van Staa, TP; Goldacre, B; Buchan, I; Smeeth, L (2016) Big health data: the need to earn public trust. BMJ (Clinical research ed), 354. i3636. ISSN 0959-8138 DOI: 10.1136/bmj.i3636

Downloaded from: http://researchonline.lshtm.ac.uk/2728972/

DOI: 10.1136/bmj.i3636
Big health data: the need to earn public trust

Failures in implementation of data sharing projects have eroded public trust. In the wake of NHS England’s decision to close down its care.data programme, Tjeerd-Pieter van Staa and colleagues examine how we can do better

Tjeerd-Pieter van Staa professor of health e-research¹ ², Ben Goldacre senior clinical research fellow³ ⁴, Iain Buchan professor of health informatics¹, Liam Smeeth professor of clinical epidemiology³

¹Farr Institute, University of Manchester, Manchester, UK; ²Utrecht Institute for Pharmaceutical Sciences, Utrecht University, Utrecht, Netherlands; ³London School of Hygiene and Tropical Medicine, London, UK; ⁴Nuffield Department of Primary Care Health Sciences, Oxford University, Oxford, UK; Correspondence to: T-P van Staa tjeerd.vanstaa@manchester.ac.uk

Better use of large scale health data has the potential to benefit patient care, public health, and research. The handling of such data, however, raises concerns about patient privacy, even when the risks of disclosure are extremely small.

The problems are illustrated by recent English initiatives trying to aggregate and improve the 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, 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, genetic information from UK Biobank, 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, drug companies, and non-healthcare industries. Care.data has now been abandoned and CPRD has stalled. The flawed implementation of care.data plus earlier examples of data mismanagement have made privacy issues a mainstream public concern. We look at what went wrong and how future initiatives might gain public support.

Why have English 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 (information security); public awareness of and engagement with how their personal data have been, or might be, used;² and data being used for high quality science.

Care.data failed to earn the trust and confidence of patients, citizens, and healthcare professionals.² An analysis of opinions reported on Twitter showed that people had concerns about informed consent and the default “opt-in”;³ trust; privacy and data security; the involvement of private companies; and legality.² The information campaign about care.data was not clear about how the system would work, including the opt-out arrangements and the sharing of personal information with commercial organisations,³ ⁴ and at times downplayed the potential benefits.

This highlights a broader problem about public perception of how data are used and managed. A recent literature review found that many people do not know how patient information is currently used or who can use it.⁵ But focus groups found that participants become more accepting of big health data uses after being given more information.⁶

Researchers currently get access to large scale healthcare data (such as CPRD) in England through copies sent to their local computers. This makes it difficult to monitor or control how the data are used, leading to stories of data mismanagement and 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.”⁷ Concerns have also been expressed by patient groups and in UK parliament about data protection being compromised by data being uploaded to the Google cloud to access more powerful analytic tools.⁹

Basic anonymisation of information (such as removing names, addresses, and other identifiable information) has been widely used to allay public concerns about use of personal data for research data. However, the challenge with linking different sources of information (such as with care.data or CPRD) is the increasing level of detail in the data and possibility of deductive disclosure. For example, this could occur if a person discloses on social media that they visited their practice on some dates and were admitted to hospital with flu.

Clearly, we need to get public support by including them in developing ways to make better use of health data. Unfortunately, so far, efforts here have been piecemeal. There are research led activities informing the public through social media such as the #datasaveslives campaign (www.datasaveslives.eu) and ad hoc media briefings by academics. Another example is the citizens’ jury in which members of the
public are provided with different perspectives to discuss. A recent jury found that when informed of both the risks and opportunities associated with health data sharing, the public believe an individual’s right to privacy should not prevent research that can benefit patients overall. It concluded that patients should be notified of information sharing schemes and have the right to opt out if they so choose. In her recent review on data security, consent, and opt outs, the UK national data guardian, Fiona Caldicott, found that the case for data sharing still needs to be made to the public.

Another key factor in gaining public support is showing that science from such projects is credible. The need to replicate findings across heterogeneous populations and settings is well recognised. However, the medical literature is plagued with spurious findings, often made from observational studies using routine healthcare data. Some studies have even reached conflicting results from the same data sources—for example, a study that found an increased risk of cancer with glucose lowering drugs using the GPRD was contradicted a few years later by another that found no effect on cancer risk. A particular barrier to replication is that algorithms and lists of clinical codes are not published alongside research papers.

What has worked elsewhere?

Large databases in other countries have managed to obtain public support. Unlike the English examples above, the Welsh Secure Anonymous Information Linkage (SAIL) system researchers go to the data rather than have the data sent to them. SAIL contains a large number of datasets and a platform for sharing knowledge about using the data. It operates a remote access system providing secure data access for approved users and data analysis tools. 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 public’s preferences, interests, and concerns about use of health data for research and their acceptance and attitudes towards the aims of the programme. This enabled SHIP to define a transparent and publicly acceptable approach to governance of research with health data.

Outside the UK, 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. A large US data source, Mini-Sentinel, collates healthcare data from around 100 million people and also uses distributed queries, 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.

What should we do now?

Public involvement is key to successful use of large scale health data. The public need to be able to access clear, high quality, up-to-date summaries of the scientific discoveries and healthcare improvements made using data from healthcare records. This would improve patient trust, reduce opt-outs, and let patients share the value of data sharing. Such summaries should be produced by the academic community in collaboration with patients and staff with skills in engaging and involving the public. Producing this resource will be a full time job and requires funders to recognise its ethical importance and practical value.

There may also be lessons 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. Developing a greater understanding of the dimensions of social acceptance seems just as relevant to use of large scale health data as it is to renewable energy.

Public trust is more likely if researchers are seen to meet high scientific standards through transparency in their methods and reproducibility of findings. The scientific community is showing increasing interest in improving reproducibility. One proposal is the e-laboratory, a shared digital laboratory supporting consistent recording, description, and sharing of data and statistical algorithms, facilitating rapid replication of findings. Registration of protocols and publications in registers may further strengthen the reliability and credibility of studies using big data.

Transparency and visible uses of data are also important for public trust. One approach could be to document where and how each person’s data have been used. Administering this is likely to be challenging from a communications perspective—for example, explaining to non-affected people why they were included (as a control) in a study of schizophrenia. A more complex approach is dynamic consent, where people can see which organisations have accessed their data, get information on data analyses, and change their consent preferences for specific uses over time. Prototypes for this are being developed. Individuals’ views on different types of data use may vary and thus imposing “all or nothing” choices on opt-out risks losing data from people who are happy with most uses but sufficiently concerned about specific uses to opt out of all data sharing.

Public confidence in 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 centralised “safe havens,” arguing that data security risks can then be managed better by segregating sensitive data, controlling data access, and monitoring data uses. In order for safe havens to operate efficiently (at low cost and rapid responsiveness) they will need to facilitate different uses of the same data. But they also need to engage with the communities and clinical teams providing the data in order to get people to relate what is happening with their data.

Many researchers prefer to download data rather than access them through safe havens. One way to improve data security and transparency for this approach is to use distributed analysis in which individual level data are analysed locally and only summary results or intermediate statistics are downloaded to and shared with researchers. A federation of local safe havens, known as Arks, is being developed, linked to the Connected Health Cities pilots in northern England.

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. 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.

Conclusion

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 people’s healthcare data. If the UK is to make use of its globally important health data assets key stakeholders in health systems must act together to properly resource meaningful, ending public involvement in big health data.

Contributors and sources: This article has been jointly written by two experts in data science and analysis (TvS and LS), an expert in public engagement (BG), and an expert in health informatics (IB). All authors contributed to the manuscript and approved the final version. TvS is the guarantor.

Competing interests: We have read and understood BMJ policy on declaration of interests and have no relevant interests to declare.

Provenance and peer review: Not commissioned; externally peer reviewed.

4 Goldacre B. Care.data is in chaos. It breaks my heart.
19 Academy of Medical Sciences. Data in safe havens. AMS, 2014.

Published by the BMJ Publishing Group Limited. For permission to use (where not already granted under a licence) please go to http://group.bmj.com/group/rights-licensing/permissions
Key messages

Success of big health data projects requires public confidence that records are held securely and anonymised appropriately

Public support requires that data use is transparent and produces credible science

The public need to be able to see and share the benefits of big data projects

Dynamic consent, enabling people to opt out of specific uses could increase support