Sharing public health data: a code of conduct

In this information age, science is advancing by leaps and bounds. What is driving the exponential growth in knowledge in areas such as genetics, astrophysics, information technology? Data sharing. Teams of scientists openly exchange information; they build on one another's work rather than duplicating it. Money is saved, learning is accelerated, science, technology and medicine move forward and humanity is the winner. Though many scientists initially feared that they would lose out by "giving away" information, it has turned out that the reverse is true. Scientists benefit because their work becomes more widely known and recognised. As data become more widely used, the value of the data rises. This in turn attracts more interest and more funding for continued work. It has also created a new norm of team-work in science, spreading the benefits of that increased funding to teams of lab technicians, data managers and other specialists whose work was previously poorly compensated.

Epidemiology and public health have been left behind in this data sharing revolution, mired in a culture that restricts access to data and information. This is in part because of a perceived need to protect the privacy of individuals involved in research. But public health is a public good; in public health research there's an ethical imperative to use information gathered from individuals to benefit the greatest possible number of people. Public health deserves to advance at the same speed as genetics, where data sharing has led to an explosion of progress. The World Health Organisation and several funders of public health research, led by the Wellcome Trust, are thus supporting the development of a code of conduct to encourage greater sharing of public health data. The code seeks to provide guidance for funders of data collection and for institutions that collect and analyse data, including those who perform secondary analysis on data collected by other people. The principles espoused by the code are universal. However because capacity for data management and analysis are unevenly distributed around the world, the code is explicit in trying to increase skills and incentives for locally-relevant data use in developing countries.

The draft code presented here is the product of initial discussions between epidemiologists and data managers from all continents. They gathered with a number of representatives from governments, international organisations and major funders of public health research in London on October 6th, 2008 to agree on the core principles in the code. The discussions of this Working Group were informed by a background paper which reviewed the major challenges to more open exchange of public health data, challenges that can be categorised broadly as incentive-related, capacity-related, ethical and technical. The draft code is structured around these four areas. The background paper has been updated to reflect the outcome of the meeting, and is appended here.

The WHO and the Wellcome Trust, who sponsored the October meeting, join with their partners in recognising the central role that Ministries of Health, Research and Technology have in supporting and guiding public health research, as well as in using its results. While we'd like to see knowledge and discovery in public health moving forward at the same speed as learning in genetics, we know that the interests of participants in research as well as those of researchers themselves must be protected. A code of conduct on data sharing is an important first step in striking the balance between the advancement of science and the rights and needs of individuals and communities. The sponsors of this initiative thus seek the support of delegates as well as their active input in developing a code of conduct on data sharing that meets the needs of governments as well as the needs of scientists, funders and the wider research community.

What does the code cover?

Many types of data have potential relevance to public health decision-making. Data from complex longitudinal studies designed to answer specific questions, demographic surveillance sites, household surveys, routine disease and risk surveillance, health service use -- all aim to increase knowledge about health and disease. And almost all collect far more data than can possibly be analysed or used, unless we change our way of doing business.

This code is relevant to all data of potential public health significance. Springing from the belief that data collected with public and philanthropic money are "public goods", it focuses particularly on research and routine data collection systems funded by taxpayers (either in the country of data collection or in a "donor" country) and foundations. This of course includes data collected, held or aggregated by international institutions.

The code recognises that in the current system, scientists in poor countries spend a disproportionate proportion of their effort collecting data that is then analysed by scientists from rich countries. The code is designed to cover data collected in any country, but it pays special attention to ensuring that the way data are shared increases the incentives and capacity to manage and analyse information in developing countries.

By "data sharing", the code does not necessarily mean that any individual should have immediate and unrestricted access to raw data arising from research. It recognises that different levels of access to data may be necessary to protect the interests of research participants, while the timing of publication of data sets may vary to protect the needs of researchers. To the extent possible, the code promotes the sharing of micro-level data -- that is, individual level records. There may occasionally be reason to restrict access to individual level data. There is rarely any reason at all to restrict access to aggregated data.

The code of conduct

Access to data -- a graduated scale

We support the maximum public access to data of public health importance compatible with the following principles:

- The protection of privacy of individuals from whom data are gathered
- Fair reward for the work of data collectors and primary investigators
- Maximum public health benefit delivered in a reasonable time frame

Respect for those principles means that data access may vary. In order to respect the principle of fair reward for the work of primary researchers, the first three of the levels of access described below may be subject to a period of

Limited-time exclusive access for primary researchers
Data are available to the research team involved in data collection and their institutional partners for a fixed period (between six and 18 months) before they are shared. This allows the research team a head start on data analysis and publication. It will generally be unnecessary in the case of routinely collected data such as surveillance or census data.

Following a period of exclusive access for primary researchers where necessary, the most common levels for access to data of public health importance will be:

Fully open access
Data (anonymised where necessary) are made available in machine-readable formats on publicly-accessible websites. This is most desirable and should be encouraged where feasible

and compatible with privacy.
Existing examples: Genetic sequencing data.

Controlled public access
Data are made available to authorised users after a screening process. This is likely to be the most common form of access for data of public health importance.
Existing examples: Demographic and Health Surveys, census data made available through the Integrated Public Use Microdata Series, studies funded by the UK's Economic and Social Research Council.

Collaborative access among scientists
Data are made available to other scientists in a collaborative network. Collaborative access may be necessary for complex datasets that include sensitive information where anonymisation is difficult (e.g. longitudinal data sets including HIV status).
Existing examples: MalariaGen study date, INDEPTH demographic surveillance site data.

Exclusive access for primary researchers
Data are only available to the research team involved in data collection and their institutional partners. This is currently the norm in public health data collection, but it is precisely this norm that the current code seeks to change. There are few cases in which this degree of exclusivity is necessary in the long term.
Existing examples: All too many, though very few of them justified.

Increasing the incentives to share data

As the genome project has shown us, sharing data and ideas greases the wheels of science and speeds up discovery and human progress. But what is best for "humanity" is not necessarily best for individual humans. In our current system, epidemiologists and health researchers are rewarded on the basis of papers published in journals. This leads to hoarding of data. Other disincentives to share data include a fear that data will be wilfully misused (so disrupting the trust of the community from whom data are collected), a fear that poor quality data will be exposed, and the loss of income associated with duplication in data collection.

Under the Code of Conduct on Data Sharing we agree to:

Put **past data sharing** performance on **a par with publication** as a criterion for evaluating the performance and job suitability of scientists, as well as evaluating grant proposals.

Reward concrete plans for data sharing when **evaluating funding proposals** for research and routine health systems functions such as surveillance.

Develop **citation standards and indices** for shared data sets; commit to using them when publishing secondary analysis.

Require **registration** of public-health related research and data collection in open access data-bases to facilitate data discovery and create demand for shared data.

Encourage **submission of micro-data to public repositories** as a condition for **journal publication** of research results.

Promote a **"creative commons" approach**, in which derived datasets and secondary analysis files based on shared data are in turn made publicly available.

Support an **ombudsman** system to oversee the fair use and proper acknowledgement by secondary users of shared data.

Increasing capacity for data management and analysis

The single biggest bottleneck to greater sharing of data is a lack of data management skills. Data management is neglected and chronically under-funded; in most countries it is not regarded as a legitimate career path for an IT worker, much less a scientist. Data are scattered, improperly cleaned or coded and poorly documented. Until that changes, data simply can't be shared. Getting data into shareable shape is a lot of work, while the immediate reward for researchers required to do that work may seem limited. Improving data management implies a huge commitment to building long-term capacity in this area, especially in developing countries. Organisations promoting more open access to data will need to ensure these resources are available if there is to be any hope of lasting progress.

Unequal distribution of skills in analysis also stands in the way of increasing access to data. Scientists in developing countries naturally want to reap maximum benefit from the data they collect, but are hampered in doing so by limited capacity for analysis. A larger number of well-trained analysts in data collection teams in developing countries would increase rapid, locally-relevant analysis of local data, without precluding secondary analysis for comparative and other purposes by users based elsewhere. Again, the creation and retention of a critical mass of well-trained analysts implies a huge, long-term commitment to improving training institutions and career paths for people with good brains, whose skills will be much sought after in the private sector.

Under the Code of Conduct on Data Sharing we agree to:

Invest substantially in the long term development of **data management skills**.

Reward data managers adequately, both financially and in terms of career prospects.

Invest in **full documentation of data sets**, to levels needed for data sharing.

Invest in **increasing analysis skills in developing countries**, including through long-term support of academic and research institutions

Encourage **secondary analysts to work with primary data collection and research teams** to develop analytic skills.

Reward senior analysts for **mentoring and documented skills transfer** to colleagues.

Commit to **funding infrastructure** needed for data curation, storage and access.

The ethical imperative to share data

Historically, the primary ethical concern of researchers and the institutions they work for, especially in the field of medical research, has been to protect the rights and privacy of individuals involved in research. An acute sensitivity to individual rights has spilled over into the field of public health, where our interest is in the welfare of communities and populations as well as that of individuals. At times, the concern with individual rights has stood in the way of larger public health goals. Restricted consent procedures have become obstacles to secondary analysis of public health data. Privacy concerns, though often easily addressed through anonymisation techniques, are used to excuse the failure to share data.

Under the Code of Conduct on Data Sharing we agree to:

Assert the **ethical imperative to maximise the use of public health data**, including through making the data widely accessible to many users.

Work with **Institutional Review Boards** to promote oversight of data accessibility and use as part of their work in ensuring ethical research.

Uphold the importance of personal privacy by investing in **anonymisation techniques** that protect individuals while increasing analysis that will improve community health.

Promote **broad consent procedures**, explaining and seeking support for secondary analysis of public health importance.

Using technology to increase data sharing

Technology represents a challenge to greater sharing of data, but more than that it represents a fantastic opportunity. The key issue is standardisation. The more we are able to use common technological standards to describe our data, to store it and to retrieve it, the easier it will be for scientists to use one another's information and to drive forward progress. Some of these standards already exist, though they may need to be adapted for public health purposes. Technology can also be used to meet some of the challenges that arose earlier in the discussion, such as in developing citation indices and ensuring traceability of shared data back to an acknowledged source.

Under the Code of Conduct on Data Sharing we agree to:

Commit to a **single metadata standard for datasets** of public health interest. Probably an extension of the existing DDI standard, this should be widely adopted after piloting.

Ensure that **metadata are open access and machine-readable**, even for data that are shared under the controlled or collaborative access standards.

Support the development of **"open source" software** for management, documentation and analysis of public health data.

Taking the code forward

It is hoped that a finalised code will meet the needs of the whole public health and research community. That means governments, the primary consumers of public health data, as well as the public and philanthropic bodies that fund data collection and research. It means the teams that collect and analyse data to answer defined research questions, as well as researchers that conduct secondary analysis to answer broader questions. It includes the journals and websites that publish data and analysis and the academic institutions that teach data management and analysis. In trying to meet the needs of this huge and varied constituency, the current draft code is vague: phrases such as "promote x" and "encourage y" predominate. As the code develops, we hope that it will become more concrete: "Funding institutions commit to investing in x", "Secondary analysts agree to provide y". The WHO, the Wellcome Trust and their many partners in this initiative actively invite further suggestions and contributions that will turn this into a collectively agreed code driven by concrete commitments.