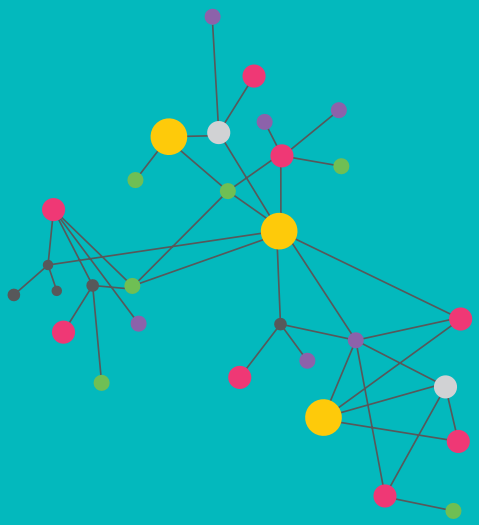
Public Health Research Data Forum

Enabling Data Linkage to Maximise the Value of Public Health Research Data: Summary



wellcome trust

Enabling Data Linkage to Maximise the Value of Public Health Research Data

Introduction

This project was commissioned by the Wellcome Trust on behalf of the Public Health Research Data Forum. It aimed to identify the gains to public health research from linking existing data sources, the opportunities in and barriers to such data linking, and how these barriers could be overcome. The objective was to deliver a set of practical recommendations for realising the gains from data linkage.

The project team (right) brought together researchers from the University of the West of England (United Kingdom), the DataFirst initiative at the University of Cape Town (South Africa), and the Centre for Injury Prevention Research, Bangladesh.

The research team used a mix of literature review, case study development and interviews with selected experts involved with data linkage. The study looked at low-, middle- and high-income countries to ensure that lessons learned would have wide applicability. Barriers to useful data linkage were analysed from statistical, operational and institutional perspectives. Given the vast amount of information on data linkage theory and practice, this project focused on useful illustrative examples as opposed to an exhaustive review of the field.

Project team

Felix Ritchie, Elizabeth Green and Don Webber, Faculty of Business and Law, University of the West of England, United Kingdom

Julie Mytton and Toity Deave, Faculty of Health and Applied Sciences, University of the West of England, United Kingdom

Alex Montgomery and Lynn Woolfrey, DataFirst, University of Cape Town, South Africa

Kamran ul-Baset and Salim Chowdhury, Centre for Injury Prevention Research, Bangladesh

The summary and full report are available to download from **wellcome.ac.uk/PHRDF**

Data linkage: the benefits and the risks

Data linkage simply means bringing together two or more sources of information which relate to the same individual, event, institution or place. This project focused on the potential to link datasets within the context of health research – including datasets collected for the purposes of research and those collected for other purposes (for example information from electronic patient records, cancer registries or socio-economic surveys).

Data linkage offers numerous benefits to public health and epidemiological research: through bringing together different pieces of information, researchers can identify factors and associations that would otherwise be difficult or impossible to determine. For example, linking health or disease outbreak data to historical information collected for other purposes – such as vital events or civil registration data – can reveal contributory factors for disease going back years into the past. Similarly, linking health data to socio-economic, geospatial or environmental datasets may provide vital insights into disease epidemiology. The ability to link data may also vastly increase the potential value that can be derived from individual datasets – which are often collected at considerable effort and expense – and reduce unnecessary or duplicative data collection efforts.

Direct identifiers in datasets (such as names and addresses) are typically of little interest to researchers; their value is in allowing the data to be linked, and so they are removed from datasets before research access is allowed. However, some data elements – for example, age, gender and ethnicity – may have considerable value to researchers, but may also potentially be combined to reveal the identity of individuals. Hence, a useful dataset is likely to have some characteristics which will in theory allow the individual to be re-identified from the data, even if this is very unlikely; this is called 'pseudonymised' (pseudo-anonymised) data.

The research community has well-established best practice protocols for managing such data safely and securely. While it must be acknowledged that the use of sensitive data for research does create a confidentiality risk – and that linked data have an increased risk – fifty years of empirical evidence suggests that it is in reality a low-level risk which can be managed effectively.

At present, although the conceptual and statistical frameworks for data linkage are well-established, researchers may face a number of significant practical challenges. These may be grouped into three broad categories:

1. **Statistical issues** – linking data, and analysing the resulting linked datasets, raises a number of distinct challenges for researchers, although well-established methodologies and tools exist.
2. **Technical and operational issues** – gaining permission to access and use datasets held by multiple organisations may often be far from straightforward for researchers, and differences in the way data are collected may sometimes limit their use.
3. **Institutional issues** – a range of legal, ethical and cultural considerations may significantly constrain the extent to which researchers can link data in practice. These may include variations and uncertainties over what is permissible, questions around consent, and concerns over public acceptability and trust.

Illustrative case studies: data linkage in different contexts

Through interviews with research practitioners in several countries, the project team compiled a resource of ten case studies which illustrate practical experiences of linking data in both high-income countries and middle- and low-income countries. Two examples are summarised briefly below.

South Africa – The Agincourt Health and Socio-Demographic Surveillance System

The Agincourt Health and Socio-Demographic Surveillance System (Agincourt HDSS) was established in 1992 and is located in the rural north-east of South Africa. It captures household roster information, pregnancy outcomes, mortality, migration, maternity history and union status, as well as a variety of social variables. The data are regularly linked to other data sources – including national South African Civil Registration systems (in collaboration with the South African Department of Home Affairs and Statistics South Africa), clinical data in the provincial primary healthcare system (in collaboration with the Department of Health and the clinics themselves) and schools data (through a pilot with the Department of Basic Education).

A key to success has been the ability of Agincourt HDSS to maintain strong relationships and build mutual trust with the government departments that act as gatekeepers to the data. This institutional relationship was useful, for example, in getting approval from the Department for Health for the clinic record linkages as their ethics processes are internal and not amenable to external argument. Legal, ethical and institutional barriers did not seriously inhibit the success of the linking project, as the projects were reviewed by the internal ethics appraisals of each of the departments and those departments had a high level of trust in the Agincourt team. However, there were operational and statistical difficulties: in particular, skills shortages in information system administration and data capturing at the clinical level. This occasionally led to poor data capture or poor maintenance of servers.

Australia – The Centre for Health Record Linkage (CHeReL)

CHeReL is a data linkage research facility which provides a record linkage system for health and human services in New South Wales and the Australian Capital Territory. It uses a master linkage key system which routinely links data between numerous health data records – including hospital admissions, emergency department datasets and disease incidence data – together with data on vital events, such as birth and death records. CHeReL also links data on an ad-hoc basis with study-specific databases, including the Australian Study of Women's Health. It holds over 93 million records based on 11 million people with an average of six links per person. Over recent years, CHeReL has seen increasing international interest and collaborations – including a highly productive data-sharing collaboration with Scotland.

The success of CHeReL built on strong support and demand from the research community, combined with advocacy from champions within data custodian agencies, at a time when funding for health and medical research was increasing. Australian research groups have conducted numerous validation studies that have helped to counter the prevailing view among grant assessors that administrative data are of poor quality. Having a master linkage key which is continually updated with routine data has also been a critical success factor. CheReL's experience has shown that data custodian agencies can be cautious and hesitant about data linkage and sharing, particularly where this involves data from more than one jurisdiction. In some Australian jurisdictions, enabling legislation is absent, or has not recently been updated, and is therefore silent about data linkage. The Australian National Health and Medical Research Council is developing a series of principles which will provide guidance to aid data custodians' decision-making about data linkage and sharing.

Key findings: six key messages

1. **The relative importance of different barriers to linkage varies between high-income countries and low- and middle-income countries.** Data quality is a major barrier to data access in low- and middle-income countries (LMICs). Although data quality is also an issue in high-income countries (HICs), institutional issues were felt to be a more significant constraint by data managers in these settings. In contrast, such issues were much less frequently raised by LMIC practitioners. The case of South Africa (see page 5) suggests a natural progression from operational problems to institutional issues as processes for data linkage are embedded. Given the longer experience of HICs in managing and linking data, there may be gains to be made from sharing information about skills, data facilities and storage models, allowing LMICs to avoid some of the problems experienced by HICs.
2. **A sole reliance on narrow-informed consent is not a good basis for epidemiological research.** Where data are collected for research purposes, broad consent – which allows for beneficial uses of data which may currently be unforeseen – is practical and acceptable to the public. Public health researchers may also gain considerable value from accessing and linking data which has been collected for administrative or statistical purposes. For these types of data, it not usually practical to obtain consent and doing so may severely compromise statistical validity. Therefore, for both types of data, a practical exemption from narrow-informed consent is essential to enable high-quality high-benefit public health studies.
3. **There is a need to change the tone of the debate:** from the assumption that nothing can be released unless it is explicitly allowed, to a position where data are expected to be available for research unless this can be shown to be unlawful, unethical, or unachievable in a manner which protects confidentiality.
4. **Policy decisions are not always evidence-based – particularly when considering how research access to sensitive data is managed.** Many years of practice in the research community suggest that, despite theoretical concerns, the use and linkage of sensitive data is a very low risk activity when well-established best practice protocols are observed. However, this may not have been communicated well enough to external interested parties such as legislators or data depositors. Hence, the data management community may have inadvertently created a climate where research data access is viewed as high-risk and difficult to manage.
5. **Maintaining good relationships and trust is key to success.** Ensuring support from the public is critical, but we are starting from a strong position: the public in general are very supportive of health research, and this is closely related to their trust in the institutions concerned. Strong relationships with data depositors and research ethics committees are also key. For HICs, strong organisational links seem to make the difference with data depositors, whereas for LMICs personal links seem to matter more. In LMICs, the level of association with governments can also prove important, as there may be a higher risk of being linked to the ideals of a particular regime rather than working for the public good.
6. **Researchers may be reluctant to share data, and incentives for data linkage are weak.** Researchers can be part of the problem – they may be unwilling to make data accessible for linkage and other uses, even though many funders require it. Researchers may have spent many years developing data resources, and such efforts are not always rewarded in funding or publications. There are also few incentives to specialise or develop expertise in data management.

Recommendations

The report's recommendations to the Public Health Research Data Forum are largely concerned with distributing useful and accurate information to change ideas about data linkage and demonstrate its potential to interested parties. A common perspective from a critical mass of funders could substantially improve the environment for and practice of data linking.

The recommendations are grouped around two topics: setting the conceptual framework, and finding solutions to practical problems.

1. Set the conceptual framework and shape the debate

There is a need to change the general language of debate to make it more supportive of data linking, and provide the conceptual basis for strategic thinking on improved data access. The provision by the Forum of appropriate materials and exemplars, robustly underpinned by evidence, could go far to achieving the required conceptual change. Specifically, this should include:

- changing the language used when discussing data access from 'default-closed' to 'default-open'
- developing and promoting high-level principles for research access to data and data linking
- encouraging practitioners to share their knowledge and experience of effective risk management in research access
- developing a toolkit of coherent cases, backed by evidence, which can be used for advocacy purposes in policy discussions
- producing guidance on best practice ethics processes to encourage collaboration and co-operation.

2. Roll out practical solutions to address barriers to the wider use of data linkage

There are a series of practical steps through which funders could support researchers in developing data linkage activities. Specifically, funders should:

- encourage and support the use of remote technology to enable knowledge transfer between HICs and LMICs, particularly through collaborative working tools
- provide dedicated funding for the creation and management of data resources as a distinct element in research grants
- support PhD training programmes focused on data linkage and re-use as a cost-effective long-term investment to develop data expertise in LMICs and HICs
- produce guidelines for research teams on addressing practical issues in enabling data access and linkage
- build up a shared resource of useful precedents, experience and exemplars of data linkage initiatives.

The Wellcome Trust
The Wellcome Trust is a global charitable foundation dedicated to improving health. We support bright minds in science, the humanities and the social sciences, as well as education, public engagement and the application of research to medicine.

Our investment portfolio gives us the independence to support such transformative work as the sequencing and understanding of the human genome, research that established front-line drugs for malaria, and Wellcome Collection, our free venue for the incurably curious that explores medicine, life and art.

Wellcome Trust
Gibbs Building
215 Euston Road
London NW1 2BE, UK
T +44 (0)20 7611 8888
F +44 (0)20 7611 8545
E contact@wellcome.ac.uk
wellcome.ac.uk