

NCRIS Capability 5.7:

Population Health and Clinical Data Linkage

**National Collaborative
Research Infrastructure Strategy**

Issues Paper
July 2007

The intent of the National Collaborative Research Infrastructure Strategy (NCRIS) is to support the development of national research infrastructure in areas of strategic importance for Australia. Population health and clinical data linkage has been identified as an area of strategic importance.

This paper is intended to frame discussion in workshops on the development of the investment plan for the Health Data Linkage Capability.

It was prepared by:

Michael Frommer
Facilitator for the Health Data Linkage Capability
Director*

with:

Christine Madronio
Research Officer*

Sarah Kemp
Research Assistant*

Rebekah Jenkin
Senior Associate*

Rosa Reitano
Administration Officer*

*Sydney Health Projects Group
School of Public Health
The University of Sydney

What is data linkage?

Data linkage means joining two or more datasets so that it becomes possible to examine relationships between variables that are present in only one of the datasets.

The purpose of data linkage may be statistical or non-statistical. Statistical data linkage is done for the purpose of enhancing statistical analysis of datasets, e.g. the linkage of cancer registration data with mortality data to determine cancer survival rates. Non-statistical data linkage is done to increase the amount of data that is available on an individual, e.g. for clinical or administrative purposes.

The NCRIS Population Health and Clinical Data Linkage Capability is concerned only with statistical data linkage.

In principle, any datasets that contain information about individuals can be linked. The information about the individuals may include their identity, or they may be partially or fully de-identified. Linkage may be deterministic or probabilistic. Deterministic linkage relies on an exact matching of individuals, and depends on unique identifiers or full identification data. Probabilistic linkage uses algorithms that determine how likely it is that two or more records in different datasets refer to the same individual.

What is the value of data linkage?

Data linkage can give us a vast amount of information that would not be available by any other means.

Through data linkage, it is possible to analyse relationships among variables without collecting additional data. Linkage thus increases the utility of data that are already collected and reduces the need for duplicate data collection.

The yield from data linkage ultimately depends on the range and quality of the source datasets.

Australia has more long-term, stable, comprehensive, population-level datasets on health and health-related matters than almost any other country in the world. By making full use of these datasets through linkage, Australia is in a unique position internationally to make a major contribution to health, medicine and related aspects of human development and wellbeing.

What is the scope of the NCRIS Health Data Linkage Capability?

The following points indicate the proposed scope of the NCRIS Population Health and Clinical Data Linkage Capability.

National: an investment that has benefits across Australia; data linkage systems and methods that could be applied in different parts of Australia; linkage of national datasets; national standardisation of dataset structures, including meta-data; national resolution of legal, policy, confidentiality, privacy and ethical considerations that affect data linkage.

Collaborative: an investment that fosters collaboration among jurisdictions, institutions, researchers and the community; involvement of the public and private sectors; multi-sectoral and multi-disciplinary approaches to solutions for system and methodological problems in data linkage; multi-sectoral, multi-jurisdictional and multi-disciplinary approaches to the resolution of legal, policy, confidentiality, privacy and ethical considerations that affect data linkage.

Research: an investment in infrastructure for a wide range of different types of research conducted in academic, clinical, government and other institutional settings. The research infrastructure will apply to: academic population health and clinical research; population health monitoring and surveillance; inquiry for the purpose of policy and operational decisions in the health system; and research on relationships between health and other fields such as education, social services and the justice system.

Infrastructure: the physical and other resources that are needed to make research possible. These include: information and communication technology (ICT) and ICT support; acquisition and maintenance of research equipment; an appropriately trained research workforce; training programs for workforce development and renewal; data management and analytical capacity; guidelines for research ethics and for privacy and confidentiality; and effective mechanisms for governing and managing data linkage capacity. Infrastructure excludes resources devoted exclusively to specific research projects.

Strategy: a national investment plan that provides a conceptual framework, specifies a developmental program, estimates the resources needed to implement the plan, describes how these resources might be obtained, persuades relevant agencies to commit to implementation and work towards sustainable support. The strategy should also outline a proposal for evaluation.

Population health and clinical research: research content areas include health determinants, prevention, the full spectrum of clinical management of patients, the organisation and delivery of health services, health status and health outcomes, and relevant fields outside the health system.

What problems have to be solved on a national scale?

Many institutions and government agencies in Australia regularly undertake statistical data linkage. The methodological feasibility of data linkage is well established, and new developments have the potential to improve timeliness and efficiency.

The main issues to be resolved in realising the potential of data linkage are outlined below. Some of these issues overlap.

Legislation: many datasets that could make a major contribution to health and wellbeing are subject to specific legislation that authorises or mandates the collection of data from individuals and defines the conditions of data release and/or use. Health-related datasets are also covered by general legislation on information and privacy. The intent of legislation is: (1) to ensure the integrity and availability of high-quality data for a wide range of purposes that benefit individuals and the community; and/or (2) to protect the privacy and confidentiality of the individuals whose data are collected. Some of the legislation is Commonwealth, and some is in the province of the States and Territories. Many national datasets are compiled from jurisdictional datasets, and approval of the contributing jurisdictions must be sought for release of their own components from national data. The extent of the legislation and its complexity can create difficulties of interpretation with regard to the release of data for linkage projects. In some instances, the legislation was drafted at a time when data release from government agencies was not often required, and the concept of data linkage was not yet developed.

Policy: many datasets that could make a major contribution to health and wellbeing are also subject to government and/or institutional policy that defines the conditions of data release and/or use. Policy is of course consistent with legislation, and its intent parallels that of legislation. In addition, policy may reflect the capacity of an agency or institution to manage the processing of requests and carry out preparation of datasets for release. Policy tends to have a local rather than national purpose.

Privacy and ethics: in addition to legal and policy requirements, access to many datasets may also be subject to clearance by a registered human research ethics committee. Ethics committee review is invariably necessary for the release of identifiable data on individuals. Research projects that use identifying data must have ethics committee approval prior to submission of requests to data custodians. Because the datasets for linkage often come from two or more institutions or agencies, linkage projects are often reviewed by two or more ethics committees, which may follow different processes and reach different conclusions. While initiatives to streamline multi-site ethical clearance are being introduced, the ethics committee system is often cumbersome and introduces delays.

Custodianship: data custodians are required to implement legislation and policy in overseeing the datasets for which they are responsible. When faced with a request for data release, custodians have to interpret the request in the context of the legislation and policy, and manage the workload of preparing the data for release. Not surprisingly, custodians are sometimes conservative in their interpretation. The first priority of custodians is to satisfy the obligations of their employing institution or agency. They may also lack resources for processing applications and for preparation

of requested datasets. These factors tend to make access to the datasets seem difficult and perhaps restrictive for potential users.

Organisational capacity for access to data: as mentioned above, release of the requested parts of a dataset in accordance with a potential user's specification can pose a substantial workload. This consists of: (1) reviewing the request, clarifying the specifications, ensuring that the request is consistent with legislative and policy provisions, and determining whether ethical conditions have been met; (2) briefing the senior officer who has delegation for authorising release, and (3) preparing the data for release, with accompanying documentation. Institutions and agencies that hold datasets, and their custodians, often lack the resources for these tasks.

Expertise: data linkage requires expertise in three broad areas – a knowledge of the datasets to be linked and their limitations and idiosyncrasies, skills in data linkage methods and the use of linkage software or programs, and skills in statistical analysis and interpretation. By itself, a basic-level ability to use commercially-available linkage software is insufficient, because correct interpretation of a linked dataset depends on an understanding of the structure and content of and variations within the component datasets. Without this knowledge, the axiom 'garbage in, garbage out' applies. The expertise does not have to reside in one person, and teamwork can be employed. It is generally acknowledged that, Australia-wide, relatively few people have the requisite expertise, particularly at a level high enough to supervise or teach others.

Researcher engagement: researchers who request access to datasets are often unaware of the legislative and policy issues that determine access to data. Researchers are also often unaware of the organisational resources needed to assess and process a request for data release in connection with a linkage project. Their frustration is sometimes directed at custodians, and in some situations this has tended to impede cooperation.

What should the investment plan achieve?

The broad objective of the investment plan is to promote and facilitate data linkage as an effective mechanism for obtaining information from existing data collections that can be applied to improve health and wellbeing.

Specific objectives are as follows.

- Establish an organisational focus to coordinate national initiatives in data linkage (the 'coordinating unit'), and establish a governance framework that represents interested institutions, agencies and jurisdictions.
- Within this organisational focus, establish an initiative with appropriate expertise and resources, to resolve the governance issues that impede worthwhile data linkage projects which are likely to benefit individuals and/or the community. The resolution of these issues requires action at national, State and Territory, and institutional levels.

- Provide resources to enhance the capacity of institutions and agencies to process requests for access to datasets that they keep, prepare data for release, and monitor compliance with conditions of release.
- Support the further development of national and jurisdictional centres that have existing expertise in data linkage, building on their specific areas of expertise and experience.
- Foster and support the development of new jurisdictional and institutional data linkage units, encouraging new units to work with established units where feasible and efficacious.
- Establish a network of data linkage centres by offering formal affiliation with the ‘coordinating unit’ for the existing centres and new units.
- Promote the standardisation of components of datasets, such as meta-data, that facilitate data linkage.
- Support the development and evaluation of new or emerging data linkage systems, technologies and methods, and ensure that Australian data linkage practice accords with or exceeds international standards.
- Support training programs and educational forums in data linkage methods and systems, with the intent of developing a data linkage workforce that is adequate for national needs.

What should a national model for data linkage provide?

- A ‘coordinating unit’ that can foster collaboration and cooperation among data linkage units throughout Australia, resolve the governance issues that impede access to datasets for worthwhile data linkage projects, promote the standardisation of relevant components of datasets, foster effective communications among the various parties involved in data linkage, determine strategic directions for data linkage work in Australia, and receive, disburse and account for NCRIS funds.
- A network of affiliated data linkage units, building on existing centres of data linkage expertise and establishing new data linkage units where needed.
- An institutional legislative framework, or access to an existing such framework, that defines conditions for release and use of datasets containing information on individuals, and that is accepted and respected by all Australian jurisdictions.
- Streamlined systems for obtaining access to important national, State/Territory and institutional datasets. These systems should meet legislative, policy and ethical requirements, be appropriate for data custodians, and be ‘user-friendly’ for researchers.

- A focus of capability to obtain trusted access to national datasets and carry out data linkage involving these datasets.
- The development of rules and guidelines regarding ‘ownership’ and archiving of linked datasets, access to linked datasets, and intellectual property created through the analysis of linked datasets.
- A capability to foster and support inter-jurisdictional and/or inter-institutional collaboration to achieve particular data linkage program objectives, such as linkage involving Indigenous health data.
- Collaboration with major institutions and agencies that have an interest in health information and health research, e.g. the Australian Health Ministers’ Conference, the Australian Health Ministers’ Advisory Council, the Australian Bureau of Statistics, the National Health and Medical Research Council, the Australian Institute of Health and Welfare, and jurisdictional data linkage units, in promoting the application of data linkage to improve health and wellbeing.
- Engagement with major Australian initiatives relating to the data linkage systems and technologies, national data definitions and standards, privacy of personal information, authorisation and authentication standards for data access.
- The development of guidelines for data custodians and researchers on the data linkage process, encompassing all aspects from access to analysis to disposal of linked datasets, and enhancement of researchers’ awareness of the legislative, policy and ethical aspects of data linkage.

Possible points for discussion in workshops

- Are the proposed objectives appropriate and sufficiently comprehensive, given the scope of the NCRIS Population Health and Clinical Data Linkage Capability?
- Are the proposed elements and functions of the national model appropriate and sufficiently comprehensive?
- Is there a need to highlight any gaps, from the perspective of the jurisdiction in which the workshop is taking place?
- Does the jurisdiction have any specific aspirations that should be highlighted in the investment plan?