Sharing health data: good intentions are not enough

Elizabeth Pisani[a] & Carla AbouZahr[b]

Abstract Epidemiologists and public health researchers are moving very slowly in the data sharing revolution, and agencies that maintain global health databases are reluctant to share data too. Once investments in infrastructure have been made, recycling and combining data provide access to maximum knowledge for minimal additional cost. By refusing to share data, researchers are slowing progress towards reducing illness and death and are denying a public good to taxpayers who support most of the research.

Funders of public health research are beginning to call for change and developing data sharing policies. However they are not yet adequately addressing the obstacles that underpin the failure to share data. These include professional structures that reward publication of analysis but not of data, and funding streams and career paths that continue to undervalue critical data management work. Practical issues need to be sorted out too: how and where should data be stored for the long term, who will control access, and who will pay for those services? Existing metadata standards need to be extended to cope with health data.

These obstacles have been known for some time; most can be overcome in the field of public health just as they have been overcome in other fields. However no institution has taken the lead in defining a work plan and carving up the tasks and the bill. In this round table paper, we suggest goals for data sharing and a work plan for reaching them, and challenge respondents to move beyond well intentioned but largely aspirational data sharing plans.

Une traduction en français de ce résumé figure à la fin de l'article. *Al final del artículo se facilita una traducción al español.* الترجمة العربية لهذه الخلاصة في نهاية النص الكامل لهذه المقالة.

Introduction

As they prepare for careers in science, today's students doubtless hear the same clichés as we did a generation ago: science advances collaboratively; we reproduce and extend the work of others; we stand on the shoulders of giants. In some fields, such as genomics, these axioms are becoming true. In epidemiology and public health, however, data sharing and collaboration remain more aspirational than real.

Students embark on a career in health research in the spirit of sharing; they want to help improve the well-being of others. For all the talk of collaboration, they will enter a world in which another axiom dominates: "publish or perish". That system puts the interests of public health researchers in direct conflict with the interests of public health.

Benefits of sharing

The situation was not so different in genomics less than 15 years ago. Then, after years of hoarding their findings in individual laboratories and progressing at an expensive snail's pace, in 1996 researchers agreed to share all their data openly.[1] Now laboratories sequence during the day and post their results that same night; other researchers can begin to stand on their shoulders the very next day. As a result, genetic research is advancing faster than any other area of biomedicine.[2]

Genomics has taught us that sharing data with other scientists is a way to add value without costing a lot. It allows the same data to be used to answer new questions that may be relevant far beyond the original study. And it allows for meta-analyses that are free from the distortions introduced when only summary results are available.[3,4] We could get far more out of public health research if we followed a similar path, if we squeezed more scientific and policy insights out of data that have already been collected.

Routine health and service use statistics can be just as useful for policy analysis as research data. Many countries are reluctant to release detailed service use data because analysis by disinterested outsiders may contradict politically acceptable interpretations. Most countries do, however, contribute aggregate statistics freely to large international databases maintained by multilateral organizations, although they are not always granted free access to those databases when they want to use them. Such restrictions on access, imposed unnecessarily by agencies wanting to protect their institutional mandates, cripple the potential utility of these expensive resources. Researchers and governments are also reluctant to see the data they provide used and manipulated by others in ways they don't understand because secondary users (including international agencies) do not always publish their methods.

Research data are desperately underused too, in part because of a critical shortage of competent data managers.[5] In other fields – genetics, banking and retailing – data management is a valuable skill. People are trained and develop careers in the field. In public health research, data management is the poor cousin of analysis. Undervalued and underfunded, inadequate data management undermines the rest of the scientific enterprise. One review in the United Kingdom of Great Britain and Northern Ireland found that many of the variables collected in epidemiological studies were never cleaned and coded, so they could not be used even by the primary researchers, let alone shared.[6] In complex population-based surveys in developing countries, data management and analysis skills are in even shorter supply, so a higher proportion of data probably goes to waste.[7]

When we're dealing with public health research, wasted data can translate into shorter, less healthy lives. Improving data management so that data can be shared is a first step to reducing that waste. But it will not be enough. We need to change the in-

[a] London School of Hygiene and Tropical Medicine, Keppel Street, WC1E 7HT, London, England.
[b] Department of Health Statistics and Informatics, World Health Organization, Geneva, Switzerland.
Correspondence to Elizabeth Pisani (e-mail: pisani@ternyata.org).
(*Submitted: 18 November 2009 – Revised version received: 5 January 2009 – Accepted: 7 January 2010*)

centives that pit the interests of individual researchers against the interests of public health, that pit institutional interests against the more rapid advancement of knowledge and understanding. Governments may hold micro-data back from international organizations, but there's no excuse for international organizations to limit access to the aggregate data that governments do provide.

It's easier to understand why individual researchers are reluctant to share data they have collected. That reluctance will certainly remain entrenched as long as their employers – research councils, foundations and universities – regard publication of research papers in peer-reviewed biomedical journals as the main yardstick of success.[8] If, however, "publish [papers] or perish" were to be replaced by "publish [data] or perish", the picture might change rapidly, as it did in genomics.

What did that experience teach us? That a change in the culture of science requires the buy-in of key research teams, yes, but that it also requires considerable and very concrete commitments from funders. The two largest funders of the Human Genome Project, the Wellcome Trust and the National Institutes for Health, invested massively in the infrastructure needed to share data on a large scale for the long term. They also changed funding mechanisms to emphasize team work and the value of roles such as data management, rather than just looking at publication and citation records. Inevitably the rapid change of culture raised some tensions, but those have now largely been resolved.[2] It would be perfectly feasible for research funders to take similar steps in other fields so that personal and professional incentives are aligned rather than in conflict.

Genomics and the social sciences (which have a dramatically better record of sharing data than most biomedical sciences) have developed techniques to deal with two of the other main obstacles to sharing of public health research data – confidentiality and consent. In part because of the development of research tissue banks (biobanks), broad consent procedures are increasingly becoming a norm.[9] Anonymization removes some of the obstacles associated with consent, and techniques for protecting identities are improving constantly. Despite concerns about the theoretical possibility of identifying individuals in shared data sets, no breaches of confidentiality have yet been recorded in anonymized data sets.[10] Social and economic sciences have also gone further in making the sharing of data sets easy through standard metadata, both for aggregate data through Statistical Data and Metadata Exchange (SDMX) standards and for individual data using Data Documentation Initiative (DDI 3.0) standards. A further lesson from other fields: it is possible to make data widely available to the research community while still safeguarding integrity, through the use of standardized data use agreements and licences.[11,12] These define who may use data and how, and may require secondary analysts to contribute both derived data and a record of their analytic methods back to the database, so that primary and other users can both verify and benefit from their work.

The data that we collect and don't make full use of do not come free. The collection of routine health statistics is paid for by our tax money. Most research aiming to reduce ill-health in the developing world is also funded either from the public purse or by charitable foundations. It is irrational to invest so much in collecting data and yet so little in ensuring that we make the best use of it.[13] It is also ethically unsound; people who participate in research have a right to expect that the results will be used to improve life for them and/or for their communities.

Funders and standard-setters have been aware of this for some time. Gradually, they are urging or adopting policies that aim to increase the use and recycling of data. Although they don't all yet practice what they preach, several international organizations, including the Organisation for Economic Co-operation and Development and the World Health Organization, have issued statements calling for increased access to routine statistics and other publicly-funded data.[14,15] Many biomedical journals have recently addressed the importance of data sharing in editorials and commentary articles.[16–18] A few biomedical journals expect researchers to make the data that underlie research articles available to others on request. An even smaller number of journals have followed the lead of *Annals of Internal Medicine* and now require authors to state whether and how they will make protocols, analysis tools and data available to others. But even *Annals* stops short of requiring authors to publish data sets along with their articles. "If we did that, we'd have a very thin journal," commented editor Christine Laine at a recent conference on biomedical publication.[19]

There are indications that public and foundation funders of public health research wish to strengthen data sharing policies, shepherding epidemiologists down the road already travelled by geneticists.[20–23] Many field researchers who have battled difficult climates, erratic electricity supplies, fuel shortages and recalcitrant local authorities will doubtless resent increasing pressure to "give data away". Some are also apprehensive that people looking at the data in the comfort of some distant, well resourced office will spot the errors that are the inevitable by-product of research in the real world.

Governments are equally reluctant to expose their data to interpretations other than those published by their official statisticians. There is a fear, too, that data may be used by others not just for professional but for economic gain. This is sometimes cast as a "north–south" divide; one spectre raised is of pharmaceutical companies exploiting data from developing countries to develop products that those countries then can't afford.[24]

Feelings of ownership over hard-won data, viscerally held even by researchers who support the idea of data sharing in principle, are understandable. And peer reviewers, mostly researchers themselves, are reluctant to approve funding for data management if it cuts into budgets for data collection. But funders of science are themselves under pressure to get the most out of expensive research studies. They have to wrestle with two important questions: how much data sharing is desirable and how much is feasible?

Researchers sometimes argue that interpretation of their data is so dependent on understanding local conditions that the data would be worthless to other scientists. This is often a reflection of inadequate documentation, but also a necessary failure of imagination. Sailors keeping log books on whaling boats in the 1600s could not have predicted that, centuries later, the data would be an important source of information for climate change scientists.[25] Most funders have stringent peer-review procedures; few invest in research that they believe is of only very localized importance, and few wish to support research that produces data of such poor quality that it has no further value. Publicly-funded data can also be invaluable to students learning data management and analysis skills. It thus seems

fair to expect that almost all public health research funded by taxpayers or charities might be useful to secondary analysts. If a piece of research is considered worthy of publication in a peer-reviewed journal, the underlying data should also be worth publishing.

How feasible would it be to make these data available to the scientific community? Technically, the challenges are not trivial, but they have been overcome in several other fields; they are broken down here into manageable parts. We maintain that the major constraints to feasibility are a cultural resistance to change from within our own scientific community, and a reluctance of any institution to take leadership of the data sharing agenda. We also believe, however, that the imperative to share data will only grow stronger. The research community should look at this pressure from funders as an opportunity rather than an imposition.

Goals for funders and researchers

Here we propose several goals to which funders and researchers can jointly aspire and towards which progress can be measured: (i) all data of potential public health importance funded by taxpayers or foundations will be appropriately documented and archived in formats accessible to the wider scientific community; (ii) all data provided by governments to databases developed by publicly-funded organizations will be freely available to any user, at the level of detail at which it was provided; (iii) the publication of a research article in a biomedical journal will be accompanied by the publication of the data set upon which the analysis is based; (iv) funders and employers of researchers will consider publication of well managed data sets as an important indicator of success in research, and will reward researchers professionally for sharing data; and (v) all planned research will budget and be funded to manage data professionally to a quality adequate for archiving and sharing.

Plan of work

These goals can only be achieved with considerable investment in several practical areas. We propose the following plan of work, necessary to underpin progress towards our stated goals.

Fill the gaps in data management

There is a need to develop metadata standards, which will lead to improved documentation and allow data to be combined more easily across time, locations and sources. This will probably require the extension of DDI and SDMX standards to encompass areas of public health interest. Agreement is also required on standards for anonymization and safeguarding of confidentiality.

We need to develop a search portal that will allow data to be discovered across a range of repositories, and standards for repositories similar to those used for registries of clinical trials.[26] We also need to invest in training in data management for public health, especially in developing countries, and the development of career paths in bioinformatics.

Increase incentives to share data

We need to further develop and adopt reliable citation standards for data sets, such as those proposed by DataCite collaboration,[27] and ensure they are indexed in databases such as PubMed. Standards and procedures for peer review or quality control of data sets are also needed. Digital fingerprinting of data would allow tracing of secondary use [28,29] and we should develop methods and measures to track the value that sharing data adds to the work of both primary researchers and funders of research. There is a need to agree on norms and standards governing fair use periods for primary researchers, data access policies and data use agreements.

Data libraries

To underpin the long-term viability of data libraries, we need to invest in expanding existing infrastructure to cover curation and access of data of public health importance. This calls for a business or funding model that assures the long-term viability of data archives.

Conclusion

All of these areas have already been identified as critical to promoting data sharing, often repeatedly so.[5,30–32] Funders, governments, publishers and many researchers want these things to happen, it seems. Some of the organizations calling for greater sharing of public health research data have expressed willingness to pay for parts of the work. But none are willing to take charge of the agenda, committing themselves to orchestrating the dull, messy but essential work of developing the norms and standards that will allow data sharing to revolutionize public health research.

It is time to move beyond expressions of good intentions and to get on with the practical work that will allow data to be shared. The first thing that is needed is leadership. We challenge other participants in this round table to commit to coordinating, funding or carrying out the work described in this paper. Only after someone takes the lead in tackling these issues will today's students of public health be able to climb onto the shoulders of the current giants in our field. ■

Competing interests: None declared.

ملخص

تبادل البيانات الصحية: النوايا الحسنة وحدها لا تكفي

إن تحرك علماء الوبائيات والباحثين في الصحة العمومية مازال بطيئاً في مجال ثورة تبادل البيانات، ومازالت أيضاً الوكالات المعنية بحفظ قواعد البيانات الإحصائية مترددة في تبادل البيانات. فحالما يكتمل الاستثمار في البنية الأساسية، سيتيح إعادة تدوير ودمج البيانات الوصول إلى أكبر قدر من المعارف بأقل قدر من التكلفة الإضافية. ويؤدي رفض تبادل البيانات إلى إبطاء التقدم المحرز نحو الحد من المراضة والوفيات ويمنع تحقيق النفع العام لدافعي الضرائب الداعمين لغالبية هذه البحوث.

وقد بدأ المانحون المطالبة بالتغيير وإعداد سياسات لتبادل البيانات. إلا أنهم حتى الآن لم يتصدوا على النحو الكافي للعقبات المؤدية لفشل تبادل البيانات. والتي تتضمن النظام المهني الذي يكافئ نشر التحليلات بدلاً من البيانات، ومسارات التمويل والتطور الوظيفي التي تواظب على التقليل من قيمة العمل الخاص بإدارة البيانات الهامة. كما أن المواضيع العملية ينبغي تخزينها أيضاً: ولكن كيف وأين يجب تخزينها لأمد طويل، ومن سيتحكم في الوصول إليها، ومن سيدفع نفقات هذه الخدمات؟ ومن الضروري توسيع

نطاق المعايير الحالية الخاصة بالبيانات الممكن الوصول إليها حتى تتواءم مع البيانات الصحية.

إن هذه العقبات معروفة منذ زمن؛ وأغلبها يمكن التغلب عليه في مجال الصحة العمومية كما أمكن التغلب عليها في مجالات أخرى. ولكن لم تتبوأ أي مؤسسة مكان الصدارة في تحديد خطة العمل ولم تحدد المهام وقائمة التكاليف. ومن خلال هذه المائدة المستديرة، سنقترح المرامي المنشودة من تبادل البيانات وخطة العمل لبلوغ هذه المرامي، والتصدي لهذا التحدي للوصول لما هو أبعد من خطط تبادل البيانات ذات المقاصد الجيدة والمثيرة للحماس البالغ.

Résumé

Partage des données sur la santé : les bonnes intentions ne suffisent pas

Les épidémiologistes et les chercheurs en santé publique s'engagent très lentement dans la révolution que subit le partage des données et les agences chargées d'entretenir les bases de données mondiales sur la santé sont réticentes à ce partage. Une fois certains investissements consentis dans les infrastructures, le recyclage et la combinaison des données peuvent donner accès à un maximum de connaissances pour un coût additionnel minimal. En refusant le partage des données, les chercheurs ralentissent les progrès vers la réduction de la morbidité et de la mortalité et interdisent l'accès à l'information à un public tout juste bon à payer les impôts qui financent la plupart de leurs recherches.

Les apporteurs de fonds pour la recherche en santé publique commencent à appeler au changement et à développer des politiques de partage des données. Cependant, ils n'ont pas encore trouvé de moyens adéquats pour aplanir les obstacles responsables de l'échec de ce partage. Il s'agit notamment de structures professionnelles qui récompensent la publication d'une analyse, mais pas celle des données, et de flux de financement et d'évolutions de carrière qui continuent de sous-évaluer le travail essentiel de gestion des données. Il convient aussi de sérier les problèmes pratiques: où et comment les données doivent-elles être stockées sur le long terme, qui exercera un contrôle sur les accès et qui paiera pour ces services ? Les normes existantes pour les métadonnées doivent être étendues pour couvrir les données relatives à la santé.

Ces obstacles sont connus depuis un certain temps ; la plupart d'entre eux peuvent être surmontés dans le domaine de la santé publique tout comme ils l'ont été dans d'autres secteurs. Néanmoins, aucune institution n'a pris la direction des opérations pour définir un plan de travail et répartir les tâches et la facture. Dans cet article destiné à une table ronde, nous proposons des objectifs pour le partage des données et un plan de travail pour les atteindre et nous sollicitons des réponses pour aller au-delà de plans de partage des données bien intentionnés, mais largement utopistes.

Resumen

Intercambio de datos sanitarios: las buenas intenciones no son suficientes

Los epidemiólogos e investigadores en salud pública están avanzando muy lentamente en la revolución del intercambio de datos, y además los organismos que mantienen las bases de datos mundiales sobre salud se muestran reacios a compartir su información. Una vez realizadas las inversiones en infraestructuras, la reutilización y combinación de datos brindan acceso a un máximo de conocimientos con un costo adicional mínimo. Al negarse a compartir los datos, los investigadores están frenando los progresos hacia la reducción de la morbilidad y la mortalidad y están negando un bien público a contribuyentes que apoyan la mayor parte de las investigaciones.

Los agentes de financiación de las investigaciones en salud pública están empezando a exigir cambios y a elaborar políticas de intercambio de datos. Sin embargo, aún no están abordando adecuadamente los obstáculos que impiden compartir esos datos. Entre ellos cabe citar unas estructuras profesionales que recompensan la publicación de análisis, pero no de datos, y unas fuentes de financiación y unas perspectivas de carrera que siguen sin reconocer el carácter crucial de la gestión de datos. Es preciso esclarecer también algunos aspectos prácticos: cómo y dónde deben conservarse los datos a largo plazo, quién controlará el acceso y quién pagará esos servicios. Las normas existentes sobre metadatos deben ampliarse para poder manejar los datos sanitarios.

Estas dificultades son conocidas desde hace ya algún tiempo, pero la mayoría pueden ser superadas en el campo de la salud pública al igual que han sido superadas en otros campos. Sin embargo, ninguna institución ha tomado la iniciativa para definir un plan de trabajo y repartirse las tareas y los costos asociados. En este artículo de la mesa redonda proponemos metas para el intercambio de datos y un plan de trabajo para su consecución, y alentamos a los encuestados a trazar algo más que unos planes de intercambio de datos bienintencionados pero demasiado ambiciosos.

References

1. Smith D, Carrano A. International large-scale sequencing meeting. *Human Genome News* 1996;7. Available from: http://www.ornl.gov/sci/techresources/Human_Genome/publicat/hgn/v7n6/19intern.shtml [accessed 26 February 2010].
2. Kaye J, Heeney C, Hawkins N, de Vries J, Boddington P. Data sharing in genomics – re-shaping scientific practice. *Nat Rev Genet* 2009;10:331–5. doi:10.1038/nrg2573 PMID:19308065
3. Nüesch E, Trelle S, Reichenbach S, Rutjes AWS, Bürgi E, Scherer M et al. The effects of excluding patients from the analysis in randomised controlled trials: meta-epidemiological study. *BMJ* 2009;339:b3244. doi:10.1136/bmj.b3244 PMID:19736281
4. Elobeid MA, Padilla MA, McVie T, Thomas O, Brock DW, Musser B et al. Missing data in randomized clinical trials for weight loss: scope of the problem, state of the field and performance of statistical methods. *PLoS ONE* 2009;4:e6624. doi:10.1371/journal.pone.0006624 PMID:19675667
5. Lord P, MacDonald A, Sinnot R, Ecklund D, Westhead M, Jones A. *Large-scale data sharing in the life sciences: data standards, incentives, barriers and funding models (The "Joint data standards study").* Glasgow & Edinburgh: National e-Science Centre; 2006. Available from: http://www.nesc.ac.uk/technical_papers/uk.html [accessed 26 February 2010].
6. Corti L, Wright M. *MRC Population data archiving and access.* London: Medical Research Council; 2002.
7. Chandramohan D, Shibuya K, Setel P, Cairncross S, Lopez AD, Murray CJ et al. Should data from demographic surveillance systems be made more widely available to researchers? *PLoS Med* 2008. 5:e57. doi:10.1371/journal.pmed.0050057 PMID:18303944
8. Field D, Sansone SA, Collis A, Booth T, Dukes P, Gregurick SK, Kennedy K et al. Omics data sharing *Science* 2009. 326:234–236. doi doi:10.1126/science.1180598

9. Mascalzoni D, Hicks A, Pramstaller P, Wjst M. Informed consent in the genomics era. *PLoS Med* 2008;5:e192. doi:10.1371/journal.pmed.0050192 PMID:18798689
10. Homer N, Szelinger S, Redman M, Duggan D, Tembe W, Muehling J et al. Resolving individuals contributing trace amounts of DNA to highly complex mixtures using high-density SNP genotyping microarrays. *PLoS Genet* 2008;4:e1000167. doi:10.1371/journal.pgen.1000167 PMID:18769715
11. *Application to use restricted microdata.* Minneapolis: IPUMS International. Available from: https://international.ipums.org/international/ [accessed 26 February 2010].
12. UK Data Archive. End user licence. Colchester: University of Essex; 2008. Available from: http://www.data-archive.ac.uk/aandp/access/licence.asp [accessed 26 February 2010].
13. Pisani E, Whitworth J, Zaba B, AbouZahr C. Time for fair trade in research data. *Lancet* 2010;375:703–5. doi:10.1016/S0140-6736(09)61486-0 PMID:19913902
14. *OECD Principles and guidelines for access to research data from public funding.* Paris: Organisation for Economic Co-operation and Development; 2007.
15. *Global strategy and plan of action on public health, innovation and intellectual property.* Geneva: World Health Organization; 2008.
16. How to encourage the right behaviour. *Nature* 2002;416:1. doi:10.1038/416001b
17. Data's shameful neglect. *Nature* 2009;461:145. doi:10.1038/461145a
18. PLoS Medicine Editors. Next stop, don't block the doors: opening up access to clinical trials results. *PLoS Med* 2008;5:e160. doi:10.1371/journal.pmed.0050160 PMID:18630986
19. Laine C, Berkwits M, Mulrow C, Schaeffer MB, Griswold M, Goodman S. Reproducible research: biomedical researchers' willingness to share information to enable others to reproduce their results. In: *Sixth International Congress on Peer Review and Biomedical Publication, Vancouver, Canada, 10–12 September 2009.* Available from: http://www.ama-assn.org/public/peer/abstracts-0910.pdf [accessed 26 February 2010].
20. *NIH guide: final NIH statement on sharing research data. Bethesda:* National Institutes of Health; 2003. Available from: http://grants.nih.gov/grants/oer.htm [accessed 26 February 2010].
21. *MRC Policy on data sharing and preservation.* London: Medical Research Council; 2008. Available from: http://www.mrc.ac.uk/PolicyGuidance/EthicsAndGovernance/DataSharing/PolicyonDataSharingandPreservation/index.htm [accessed 26 February 2010].
22. *Policy on data management and sharing.* Wellcome Trust; 2007. Available from: http://www.wellcome.ac.uk/About-us/Policy/Policy-and-position-statements/WTX035043.htm [accessed 26 February 2010].
23. *Sharing public health data: a code of conduct.* London: Wellcome Trust; 2008. Available from: http://www.wellcome.ac.uk/About-us/Policy/Spotlight-issues/Data-sharing/Public-health-and-epidemiology/index.htm [accessed 26 February 2010].
24. Supari SF. *Saatnya dunia berubah: tangan Tuhan di balik virus flu burung / Siti Fadilah Supari* [in Indonesian]. Jakarta: Sulaksana Watinsa Indonesia; 2008.
25. *International Comprehensive Ocean-Atmosphere Data Set.* Washington, DC: National Oceanic and Atmospheric Administration; 2009. Available from: http://icoads.noaa.gov/ [accessed 26 February 2010].
26. *International Clinical Trials Registry Platform, WHO Registry Criteria, version 2.1.* Geneva: World Health Organization; 2009. Available from: http://www.who.int/ictrp/network/criteria_summary/en/index.html [accessed 26 February 2010].
27. DataCite - International initiative to facilitate access to research data. Hannover: German National Library of Science and Technology; 2009. Available from: http://www.datacite.org/ [accessed 26 February 2010].
28. Paskin N. Digital Object Identifier (DOI) System. In: *Encyclopedia of library and information sciences.* New York: Taylor & Francis; 2008.
29. Altman M, King G. A proposed standard for the scholarly citation of quantitative data. *D-Lib* 2007. 13.
30. Lowrance W. *Access to collections of data and materials for health research.* London: Medical Research Council; 2006.
31. Pisani E. *OpenEpi: a new culture for public health data?* London: Wellcome Trust; 2008.
32. National Academy of Sciences. *Ensuring the integrity, accessibility and stewardship of research data in the digital age.* Washington, DC: National Academy Press; 2009.

Round table discussion

Publishing research data on a professional basis

Toby Green[a]

As Pisani & AbouZahr have identified, there are many obstacles to the publishing of data: social (incentives for researchers to make the effort to publish), financial (having adequate financing to cover short-term publishing and long-term curation costs), and technical (standards and systems).[1] This paper looks at some of the technical challenges of publishing data professionally and describes the discoverability and citability benefits that follow.

Let's take it as read that publishing research data is a "good thing," that researchers are as willing to publish data as they are research papers and funding is in place to make them available online in the long run. Job done? Well, no, not by a long chalk.

Just as loading a journal article onto a web site somewhere isn't the same as publishing it properly, so the same is true for data. To be as discoverable and as citable as research articles, data sets need to be published using an infrastructure that is compatible with research articles. It is not enough that data sets hang like dongles off a research article; they need to be discoverable and citable in their own right – just like a journal article. This means the metadata must be compatible with existing bibliographic management and citation systems like Ref Works® and CrossRef®. Users will expect search engines, abstracting and indexing services and library catalogues to reference data sets, so, for example, librarians will need MARC (MAchine-Readable Cataloging) records.

Is this overkill? Well, the Organisation for Economic Co-operation and Development (OECD) doesn't think so. OECD publishes more than 390 data sets as stand-alone objects, as well as thousands of data sets as supplemental data to its books and journal articles. Sub-sets of the data sets are also posted on the web as stand-alone objects too. So it is no surprise that, in the absence of good discovery metadata and systems, the number one complaint from users is the challenge of finding a relevant data set. They know the data is there, but they can't find it – even with Google's help.

To solve this problem, OECD's Publishing Division has spent the past three years grappling with the challenge of how to publish these many thousands of data objects so that users can not only find the data they need, but can then cite and manage the data sets using the same tools that they already use to manage journal articles or book chapters. The first result was a white paper,[2] first released in March 2009, which described this challenge and proposed a set of metadata schema for databases in their own right, sub-sets of databases and supplemental data.

More significantly, was the launch of OECD iLibrary, OECD's new publishing platform, in July 2009. OECD iLibrary[3] hosts all OECD books, working papers, journals and data sets in a seamless manner. OECD iLibrary puts the white paper's proposed bibliographic schema for data objects into practice. Search for "health data" and the search results include data sets, book chapters – even individual tables found inside books.

[a] Organisation for Economic Co-operation and Development, 2 rue André Pascal, 75775 Paris Cedex 16, France.
Correspondence to Toby Green (e-mail: toby.green@oecd.org).

OECD's data sets can now be discovered more easily and they can be cited as simply and as easily as a research article using the downloadable citation provided. Later in 2010, librarians will be supplied with MARC records and the bibliographic records for OECD data sets will be shared with discovery platforms like RePEc (Research Papers in Economics)[4] – the world's largest collection of economics grey literature – enabling visitors to find data objects alongside working papers and journal articles. Imagine being able to discover and cite data sets as easily and as simply as journal articles. Imagine no more. ■

Competing interests: None declared.

References

1. Pisani E, AbouZahr C. Sharing health data: good intentions are not enough. *Bull World Health Organ* 2010;88:462–6.
2. Green T. *We need publishing standards for datasets and data tables.* Paris: Organisation for Economic Co-operation and Development; 2010. Available from: http://dx.doi.org/10.1787/603233448430 [accessed 1 March 2010].
3. OECD iLibrary [Internet site]. Paris: Organisation for Economic Co-operation and Development; 2009. Available from: www.oecd-ilibrary.org [accessed 19 May 2010].
4. RePEc (Research Papers in Economics) [Internet site]. Available from: www.repec.org [accessed 1 March 2010].

Sharing data for public health: where is the vision?

Alan D Lopez[a]

"By refusing to share data, researchers are slowing progress towards reducing illness and death." Pisani & AbouZahr are making a big claim in this round table.[1] Is this claim sensationalist or does it have some basis? Can we argue that data from public health research really affect the ways prevention and control programmes are designed? Lives have become longer and healthier in the past 50 years, despite an arguably poor evidence base for health and an even poorer appreciation by policy-makers of the value of reliable health information.[2,3] Pisani & AbouZahr are arguing that such gains would have been bigger, faster and more equitable had the world had better information about what works and does not work in public health; lost ground is partly due to widespread hoarding of research findings, particularly primary data.

They have a point. Restricting access to data to only those scientists directly engaged in a research project limits the scope of legitimate scientific enquiry and the potential for research to influence policy and practice. No individual scientist who collects or collates data has all the possible analytic methods, expertise and time to extract key public health messages from research or routine data sets.[4–7] Lost opportunity for analysis is the main consequence of poor data sharing practices.

Yet, as Pisani & AbouZahr argue, it is unreasonable to expect data collectors to share without adequate incentives. Incentives could include professional recognition for well collected and documented data, appropriately disseminated using good data management practices. Data collectors too need assurance that their efforts will be respected and that errors in data are inevitable and rarely disastrous. Experienced researchers are aware of these risks and can use a range of quality assessment techniques to deal with errors.

Mentoring is one incentive that is missing from the otherwise excellent set proposed by Pisani & AbouZahr. Partnerships between researchers and data collectors, including intensive methodological workshops, are feasible and can help ensure that those who collect data realize the public health potential and value of their efforts. Such an approach could rapidly increase analytical capacity and diversify the analysis of rich, but underutilized, data sets. Funding such collaborations would be an innovative and constructive use of research funds. Competent analysts should be able to resolve potential challenges in interpreting data because of specific local conditions surrounding their collection. Restricting access on this basis reflects a lack of confidence, imagination or trust by those who collect data and should be questioned when used to preclude further analysis.

The authors propose an urgent agenda for action to improve data sharing practices that will benefit all stakeholders – data collectors, analysts, the policy community and, ultimately, the public. This is admirable but, for such a plan to succeed, funders, researchers and data collectors alike need to understand its benefits. That will only happen with effective and committed leadership. What better role for the World Health Organization? ■

Competing interests: None declared.

References

1. Pisani E, AbouZahr C. Sharing health data: good intentions are not enough. *Bull World Health Organ* 2010;88:462–6.
2. Setel PW, Macfarlane SB, Szreter S, Mikkelsen L, Jha P, Stout S et al.; Monitoring of Vital Events. A scandal of invisibility: making everyone count by counting everyone. *Lancet* 2007;370:1569–77. doi:10.1016/S0140-6736(07)61307-5 PMID:17992727
3. Horton R. Counting for health. *Lancet* 2007;370:1526. doi:10.1016/S0140-6736(07)61418-4 PMID:17992726
4. Murray CJL, Lopez AD. The utility of DALYs for public health policy and research: a reply. *Bull World Health Organ* 1997;75:377–81. PMID:9342897
5. Rajaratnam JK, Tran LN, Lopez AD, Murray CJL. Measuring under-five mortality: validation of new low-cost methods. *PLoS Med* 2010;7:e1000253. doi:10.1371/journal.pmed.1000253 PMID:20405055
6. Obermeyer Z, Rajaratnam JK, Park CH, Gakidou E, Hogan MC, Lopez AD et al. Measuring adult mortality using sibling survival: a new analytical method and new results for 44 countries, 1974–2006. *PLoS Med* 2010;7:e1000260. doi:10.1371/journal.pmed.1000260 PMID:20405004
7. Murray CJL, Rajaratnam JK, Marcus J, Laakso T, Lopez AD. What can we conclude from death registration? Improved methods for evaluating completeness. *PLoS Med* 2010;7:e1000262. doi:10.1371/journal.pmed.1000262 PMID:20405002

Data sharing: reaching consensus

Jimmy Whitworth[b]

Pisani & AbouZahr write passionately about the need to change the culture of data sharing in public health research.[1] They explain why this is in everybody's best interests and outline ways in which the main obstacles can be overcome. This is laudable and much appreciated; it is time for a change, the current situation is unacceptably inefficient in terms of scientific progress and value for money from research.

[a] School of Population Health, University of Queensland, Herston Road, Herston, Qld., 4006, Australia (e-mail: a.lopez@sph.uq.edu.au).
[b] Wellcome Trust, Gibbs Building, 215 Euston Road, London, NW1 2BE, England (e-mail: j.whitworth@wellcome.ac.uk).

The two authors challenge institutions, in particular research funders, to take charge of the agenda to make these changes happen. They call for leadership but, while funding agencies are clearly influential and can certainly facilitate changes in scientific behaviour and culture, they are unlikely to be able to effect all the changes called for by Pisani & AbouZahr. While funders might support and encourage, we are not in a position to dictate changes to professional structures, to create career paths for scientific disciplines at academic institutions, nor to determine scientific reward mechanisms.

What is required as a first step is the facilitation of dialogue and the building of consensus between all interested parties, including funders, researchers, institutions, journal editors, ethics committees, multilateral agencies and governments. No one agency has the mandate or the legitimacy to take this whole agenda forward unilaterally. A more sustainable and palatable pathway be will to build consensus and to create a broad coalition.

It is worth reflecting on why data sharing is not more commonly practiced among epidemiologists and public health researchers. Pisani & AbouZahr point out many of the constraints, such as the lack of appropriate incentives from employers such as research councils, foundations and universities, the short supply of data managers especially in low- and middle-income countries, and concerns over the control and ownership of data. There are also technical issues, data sets for cohort studies are more complicated than standard genetic data sets because of their longitudinal nature, and there are no off-the-shelf tools available for managing and curating standardized and interoperable longitudinal data sets.

Overcoming these constraints requires a broad consensus among stakeholders. Indeed Pisani & AbouZahr seem to acknowledge this. When they write that "we" need to develop a search portal, invest in training in data management, develop reliable citation standards, develop methods to track the value of data sharing, and so on, these are clearly tasks for the wider scientific community.

Of course, individual institutions – and funders – can take the initiative over certain aspects of the agenda and form alliances with those agencies that can help in other domains. Indeed, the Wellcome Trust has already led various initiatives in this field, including convening international meetings of public health researchers and funding agencies, and has raised these issues at meetings of public health policy-makers and international journal editors. The Trust is currently revising its grant conditions about data sharing, which will be strengthened and, importantly, will provide more guidance about the technicalities of *how* to share data more effectively. We are ready to take the lead in those areas, where it is appropriate for us to do so, and we are open to the formation of alliances with other agencies that can help to facilitate progress in other areas. ■

Competing interests: Jimmy Whitworth is employed by the Wellcome Trust, which commissioned Elizabeth Pisani to work on its data-sharing project.

References

1. Pisani E, AbouZahr C. Sharing health data: good intentions are not enough. *Bull World Health Organ* 2010;88:462–6.

Sharing health data: developing country perspectives

Viroj Tangcharoensathien,[a] Jirawan Boonperm[b] & Pongpisut Jongudomsuk[c]

Not only is it difficult to change the "publish or perish" mindset among health researchers, there are other fundamental barriers in data sharing that Pisani & AbouZahr's paper should have addressed.[1] Sharing data is not only about the technical dimension such as data management, repositories and libraries; developing countries are concerned about factors that impede data sharing, in particular, fairness. Pisani & AbouZahr provide clear analyses on barriers but their proposed solutions will not be effective unless they address the fundamental problems.

From the perspective of developing countries, the goal of data sharing is beyond national interests and is for the benefit of all mankind. Without this explicit goal, data sharing more often helps scientists in developed countries get published. While these scientists may have higher analytical capacities, they have neither shared the "legwork" in collecting routine administrative data nor made intellectual contributions to designing and solving problems in conducting field work with scientists in developing countries.

Developing countries need to strengthen capacities in survey design, data management and analysis and policy use. There is clearly an unlevel playing field that impedes data sharing. Scientists from developed countries often take the following approach with researchers in developing countries: "Share your data with me, you do not have analytical capacities. I will analyse and publish papers for global public good." Instead, their approach should be: "We can analyse the data together and learn from each other for the benefit of all people." This approach would gradually create equal partnerships, a level playing field, goodwill and trust for collaborations beyond simply sharing data.[2–4] International data sharing cannot be achieved through forced marriage; as shown by the defeat of the policy proposed by the Annals of Internal Medicine of a publicly accessible database as a condition for journal publication.[5]

The recent sharing of avian flu virus specimens by developing countries through the World Health Organization resulted in the production of avian influenza vaccines at a price of US$ 10–20 per dose. This is unaffordable in low-income countries where total health expenditure is less than US$ 30 per person. Should an avian flu pandemic occur, there would be huge death tolls in countries without access to vaccines; while rich countries' populations would be fully protected, literally from any moral obligations to countries that shared their specimens. Such unilateral benefit inhibits data sharing.

[a] International Health Policy Program, Ministry of Public Health, 376 Mooban Panya, Patanakan Road, Bangkok, 10250, Thailand.
[b] National Statistical Office, Laksi, Thailand.
[c] Health Systems Research Institute, Bangkok, Thailand.
Correspondence to Viroj Tangcharoensathien (e-mail: viroj@ihpp.thaigov.net).

It is important to have evidence on the benefits that populations receive directly as a result of sharing, beyond publications by secondary users. Success in international data sharing may start with efforts at country level or through multi-country research partnerships. Undeniably, multi-country studies provide huge benefit in supporting evidence-based policy. Collaborative partnerships among a number of developed and developing countries, such as for maternal and perinatal health, are foundations for building long-term trust.[6] In research partnerships, there is equitable access to and use of data sets, beyond the conventional practice of passive data sharing without partnership.

In Thailand, rules and procedures for data sharing were developed through a research funding agency and the National Statistical Office. Primary users were granted a reasonable-use period of two years after complete data collection prior to access by secondary users. Good practices are emerging. With the aim of capacity building and mutual benefit, the National Statistical Office grants approval to international secondary users to access nationally representative household data sets, on the condition that they develop partnerships with local scientists. Such engagement gradually builds trust and longer-term partnerships between scientists from developed and developing countries. ■

Competing interests: None declared.

References

1. Pisani E, AbouZahr C. Sharing health data: good intentions are not enough. *Bull World Health Organ* 2010;88:462–6.
2. Pitayarangsarit S, Tangcharoensathien V. Sustaining capacity in health policy and systems research in Thailand. *Bull World Health Organ* 2009;87:72–4. doi:10.2471/BLT.07.044479 PMID:19197407
3. Pitayarangsarit S, Tangcharoensathien V. Capacity development for health policy and systems research: experience and lessons from Thailand. In: Green A, Bennett S, eds. *Sound choices: enhancing capacity for evidence-informed health policy.* Geneva: World Health Organization; 2007.
4. Mayhew SH, Doherty J, Pitayarangsarit S. Developing health systems research capacities through north-south partnership: an evaluation of collaboration with South Africa and Thailand. *Health Res Policy Syst* 2008;6:8. doi:10.1186/1478-4505-6-8 PMID:18673541
5. Laine C, Berkwits M, Mulrow C, Shaeffer MG, Griswold M, Goodman S. Reproducible research: biomedical researchers' willingness to share information to enable others to reproduce their results. In: *Sixth International Congress on Peer Review and Biomedical Publication, Vancouver, 10–12 September 2009.* Available from: http://www.ama-assn.org/public/peer/abstracts-0910.pdf [accessed 26 April 2010]
6. Lumbiganon P, Laopaiboon M, Gülmezoglu AM, Souza JP, Taneepanichskul S, Ruyan P et al.; World Health Organization Global Survey on Maternal and Perinatal Health Research Group. Method of delivery and pregnancy outcomes in Asia: the WHO global survey on maternal and perinatal health 2007–08. *Lancet* 2010;375:490–9. doi:10.1016/S0140-6736(09)61870-5 PMID:20071021