Big Health Data: The need to earn public trust

Tjeerd-Pieter van Staa[1], Ben Goldacre[2],[3], Iain Buchan[1], Liam Smeeth[2]

1 Farr Institute, University of Manchester, Manchester, United Kingdom

2 London School of Hygiene & Tropical Medicine, London, United Kingdom

3 Nuffield Department of Primary Care Health Sciences, Oxford University, Oxford, Unite

Summary

There are great potential gains for day-to-day patient care, public health and research from better uses of large-scale and large volume health data (also known as ‘big data’). The handling of such data, however, raises concerns over preserving patient privacy, even where the risks of disclosure are extremely small. In England, the flawed implementation of recent big data projects using healthcare records coupled with historic examples of data mismanagement have made this a mainstream public concern. We explore whose responsibility it was (and is) to make the case to the public for large-scale linkage and analysis of health data. Looking to the future, we discuss what could be done better in future in countries or regions seeking to develop data- and research-intensive health systems.

Introduction

Recent English initiatives sought major expansion of the aggregation and accessibility of routinely collected healthcare and related records, sometimes loosely referred to as “big data”. One such initiative, care.data, was set to link and provide access to health and social care information from different settings, including primary care, in order to facilitate the planning and provision of healthcare, and to advance health science [1]. Data were to be extracted from all primary care practices in England. A related initiative, the Clinical Practice Research Datalink (CPRD), evolved from the General Practice Research Database (GPRD). CPRD was intended to build on GPRD by linking patients’ primary care records to hospital data, around 50 disease registries and clinical audits, UK Biobank with genetic information, and even the loyalty cards of a large supermarket chain, creating an integrated data repository and linked services for all of England that could be sold to universities, pharmaceutical companies and non-healthcare industries. However, these initiatives have stalled.

Why have English health big data initiatives not worked?

Key elements for success of big health data projects include: public confidence that records are held securely and anonymised appropriately (i.e. information security) [2]; public awareness of and engagement with how their personal data have been, or might be, used (i.e. citizen-visible data uses) [2]; and high quality science with the data. An analysis of opinions expressed on Twitter reported concerns about: informed consent and the default 'opt-in'; trust; privacy and data security; the involvement of private companies; legal issues; GPs' concerns; and communication failures resulting in confusion about care.data [3]. The public information campaign and the information leaflets of care.data gave no clarity on how the system would work, including the opt-out arrangements and the sharing of personal information with commercial organisations [4,5]. Failure to earn the trust and confidence of patients, citizens and healthcare professionals in the use of big data was a key failing of care.data [2]. The leaflet on care.data delivered to every household in England did not mention research benefits.

A recent literature review found that members of the public often have little awareness of the ways in which patient information is currently used and to whom it is available [6]. But focus groups found that participants become more accepting of big health data uses after being given more information [7]. It is clear that public involvement in bettering uses of health data is crucial. Unfortunately, so far, efforts here have been piecemeal. There are research-led activities via social media (such as the #datasaveslives campaign [http://www.datasaveslives.eu/]), and ad-hoc media briefings by academics. Overall, however, this important area has received very little investment.

Research access in England to large-scale healthcare data (such as CPRD) are currently based on centralised national approaches in which with copies of data are distributed to local computers of researchers. The disadvantage of this is that the uses of the data are not easily audited or controlled and there is limited transparency in data uses. It has created stories in the public domain of data mismanagement. A recent review of historic releases of health care data by the government [8] triggered newspaper headlines such as “Millions of patient records were sold to insurance firms who used it to set their critical illness premiums in a series of 'unacceptable lapses”. Concerns have also been expressed that data had been uploaded to the Google cloud for more powerful analytics, which was seen to compromise UK and EU data protection.

Credible science is another key foundation for meaningful public engagement. The need to replicate findings across heterogeneous populations and settings is well recognised [9]. The medical literature is plagued with specious findings, often made from observational studies using routine healthcare data [10], even reaching opposite conclusions with the same data sources. An example is the cancer effects of different diabetes medications [11,12]. Particular barrier to replication has been the lack of publication of algorithms and clinical code lists alongside manuscripts.

Basic anonymisation of information (i.e. removal of obvious identifiable information such as names and addresses) has been widely used to address public concerns with research uses of personal data. However, the challenge with linking different sources of information (such as with care.data or CPRD) is the granular nature of the data, increasing the risk of deductive disclosure (in which an individual can be identified indirectly through data analysis).

What has worked elsewhere?

The Welsh Secure Anonymous Information Linkage System (SAIL) is a single safe haven (i.e. researchers go to the data rather the data to them) of a large number of datasets and a platform for sharing knowledge about using the data (e.g. data dictionaries). It operates a remote access system providing secure data access for approved users and data analysis tools [13]. The Scottish Health Informatics Programme (SHIP) also developed ways for researchers to manage and analyse electronic patient records, and associated linked data. SHIP ran a substantial public engagement programme aimed at understanding the publics' preferences, interests and concerns relating to the sharing of health data for research and their acceptance and attitudes towards the aims of the programme [14]. The Canadian Network for Observational Drug Effect Studies (CNODES) uses a system of sending analysis queries to local data repositories across the country with the results combined centrally in a meta- analysis [15]. A large US data source, Mini-Sentinel, collates healthcare data from around 100 million individuals and which also uses distributed queries [16], and PCORnet (www.pcornet.org) marks a ramping up of US investment in this area. The Nordic countries routinely extend their health data linkage to income and educational attainment records [17].

What should we do now?

Public confidence and transparency in the information security is pivotal. A workshop organised by the Academy of Medical Sciences (among others) proposed that sensitive data should be stored and analysed in safe havens. Data security risks can then be managed better by segregating sensitive data, controlling data access and monitoring data uses [18]. In order for safe havens to operate efficiently (at low cost and rapid responsiveness) they will need to combine different uses of the same data in an economy of scale. But they also need to engage with their communities over data uses. If the population covered is too large to be considered a natural community then interacting with citizens over using their personal data for the public good might be difficult, as it is often easier for a citizen to relate to people from their region than outside [19]. Over the past decade, progress has been made toward automating more distributed analysis with individual-level data and federated datasets [20]. With distributed approaches individual-level data are analysed locally and only summary results or intermediate statistics are shared. Different computational models for distributed analyses are being developed, such as meta-analysis of individual node results, distributed securemultiparty computation platforms, or distributed queries. An example of a federation of local safe havens, known as Arks, is being developed, linked to the Connected Health Cities pilots in North England [21]. The ultimate solution, however, must combine new technologies with clear accountability, transparent operations and public trust. In addition, data stewardship is not just about physical and digital security: staff training, standard operating procedures, and the skills and attitudes of staff are also important [22]. This combination of data protection (safe havens) and culture of best practice not only underpins a 'trustworthy research environment' but also a 'learning health system' [23,24]. But many researchers much prefer to download data rather than access them through safe havens [25]. But this approach of data handling in many local systems may pose challenges to both data security and transparency in use as the data are not easily audited or controlled.

Transparency and citizen-visible data uses are another ingredient for public trust [2]. One approach could be to document for each individual where and how their data have been used. Administering this is likely to be challenging from a communications perspective, for example explaining to a non-affected person why they were included (as a control) in a study of schizophrenia. A more complex approach is dynamic consent, linking audit to granular opt-out, where people can see which organisations have accessed their data, get information on data-analyses such as research findings, and change their consent preferences for specific uses over time [26] – uses such as research into adverse drug-outcomes or audits of clinical services. Prototypes for this are being developed [27]. The concept of dynamic consent views sees public acceptability of data sharing as varying with the types of use. There may be a range of public views on such re-use and decisions at a national level may struggle to reflect polarised opinions among individuals, either excluding uses which many find acceptable, or imposing 'all or nothing' choices around opt-out on individuals who are happy with the bulk of data uses, but sufficiently concerned about a small minority of uses that they consider opting out of all data sharing.

Public involvement is key to the success of large-scale health data uses [28]. There is a need for a resource where the public can access clear, high quality, up-to-date summaries of the scientific discoveries and healthcare improvements being made with the healthcare records in their community. We believe such a resource, embedded in health systems, would improve patient trust, reduce opt-outs, and reciprocate the value of data sharing by patients. It is best delivered by the academic community, in co-production with patients/citizens and staff with specialist skills in engaging and involving the public. This is a full-time labour, and such a resource will only exist if funders recognise its ethical importance and practical value. There may also be lessons to learn from wider policy arenas where public acceptance is crucial to success. Renewable energy is one such contentious area, with apparent contradictions in public opinion; for example, the apparent general public support for renewable energy and simultaneous difficulty in implementing specific local projects [29]. Developing a greater understanding of the dimensions of social acceptance seems just as relevant to uses of largescale health data as it is to renewable energy.

Conclusion

Public trust is likely to be more easily earned when researchers are seen to meet high scientific standards with transparency in methods and reproducibility of findings. There is now increasing interest in reproducibility of research [30] and open access to the statistical and data management algorithms used for complex analyses [31], to improve not only the analytics but also the interpretation of results. One proposal is the e-lab, a shared digital laboratory supporting consistent recording, description and sharing of data and statistical algorithms, facilitating rapid replication of findings [32]. Registration of protocols and publications in registers may further strengthen the reliability and credibility of studies with big data [33].

Most people would expect a health service to monitor clinical outcomes so that quality of care and the effects of interventions can be assessed. Such activities, by definition, need peoples' health care data, just as running schools requires data on individual pupils. The UK has globally important health data assets which, when analysed together, can improve health systems. Harnessing the data for patient and public benefit, however, has been set back by the flawed implementation of a national ‘big data’ project. This pause has revealed a bigger picture of the need for large-scale involvement of citizens in advancing the uses of their communities’ health data. The time is now for the key stakeholders in health systems to act in concert and properly resource meaningful, enduring public involvement in big health data. Public trust can only be earned if there is: transparency in information security; dynamic consent with the ability to opt-out of specific uses of data; scientific transparency and reliability; and systematic public engagement.

Key Messages

- Big health data projects can only succeed if the public has confidence that their records are held securely and anonymised appropriately
- Meaning public engagement requires a ‘trustworthy research environment’ and public transparency in data use
- It also requires credible science with the need to replicate findings across heterogeneous populations and settings and the need for an e-lab, a shared digital laboratory supporting consistent recording, description and sharing of data and statistical algorithms

Author Contributions

TvS is the guarantor. Contributed to the writing of the manuscript: TvS BG IB LS. Wrote the first draft of the manuscript: TvS. Agree with the manuscript’s results and conclusions: TvS BG IB LS. Conceptualized the project and interpreted data for the manuscript: TvS BG LS. Critically revised the manuscript for intellectual content: BG IB LS. Final approval of the version to be published: TvS BG IB LS. Agreement to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved: TvS BG IB LS.

Author Experiences

This article has been jointly written by two experts in data science and analysis (TvS and LS), one expert in public engagement (BG) and one expert in Health Informatics (IB)

1. NHS England. England. NHS England » The care.data programme – collecting information for the health of the nation. https://www.england.nhs.uk/ourwork/tsd/care-data/
2. Carter P, Laurie GT, Dixon-Woods M. The social licence for research: why care.data

ran into trouble. J Med Ethics 2015;41:404–9. doi:10.1136/medethics-2014-102374

3. Hays R, Daker-White G. The care.data consensus? A qualitative analysis of opinions expressed on Twitter. BMC Public Health 2015;15:838. doi:10.1186/s12889-015-2180-9
4. Goldacre B. Care.data is in chaos. It breaks my heart. Guard. 2014.
5. Goldacre B. The NHS plan to share our medical data can save lives – but must be done right. Guard. 2014.
6. General Medical Council. Review of public and professional attitudes towards confidentiality of healthcare data. http://www.gmc-uk.org/Review_of_Public_and_Professional_attitudes_towards_confidentiality_of_Healthcare_data.pdf_62449249.pdf (accessed 18 Feb2016).
7. Hill EM, Turner EL, Martin RM, et al. 'Let's get the best quality research we can': public awareness and acceptance of consent to use existing data in health research: a systematic review and qualitative study. BMC Med Res Methodol 2013;13:72. doi:10.1186/1471-2288-13-72
8. Review of data releases by the NHS Information Centre. https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/367788/Sir_Nick_Partridge_s_summary_of_the_review.pdf (accessed 18 Feb2016).
9. Thompson S, Kaptoge S, White I, et al. Statistical methods for the time-to-event analysis of individual participant data from multiple epidemiological studies. Int J Epidemiol 2010;39:1345–59. doi:10.1093/ije/dyq063
10. Ioannidis JPA. Why most published research findings are false. PLoS Med 2005;4:e124. doi:10.1371/journal.pmed.0020124
11. Currie CJ, Poole CD, Gale E a M. The influence of glucose-lowering therapies on cancer risk in type 2 diabetes. Diabetologia 2009;52:1766–77. doi:10.1007/s00125-009-1440-6
12. Van Staa TP, Patel D, Gallagher AM, et al. Glucose-lowering agents and the patterns of risk for cancer: a study with the General Practice Research Database and secondary care data. Diabetologia 2012;55:654–65. doi:10.1007/s00125-011-2390-3
13. Jones KH, Ford D V, Jones C, et al. A case study of the Secure Anonymous Information Linkage (SAIL) Gateway: a privacy-protecting remote access system for health-related research and evaluation. J Biomed Inform 2014;50:196–204. doi:10.1016/j.jbi.2014.01.003
14. Scottish Health Informatics Programme. http://www.scotship. ac.uk/sites/default/files/Reports/SHIP_BLUEPRINT_DOCUMENT_final_100712.pdf (accessed 18 Feb2016).
15. Suissa S, Henry D, Caetano P, et al. CNODES: the Canadian Network for Observational Drug Effect Studies. Open Med 2012;6:e13440.http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3654509&tool=pmcentrez&rendertype=abstract (accessed 18 Feb2016)
16. Platt R, Carnahan RM, Brown JS, et al. The U.S. Food and Drug Administration's Mini-Sentinel program: status and direction. Pharmacoepidemiol Drug Saf 2012;21
17. Furu K, Wettermark B, Andersen M, et al. The Nordic countries as a cohort for pharmacoepidemiological research. Basic Clin Pharmacol Toxicol 2010;106:86–94. doi:10.1111/j.1742-7843.2009.00494.x
18. Academy of Medical Sciences. Data in Safe Havens | Academy of Medical Sciences. 2014.
19. Ainsworth J, Buchan I. Combining Health Data Uses to Ignite Health System Learning. Methods Inf Med 2015;in press.

20. Boyd S. Distributed Optimization and Statistical Learning via the Alternating Direction Method of Multipliers. Found Trends® Mach Learn 2010;3:1–122. Doi.10.1561/2200000016
21. UK Govt. Chancellor George Osborne’s Budget 2015 speech - Speeches - GOV.UK. 2015.
22. Mackenzie IS, Mantay BJ, McDonnell PG, et al. Managing security and privacy concerns over data storage in healthcare research. Pharmacoepidemiol Drug Saf 2011;20:885–93. doi:10.1002/pds.2170
23. Ainsworth J, Buchan I. eLabs and Work Objects: Toward Digital Health Economies. Lecture Notes of the Institute for Computer Sciences. Soc Informatics Telecommun Eng 2009;:205–16.
24. Friedman C, Rubin J, Brown J, et al. Toward a science of learning systems: a research agenda for the high-functioning Learning Health System. J Am Med Inform Assoc 2015;22:43–50. doi:10.1136/amiajnl-2014-002977
25. Wellcome Trust. Enhancing Discoverability of Public Health and Epidemiology Research Data. 2014.http://www.wellcome.ac.uk/stellent/groups/corporatesite/@policy_communications/documents/web_document/wtp056925.pdf (accessed 18 Feb2016)
26. Williams H, Spencer K, Sanders C, et al. Dynamic consent: a possible solution to improve patient confidence and trust in how electronic patient records are used in medical research. JMIR Med informatics 2015;3:e3. doi:10.2196/medinform.3525
27. EnCoRe - Ensuring Consent and Revocation. http://www.hpl.hp.com/breweb/encoreproject/index.html (accessed 18 Feb2016).
28. Nuffield Council on Bioethics. The collection, linking and use of data in biomedical research and health care: ethical issues. Nuff. Bioeth. 2015.
29. Wüstenhagen R, Wolsink M, Bürer MJ. Social acceptance of renewable energy innovation: An introduction to the concept. Energy Policy 2007;35:2683–91. doi:10.1016/j.enpol.2006.12.001
30. The Academy of Medical Sciences. Reproducibility and reliability of biomedical research: improving research practice.http://www.acmedsci.ac.uk/policy/policyprojects/reproducibility-and-reliability-of-biomedical-research/
31. Check Hayden E. Journal buoys code-review push. Nature 2015;520:276–7. doi:10.1038/520276a
32. Custovic A, Ainsworth J, Arshad H, et al. The Study Team for Early Life Asthma Research (STELAR) consortium ‘Asthma e-lab’: team science bringing data, methods and investigators together. Thorax 2015;70:799–801. doi:10.1136/thoraxjnl-2015-206781
33. Swaen GMH, Carmichael N, Doe J. Strengthening the reliability and credibility of observational epidemiology studies by creating an Observational Studies Register. J Clin Epidemiol 2011;64:481–6. doi:10.1016/j.jclinepi.2010.04.009