Privacy: Responding to the Surprising Failure of Anonymization. *UCLA Law Review*, *57*, 1701. Retrieved from http://papers.ssrn.com/abstract=1450006.

Open Knowledge. (2013). Open Definition. *Open Knowledge*. Retrieved August 04, 2014, from http://opendefinition.org/od/.

Open Knowledge. (2014a). Bad Data. Retrieved August 06, 2014, from http://okfnlabs.org/bad-data/.

Open Knowledge. (2014b). Open Data Index. Retrieved August 06, 2014, from https://index.okfn.org/about/#criteria.

Park, T., & VanRoekel, S. (2013). Introducing: Project Open Data. *The White House*. Retrieved August 04, 2014, from http://www.whitehouse.gov/blog/2013/05/16/introducing-project-open-data.

Parry, M. (2011). Harvard's Privacy Meltdown. *The Chronicle of Higher Education*. Retrieved August 05, 2014, from http://chronicle.com/article/Harvards-Privacy-Meltdown/128166/.

The Economist. (2013, October). Unreliable research: Trouble at the lab. *The Economist*. Retrieved from http://www.economist.com/node/21588057/print.

The Royal Society. (2012). *Science as an open enterprise: Final report*. London. Retrieved from https://royalsociety.org/policy/projects/science-public-enterprise/report/.

UK Data Archive. (2014). Data Management Checklist. Retrieved August 05, 2014, from http://www.data-archive.ac.uk/create-manage/planning-for-sharing/data-management-checklist.

UK Parliament. Data Protection Act 1998. Her Majesty's Stationery Office (1998). United Kingdom of Great Britain and Northern Ireland.

United States Congress. Health Insurance Portability and Accountability Act (1996). United States.

Wartenberg, D., & Thompson, W. D. (2010). Privacy versus public health: the impact of current confidentiality rules. *American Journal of Public Health*, *100*(3), 407–12. doi:10.2105/AJPH.2009.166249.

Wellcome Trust. (2014). *Establishing incentives and changing cultures to support data access*.

Zhao, M., Wang, P., Guan, Y., Cen, Z., Zhao, X., Christner, M., … Wang, J. (2011). Open-Source Genomic Analysis of Shiga-Toxin–Producing.