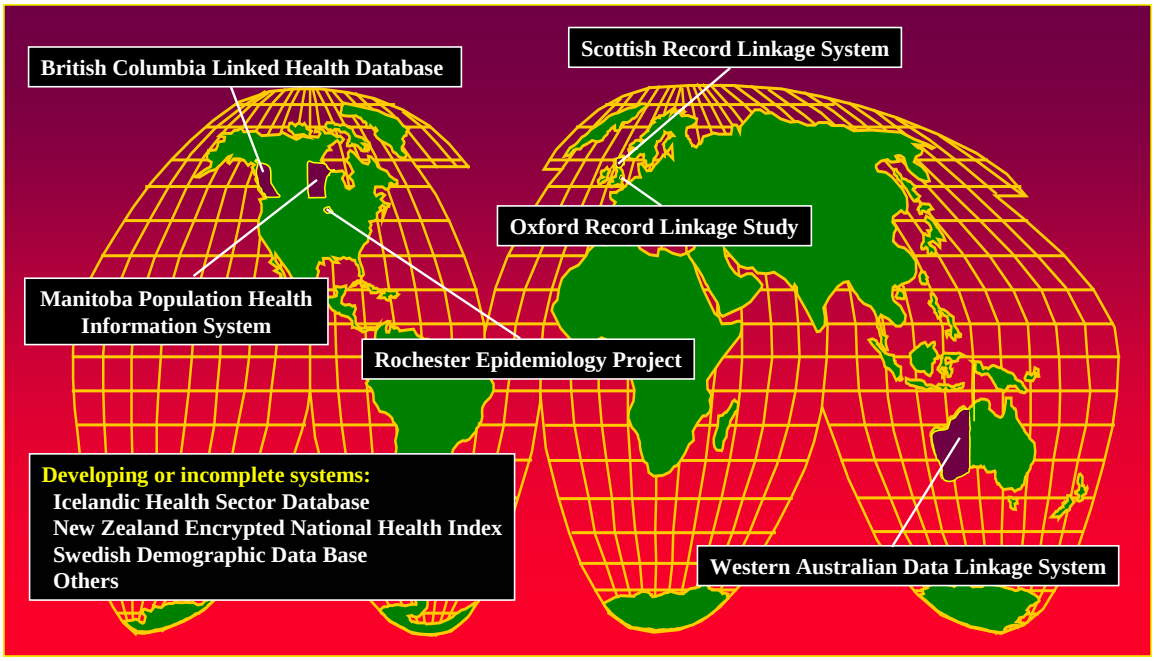
Record Linkage and Health Data – the Western Australian Data Linkage System (WADLS) and National Initiatives

The *Data Linkage Australia (DLA)* Project

Record linkage of health service data to allow the development of models to evaluate health service outcomes, particularly at the population level, is a priority of the Australian Government. The Western Australian Data Linkage System (WADLS) was established in 1995 and consists of over 30 population health datasets including the hospital morbidity discharge data, birth and death records, mental health services data, cancer registrations and midwives' notifications dating back to 1960s. Recent extensions of the record linkage project include data from the State Electoral Roll, and Commonwealth data sets such as Medicare, the Pharmaceutical Benefits Scheme and Aged Care, linked back to 1990 for all Western Australians.

Of the handful of comprehensive data linkage systems in existence today, as shown in the figure below, the WADLS offers the world's most comprehensive collection of data sets and robust system architecture available to researchers within the one system.



Through computerised probabilistic matching, the WADLS creates a dynamic master linkage key between over 30 population-based administrative and research health data collections. This means that the total historical population (more than three million individuals over 25 years of data collection) can be researched for all major diseases, risk factors, and outcomes of using health services. In its ten years of operation since 1995, WADLS has provided data linkage and related services to researchers from tertiary academic, community, health industry and government organisations, whose work has yielded such contributions to science as:

- over 600 distinct studies of disease aetiology, clinical needs analysis, patterns and costs of care, and outcomes of health services;
- over 800 scientific journal publications based on WA data linkage research; and
- Around 90 higher degrees, completed or in progress.

The WADLS has been the result of collaboration between the School of Population Health at The University of Western Australia, the Telethon Institute for Child Health Research, Curtin University of Technology, and the WA Department of Health. In 2004, these four agencies consolidated their partnership through the formation of a state-sponsored centre of excellence known as **Data Linkage Australia**. The **DLA** Centre brings together a multidisciplinary group of science and technology leaders with over 100 person-years of experience in developing and managing data linkage infrastructure to support health research.

Overview of Data Linkage

In 1946, Dr Halbert Dunn proposed the idea that each person on earth creates a 'Book of Life'. Starting with birth and ending in death, this book is composed of all the records of principal events in a person's life – birth, marriage, divorce, death and medical, pharmaceutical and educational records. Dunn developed the concept of collating the most important records of a person's life into a personal file and defined the process as record linkage. Predicting that record linkage would be of interest to many organisations in obtaining knowledge about their service programs, Dunn nevertheless surmised that linked statistical analysis of the records would be of most use to health and welfare organisations. He suggested also that linking files of data would establish the accuracy of the records. All of Dunn's predictions have been proven correct.

Definition and applications of data linkage

According to **DLA**, data linkage is the bringing together from two or more different sources, data that relate to the same individual, family, place or event. This definition has evolved somewhat from earlier versions. Substitution of the word 'data' for 'record' embraces a broader concept of information sources that goes well beyond old fashioned paper records and even electronic records. The data sources may include, for example, spatially referenced geographic information systems, where data are not necessarily structured as distinct 'records'. The increased utility of family linkage is also taken into account, in view of recent developments in genetic and molecular epidemiology.

Linkage of two or more data sets requires the use of unique or partial identifiers that are common to each set. In practice, this may involve taking each new record and comparing it to a master file of records or the linkage of two large data files *de novo*. Any of three general techniques are employed in the matching phase of data linkage:

i. *Unique matching* : Data are linked using unique identifiers, such as an insurance number. The files are sorted into the same order, and matched within blocks. This merging process is also known as *deterministic matching*, a term that belies the fact that the process may only identify 80-85% of the true matches due to recording errors.

ii. *Fuzzy Matching*: Partial identifiers such as names, date of birth, sex, postcode or place of birth (for individuals) are used for matching, allowing for a margin of error by linking data that are almost the same. Computer programs either present a choice of matches to the user or rely on a scoring system. Typically this technique identifies 85-90% of true matches. Another form of fuzzy matching involves the use of cyphers, taking specified characters from partial identifiers.

iii. *Probability Matching*: Also known as probabilistic matching, this is the most reliable technique, in which decision rules are built into a software package (i.e., automated computerised matching) based on the probability of two records being from different subjects having the same identifier. These probabilities are aggregated into a score and checked against a threshold to determine whether a link should be made. Such techniques typically identify 95-99% of true matches, while 1-2% of matches are false positives.

Increasingly, epidemiologists are studying diseases associated with behavioural and environmental exposures that occur over long periods of time and necessitate long term follow-up of individuals. Health services researchers are concerned with the evaluation of patterns of service provision or outcomes of interventions based on very large administrative data sets. Data linkage provides a valuable technology to facilitate these research activities. More specifically, the applications of data linkage include, but are not limited to:

i. *Disease surveillance*: e.g., to collate information on incident events from multiple sources or to identify first-time incident events.
ii. *Aetiological research*: e.g., identification of health events during the longitudinal follow-up of occupational cohorts.
iii. *Studies of health services utilisation*: e.g., patterns of care in clinical populations to identify differences in utilisation between population subgroups; or relationships between different characteristics of health services, such as between length of stay and readmissions.
iv. *Studies of health services outcomes*: e.g., safety research such as post-marketing surveillance for adverse effects of therapeutic substances or new procedural technology; or economic outcomes such as the impact of survival of low birth weight neonates on health care costs; or efficacy and prognostic research such as improvements in survival after cancer treatments. By enabling the evaluation of clinical outcomes, data linkage can be used to promote clinical best practice, and provides an opportunity to undertake analyses of new and existing treatments involving medical devices, drugs and surgical procedures.

Thus data linkage has applications in epidemiological research, health services research and health service management. It enables a 'wide-angle' view of population health and the health system, spanning across questions of disease causation and occurrence to matters concerning use of health services by different groups of patients and the long- term health outcomes that are achieved.

Benefits of data linkage

The advantages and social benefits of data linkage fall into several categories, some of which (e.g., the conservation of privacy) have come to be understood only in relatively recent times:

i. *Cost-efficiency of research*: Linking existing data can be a relatively cheap and effective alternative to performing *de novo* longitudinal studies and other more traditional approaches to epidemiologic and health services research. This conserves the limited resources available to a nation to support medical and health research, and enables more research to be performed with a given research budget.
ii. *Adding value to existing information assets*: Data linkage is a non-invasive and cost-effective means to generate a greater return on the substantial existing

investment in routine administrative data sets, which are often perceived simply as record-keeping systems. It also adds value to existing data sets through the quality improvement of data. Duplications errors in records are uncovered and other technical issues with the data often come to light, leading to measures to ensure greater accuracy of data recording at an administrative level.

iii. *Conservation of patient privacy*: The privacy of individual patients is conserved by reducing the need for release of names and other personal identifiers to researchers, because a major reason for the release of identifiers to researchers in the past (to find and clerically collate information on patients) is removed. Contrasted with consent-based responses to privacy concerns, data linkage systems are superior by virtue of conserving the privacy of all patients, regardless of whether they would have given consent to the use of their information. A consent-based approach can conserve the privacy only of those who decline to participate, and even then, this may be at the cost of irreconcilable impracticability of the research.

iv. *Community development*: A data linkage system enhances the levels of quality of interactions between researchers, clinicians, administrators, community and consumer groups and the mass media. It provides a focal point for envisioning future possibilities, for improved cooperation and for rigorous debate about the uses of data and the results of subsequent research. This strengthens a community by exercising the machinery necessary for teamwork and other organised responses to a community's problems.

v. *Contributions to medical and scientific knowledge*: A very considerable volume of published, peer-reviewed research results can be forthcoming when access to a population-based data linkage system is made readily available to *bona fide* researchers. The contributions to knowledge take the form of scientific publications and conference presentations, as well as theses and dissertations prepared by research trainees.

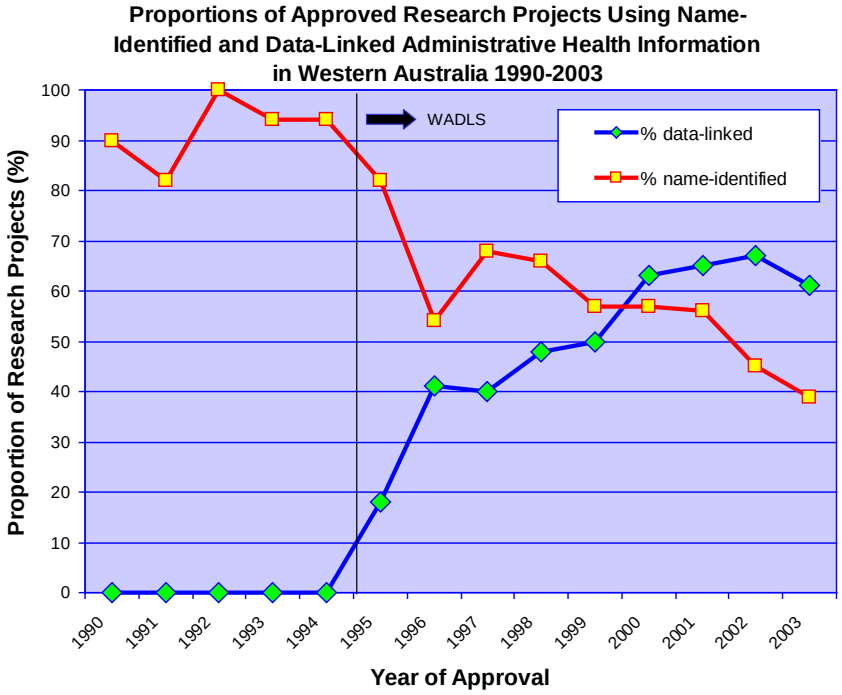
vi. *Improvements in population health*: These occur at least at two different levels. Contributions to knowledge published in peer-reviewed journals arising from the use of data linkage add to worldwide medical and scientific knowledge, which is eventually translated into better programs of preventive services, treatment and care for the general public. Specific local improvements in population health can also be traced directly to data linkage-based research, especially when conducted in the context of a strong framework for community development.

Commercial and other competitive benefits: The development of a data linkage system, when used effectively by a jurisdiction, can endow its research community with a considerable competitive advantage in attracting research funds relative to communities with no data linkage system. The resultant flows of income into the jurisdiction from outside sources can be substantial by science and technology industry standards. This revenue provides employment and stimulates the economy. A jurisdiction with a data linkage system may also be relatively more successful in attracting and retaining individual research leaders of particularly high quality. This enhances the intellectual life of the community and has positive flow-on benefits in the local capacities for policy development and quality improvement in the health system.

Privacy and confidentiality

The Australian community is now sensitised to privacy concerns, and legislative and regulatory impediments to medical and health research are increasing. A broader system of protocols is being developed progressively to address the concerns of consumers and data custodians with respect to privacy and data release. The WA Data Linkage Management Committee and the WA Cross-Jurisdictional Data Linkage Steering Committee both include consumer representation and protocols for best practice linkage, access and delivery are supported by health consumers, researchers and privacy professionals. A full-time consumer research liaison officer has also been employed to provide input on research agendas, community concerns and knowledge dissemination.

It has also recently been shown that the WADLS has significantly reduced the exposure of private and confidential person health information in WA, by confining access to personal details to a small linkage group who adhere to rigorous, strict privacy and confidentiality requirements. Research projects using named data provided by the DLU fell from 90% in 1990 to 36% in 2003. Personal details can therefore be removed from data released to researchers, while allowing research based on multiple data streams to be performed.



<u>National Collaborative Research Infrastructure Strategy (NCRIS) – Population Health Research Network (PHRN)</u>

The success of the WA DLS since 1996 has highlighted the need for essential national record linkage infrastructure to support health research. This need has been recognized in a recent national initiative within Australian Government's National Collaborative Research Infrastructure Strategy (NCRIS). Within NCRIS, the Population Health Research Network (PHRN) is developing data linkage infrastructure for health research, policy and planning in Australia. $A30 million has been provided by the NCRIS and a related initiative, with a further $A32 million in cash and in-kind provided by states/territories and academic partners. Data linkage units (nodes) are being established for all states/territories and will link together health-related information within their jurisdictions. A separate dedicated linkage

unit will facilitate linkages across jurisdictions and a national data access and delivery regime is being developed to streamline provision of information to support research, monitoring and policy evaluation. Key elements of the national initiative include:

1. The *PHRN Program Office* is leading implementation of the new data linkage infrastructure throughout Australia. This includes establishment and ongoing management of PHRN governance, contractual, policy and client services processes. The Program Office is managed by The University of Western Australia and located at the Telethon Institute for Child Health Research.
2. The *PHRN Centre for Data Linkage (CDL)* has been established to build a secure data linkage facility to facilitate linkage between jurisdictional datasets, and between these datasets and research datasets, using demographic data. The Centre will not hold these datasets, but will link the demographic data that has been separated from the remainder of each dataset to create 'linkage keys'. The CDL is managed by the Curtin University of Technology.
3. *A Proof of Concept Collaboration* has also been created to test the ability of the new linkage infrastructure to perform cross-jurisdictional linkages and provide linked de-identified data for research studies. The collaborative process and selected research projects are managed through the WA Node based at the Department of Health, Western Australia. "In-hospital and 30-day post-discharge mortality: Learning about quality of care using national data linkages" has been chosen as the first topic to investigate.

Australian National Disability Data

A key interest in the future direction of the WADLS and the PHRN initiative is to include disability data at a greater level than it currently exists within Australia for research.

There are two sources of Australian national disability data: the Australian Bureau of Statistics (ABS) through both census and disability surveys, and the Australian Institute of Health and Welfare (AIHW) which is the national body responsible for the collection of health and welfare statistics. National disability data collection in Australia serves particular descriptive, policy, and administrative objectives. The activities of the ABS and the AIHW make considerable use of the ICF in functional description. As yet, there appears to be little exploration of the use of datasets in research or longitudinal studies, nor in analysing impacts on identified groups of people with a disability, although the AIHW NMDS described below does have a data linkage key that lends itself to longitudinal studies of cohorts, but there is limited evidence of its use for that purpose. Existing national datasets focus on variables around diagnosis, functional capacity, disease states, and use of specialist disability services rather than broader health variables or access to health services. There are important initiatives in some state jurisdictions.

ABS Data sources

In addition to some questions relating to disability that are included in periodic national census, the ABS has carried out five national surveys between 1981 and 2003 that gathered sociodemographic data on people with a disability and also older people and carers of people with a disability. The most recent survey (ABS, 2004a, b) provides some comparison with surveys from previous years and with data from other groups. Data were gathered through interviews with a "responsible adult" within a

sample of households and also in establishments that provided accommodation care. ABS publications referenced below provide details on the survey methodology including sampling information. The definition of disability was consistent with the International Classification of Functioning, Disability and Health (ICF). Six areas of data were collected:

1. Impairments, long-term health conditions, and causes of disabling conditions.
2. Difficulties experienced in five areas of activities and the help required.
3. Five additional areas of activities for people aged over 60 years and people with a disability.
4. Type of assistance received for each activity, providers of assistance, and extent of need that was met.
5. Use of aids and equipment.
6. Access and use of computers and the Internet.

Both 1998 and 2003 surveys reported a disability rate in Australia of around 20% with a rate of profound or severe core-activity limitation between 6.4% and 6.3%. Physical conditions were the most common reported health conditions of people with a disability (84%) with the remaining 16% having a mental or behavioural disorder. Diagnostic groups were not identified or reported in the two ABS documents consulted.

AIHW

Within a broad range of responsibilities for the collection of health and welfare statistics in Australia, the AIHW provides comprehensive descriptive data on disability and diagnostic groups that include intellectual disability. Four focuses of AIHW are described briefly.

1. Service utilisation data

The AIHW has the brief to provide annual data on people with a disability that describe all use of specialist disability services funded by the Commonwealth, States, and Territories. Since 1991, a number of Commonwealth State/Territory Disability Agreements (CSTDA) have determined a range of government disability policies and practices, including the responsibilities of levels of government for funding of services. Since 1995, funded service providers have been required to provide annual data on service usage that is collected by the AIHW (the National Minimum Data Set (NMDS)). The most recent NMDS was reported in 2009 and referred to the 2007-08 fiscal year (AIHW, 2009). NMDS data items include service types (including accommodation, community support, community access, respite, and employment), the agency sector and location, hours worked by staff, times of operation, number of service users, selected service user identifiers, aspects of disabilities, and start and exit information by service users in specific services. Data are also provided on government disability service expenditure (not including income support). In 2007-08, for example, CSTDA funded services received $4.8b. Data are provided for each jurisdiction, for each service type, and indicate funding source. NMDS reports include detailed accounts of service activity for each jurisdiction. Disability type is reported. Over a five-year collection period, intellectual disability has been the most commonly reported primary disability, comprising nearly a third (32%) of service users in 2007-08. A substantial proportion of all service users aged 15-64 years (56.6%) were not participating in the labour market, reflecting a major source of concern across OECD countries. People with an intellectual disability were more likely to receive income support (the Disability Support Pension) (88%) than all disability groups (50-61%).

The NMDS defines support needs in a manner consistent with the ICF. People with an intellectual disability have higher support needs compared to other disability groups – 60% have severe communication limitations and are highly likely to have severe limitations in the three core activities of self-care, mobility and communication. The 2009 report stated that it is also important to consider the level of support needed in non-core activities that are associated with work and education.

2. Disability and health conditions

In 2004, the AIHW published a report on the relationships between health and disability (AIHW, 2004). The report examined existing Australian data and provided prevalence estimates of significant diseases and health conditions associated with disability, their severity, and the impacts on personal and environmental factors. It drew on data from three studies, an ABS 2001 National Health Survey, an AIHW 1999 study on the burden of disease and injury in Australia, and the 1998 ABS Survey of Disability, Ageing and Carers.

3. Intellectual disability in Australia

The AIHW reports on the status of various groups in Australia, including a report in 2008 on intellectual disability (AIHW, 2008). The report drew on ABS and AIHW NMDS data and focused particularly on transition from home to school and from school to adult life, prevalence of intellectual disability, and unmet demand for services. Population estimates of unmet demand underlined the relatively low level of participation by people with an intellectual disability in inclusive schooling and in employment.

4. Health and community services information systems

In 2006, the AIHW reported on the development of a module that:

- "Can be used to describe health status, outcomes of health interventions, and the need for assistance in areas of human functioning, and
- Enables the efficient and effective capture, storage and transmission of data on human functioning in a wide range of human service systems." (p. xi)

The module addressed the aim of establishing a measure of functioning that could monitor status at different points within the health care system.

The work focused on cardiovascular disease and acquired brain injury and carried out a mapping exercise of a range of existing functional assessment tools using the ICF as the mapping framework. This is consistent with the intentions of the WHO and international initiatives to incorporate the ICF framework in monitoring and measurement of outcomes for many groups including people with a disability.

The AIHW study was unable to reliably map the tools in a single data capture framework because of the many sources of variation. They developed a draft health outcomes module using the ICF framework that consisted of a four-matrix table for data capture with information on:

- "Body functioning qualified by extent of impairment.
- Body structure qualified by extent, location and nature.
- Performance in life areas qualified by support needed and satisfaction with participation.
- Environmental factors qualified by extent of influence." (P. 4)

The report anticipated further development and testing of the module.

References

ABS (2004a). *Disability, ageing and carers: Summary of Finding (ABS Catalogue No. 4430.0)*. Belconnen, ACT: ABS.

ABS (2004b). *Disability, ageing and carers, Australia: User Guide (ABS Catalogue No. 4431.0.001)*. Belconnen, ACT: ABS.

AIHW (2004). *Disability and its relationship to health conditions and other factors (Cat. No. DIS 37)*. Canberra, ACT: AIHW.

AIHW (2008). *Disability in Australia: intellectual disability (Bulletin 67, November 2008*). Canberra, ACT: AIHW.

AIHW (2009). *Disability support services 2007-08. National data on services provided under the Commonwealth State/Territory Disability Agreement (Cat. No. 56).* Canberra, ACT: AIHW.

AIHW (2006). *Â functioning and related health outcomes module. The development of a data capture tool for health and community services information systems (Cat. No. WP53).* Canberra, ACT: AIHW.

Errol Cocks MPsych PhD
Director
Centre for Research into Disability and Society
Curtin Health Innovation Research Centre

School of Occupational Therapy and Social Work
Curtin University of Technology

e.cocks@curtin.edu.au

James Semmens
Director
Centre for Population Health Research
Curtin Health Innovation Research Institute
Curtin University of Technolgy
james.semmens@curtin.edu.au

23 February 2010