Is data sharing really good for health?

Journal:	*BMJ*
Manuscript ID	BMJ.2016.034227
Article Type:	Analysis
BMJ Journal:	BMJ
Date Submitted by the Author:	25-Jun-2016
Complete List of Authors:	Pisani, Elizabeth; King\'s College London, Policy Institute Aaby, Peter; Bandim Health Project, Breugelmans, J. Gabriella; European & Developing Countries Clinical trials Partnership Carr, David; Wellcome Trust Groves, Trish; BMJ, BMJ Editorial Guerin, Philippe; University of Oxford, Infectious Diseases Data Observatory; University of Oxford, WorldWide Antimalarial Resistance Network Helinski, Michelle; European & Developing Countries Clinical trials Partnership Kamuya, Dorcas; KEMRI Wellcome Trust Research Programme; University of Oxford Ethox Centre Kern, Steven; Bill and Melinda Gates Foundation Littler, Katherine; Wellcome Trust Marsh, Vicki; KEMRI Wellcome Trust Research Programme; University of Oxford Mboup, Soulaymane ; Universite Cheikh Anta Diop Merson, Laura; University of Oxford, Infectious Diseases Data Observatory; University of Oxford, Centre for Tropical Medicine and Global Health Sankoh, Osman; INDEPTH Network; University of the Witwatersrand, School of Public Health Serafini, Michaela; Medecins Sans Frontieres Schneider, Martin; Universite de Geneve Institut de sante globale Schoenenberger, Vreni; International Federation of Pharmaceutical Manufacturers and Associations
Keywords:	data sharing, research ethics, public health, clinical trials, LMIC

Is data sharing really good for health?

It's only worth sharing data if we invest enough to make shared data useful

Abstract

The sharing of individual-level health research data has been much talked about but little practiced, especially when these data are collected in low and middle income countries. Commonly cited barriers relate to the possibility of data grabs by well-resourced northern analysts, to worries about patient consent and to inadequate infrastructure. As researchers who share the data we gather in low and middle income countries, we here examine the extent to which these fears have been realised in our own work. Our group includes researchers working for academic and humanitarian organisations, as well as public, charitable and industry funders of data sharing efforts.

In our experience, data sharing has resulted in health benefits principally where data are well documented and standardised, so that analyses can be conducted across studies conducted in different places and at different times. This requires substantial investment in data management. But better technology will not by itself wring more knowledge out of shared data. To share data usefully, we must start sharing science more equitably throughout networks that include those who are collecting data in lower income countries. That means sharing study protocols, models of governance and the tools, technology and analytic skills that turn shared data into better health.

(202 words)

Introduction

As little as a decade ago, many researchers working in global health recoiled at the idea that they should openly share individual patient level data with one another. Now, data sharing is being herded into the mainstream by pioneering researchers, with added pressure from funders, medicine regulatory authorities, public health agencies and medical journals.[1–6] But even those researchers most willing to share data are given very little guidance on how that should happen. This enables the less willing to continue to sing the same anxious songs about data sharing: data sharing will lead to data grabs by rich northern institutions, goes the refrain, while southern researchers will lose control of their data and get little in return. Data sharing might harm patients and communities by breaching confidentiality, some fear. And the infrastructure's not up to it; many feel there's nowhere safe to put shared data.[7]

We have ourselves been among those who have raised these concerns over the years.[8–13] But we are also among those who have been most actively involved in sharing information collected in low and middle income settings, including demographic surveillance data and the records of individual patients in clinical trials. Our discussion refers largely to these types of data, rather than to genomics data, data from well-resourced clinical trials in the global north, or surveillance data from routine, government-led systems. Representing academia, industry, global health organisations, non-governmental organisations, research funders and medical

publishers, we've taken very different approaches and been down some dead-end streets. Between us, we now have enough experience to examine the truth behind the many tropes around data sharing in global health. In a workshop in April 2016 supported by the Geneva Health Forum and two major funders of health research (the Wellcome Trust and the Bill and Melinda Gates Foundation) we came together to discuss the implications of our experience for data sharing in the future. Table 1 lists some of the data sharing models represented and/or discussed in the workshop.

Table 1: Example of data sharing platforms discussed

INDEPTH Network
An investigator-led network of 49 health and demographic surveillance sites in 20 low and middle income countries. Core data from each site are standardised and made available to other researchers through a web-based platform. Based in Accra, Ghana.
http://www.indepth-network.org/

WorldWide Antimalarial Resistance Network
A investigator-led network of 260 collaborators, most performing clinical trials related to malaria drug efficacy and resistance in endemic countries. Data are standardised by platform staff and shared in order to answer specific research questions, with the approval of data contributors. Based at Oxford University, UK.
http://www.wwarn.org/

Clinical Study Data Request
An on-line repository of clinical trial data contributed by 13 major pharmaceutical firms. Data are not standardised; individual study data are made available to researchers on request, after research proposals are approved by an independent data access panel.
https://clinicalstudydatarequest.com/

West Africa Network of Excellence for TB, AIDS and Malaria
A regional collaboration between research institutions that aims to build skills and structures to generate shareable clinical research data through use of common protocols for research, analysis and data management.
Co-ordinated from Dakar, Senegal.
http://orlysoft.com/sites/wanetam/

Yale University Open Data Access
A platform for access to patient-level data from clinical trials, currently mostly industry-sponsored. Platform staff provide some standardisation and curation services. Data are made available to researchers on request, after research proposals are approved by an independent data access panel. Based at Yale University, USA
http://yoda.yale.edu/

Figshare
A repository which allows individual researchers to upload datasets in any format at no charge. Datasets are assigned citable DOIs. Though minimal metadata must be supplied, data are not standardised or quality-assured. Data published on Figshare are reusable by anyone with internet access under Creative Commons CC0 licence. Based in London, UK.
https://figshare.com/

Infectious Diseases Data Observatory
A collection of data sharing platforms focused on emerging and infectious diseases. Centralised data curation and standardisation produce pooled databases from clinical trials, surveillance and/or treatment records. Data are accessible to requestors through an independent data access committee. The expanding portfolio of disease platforms currently includes Ebola, malaria and visceral leishmaniasis. Based at the University of Oxford, UK.
http://www.iddo.org

Getting more health out of the same data

We examined in turn four of the contentions most frequently raised in discussions about sharing data collected in lower-income settings, beginning with the oft-repeated assertion that data sharing really is good for health, that it generates new information that can save lives.[14] We found many examples where this was demonstrably true, where analyses of data pooled from different studies in different locations allowed for new information relevant to appropriate dosing, improved treatment of sub-groups and the development of new therapies.[15–19] We also identified areas where the failure to share data has disrupted efforts to respond rapidly to outbreaks, or foreclosed more detailed evaluation of interventions which may undermine child survival.[20–22] In these cases, not sharing data has been bad for science, and probably also bad for health.

We believe that it is no coincidence that most of the people involved in our experience-based workshop were members of collaborative networks. Our investigations suggest that in lower income settings, such collaborations account for most of the examples in which new knowledge was derived from shared data. These networks are characterised by significant investment in the sometimes difficult work of building trust and relationships between investigators and in developing institutional capacity, as well as in managing and standardising data.[23]

In discussing data sharing policies, we propose classifying shared data as accessible, useable or useful, as shown in Table 2.

Table 2: Costs, benefits and characteristics of different levels of data sharing

	Research Transparency	Potential health benefit	Curation costs	Characteristics
Accessible	✓	?	¢	Online repository
Useable	✓	✓	$	Online repository, extensive, documented metadata, discoverable
Useful	?	✓✓✓	$$$	Curated, standardised, comparable across time/place

Developing and maintaining platforms for 'usefully' shared data tends to be expensive, because data from different sources, often collected in different formats

using different protocols and end-points must be quality controlled and standardised so that analysis can be performed across studies.[24] The up-front costs in developing community standards and networks of collaboration can also be high. However once these investments have been made, the cost in time and effort to potential users is relatively low, and the potential for data to be re-used in ways that benefit public health is high. Currently, most efforts to standardise clinical data in this way occur within consortia or networks of people with similar interests who work together to formulate new questions and to answer them in contextually appropriate ways. Data shared in these networks may thus not always tick the 'transparency' box increasingly required by journals that wish to allow for re-analysis of individual datasets.

'Useable' datasets have in some cases been used for replication of analyses, and their open availability promotes transparency in research. Pharmaceutical firms have recently taken a lead in making data from individual clinical trials available in increasingly usable forms. [25,26] Although it is early days, data requests are on the rise.[27] However the evidence suggests that neither industry data nor unstandardised data made 'available' in academic repositories have yet resulted in the publication of pooled analyses that contribute to public scientific discourse.[28] This is likely because in this case the hard work of harmonising datasets lies with the potential secondary analyst; they may be especially reluctant to invest heavily in data management because secondary analysis is widely perceived to be difficult to publish.

The power of technology

A second contention we examined was that data sharing is hampered by a lack of appropriate platforms. Datasets and even data repositories have multiplied so rapidly and chaotically that one of our group likened them to an asteroid field. We agree with those who maintain that better technology would facilitate data sharing. Common search portals, improved discoverability, and tools to help with reliable anonymisation and the standardisation of heterogeneous data might even begin to reshape the asteroid field into an organised solar system.

Developing that solar system and keeping the planets in orbit will require significant, long-term investment. In recent years, the pharmaceutical industry has expanded efforts in data transparency by way of platforms such as clinicalstudydatarequest.com, and has begun the process of transforming "useable" data into something more "useful" through data standardisation and curation in fields such as oncology. In some cases they are outsourcing this work to academic institutions -- the YODA platform held at Yale is an example. There's scope to expand these public-private partnerships, using fees from well-resourced diseases to subsidise curation of data for conditions with less commercial appeal. Realistically, however, grants from development institutions are likely to remain a key source of funding for data platforms for the most neglected diseases. Currently, few such institutions provide long-term funding for data infrastructure and curation. In addition, the groups most plugged in to those funding sources tend to be academic, and academic researchers may not be best-placed to design or build the data solar system. Initiatives such as the Clinical Data Interchange Standards Consortium are crowd-sourcing metadata standards from scientists, but we probably need to draw on data management expertise from the vast data management industry that

flourishes outside academia to do much of the heavy lifting of platform development most efficiently, not at least in order to reduce unnecessary reinvention and duplication.

Do no harm

The third contention that we examined was that data sharing poses a risk to individuals and communities. Concerns that patient confidentiality and consent may be breached are often cited by researchers as a reason for not sharing data.[13] Several of us have been sharing data for a decade or more, including around illicit behaviours and stigmatised diseases;[29] between us we could find very few examples of harm -- certainly far, far fewer compared with examples of benefits. That is in part because many of us have worked hard to develop strong governance structures. We have also consulted with patients and communities among whom research is conducted about sharing the information they provide to us, because we believe efforts to expand data sharing can only succeed as long as there is a broad social consensus supporting the practice.[30] While governance structures for secondary analysis should be simplified in ways that are proportionate to the often more limited risks of data re-use, they must remain robust. These governance protocols should be shared much more widely as we gain experience in how to maximise useful sharing while minimising risks. Collaboration around governance also reduces the hurdles to contributing data to repositories for pooled analyses.

Equity in research: the threat of data parasites

Finally, we looked at a fourth common trope in discussions of data sharing: the contention that sharing data undermines the career prospects for researchers, especially in low and middle income countries, exposing them to 'research parasites' who will suck their data into the maw of far-off computers and spit out papers in high impact journals.[31] We could find no evidence for this. It's just not that easy to pick poorly-documented data out of scattered repositories and make coherent, publishable sense of it. Where well-documented data are shared more 'usefully' in professional networks, our experience is that sharing data has increased our own work's visibility and expanded our collaborations.[13,32] Clearly, we're hardly an unbiased sample; we're setting out our ideas here precisely because we have found data sharing beneficial. Those of us who work in Africa are all involved in the sorts of investigator-led networks mentioned earlier, in which secondary users work collaboratively with the researchers who are on the front line of data collection to define and answer questions. That's an important start in moving towards a 'fair trade' culture in health research, though it is still only a start. Where authors are named, first and last authors on secondary analysis are still very often northern.

In global health, researchers on the front line are likely to be from low and middle income countries where disease burdens are high. Conducting clinical trials and other health research in those settings is time-consuming, challenging and often financially insecure; it leaves investigators with little time to build up, let alone exercise the skills needed for large-scale secondary analysis of pooled datasets.[8] Data sharing collaborations have the potential to introduce greater equity in global health research, but that will require long-term investments in both skills and career pathways for researchers from high-burden countries. Changing the incentive

system to reward the publication of quality-assured datasets with standardised metadata in the same way that we reward the publication of research papers in high impact journals would also go a long way to damping down the panic about 'data parasites'.

Towards a data sharing solar system

In our experience, sharing data from demographic surveillance and health research, including clinical trial data at the individual patient level, can indeed lead to advances in knowledge that wouldn't have been possible without bringing those data together. To that extent, data sharing is good for health. But knowledge only improves health if it leads to changes in policy and practice; one of the most important determinants of the translation of research results into health policy in low and middle income settings is active collaboration between local researchers and policy makers in shaping research questions and interpreting results.[33]

To our knowledge, most examples of policy change based on analysis of shared data in low and middle income settings involve compendia of datasets that are quality controlled, standardised and otherwise highly curated. In general, the analyses are performed in collaborations between global disease experts and local researchers who know their contexts well and who help formulate questions and answer them. These researchers can also act as a bridge to national policy makers, ultimately delivering changes that benefit the populations from whom data were collected.

This sort of 'sharing' requires far more effort than the dumping of a dataset in an online repository. Useful scientific collaborations are expensive to develop and require a shift in attitudes, incentives and investment patterns. A degree of technical and economic efficiency may have to be sacrificed in the interests of fostering collaboration, for example by investing in building skills in high disease burden countries rather than simply using skills already available in northern universities. The peer-reviewed research results paper must lose its supremacy as the major metric of scientific productivity; and funders must commit to long-term investments in both technical and human infrastructure if they want to promote data sharing that is useful, used and likely to change policies for the greater benefit of patients.

This cannot happen for all diseases or all types of data at once -- it's just too expensive. The alternative is not, however, to downgrade to a 'useable' (but not used) or 'accessible' (and not even useable) model of data sharing. Rather, we must think in fresh ways about how existing structures can be made more useful, in ways that maximise health gains in poorer countries. We need to figure out which platforms and technological structures can be shared across diseases, which diseases would most benefit from the sort of pooled analysis that has proven useful and been used so far. Obvious candidates include infectious diseases in poor regions with only sparse data and small sample sizes, emerging infections about which little is known, and diseases such as TB and malaria that face changes in disease burden and spreading drug resistance. For better and worse, the utility of investing in a platform is also likely to be affected by many other factors, including the potential for data standardisation, the institutional politics in which the disease is embedded, and the degree to which research is financed by public or charitable bodies.

We need to stop thinking of data sharing as an afterword to the scientific enterprise: it is relevant to every stage of the research cycle. Dumping decontextualised results

into a growing asteroid field may tick a transparency box, but it is otherwise wasteful. To be useful in the low and middle income settings which shoulder high burdens of disease and a legacy of under-investment in research infrastructure, data sharing must be treated as an integral part of the larger scientific solar system. We favour sharing data, certainly, but only as one part of a research collaboration which also fairly shares models of governance and the tools, technology and analytic skills that turn shared data into better health.

(2300 words, excluding tables)

Key Messages:

Our experience over a decade sharing research data suggests different models of data sharing are valuable in different ways: simple accessibility of data is enough to promote research transparency, but public health gains require more complex models.

To derive public health benefit from shared data in low and middle income countries, meaningful and equitable collaboration with local researchers and policymakers is needed. Without it, the right questions don't get asked, and research results don't get used.

Useful data sharing requires long term investment in infrastructure, networks and scientific careers, including in the data sciences.

It is not enough to share data: we need to share governance structures, scientific questions and ideas, and interpretation. More sharing of science will lead to more useful sharing of data

Contributors and Sources

This article is the result of an extremely lively workshop held under the auspices of the Geneva Health Forum in April 2016, and supported also by the Wellcome Trust and the Bill and Melinda Gates Foundation. All the authors were invited to participate in the workshop because they have themselves shared health research data, they have funded or supported data sharing, or they actively advocate it through their professional position. All participated actively in the workshop; we determined from the start that we would not revisit theoretical discussions but would focus exclusively on what we could learn from our shared experience. Our own (often unpublished) experience of data sharing thus forms the basis for our assertions and conclusions.

Elizabeth Pisani wrote the first draft of this paper and acts as its guarantor.

Conflict of interest statement

Steven Kern is employed by the Bill and Melinda Gates Foundation, David Carr and Katherine Littler are employed by the Wellcome Trust. Both organisations supported the workshop financially. Vicki Marsh and Dorcas Kamuya work for organisations supported by the Wellcome Trust. Philippe Guerin and Laura Merson are supported by the Gates Foundation and the Wellcome Trust. Elizabeth Pisani

received consultancy fees from Oxford University for her participation. Trish Groves is employed by the British Medical Journal group and is a deputy editor of the BMJ. Vreni Schoenenberger is employed by the International Federation of Pharmaceutical Manufacturers and Associations. No other conflicts of interest have been declared.

References

1. Wellcome Trust. Sharing public health data: a code of conduct [Internet]. 2008. Available from: http://www.wellcome.ac.uk/About-us/Policy/Spotlight-issues/Data-sharing/Public-health-and-epidemiology/index.htm

2. Walport M, Brest P. Sharing research data to improve public health. The Lancet. 2011 Feb 18;377(9765):537–9.

3. Taichman DB, Backus J, Baethge C, et al. Sharing clinical trial data: A proposal from the international committee of medical journal editors. JAMA. 2016 Feb 2;315(5):467–8.

4. European Medicines Agency. European Medicines Agency policy on publication of clinical data for medicinal products for human use. European Medicines Agency; 2015.

5. European Union. Commission Implementing Regulation (EU) 2016/9 of 5 January 2016 on joint submission of data and data-sharing in accordance with Regulation (EC) No 1907/2006 of the European Parliament and of the Council concerning the Registration, Evaluation, Authorisation and Restriction of Chemicals (REACH). 2016/9 Jan 5, 2016.

6. Sharing Clinical Trial Data: Maximizing Benefits, Minimizing Risk. Washington, D.C.: National Academies Press; 2015

7. van Panhuis WG, Paul P, Emerson C, Grefenstette J, Wilder R, Herbst AJ, et al. A systematic review of barriers to data sharing in public health. BMC Public Health. 2014;14:1144.

8. Sankoh O, Ijsselmuiden C. Sharing research data to improve public health: a perspective from the global south. Lancet Lond Engl. 2011 Jul 30;378(9789):401–2.

9. Pisani E, Whitworth J, Zaba B, Abou-Zahr C. Time for fair trade in research data. The Lancet. 2010 Mar;375(9716):703–5.

10. Pisani E. Sharing health data: good intentions are not enough. Bull World Health Organ. 2010;88(6):401–80.

11. Merson L, Gaye O, Guerin PJ. Avoiding Data Dumpsters - Toward Equitable and Useful Data Sharing. N Engl J Med. 2016 May 11;

12. Jao I, Kombe F, Mwalukore S, Bull S, Parker M, Kamuya D, et al. Involving Research Stakeholders in Developing Policy on Sharing Public Health Research Data in Kenya: Views on Fair Process for Informed Consent, Access Oversight, and Community Engagement. J Empir Res Hum Res Ethics JERHRE. 2015 Jul;10(3):264–77.

13. Jao I, Kombe F, Mwalukore S, Bull S, Parker M, Kamuya D, et al. Research Stakeholders' Views on Benefits and Challenges for Public Health Research Data Sharing in Kenya: The Importance of Trust and Social Relations. PloS One. 2015;10(9):e0135545.

14. The Chatham House data sharing advisory group. Public Health Surveillance: A Call to Share Data. International Association of National Public Health Institutes; 2016 [cited 2016 May 30]. Available from: http://ianphi.org/news/2016/datasharing1.html

15. Ekouevi DK, Balestre E, Coffie PA, Minta D, Messou E, Sawadogo A, et al. Characteristics of HIV-2 and HIV-1/HIV-2 Dually Seropositive Adults in West Africa Presenting for Care and Antiretroviral Therapy: The IeDEA-West Africa HIV-2 Cohort Study. PloS One. 2013;8(6):e66135.

16. Arthur SS, Nyide B, Soura AB, Kahn K, Weston M, Sankoh O. Tackling malnutrition: a systematic review of 15-year research evidence from INDEPTH health and demographic surveillance systems. Glob Health Action. 2015;8:28298.

17. Poespoprodjo JR, Fobia W, Kenangalem E, Lampah DA, Sugiarto P, Tjitra E, et al. Treatment policy change to dihydroartemisinin–piperaquine contributes to the reduction of adverse maternal and pregnancy outcomes. Malar J. 2015 Dec;14(1).

18. The WorldWide Antimalarial Resistance Network (WWARN) DP Study Group. The Effect of Dosing Regimens on the Antimalarial Efficacy of Dihydroartemisinin-Piperaquine: A Pooled Analysis of Individual Patient Data. Garner P, editor. PLoS Med. 2013 Dec 3;10(12):e1001564.

19. Diouara AAM, Ndiaye HD, Guindo I, Bangoura N, Cissé M, Edmond T, et al. Antiretroviral treatment outcome in HIV-1-infected patients routinely followed up in capital cities and remote areas of Senegal, Mali and Guinea-Conakry. J Int AIDS Soc. 2014;17:19315.

20. Aaby P, Benn CS. Should we introduce a malaria vaccine which may increase child mortality? Reader response. PLOS Med. 2014 Jul 29;11(7):e1001685.

21. Klein SL, Shann F, Moss WJ, Benn CS, Aaby P. RTS, S Malaria Vaccine and Increased Mortality in Girls. mBio. 2016;7(2):e00514–6.

22. Yozwiak NL, Schaffner SF, Sabeti PC. Data sharing: Make outbreak research open access. Nature. 2015;518(7540):477.

23. Jumbe NL, Murray JC, Kern S. Data Sharing and Inductive Learning—Toward Healthy Birth, Growth, and Development. N Engl J Med. 2016

24. Haug CJ. From Patient to Patient — Sharing the Data from Clinical Trials. N Engl J Med. 2016 Jun 23;374(25):2409–11.

25. Krumholz HM, Gross CP, Blount KL, Ritchie JD, Hodshon B, Lehman R, et al. Sea Change in Open Science and Data Sharing: Leadership by Industry. Circ Cardiovasc Qual Outcomes. 2014 Jul 1;7(4):499–504.

26. PhRMA. Principles for responsible clinical trial data sharing. PhRMA; 2013 [cited 2016 Jun 17]. Available from: http://www.phrma.org/sites/default/files/pdf/PhRMAPrinciplesForResponsibleClinicalTrialDataSharing.pdf

27. Clinical Study Data Request. Metrics. Clinical Study Data Request. [cited 2016 Jun 17]. Available from: https://clinicalstudydatarequest.com/Metrics.aspx

28. Navar A, Pencina MJ, Rymer JA, Louzao DM, Peterson ED. Use of open access platforms for clinical trial data. JAMA. 2016 Mar 22;315(12):1283–4.

29. Republic of Indonesia M of H, Statistics Indonesia, Family Health International. Behavioural and drug-taking risk behaviour among female sex workers and men in mobile occupations in Indonesia, 2002-2004 [Internet]. Harvard Dataverse; 2010. Available from: http://hdl.handle.net/1902.1/15047

30. Carter P, Laurie GT, Dixon-Woods M. The social licence for research: why care.data ran into trouble. J Med Ethics. 2015 Jan 23;medethics – 2014–102374.

31. Longo DL, Drazen JM. Data Sharing. N Engl J Med. 2016 Jan 21;374(3):276–7.

32. Binka F. Editorial: North–South research collaborations: a move towards a true partnership? Trop Med Int Health. 2005 Mar 1;10(3):207–9.

33. Kok MO, Gyapong JO, Wolffers I, Ofori-Adjei D, Ruitenberg J. Which health research gets used and why? An empirical analysis of 30 cases. Health Res Policy Syst. 2016;14:36.