BMJ 2016;355:i5295 doi: 10.1136/bmj.i5295 (Published 10 October 2016)

ANALYSIS

Beyond open data: realising the health benefits of sharing data

Accessible data are not enough. We need to invest in systems that make the information useful, say **Elizabeth Pisani and colleagues**

Elizabeth Pisani *visiting senior research fellow*[1], Peter Aaby *professor*[2], J Gabrielle Breugelmans *networking manager*[3], David Carr *programme manager*[4], Trish Groves *director of academic outreach*[5], Michelle Helinski *project officer*[3], Dorcas Kamuya *researcher*[6], Steven Kern *deputy director*[7] *quantitative sciences*[7], Katherine Littler *senior policy adviser*[4], Vicki Marsh *associate professor*[6], Souleymane Mboup *professor*[8], Laura Merson *researcher*[9], Osman Sankoh *executive director*[10], Micaela Serafini *medical director*[11], Martin Schneider *PhD candidate*[12], Vreni Schoenenberger *manager, policy, ethics and compliance*[13], Philippe J Guerin *director*[9]

[1]Policy Institute, King's College London, London, UK; [2]Bandim Health Project, Guinea-Bissau; [3]European and Developing Countries Clinical Trials Partnership, The Hague, Netherlands; [4]Wellcome Trust, London, UK ; [5]BMJ, London, UK; [6]KEMRI Wellcome Trust Research Programme, Kilifi, Kenya; [7]Bill and Melinda Gates Foundation, Seattle, Washington, USA; [8]Universite Cheikh Anta Diop, Dakar, Senegal; [9]Infectious Diseases Data Observatory, University of Oxford, Oxford, UK ; [10]INDEPTH Network, Accra, Ghana; [11]Médecins Sans Frontières, Geneva, Switzerland; [12]Institute of Global Health, University of Geneva, Geneva; [13]International Federation of Pharmaceutical Manufacturers and Associations, Geneva;
Correspondence to: E Pisani pisani@ternyata.org

As little as a decade ago, many researchers working in global health recoiled at the idea that they should openly share individual patient 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 little guidance on how that should happen, and the practice is still unusual, especially in low and middle income countries.

Concerns continue to be raised that data sharing will lead to data being analysed by rich institutions in industrialised countries while researchers in poorer countries with the highest burdens of infectious disease will lose control of their data and get little in return. Some fear that data sharing might harm patients and communities by breaching confidentiality, that the infrastructure is not up to it, and there is nowhere safe to put shared data.[7]

Our group includes researchers working for academic and humanitarian organisations, as well as public, charitable, and industry funders of data sharing efforts. Although we have raised concerns in the past,[8-13] we are now involved in sharing information collected in low and middle income settings, including demographic surveillance data and the records of individual patients in clinical trials. We examine the extent to which the fears about data sharing have been realised in our work and what is needed to get the most value out of shared data.

Getting more health out of the same data

Data sharing is often asserted to be good for health and to generate new information that can save lives.[14] We found many examples where this was demonstrably true, with analyses of data pooled from different studies in different locations providing new information relevant to appropriate dosing, improved treatment of subgroups, and the development of new treatments.[15-18]

Box 1 lists some of the better known data sharing models. One example is a meta-analysis of individual patient data from a large and diverse population of patients shared through the WorldWide Antimalarial Resistance Network (WWARN). This provided the power to determine the efficacy of antimalarial drug dihydroartemisinin-piperaquine across a wide range of age groups and settings.[19] The meta-analysis revealed that treatment failure associated with a lower dose of piperaquine was particularly dangerous in young children, suggesting potential for further dose optimisation. The results contributed to a revision of the World Health Organization's guidelines for treating malaria.[20]

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 that may be harmful.[21 22] In these cases, not sharing data has been bad for science and almost certainly bad for health. In the 2014 Ebola outbreak in west Africa some researchers made genomic data immediately available for further study, confirming that the virus had spread from Guinea to Sierra Leone, that it was

Box 1: Examples of data sharing platforms

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—An 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 online repository of clinical trial data contributed to by 13 major drug companies. 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. Coordinated 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 that allows individual researchers to upload datasets in any format at no charge. Datasets are assigned a citable doi. Though minimal metadata must be supplied, data are not standardised or quality assured. Data published on Figshare are reuseable 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 (https://www.iddo.org/)

sustained by human-to-human transmission, and that it was mutating rapidly in certain areas. However, the researchers subsequently reported that "What followed was three months of stasis, during which no new virus sequence information was made public [even though] thousands of samples were transferred to researchers' freezers across the world."[23] They called for greater data sharing through collaborative networks.

Our investigations suggest that in lower income settings such networks account for most of the examples in which new knowledge was derived from shared data. These networks are characterised by substantial 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.[24]

In discussing data sharing policies, we propose classifying shared data as accessible, useable, or useful, as shown in table 1⇓.

Developing and maintaining curated platforms for "useful" sharing of data tends to be expensive. Data from different sources, often collected in different formats using different protocols and endpoints, must be quality controlled and standardised so that analysis can be performed across studies.[25] The upfront costs of developing community standards and networks of collaboration can be high. However, once these investments have been made, the time and effort required by potential users is relatively low, and the potential for data to be reused in ways that benefit public health is high, making the investments cost effective.

Currently, most efforts to standardise clinical data in this way occur within consortiums 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 meet the transparency criteria increasingly required by journals to allow for independent reanalysis of individual datasets.

Replicate analyses have been done with useable datasets, and their open availability promotes transparency in research. Drug companies have recently taken a lead in making data from individual clinical trials available in increasingly useable forms.[26 27] The first evaluation of prominent platforms for sharing clinical trial data found that, although individual patient data from more than 3000 trials had been made available to investigators over the past two years, only 15.5% of the trial datasets had ever been requested.[28] Most proposals focused on subgroups or predictors of response not prespecified in the original analysis rather than validation of study results, and only one of the proposals examined had led to a published pooled analysis and contributed to public scientific discourse.[29] This is probably because the hard work of harmonising datasets lies with the secondary analyst, who may be reluctant to invest heavily in data management because secondary analysis is widely perceived to be difficult to publish. These repositories are only recently established, however, and data requests are on the rise.[30]

Power of technology

Datasets and even data repositories have multiplied so rapidly and chaotically that one of our group likened them to an asteroid field. Better technology and metadata standards—especially common search portals, improved discoverability, and tools for reliable anonymisation and standardisation of heterogeneous data—could begin to reshape the asteroid field into an organised solar system.

Developing that solar system and keeping the planets in orbit will require substantial long term investment. In recent years, the pharmaceutical industry has expanded efforts in data transparency through 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 it is outsourcing this work to academic institutions—for example, the YODA platform held at Yale. There is 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 neglected diseases. Currently, few such institutions provide long term funding for data infrastructure and curation. In addition, the groups best connected 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 crowdsourcing metadata standards from scientists, but we need to draw on data management expertise from the vast data industry outside academia to develop data sharing platforms most efficiently, not least in order to reduce unnecessary reinvention and duplication.

Do no harm

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.[31] Between us we could find few examples of harm—certainly far fewer than examples of benefits—partly because we have worked hard to develop strong governance structures. We have also consulted with patients and communities about sharing the information they provide to us, because we believe that efforts to expand data sharing can succeed only with broad social support.[32] While governance structures for secondary analysis should be simplified so that they are proportionate to the often more limited risks of data reuse, they must remain robust. These governance protocols should be shared 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

A common generalisation in discussions of data sharing is that it undermines the career prospects for researchers, especially in low and middle income countries, exposing them to "research parasites" who will ingest their data into far-off computers and beget papers for high impact journals.[33 34] We could find no evidence for this. It is difficult to pick poorly documented data out of scattered repositories and make coherent, publishable sense of it. When well documented data are shared usefully in professional networks, our experience is that sharing has increased our work's visibility and expanded our collaborations.[13 35] Investigator led networks in which secondary users work collaboratively with the researchers collecting the data to define and answer questions are an important start in moving towards a "fair trade" culture in health research, though it is still only a start. In journal publications of secondary analyses, first and last authors are still most often from wealthier countries.

Conducting clinical trials and other health research in low and middle income countries 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 countries with high disease burden. 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 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 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 improves health only 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 collaboration between local researchers and policy makers in shaping research questions and interpreting results.[36]

Most examples of policy change based on analysis of shared data in low and middle income settings involve compendiums of datasets that are quality controlled, standardised, and otherwise highly curated.[15-20] 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 which data were collected.

This sort of sharing requires far more effort than simply uploading a dataset to 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 and equity—for example, by investing in building skills in high disease burden countries rather than simply using skills already available in universities in industrialised countries. 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 is just too expensive. The alternative is not, however, to downgrade to a useable (but not used) or accessible (and not useable) model of data sharing. Rather, we must think in fresh ways about how existing structures can be made more useful to maximise health gains. We need to figure out which platforms and technological structures can be shared across diseases and which diseases would most benefit from the sort of pooled analysis that has already proved useful. Obvious candidates include neglected tropical diseases and other infectious diseases in poor regions with only sparse data and small sample sizes; emerging infections about which little is known; and diseases such as tuberculosis and malaria that face changes in disease burden and spreading drug resistance. The value 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. Depositing 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 that also fairly shares models of governance and the tools, technology, and analytical skills that turn shared data into better health.

Contributors and sources: The authors of this paper all participated in a workshop held under the auspices of the Geneva Health Forum in April 2016, supported by the Wellcome Trust and the Bill and Melinda Gates Foundation. All the authors were invited to participate in the discussions because they have shared health research data, funded or supported data sharing, or advocate it through their professional position. EP wrote the first draft and is the guarantor.

Key messages

- Simple accessibility of data is enough to promote research transparency, but public health gains require more complex models
- Meaningful and equitable collaboration with local researchers and policy makers in low and middle income countries is needed to ensure the right research questions get asked and research results are 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

Competing interests: We have read and understood BMJ policy on declaration of interests and declare the following interests: SK is employed by the Bill and Melinda Gates Foundation, DC and KL are employed by the Wellcome Trust. Both organisations supported the workshop financially. VM and DK work for organisations supported by the Wellcome Trust. PG and LM are supported by the Gates Foundation and the Wellcome Trust. EP received consultancy fees from Oxford University for her participation. TG is employed by BMJ and is a deputy editor of *The BMJ*. VS is employed by the International Federation of Pharmaceutical Manufacturers and Associations.

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

1 Wellcome Trust. Sharing public health data: a code of conduct. 2008. 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. *Lancet* 2011;377:537-9. doi:10.1016/S0140-6736(10)62234-9 pmid:21216456.
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;315:467-8. doi:10.1001/jama.2015.18164 pmid:26792562.
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.
6 Sharing Clinical Trial Data. *Maximizing benefits, minimizing risk*. National Academies Press, 2015.
7 van Panhuis WG, Paul P, Emerson C, et al. A systematic review of barriers to data sharing in public health. *BMC Public Health* 2014;14:1144. doi:10.1186/1471-2458-14-1144 pmid:25377061.
8 Sankoh O, Ijsselmuiden C. Sharing research data to improve public health: a perspective from the global south. *Lancet* 2011;378:401-2. doi:10.1016/S0140-6736(11)61211-7 pmid:21803205.
9 Pisani E, Whitworth J, Zaba B, Abou-Zahr C. Time for fair trade in research data. *Lancet* 2010;375:703-5. doi:10.1016/S0140-6736(09)61486-0 pmid:19913902.
10 Pisani E, AbouZahr C. Sharing health data: good intentions are not enough. *Bull World Health Organ* 2010;88:462-6. doi:10.2471/BLT.09.074393 pmid:20539861.
11 Merson L, Gaye O, Guerin PJ. Avoiding data dumpsters—toward equitable and useful data sharing. *N Engl J Med* 2016;374:2414-5. doi:10.1056/NEJMp1605148 pmid:27168351.
12 Jao I, Kombe F, Mwalukore S, 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* 2015;10:264-77. doi:10.1177/1556264615592385 pmid:26297748.
13 Jao I, Kombe F, Mwalukore S, 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:e0135545. doi:10.1371/journal.pone.0135545 pmid:26331716.
14 Chatham House Data Sharing Advisory Group. Public health surveillance: a call to share data. International Association of National Public Health Institutes, 2016
15 Ekouevi DK, Balestre E, Coffie PA, et al. IeDEA West Africa Collaboration. 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:e66135. doi:10.1371/journal.pone.0066135 pmid:23824279.
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. doi:10.3402/gha.v8.28298 pmid:26519130.
17 Poespoprodjo JR, Fobia W, Kenangalem E, et al. Treatment policy change to dihydroartemisinin-piperaquine contributes to the reduction of adverse maternal and pregnancy outcomes. *Malar J* 2015;14:272. doi:10.1186/s12936-015-0794-0 pmid:26169783.
18 Diouara AAM, Ndiaye HD, Guindo I, 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.pmid:25527333.
19 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. *PLoS Med* 2013;10:e1001564, discussion e1001564. doi:10.1371/journal.pmed.1001564 pmid:24311989.
20 World Health Organisation. *Guidelines for the treatment of malaria.* 3rd ed. WHO, 2015.
21 Aaby P, Benn CS. Should we introduce a malaria vaccine which may increase child mortality? Reader response. *PLoS Med* 2014;11:e1001685.pmid:25072396.
22 Klein SL, Shann F, Moss WJ, Benn CS, Aaby P. RTS,S malaria vaccine and increased mortality in girls. *MBio* 2016;7:e00514-6. doi:10.1128/mBio.00514-16 pmid:27118593.
23 Yozwiak NL, Schaffner SF, Sabeti PC. Data sharing: make outbreak research open access. *Nature* 2015;518:477-9. doi:10.1038/518477a pmid:25719649.
24 Jumbe NL, Murray JC, Kern S. Data sharing and inductive learning—toward healthy birth, growth, and development. *N Engl J Med* 2016;374:2415-7. doi:10.1056/NEJMp1605441 pmid:27168111.
25 Haug CJ. From patient to patient—sharing the data from clinical trials. *N Engl J Med* 2016;374:2409-11. doi:10.1056/NEJMp1605378 pmid:27168009.
26 Krumholz HM, Gross CP, Blount KL, et al. Sea change in open science and data sharing: leadership by industry. *Circ Cardiovasc Qual Outcomes* 2014;7:499-504. doi:10.1161/CIRCOUTCOMES.114.001166 pmid:24891590.
27 PhRMA. Principles for responsible clinical trial data sharing. 2013. http://www.phrma.org/sites/default/files/pdf/PhRMAPrinciplesForResponsibleClinicalTrialDataSharing.pdf
28 Navar AM, Pencina MJ, Rymer JA, Louzao DM, Peterson ED. Use of open access platforms for clinical trial data. *JAMA* 2016;315:1283-4. doi:10.1001/jama.2016.2374 pmid:27002452.
29 Le Noury J, Nardo JM, Healy D, et al. Restoring Study 329: efficacy and harms of paroxetine and imipramine in treatment of major depression in adolescence. *BMJ* 2015;351:h4320. doi:10.1136/bmj.h4320 pmid:26376805.
30 Clinical Study Data Request. Metrics. 2016. https://clinicalstudydatarequest.com/Metrics.aspx
31 Republic of Indonesia Ministry of Health, Statistics Indonesia, Family Health International. Behavioural and drug-taking risk behaviour among female sex workers and men in mobile occupations in Indonesia, 2002-2004. Harvard Dataverse 2010. http://hdl.handle.net/1902.1/15047
32 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 pmid:25617016.
33 Longo DL, Drazen JM. Data sharing. *N Engl J Med* 2016;374:276-7. doi:10.1056/NEJMe1516564 pmid:26789876.
34 Devereaux PJ, Guyatt G, Gerstein H, Connolly S, Yusuf S. International Consortium of Investigators for Fairness in Trial Data Sharing. Toward fairness in data sharing. *N Engl J Med* 2016;375:405-7. doi:10.1056/NEJMp1605654 pmid:27518658.
35 Binka F. Editorial: north-south research collaborations: a move towards a true partnership? *Trop Med Int Health* 2005;10:207-9. doi:10.1111/j.1365-3156.2004.01373.x pmid:15730502.
36 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. doi:10.1186/s12961-016-0107-2 pmid:27188305.

Accepted: 28 09 2016

Table

Table 1| Benefits and costs of different levels of data sharing

	Research becomes transparent	Potential health benefit	Upfront curation costs
Accessible—online repository	Yes	Uncertain	Cheap
Useable—repository with discoverable, well documented metadata	Yes	Possible with extensive user effort	Moderate
Useful—data are curated, standardised, and comparable across time and place	Sometimes	Great	Expensive