**Community Services**
**Ministers' Advisory Council**

# Statistical Data Linkage in Community Services Data Collections

*A report prepared by the*
*Statistical Linkage Key Working Group*

May 2004

National Community Services Information Management Group
NCSIMG

A complete list of the Institute's publications is available from the Media and Publishing Unit, Australian Institute of Health and Welfare, GPO Box 570, Canberra ACT 2601, or via the Institute's web site at <http://www.aihw.gov.au>.

ISBN 1 74024 211 4

**Suggested Citation**

NCSIMG 2004. Statistical Data Linkage in Community Services Data Collection. Canberra: Australian Institute of Health and Welfare



*Any enquiries about or comments on this publication should be directed to:*

**Margaret Fisher**
Australian Institute of Health and Welfare
GPO Box 570, Canberra ACT 2601
**Telephone:** (02) 6244 1033

Published by Australian Institute of Health and Welfare

Designed by Spectrum Graphics

# Contents

# List of tables



# List of figures

# Preface and Acknowledgments

*This report is the outcome of a major project conducted by the National Community Services Information Management Group with the financial support of the Community Services Ministers' Advisory Council. The potential benefits of statistical linkage for policy, research and planning purposes are of increasing interest across jurisdictions and in the health and community services sectors. The report raises key issues for wider consideration and discussion concerning the appropriate conduct of statistical linkage projects using community services data and the appropriate mechanisms for the protection of individual privacy.*

*The report was endorsed by the Community Services Ministers' Advisory Council in March 2002 and published electronically, inviting further comment. This version updates the initial version and was endorsed in July 2003 by NCSIMG.*

*The main authors of the report were Mr D'Arcy Jackson and Dr John Bass from the Department of Health and Ageing. Other contributors from the Statistical Linkage Key Working Group included Dr Ching Choi from the Australian Institute of Health and Welfare, Mr Paul Basso from the Department of Health Services in South Australia and Mr Andrew Stuart and Mr Mark Thomann from the Department of Health and Ageing. Thanks also go to all members of (and contributors to) the Statistical Linkage Key Working Group and the wider National Community Services Information Management Group for comments received on the report.*

# Statistical Data Linkage in Community Services Data Collections

## Executive summary

### Background

This report explores issues associated with developing and implementing a statistical linkage key (SLK) process and mechanism for the analysis of linked data across community service sector data sets.

This project was conducted on behalf of the National Community Services Information Management Group (NCSIMG) by the Statistical Linkage Key Working Group (SLKWG). The SLKWG was chaired by the Commonwealth Department of Health and Ageing and included representatives from the Australian Institute of Health and Welfare, Department of Family and Community Services, Australian Bureau of Statistics, Ageing, Disability, and Home Care Department (NSW) and Department of Health Services (SA).

The report deals with the following issues:

- *the need for and possible uses of a statistical linkage key;*
- *the current status of the use of statistical linkage keys in statistical and policy analysis work;*
- *current Commonwealth and State/Territory privacy laws in relation to data collection, storage/dissemination and statistical linkage;*
- *the privacy issues and the protocols that might be required to address them;*
- *the operation of linkage keys currently in use in the Supported Accommodation Assistance Program (SAAP) National Data Collection, the Commonwealth/State Disability Agreement (CSDA) Data Collection, and the Home and Community Care (HACC) National Minimum Data Set;*
- *the issues, including possible protocols, involved in linking data across community service programs and sectors; and*
- *recommendations on the feasibility of adopting an SLK process (with the associated framework to do so) across the community services sector.*

## Main findings and recommendations

### Framework for data linkage

Data linkage can be undertaken for two main purposes:

- *linkage for individual client management purposes; or*
- *linkage for statistical, research and policy purposes.*

Linkage for statistical research, planning or policy purposes is conceptually distinct from linkage for individual client management purposes. This report only addresses linkage for statistical research, planning or policy. Data that are statistically linked for research and policy purposes should not be used subsequently for individual client management purposes, especially where this might deprive an individual of a service or benefit. If a statistical linkage project were to identify possible individual client management issues, they would need to be followed up by other more precise and appropriate methods.

The use of statistical linkage keys must be accompanied by adequate safeguards and protocols (such as the proposed protocol identified in Section 7.3) to ensure the individual client's privacy is protected during the statistical linkage and analysis process.

**Recommendation 1:**
***The NCSIMG endorse the use of statistical linkage methodologies for research, planning and policy analysis.***

**Recommendation 2:**
***The NCSIMG endorse the principle that data collections produced by linkage for statistical and research purposes should not be used subsequently for client management purposes.***

### Statistical linkage methods

The statistical linkage of data can be conducted using two main methods:

- *'deterministic' linkage — involving the exact, one-to-one character matching of linkage variable(s) across two or more data collections; or*
- *'probabilistic' linkage — involving the researcher making assumptions on the basis of probability as to which records should be included in the combined data file and attributed to one individual record.*

Existing community service linkage keys (for example, HACC, SAAP) are linked using deterministic methods, based on an SLK derived at the point of data collection. Probabilistic methods can be used where more detailed demographic information on individual clients is available from each agency participating in the linkage project, and can lead to a much better linkage of disparate data records of varying data quality that potentially relate to the same person.

Measures of the effectiveness of SLKs have tended to focus on how well the target population is represented by the key (that is, levels of participation or consent to linkage) or measures of the accuracy of the key (that is, in relation to the number of duplicate keys created). A comparison of the effectiveness of SLKs using both deterministic and probabilistic linkage methods has been undertaken, including the effect these keys have on the analysis of results of linked data.

The results show that the type of SLK used can significantly affect the results obtained through the analysis of linked data, especially where either longitudinal data are used or small client groups are subject to analysis. Every statistical linkage proposal therefore needs to consider whether linkage using a particular key is sufficiently accurate for the data analysis envisaged, as some linkage/analysis combinations lead to results that are less accurate than matching on full demographic data.

It is sometimes assumed that the existing SLKs in the community services sector by themselves provide adequate protection for the personal information of clients. While this is true to a limited extent, the main protections and safeguards to personal information are offered through protocols that specify in detail the use and applications of the linked data. The proposed protocol outlined in Section 7.3 of this report is therefore of primary importance, and covers issues relating to client consent, purpose and usage of linked data, role of the data repository, access and sharing of data between agencies and data security issues (that is, use of encryption).

The protocol outlines a possible framework for conducting statistical linkage projects in the community services sector. It has been developed around three main stages of the life of a statistical linkage project, namely the:

- *pre-linkage phase;*
- *statistical linkage of data; and*
- *post-linkage phase (research applications).*

The protocol employs a number of assumptions regarding the agencies participating in the statistical data linkage, the type of linkage and encryption algorithms used, the use of the data repository to perform the analyses, staff skills and choice of linkage methodology. Some of these assumptions may not be relevant to some agencies participating in a linkage project, especially where they may already have robust privacy and ethics processes/structures. The SLKWG intends the proposed protocol to be used as a framework by which specific agencies may guide their development of relevant and appropriate statistical linkage management processes and protocols to suit their particular circumstance.

The basic scenario underpinning the protocol discussed in Section 7.3 is one-off linkage of administrative by-product data across government agencies either within or between jurisdictions. Linkage between administrative and population survey data sets has not been considered and is outside the scope of this report.

**Recommendation 3:**

***The NCSIMG acknowledge the need for a statistical data linkage protocol and:***

*(a) notes the proposed draft protocol outlined in Section 7.3 provides a proposed framework for statistical linkage projects in the community services sector and is intended to guide the development of SLK projects, rather than to prescribe a set methodology and process for undertaking such projects;*

*(b) request the jurisdictions represented to assess the impact of the proposed protocol and report to enable their finalisation;*

*(c) refer the protocol and the report to the National Community Services Data Committee for its consideration.*

**Recommendation 4:**

***The NCSIMG note that in some instances the use of a third party data repository in community services sector statistical linkage projects may be desirable (for example, for cross-jurisdictional statistical data linkages) and their use should be formally considered by each statistical linkage project.***

**Recommendation 5:**

***The NCSIMG recognise that the linkage of data is context specific, and there is no one preferred method for statistical data linkage. Where possible, the use of full demographic data is appropriate for statistical linkage, but this does not preclude the use of more limited linkage methods.***

**Recommendation 6:**

***The NCSIMG recognise that security of data in transmission between agencies and any third party data repository is essential, and that the encryption of an SLK provides one option to ensure this security.***

**Privacy and legal considerations**

The investigation of statistical linkage key methodologies requires the issues surrounding client confidentiality and privacy to be addressed carefully. Existing privacy legislation, safeguards and protocols have been investigated to determine how well they provide protection of the rights of community service program clients and affect the development and use of a statistical linkage process.

While complex, the privacy and legislative issues identified in this report can be addressed successfully by agencies considering implementing linkage key methodologies for statistical and research purposes. The key issue for each agency will be in accepting responsibility to ensure that linkage project(s) are implemented using the safeguards and protocols necessary to minimise the chances that individual information can be identified, and to meet the requirements of relevant privacy legislation and principles.

It is not possible to provide a definitive statement of the privacy and confidentiality implications of all statistical linkage projects across all community services sector data collections. Each project will need to be considered on a case-by-case basis by the participating agencies, seeking specific legal advice as required. The suggested protocol aims to assist agencies to ensure that relevant privacy and confidentiality issues are addressed. Responsibility, however, rests with each agency to ensure that record linkage is undertaken in a manner consistent with existing legislation, including privacy legislation.

**Recommendation 7:**
***The NCSIMG recognise that the privacy, client consultation and legal implications of each statistical linkage project will have to be identified, assessed and resolved on a case-by-case basis by the relevant steering committee (and ethics committee) involved in each project.***

**Recommendation 8:**
***The NCSIMG recommend to member agencies considering participating in statistical linkage projects that they review the purposes under which clients contribute data to their agency. The review should seek to ensure that the potential use of information for research and planning purposes (based on statistical linkage) is made clear to clients.***

**Recommendation 9:**
***The NCSIMG recommend to agencies currently using SLK methodologies that the privacy and legal implications of their projects are considered in the light of the issues raised in this report, and in a manner consistent with relevant legislation, such as the Privacy Act 1988***

### Engagement with the community

One of the most important issues considered by the SLKWG in relation to privacy issues involved the need for any future statistical linkage methodology to be undertaken with the greatest possible degree of transparency and openness. This should involve the active involvement and representation of community services consumer groups in the development and implementation of statistical linkage projects as a prerequisite to any project going forward. The roles of these groups, and the opportunities for their input to inform the development of a statistical linkage project, have been outlined in the proposed protocol.

Consumer representation may not be required at some of the more specific or detailed stages of the linkage project (for example, the technical or statistical aspects of potential linkage projects). However, consumer representation is suggested at the more strategic or management level of each project. The potential benefits to a statistical linkage project of consumer involvement far outweigh any initial difficulties which may be encountered in engaging consumer interest and representation in the management of such a project.

**Recommendation 10:**
***The NCSIMG endorse the involvement of relevant community sector consumer representatives in the development and implementation of statistical linkage projects. The appropriate level of involvement will be determined by the relevant steering committee, and mechanisms built into each project's work program (for example, memorandum of understanding).***

**Recommendation 11:**
***The NCSIMG acknowledge that participation and education of both community services sector agencies and consumers are important to the successful implementation of statistical linkage in the sector.***

### Coordination with the health sector

Considerable progress has been made in recent years between the Commonwealth and some States involving direct linkage of disparate health data collections to improve the information available for research for both administrative and statistical purposes.

Health data linkage projects are well developed and have been implemented across both health and community services sectors (for example, hospital, disability services and aged care services).

Protocols and draft guidelines are currently being developed within the health sector that can further inform the development of these methods within the community service sector.

**Recommendation 12:**
***The NCSIMG acknowledge that the issues in implementing statistical linkage projects for research purposes in the community services sector are in many cases the same as those being considered by the health sector.***

**Recommendation 13:**
***The NCSIMG considers that the further development of statistical linkage methodology for the community services sector should occur in close consultation with similar developments in the health sector.***

**Recommendation 14:**
***The NCSIMG seek to cooperate with the health sector (possibly through the Statistical Information Management Committee) where relevant infrastructure (for example, ethics committees, data repositories) or expertise can be shared, to facilitate efficient and appropriate linkage implementation across both sectors.***

# 1 Introduction

This report explores the issues associated with developing and implementing a statistical linkage key (SLK) process and mechanism for the analysis of data across the community services sector. The goal of the project has been to outline for the National Community Services Information Management Group (NCSIMG) a series of recommendations regarding the feasibility of implementing an SLK process for statistical and policy analysis across different data sets and jurisdictions within the sector. These recommendations are supported by protocols and strategies to ensure that the relevant privacy issues, legislative frameworks and client consultative mechanisms are addressed in any recommended SLK process.

The objective of the report is to provide recommendations on the feasibility of adopting statistical linkage methodologies (including protocols). While this necessitates an analysis and assessment of existing statistical linkage keys used in the community care sector, the report does not provide a recommendation on a single key for use across all community sector agencies. This approach received wide support from the community services sector during preparation of the report.

## 1.1 Project administration

In March 2000 the Community Services Ministers' Advisory Committee (CSMAC) agreed to fund an NCSIMG project to investigate the issues surrounding the development of a statistical data linkage key for the community services sector.

The project has been managed on behalf of NCSIMG by the Statistical Linkage Key Working Group (SLKWG). Members of the SLKWG are identified at Appendix A. The Working Group included representatives from the Australian Institute of Health and Welfare, Department of Family and Community Services, Australian Bureau of Statistics, Ageing, Disability & Home Care Department, (NSW), the Department of Health Services (SA) and was chaired by the Commonwealth Department of Health and Ageing.

The role of the SLKWG has been to investigate and report on each of the tasks identified in the CSMAC project budget brief relating to the development of an SLK process. During the course of this work, the SLKWG has also contracted an independent data linkage expert to provide it with advice, information and guidance in relation to related developments in the health sector and to conduct detailed analyses of the effectiveness of SLKs currently in use in the community services sector. The final report for the project has been prepared on behalf of the SLKWG by Mr D'Arcy Jackson and the contracted consultant, Dr John Bass.

## 1.2 Project rationale

Most government agencies[1] are exploring opportunities to better coordinate program delivery across programs and jurisdictional boundaries. There are significant policy development opportunities available through statistical analysis of the client

[1] Throughout this report, the term 'agency' or 'agencies' is used to refer primarily to Commonwealth or State/Territory Government departments. While most of the issues raised apply universally to non-government community sector organisations and educational/research institutions (for example, Universities), the report is based principally from the perspective of a government agency.

groups assisted by different community service programs. This includes assessing the extent and pattern of service use across programs and the identification of gaps and/or overlaps in service provision.

The project has developed due to a recognition from the NCSIMG that there is great potential for better quality information to be derived from data collections currently held across separate jurisdictions through data linkage for statistical analysis. The information that could be gained from quantifying the flow and patterns of use of client groups accessing community services across the sector would provide NCSIMG members with a better basis for policy analysis, planning and evaluation.

Groups of community service clients often use more than one service over time, or use different community services concurrently. Government and non-government community service providers often supply more than one service and frequently receive funding from more than one government source.

Currently, community service agencies hold discrete administrative data collections that cover services provided to their clients within the boundaries of their program of responsibility. This leaves policy makers and researchers from each of the different agencies with a fragmented and incomplete picture of an individual's overall service utilisation across the sector. This hinders the development of public policy to improve services over time and across programs offered by different agencies. The formulation of policy to provide better 'joined up' assistance for individuals requires as a starting point the availability of better 'joined up' information on service usage. Statistical linkage key methodologies provide one way of joining up de-identified data across data collections (and across programs/jurisdictions) to provide better statistical information and inform public policy decisions and analysis.

Statistical linkage for policy and research purposes is a distinct and separate process from other uses of data linkage, such as data matching for administrative or case management purposes. Data linkage involving one-to-one 'data matching' methods for individual client management purposes is primarily focused on sharing an (identified) individual client service information across different providers to ensure better coordination and continuity of care for that particular client. This form of data linkage is not the focus of this project.

The investigation of the feasibility of developing and implementing a statistical linkage key mechanism or process needs to address carefully the issues surrounding client confidentiality and privacy. The investigation of these issues is extremely important in determining the feasibility of any proposal to adopt a linkage key across community services data collections. Existing privacy legislation, safeguards and protocols are investigated to determine how they provide protection of the rights of community service program clients and affect the development and use of a statistical linkage process.

## 1.3 Objective of the report

The objective of the project has been to provide to the NCSIMG a consolidated report that:

- *analyses the need for and possible uses of statistical linkage keys;*
- *identifies the current status of the use of statistical linkage keys in statistical and policy analysis work;*
- *reviews current Commonwealth and State/Territory privacy laws in relation to data collection, storage/dissemination and statistical linkage;*
- *analyses the privacy issues and the protocols that might be required to address these issues;*
- *examines the operation of linkage keys currently in use in the Supported Accommodation Assistance Program (SAAP) National Data Collection, the Commonwealth/State Disability Agreement (CSDA) Data Collection, and the Home and Community Care (HACC) National Minimum Data Set;*
- *analyses the issues, including possible protocols, involved in linking data across community service programs and sectors; and*
- *provides a recommendation on the feasibility of adopting an SLK process (with the associated framework to do so) across the community services sector.*

The SLKWG would like to emphasise that the objective of the report is to provide recommendations on the feasibility of adopting statistical linkage methodologies (including protocols). While this work necessitates an analysis and assessment of existing statistical linkage keys used in the community services sector, the SLKWG does not intend to provide a recommendation on a single key for use across all community services agencies. As will be seen in later sections, the statistical linkage key itself is of secondary importance relative to the structures and processes outlined in the protocols governing the statistical linkage methodology.

The review of the Commonwealth and State privacy laws included in this report is also not intended to provide a definitive overview of all the legislative requirements relating to privacy and statistical data linkage by and between Commonwealth agencies, State and Territory agencies and private sector organisations. The evolution of both Commonwealth and State legislation regarding privacy issues necessitates that any review will be limited to the environment as it was at the time of the review. Agencies considering linkage projects in the future will need to consider in detail the implications of the legislation that is current at the time of the project's development.

The review of the Commonwealth and State and Territory privacy laws included in this report is also not intended to provide a definitive overview of all the legislative requirements relating to privacy and statistical data linkage by and between Commonwealth agencies, State and Territory agencies and private sector organisations. The evolution of both Commonwealth and State and Territory legislation regarding privacy issues necessitates that any review will be limited to the environment as it was at the time of the review. Agencies considering linkage projects in the future will need to consider in detail the implications of the legislation that is current at the time of the project's development

# 2 Data linkage methods

It is important to define from the beginning the conceptual framework that the SLKWG has used to develop the issues concerning statistical linkage for the community services sector data collections. The SLKWG views data linkage methodologies as a continuum, ranging from the more generally understood and widely used methods involved in the exact 'data matching' of unique and identified records, to the less well understood statistical linkage processes the current report deals with. This continuum is briefly outlined below.

## 2.1 Purposes of data linkage

Data linkage refers to the bringing together of data from different sources in order to obtain a greater understanding of a situation or individual from the combined (or linked) data set. Data linkage is usually undertaken for two main purposes as described below. The current work of the SLKWG is centred on the second of these purposes.

### 2.1.1 Linkage for client management purposes

The first purpose relates to the linkage of an individual's data records across collections to examine or analyse the details of that individual. The linkage must be as specific and accurate as possible to ensure that each linked data record belongs to the single unit being analysed.

A common example of a variable used for this form of linkage would be an individual's tax file number or an agency-specific identifier (such as a Centrelink customer reference number). For example, an individual's social security benefit details may be linked against data held by the Australian Taxation Office for audit and fraud control reasons. In terms of client management, an individual's health record may be linked between a range of identified health providers (for example, hospitals) to assist these providers to effectively treat the individual.

Linkage of data for client management purposes can also be used to inform research and statistical analysis of an individual's records as part of wider client groups. However, its primary aim is to identify and inform some aspect of an individual record for purposes relating to the provision of service to that individual.

The potential for errors to occur in such linkage of an individual's data needs to be kept to a minimum, as a missing (or incorrectly linked) record can have major significance on the case management or treatment of the individual. The informed consent of the individual to this form of data linkage is usually a necessity.

### 2.1.2 Linkage for statistical, research and policy purposes

The second main purpose for which data linkage occurs is for organisations to gain a better understanding of the patterns of service use by ***groups*** of clients for research, statistical or policy analysis, planning and evaluation purposes. The use of data for these 'statistical' linkage purposes allows organisations to make full use of the extensive data collections already held to gain new information (at relatively little extra cost) on the access and use of their services by client groups.

For 'statistical' linkage purposes, the individual unit (that is, an individual's service experience) is important only in terms of its contribution to the pattern of use of the client group overall. As such, the identity of the individual unit is unimportant for 'statistical' linkage. Within the health sector, an example could be an analysis of the service use (say, admissions) of an entity such as a hospital by all specified clients (say, emergency cases) during a specific time period (say, the last financial year).

In statistical linkage, the variable(s) used to combine separate data collections for research purposes are usually not pre-defined. The linking variables can be constructed by agencies from as much data as is available from the collections for each client to produce the best possible linkage.

This report considers in detail the data linkage of client records for statistical, research and policy uses.

### 2.1.3 Comparison of linkage for client management and statistical research

As detailed in the previous section, the data linkage activity undertaken by an organisation is usually for one or two intrinsically different purposes. With the 'statistical' linkage application, which is the focus of this report:

- *the purpose and use of the combined data is for policy analysis and research (not client management);*
- *the focus of analysis is the group of clients (not the individual);*
- *the required level of accuracy required in the linkage process is lower; and*
- *the identity of the individual or entity is unimportant and not retained in the combined data collection.*

In practical terms, this means that 'statistical' linkage can often be performed within a shorter timeframe and at a lower financial cost than linkage for client administration purposes. A comparison of these two purposes for undertaking data linkage is presented in Table 1.

**Table 1: Comparison of data linkage methods — client management and statistical**

| | Client management | Statistical |
|---|---|---|
| **Type of use** | Client management. | Statistical analysis, research, policy analysis and/or evaluation. |
| **Examples of usage** | Tracking a client through housing records to verify social security claims. | Investigating the use of assisted housing by disabled persons to improve planning for future needs. |
| | Tracking clients through a variety of services to provide optimal management. | Analysing patterns of client movements between acute care hospitals and aged care facilities. |
| **Linkage quality** | Number of errors must be minimal. | A small but known proportion of errors is acceptable. |
| **Timeliness** | Data from all sources need to be complete and up-to-date. | Stability is important with less emphasis on data being up-to-date. Data 'snapshots' are preferable for analysis. |
| **Privacy and confidentiality** | Data must be accessible by one or more personal identifiers (for example, name and address, Medicare number, or a 'healthcare personal identification number'). | Personal identifiers are unnecessary; an arbitrary number is sufficient to group the information in de-identified linked data files. |
| | A significant number of users may need to access at least some of each individual's data. | A limited number of researchers would be the only people with access to unit record data. |
| | Individual consent is required. | Individual consent to linkage is not necessary where de-identified data are involved. |
| | Extremely sensitive to privacy concerns. | De-identification and aggregation reduce concerns about privacy. |
| **Implementation** | Can be complicated, politically and administratively. | Relatively simple and inexpensive, both technically and politically. |

## 2.2 Applications of 'statistically' linked data

Data that has been produced by linkage for statistical and research purposes should *not* be used subsequently for client management purposes.

The conceptual separation of 'statistical' linkage from client management uses needs to be a fundamental tenet of the approach of the NCSIMG to data linkage in the community services sector. Data produced by linkage for statistical research purposes should never be used for client management purposes. This is not only because the combined data may be inaccurate and refer to the experiences of more than one client, but also because such linkage is permitted only for statistical and research purposes.

This distinction is particularly important in the discussion of privacy and confidentiality issues associated with statistical linkage keys. This is dealt with in later sections.

The commitment of the NCSIMG to this conceptual separation is important. If the use of data linkage projects becomes more widespread within the community services sector, this commitment will be tested when administrative matters or irregularities (that is, fraudulent client or service behaviour) become apparent during the examination

of statistically linked files. While such irregularities could be reported in the analysis of the linked data, the linked data itself must not be used to identify the individual(s) or services involved or to administratively pursue the matter any further. Once identified, the issue would need to be considered by the relevant agency and, if it were to be investigated further, followed up using other processes and means.

## 2.3 Benefits of statistical linkage

The SLKWG recognises that before any statistical linkage project begins, it is important for researchers to have clearly identified the purpose, use and potential policy benefits of undertaking the statistical linkage.

For example, the SLKWG has identified the following range of situations in which a statistical linkage key methodology could provide substantial benefits in consideration of common policy questions:

- *identification of any gaps or overlaps in service provision between programs (or across agencies);*
- *identification of the progression pathway of client groups through community services programs;*
- *ability to look at the range of government programs offered by different agencies from the client's point of view; and*
- *ability to assess the (intended or unintended) impacts of one program on another.*

Too often the perceived 'dangers' of data linkage projects are over-emphasised, with the positive and far-reaching benefits of many data linkage projects under-emphasised.

There are many instances in the health sector where data linkage projects have improved the quality of the lives of and services delivered to clients. Some examples are briefly described here from the area of perinatal health in Western Australia.

### 2.3.1 Folate and neural tube defects

Linked records in the Western Australia Birth Defects Registry and the Maternal and Child Health Research Data Base (MCHRDB) are used to monitor trends in neural tube defects in that State. Since the introduction of statewide health promotion activities and voluntary fortification of foods nationally with folic acid, aimed at increasing the folate intake of women of childbearing age for the prevention of these serious birth defects, there has been a 30% fall in neural tube defects in Western Australia.

### 2.3.2 IVF and birth defects

Linkage of four population-based registers — the Reproductive Technology Register, the Birth Defects Registry, the Midwives' Notification of Birth System, and the MCHRDB — was undertaken to compare the estimates of prevalence of birth defects in infants born following in-vitro fertilisation (IVF) techniques with naturally conceived infants. This design was able to overcome the major methodological flaws of previous studies (including small sample sizes, biased samples, inappropriate comparison data, differing classification systems, and inability to control for potential confounding), and found that IVF infants were more than twice as likely to have a major

birth defect diagnosed by one year of age compared with naturally conceived infants. These results have important implications for counselling couples embarking on assisted conception treatment.

### 2.3.3 Sudden infant death syndrome

Using linked data from the MCHRDB, the trends in mortality from sudden infant death syndrome (SIDS) by Indigenous status were documented for the first time in Australia. They showed the fall in SIDS in non-Indigenous infants following the introduction of the *SIDS Reducing the Risks* campaign (in particular encouraging parents to avoid the prone sleeping position for their infants). This analysis also documented a six-fold greater risk of SIDS for Indigenous infants, which was little affected by the *SIDS Reducing the Risks* campaign. This has led to further research investigating risk factors for SIDS in these children.

### 2.3.4 Cerebral palsy

Analyses of data from the Cerebral Palsy Register of Western Australia and the MCHRDB have documented that only a small proportion of cerebral palsy can be attributed to adverse events occurring around the time of birth. This has important implications for determining liability of obstetricians, and has contributed to the development of an international consensus statement on the relationship between acute intrapartum events and cerebral palsy.

## Summary of key points

- *Data linkage is usually undertaken for two main purposes, namely:*
  - *linkage for client management purposes; or*
  - *linkage for statistical, research and policy purposes.*
- *Linkage for statistical, research, planning or policy purposes is distinct from linkage for client management purposes.*
- *Statistically linked data should not be used for client management purposes.*
- *Statistical linkage projects have the capacity to:*
  - *identify gaps or overlaps in service provision between programs (or across agencies);*
  - *identify the progression pathway of client groups through community services programs;*
  - *look at the range of government programs offered by different agencies from the client's point of view; and*
  - *assess the (intended or unintended) impacts of one program on another.*
- *The potential benefits of data linkage projects, and statistical linkage methodologies, have been demonstrated widely in the health sector and should not be under-emphasised.*

# 3 Methods of statistical linkage

Linkage for statistical or research purposes can be undertaken using one of two distinct 'linkage' methodologies. Each methodology is described below, using mock data to highlight the differences between 'deterministic' and 'probabilistic' linkage methodologies.

**Example data**

| | | | | | |
|---|---|---|---|---|---|
| 1 | LONGFORD | BRUCE | PATRICK | 07/02/1944 | MALE |
| 2 | LONGFORD | BRUCE | PATRICK | 07/02/1944 | MALE |
| 3 | LONGFORD | PATRICK | BRUCE | 07/02/1944 | MALE |
| 4 | LONGFORD | BRUCE | PATRICK | 07/02/1943 | MALE |
| 5 | LONGFORD | BRUCE | PATRICK | 07/02/1944 | FEMALE |
| 6 | LANGFORD | BRUCE | PATRICK | 07/02/1944 | MALE |

The personal identifying data that is available in these mock records to a researcher considering a 'statistical' linkage application here includes surname, first and second given names, date of birth and gender.

## 3.1 Deterministic methodology

Linkage of these mock records by a researcher for 'statistical' purposes using a deterministic methodology would simply involve the exact matching of the available personal demographic data from each of the six records. Using the mock records, the exact matching of these variables would provide the researcher with a link for records 1 and 2 only. The records for data collections 3, 4, 5 and 6 would all remain unlinked due to minor (and realistic) differences in the name and/or demographic details from each of the data collection source records.

**Linked records — deterministic methodology**

| | | | | | |
|---|---|---|---|---|---|
| 1 | LONGFORD | BRUCE | PATRICK | 07/02/1944 | MALE |
| 2 | LONGFORD | BRUCE | PATRICK | 07/02/1944 | MALE |

**Unlinked records — deterministic methodology**

| | | | | | |
|---|---|---|---|---|---|
| 3 | LONGFORD | PATRICK | BRUCE | 07/02/1944 | MALE |
| 4 | LONGFORD | BRUCE | PATRICK | 07/02/1943 | MALE |
| 5 | LONGFORD | BRUCE | PATRICK | 07/02/1944 | FEMALE |
| 6 | LANGFORD | BRUCE | PATRICK | 07/02/1944 | MALE |

## 3.2 Probabilistic methodology

For 'statistical' linkage purposes, it can be assumed that most or all of these records belong to the same individual. The differences are typical errors that commonly occur in data collections. As the outcome of the research does not in any way affect the individual involved (Mr Bruce Longford), it is valid for a researcher conducting linkage for 'statistical' purposes to make some assumptions as to which records should be included in the combined data file and attributed to this individual.

These assumptions may not always be correct. Therefore an element of error is introduced into the 'statistical' linkage. This level of error is both expected and accepted by the researcher, as a consequence of obtaining the best 'probable' links

between the information held on the one individual across the six data collections used in the example.

Using a probabilistic methodology in the above example, each demographic variable could be compared in a pre-defined way by an operator who would generate a similarity 'score' for each variable. Again, this assumption is based on the knowledge that the identity of the individual client is not of primary importance in linkage for statistical research purposes. When comparing two records from different data collections, the scores for each variable could be added together with the total used to determine whether or not the two records are likely to belong to the same person.

The method can calculate a degree of similarity between two names (LONGFORD and LANGFORD), check whether given names have perhaps been transposed, and take into account similarities between dates (07/02/1944 and 07/02/1943). In addition, the relative frequencies of names are taken into account so that two records with the surname JONES will (depending on the data being matched) usually be given a lower weight than a pair of records with the surname ZWEILLIGERHOF. Non-matching fields could be given a negative score that also depends on the distribution of information within each field, so a mismatched gender would usually receive a greater negative score than a mismatched surname.

Using a probabilistic methodology for the statistical linkage of the mock data identified above would probably result in linkage of records 1, 2, 3, 4 and 6 (and possibly record 5). While there is a chance that one or more of the linked records do not belong to the same individual, the chance of this error is estimated within certain statistical constraints and therefore considered acceptable for the linkage of data for statistical and research purposes.

**Linked records — probabilistic methodology**

| | | | | | |
|---|---|---|---|---|---|
| 1 | LONGFORD | BRUCE | PATRICK | 07/02/1944 | MALE |
| 2 | LONGFORD | BRUCE | PATRICK | 07/02/1944 | MALE |
| 3 | LONGFORD | PATRICK | BRUCE | 07/02/1944 | MALE |
| 4 | LONGFORD | BRUCE | PATRICK | 07/02/1943 | MALE |
| 6 | LANGFORD | BRUCE | PATRICK | 07/02/1944 | MALE |

**Unlinked records — probabilistic methodology**

| | | | | | |
|---|---|---|---|---|---|
| 5 | LONGFORD | BRUCE | PATRICK | 07/02/1944 | FEMALE |

Owing to the invariable presence of errors and variations in the recording of demographic data, probabilistic methodologies can lead to a much better linkage of records from separate data collections than simple deterministic methodologies for 'statistical' linkage purposes.

## 3.3 Statistical linkage keys

As identified in Section 2.1.2, the individual variable(s) used to link data for statistical research purposes will vary from agency to agency, depending on the level of common information available in each individual record within each data collection.

Unlike linkage for administrative or client management purposes, which usually makes use of a global identifier (such as a tax file number) or agency-specific

identifier, statistical linkage often relies on a constructed 'key' for each individual to effect the linkage. This 'statistical linkage key' can be constructed from as little or as much individually specific information as is thought necessary by the agencies participating in the statistical linkage project.

The SLKWG has developed the following working definition of a statistical linkage key (SLK):

> *A derived variable used to link data for statistical and research purposes that is generated from elements of an individual's personal demographic data and attached to de-identified data relating to the services received by that individual.*

### 3.3.1 Source of the SLK

A statistical linkage key used by an agency to link data from different collections is usually generated from one of two sources:

- *construction by the agency from full demographic data already held or collected by that agency (that is, a post-collection SLK); or*
- *direct collection from clients (that is, at the point of collection' SLK).*

Both of these sources of the derived linkage variable have varying advantages and disadvantages, which are outlined below. The understanding of the source of the SLK also has a significant bearing on any analyses of data linked with the respective keys.

Where an SLK is *constructed* by an agency from full demographic data (usually drawn from existing data collections or operational systems), the SLK can be generated for all client records held by the agency. The quality of the linkage key can be affected by errors in the source data collection, however the quality of the SLK can be improved by employing probabilistic linkage methodologies

Where the SLK is *directly collected* from clients, the SLK is based on a limited amount of the individual's personal information and is only available with the consent of the client. The SLKs obtained by the Home and Community Care Minimum Data Set (HACC MDS) and the Supported Accommodation Assistance Program (SAAP) National Data fall into this category.

Both the HACC and SAAP collections use an SLK constructed from a limited number of elements of an individual's personal identifying data (that is, first name, surname, date of birth). These SLKs, when added to an individual's de-identified data, become the basic unit of linkage for statistical purposes. These keys, once generated at the data collection point and attached to the relevant service information, are usually then linked using a simple 'deterministic' linkage methodology.

### 3.3.2 Effect of the source of the SLK

Differences in the source of the SLK can significantly affect any analyses based on the SLK, especially where there is a low level of client consent to use of their data for linkage purposes. In the SAAP data collection, for example, the recent levels of client consent to participation range from less than 70% to over 90%, averaging about 80%.

A 'direct collection' SLK also rarely allows an agency to reflect any changes in the details of a specific individual, meaning that inaccurate and/or multiple

SLK records often exist for the same client (for example, surname changes). This effect is magnified over time, with the result that the analysis of data linked using a direct collection SLK over time (for example, for longer-term trends) will be affected. As long-term or trend analysis is one of the principal uses of linked data, this can seriously affect the quality of the analysis possible. This effect is demonstrated in Appendix B.

The main benefits of the 'direct collection' keys used in the HACC MDS and SAAP National Data collection have been to provide some privacy protection to clients contributing the data as well as providing a convenient linkage variable by which the data from two different collections can be combined (using deterministic methods). However, these benefits must be weighed carefully against the potential disadvantages of 'direct collection' keys, namely:

- *in a non-encrypted form, the direct collection keys provide limited privacy protection as they are composed of many personally identifying details of the client (as described in the following section); and*
- *the direct collection keys are more likely to contain inaccuracies, errors or generate multiple keys resulting in poorer linkage between discrete data collections relative to constructed keys (see Section 5.3 and Appendix B).*

The 'constructed', or post-collection, SLKs are more likely to be useful for long-term longitudinal data linkage, as agencies with full demographic data often have historical information reflecting name changes and other variations in personal details of clients.

In recognition of these differences and the effect this may have on the quality and effectiveness of the SLK, the SLKWG has considered using a statistical linkage key which is *constructed* from as much demographic information as is available to agencies, rather than utilising existing *direct collection* SLKs which may be more limited in their ability to support the best possible linkage. The 'full demographic' SLK then links the data from different collections using a probabilistic linkage methodology. As indicated above, a comparison of the existing, *direct collection* SLKs against the *constructed*, 'full demographic' SLK is presented in Section 5.3, which looks at the effectiveness of SLKs across sample data and the effects of use of each type of SLK on the analysis of results.

### 3.3.3 Statistical linkage keys as identifiers

Another important point to make about existing statistical linkage keys is that, without encryption, they contain enough personal information to allow a data custodian with access to an agency-specific collection to identify an individual with a reasonable degree of certainty.

It is a very common misconception that an SLK by itself does not allow an individual to be identified when attached to non-identifiable data. Due to the reliance on personal (identifiable) data in many existing SLKs, they are by definition highly specific to an individual record and technically could be re-constructed (by a data custodian with a relevant data collection) to allow an individual to be re-identified with some degree of accuracy in some situations.

For example, a common SLK (for example, the HACC linkage key) includes the gender and date of birth plus three characters from known positions within the surname and a further two from known positions within the first given name. Information at this level contains much that could be used to identify individuals.

The HACC linkage key is primarily a tool used to uniquely identify an individual with a high degree of reliability, without regard for that individual's identity. It is not a tool primarily designed to protect or ensure the anonymity of the individual. While the HACC linkage key does provide some protection to clients to ensure that clients are not unintentionally identified, it is not (on its own) sufficient to provide complete privacy protection. Thus, the key issue for the HACC and other existing linkage keys has been to identify and develop other appropriate safeguards and measures (for example, protocols, encryption processes) to ensure that these requirements are met in conjunction with the use of the SLK.

The re-construction of a non-encrypted SLK such as the HACC linkage key to identify an individual could be technically possible for a data custodian with a related data collection. However, such a practice would not be possible where the use of the key is governed by the safeguards such as those outlined in the proposed protocol identified in Section 7.3 (including encryption, arbitrary identification number (replacing the SLK) and nominated researchers). The protocol proposed in Section 7.3 extends the existing protocols governing the use of HACC, SAAP and the Commonwealth/State Disability Agreement (CSDA) linkage keys to provide a greater protection for the privacy of the use of an individual client's information in statistical linkage research.

### Summary of key points

- *There are two distinct methodologies by which data linkage can be undertaken.*
  - *'Deterministic' linkage methods involve the exact, one-to-one character matching of linkage variable(s) across two or more data collections.*
  - *'Probabilistic' linkage methods involve the researcher making some assumptions as to which records should be included in the combined data file and attributed to one individual.*
- *Statistical linkage using probabilistic methodologies can lead to a much better linkage of disparate data records relating to the same person than simple 'deterministic' matching.*
- *A statistical linkage key has been defined by the SLKWG as:*

  *'A derived variable used to link data for statistical and research purposes that is generated from elements of an individual's personal demographic data and attached to de-identified data relating to the services received by that individual.'*

- *The SLK can be either:*
  - *constructed by agencies after collection from whatever demographic data are available and common to clients across the relevant data collections; or*
  - *directly collected (that is, generated) from clients at the point of data collection, and sent as a pre-determined SLK to agencies for analysis.*

- *Existing SLKs in use in the sector (for example, HACC, SAAP) are directly collected from clients at the point where the individual's personal demographic data are collected.*
- *The HACC and SAAP SLKs are used primarily as data collection tools which also provide some protection to the privacy of the individual's personal information.*
- *The main protection for the personal information gathered through the use of the HACC and SAAP keys is provided by the protocols which specify in greater detail the protections and safeguards governing the use of the data.*
- *The existing 'direct collection' SLKs are compared against a fully constructed SLK (using full demographic data) in Section 5.3 and Appendix B.*
- *Existing, non-encrypted SLKs cannot be considered to be 'non-identifying', as they contain a large degree of personal information used in the construction of the key.*
- *Existing, non-encrypted SLKs could technically be re-constructed to identify individuals (in some situations) with a high degree of probability.*
- *The use of statistical linkage keys must be accompanied by adequate safeguards and protocols (such as the proposed protocol identified in Section 7.3) to ensure the individual client's privacy is protected during the statistical linkage process.*

# 4 Statistical linkage in the health sector

The following section outlines related projects in the health sector where linkage is being undertaken for administrative client management purposes (Health*Connect*) or statistical and research purposes. Considerable progress has been made in recent years between the Commonwealth and some States involving direct linkage of disparate health data collections to improve the information available for research purposes.

These projects are not presented here as representative of related developments in the health sector. Rather, they highlight the common issues and concerns that underlie statistical linkage research within both the health and community services sectors. Many of the privacy, legislative, guideline development and protocol issues are similar across both sectors, and will need to be resolved in a way that is compatible with, and meets the needs of, both sectors.

## 4.1 Health*Connect*

The National Electronic Health Records Taskforce has proposed the concept of a national health information network (Health*Connect*) that would allow personal health information to be collected, safely stored and exchanged (with the individual health consumer's permission).

The Health*Connect* initiative aims to investigate the feasibility of developing a national network of electronic health records to provide better health care for all Australians. The e-health initiative is focused on harnessing the potential information available to health care providers through linked electronic health records to improve the delivery of health care to individual clients across Australia.

Under Health*Connect*, health-related information about an individual would be collected in a standard, electronic format at the point of care (such as at a hospital or a general practitioner's clinic). This information would take the form of event summaries, not all the notes that a health care provider may choose to keep about a consultation. With the consumer's consent, these summaries would then be able to be retrieved at any time they were needed and exchanged via a secure network with those health care providers authorised by consumers to access the information.

Having more complete and up-to-date information available for each health client would mean that consumers and their providers would be in a better position to make decisions in partnership, through shared information.

The Health*Connect* initiative is a joint Commonwealth, State and Territory project investigating the safe collection, storage and exchange processes for consumer health information in Australia. Participation by both consumers and health providers in the proposed network, if implemented, would be voluntary.

### 4.2 Unique patient identifiers draft business rules

A paper (Marshall 2001) which was developed for presentation to the National Health Information Management Group (NHIMG) and endorsed by the Australian Health Ministers' Advisory Council outlines some proposed business rules for the use of unique patient identifiers in health data collections. The paper proposes a number of principles that should be followed in using unique patient identifiers for statistical linkage projects.

The paper outlines guidance for relevant health jurisdictions on the privacy principles to follow in the linkage of data for planning, research and statistical purposes. The business rules cover issues such as:

- *increasing the distinction between data collections used for client management as opposed to research uses of data;*
- *the importance of data custodians ensuring the risk of identification of individual clients is minimised;*
- *the need for ethical handling practices to be clearly specified and understood by researchers to guide appropriate handling of linked data;*
- *the importance of aggregating data;*
- *issues relating to supervision and oversight of the linkage projects;*
- *use and destruction of linked data sets, including restrictions; and*
- *the need to further develop technical standards for infrastructure and linkage between data systems.*

The SLK WG has used these general privacy principles and draft business rules to inform development of the proposed protocol described in Section 7.3. The resolution of these issues by both the health sector and the community services sector needs to be consistent and compatible to ensure compliance with relevant privacy legislation and to gain the trust and acceptance of consumers (including relevant consumer representative organisations).

### 4.3 Western Australia diabetes linkage project

A pilot project to link State hospital discharge data from Western Australia with Medicare data and the National Death Index has recently received approval. The project involves five participating agencies (Department of Health and Ageing, Australian Institute of Health and Welfare (AIHW), Health Insurance Commission (HIC), Health Department of Western Australia (HDWA) and the University of Western Australia). The signed memorandum of understanding covers the extraction and linkage of Health Insurance Commission data with hospital discharge and death data relating to diabetic patients living in Western Australia.

The data will be supplied to researchers as de-identified files for use in health services planning and research. The period of interest covers 1990 through 1999, and information will be included on individuals living in the State and identified as diabetic by means of HIC data, HDWA data (hospital discharge data) and death data from the Western Australian Registrar of Births, Deaths and Marriages. Further information on deaths will be obtained by linkage to the National Death Index held at the AIHW.

Although this is essentially a pilot project dealing with a limited cohort (diabetic residents in Western Australia from 1990 through 1999), the approach adopted is intended as a model of 'best practice' in the utilisation of administrative data for the production of de-identified linked data files using probabilistic methodologies (see Kelman, Bass & Holman 2001). The fundamental philosophy is to:

- *maximise the preservation of individual privacy;*
- *minimise access to identified data;*
- *allow data custodians full control over the dissemination and use of de-identified data files;*
- *provide linked data files only to named researchers involved in specific approved projects;*
- *provide researchers with no more than the minimal data required for their specific analyses; and*
- *ensure that all copies of named data and all linked data files are destroyed immediately after use.*

A summary of the protocol developed for this pilot is provided at Appendix C.

The two projects described demonstrate the similarity of issues and concerns around linkage for statistical research purposes across both the health and community services sectors.

### Summary of key points

- *Health data linkage projects for both client management or statistical/research/policy purposes are well developed and have been implemented across both health and community services sector industries (for example, hospital, disability services and aged care services).*
- *Protocols and draft guidelines are being developed within the health sector which can further inform the development of these methods within the community services sector.*

# 5 Statistical linkage in the community services sector

The following section outlines both the community services sector environment and recent developments in statistical linkage projects conducted within the community services sector. A description of the current status and use of statistical linkage keys (SLKs) in the sector is also provided.

## 5.1 Community services sector environment

The community services industry covers a wide range of activities including the provision of aged care services (including residential and community care), disability services, child care services (including preschools), family support services, child welfare (including juvenile justice), supported accommodation assistance and emergency relief services (AIHW 1999b).

A report on community services in Australia (ABS 2001) for 1999–2000 estimated that at June 2000 there were around 8,400 organisations involved in the provision of community services across Australia, employing around 560,000 people.

The activities of the industry and the client groups receiving assistance are often closely linked with the health sector, especially in regards to aged care and disability services. There are also strong parallels in the sensitivities between particular client groups of each sector to data linkage activities. For example, the difficulties, sensitivities and issues raised in the health sector around research activities involving HIV clients are likely to be mirrored by community sector clients accessing similarly sensitive services (for example, child protection or juvenile justice services).

The community services sector is a distinct environment from health, with a range of sensitivities and issues specific to the application and delivery of community services programs. However, the issues which underlie the implementation of data linkage methodologies across both the health and community services sectors are very similar, providing both sectors with an opportunity to learn from the experiences of the other.

## 5.2 Existing use of SLKs in the community services sector

Statistical linkage keys have been developed or used in four community services data collections — the Supported Accommodation Assistance Program (SAAP) National Data Collection, the Home and Community Care (HACC) Minimum Data Set, the Commonwealth and State Disability Agreement (CSDA) Minimum Data Set Collection, and the Reconnect program (previously the Youth Homelessness Pilot Project).

The SLKs in these collections are being used primarily as data collection tools for information on each program's client group, rather than to facilitate linkage between collections. These SLKs provide estimates of client numbers, the amount of service provided/received and measure over/under-servicing issues. In the CSDA data collection, the extent of use of both State and Commonwealth services is also measured. Multi-year longitudinal data sets are also being developed in the HACC, SAAP and CSDA data collections using statistical linkage key methods.

A short description of each of these SLKs and their application in the community services sector is provided below.

### 5.2.1 Home and Community Care (HACC) linkage key

The Home and Community Care Minimum Data Set (HACC MDS) has been developed and approved by HACC officials and includes the use of a statistical linkage key. The data set was implemented in 2000–2001 and the first collection of data (covering service provision in January–March 2001) was received in April 2001.

The data on HACC clients is collected by service providers and transmitted to an independent third party data repository. The HACC SLK itself is generated at the point of collection by HACC service providers, with the service experience data of that individual added onto the HACC SLK. The HACC SLK and service data are then transmitted to a third party data repository. Identifying information such as name and address is held by the service provider alone, and is not passed on to the data repository. Service providers require this identified information for administrative or client management purposes. As the identity of the individual is not required for statistical linkage purposes, this information is used only by service providers to generate the HACC SLK, which is then passed on to the data repository.

The development of the HACC SLK involves predetermined combinations of the following personal details of HACC clients — the second, third and fifth letters of the surname, second and third letters of the first name, sex and date of birth. Using the mock data outlined in Section 3 for the fictitious Mr Bruce Longford, a HACC SLK for this individual would be represented as ONFRUM07021944.

#### *HACC collection protocols*

HACC service providers collect the data from individual clients. Names and home addresses of the clients are not collected for minimum data set purposes.

The HACC data collection protocols require clients to be informed of the purpose of the data collection as well as the way the linkage key data are used. Clients are also assured that no identifying details will be passed to the data repository, that the information collected will be used for statistical purposes only and that it will not affect their entitlement to services.

The HACC MDS adopts an 'opt-out' system, where service providers are required to respect the client's preference for non-participation if such preference is made known. Service providers are asked to discuss the issue of refusal with clients to determine if this is caused by sensitivity to certain data items and to offer the exclusion of those sensitive data items as an alternative to total refusal.

### 5.2.2 Supported Accommodation Assistance Program (SAAP) linkage key

The SAAP National Data Collection was introduced in July 1996 with a statistical linkage key. The SAAP SLK consists of the second and third letters of the first name, the first and second letters of the surname, the last letter of the surname, sex and year of birth. Procedures have been established to handle cases where there are fewer than three letters in the surnames and given names.

Again, using the mock data outlined in Section 3, the SAAP SLK for Mr Bruce Longford would be RULODM1944.

The SAAP SLK data are collected by service providers and passed onto the National Data Collection Agency (NDCA) located at the Australian Institute of Health and Welfare (AIHW). The SLK is encrypted and the method of encryption is known only to the NDCA. In accordance with the agreed data collection protocol, the NDCA does not keep the SLK data (except sex and year of birth) after they are encrypted. The encrypted SLK allows the linkage of records that belong to an individual to a certain level of confidence, but does not identify the individual.

The method for the construction of the encrypted alpha code and its implementation was assessed by the AIHW Ethics Committee and has received ethical clearance.

*SAAP collection protocols*

Individual client data is collected by SAAP service providers and sent to the NDCA (located at the AIHW). Name and home address information on clients is not collected.

The SAAP data collection protocols require informed consent from clients for the collection of SLK data and other personal information. Informed consent takes the form of 'opting-in'. If a client has not explicitly given consent, then consent cannot be implied. SAAP clients are informed of the purpose of the data collection as well as the way in which the SLK data is used. Clients are also assured that the services provided to them will not be affected by their decision to give or not to give consent.

The SAAP data collectors' manual has a chapter on confidentiality and clients' rights and refers to the need to adhere to the Information Privacy Principles of the Commonwealth *Privacy Act 1988*.

### 5.2.3 Commonwealth/State Disability Agreement (CSDA) linkage key

The CSDA Minimum Data Set (CSDA MDS) is an annual national collection conducted by all jurisdictions that provide or fund services under the CSDA for people with a disability.

The National Disability Administrators (formerly the Disability Services Subcommittee) trialled the use of the statistical linkage key developed by HACC officials. Pilot tests of the key were conducted in New South Wales, Victoria, Queensland and the Australian Capital Territory during the August 1998 data collection. Based on these trials, the HACC SLK was adopted and the collection of information for all jurisdictions started from 1999.

*CSDA collection protocols*

Service providers are obliged under contract with their respective government departments to collect data specified under the CSDA MDS. Clients (or their advocates) are informed that the information about service users (not including full name or address) will be released to the respective government departments and to the AIHW to enable statistical research to be undertaken. Clients are also informed that the information will be used only for statistical purposes and will not affect their access to services.

The CSDA Data Collection Network Guide contains a chapter on privacy issues, and a special section on the collection of linkage key information.

Clearance by the AIHW Ethics Committee was granted for procedures to be adopted for the collection of the linkage key information and the use of linked data. The Committee's approval requires each jurisdiction collecting the linkage key data to agree to the following conditions:

1. *clients will be informed of the purpose of the collection;*
2. *data on individuals will not be matched with any other information for the purpose of identifying the client;*
3. *the jurisdictions will not disclose or grant access to the information to other persons or organisations, except as statistical information that does not identify an individual; and*
4. *the information will be used only for statistical purposes and will not be used as a basis for any legal, administrative or other purposes.*

#### 5.2.4 Reconnect program linkage key

The Commonwealth Department of Family and Community Services (FaCS) is responsible for the collection of Reconnect program (previously Youth Homelessness Pilot Project) client data and this collection includes an SLK. The SLK used in this program is the same as the SAAP SLK. Like the SAAP data collection, service providers collect the data and inform the clients about the purposes of the data collection and ask for consent to collect personal data from clients.

### 5.3 Measures of the effectiveness of existing linkage keys

Measures of the effectiveness of SLKs have usually focused on how well the linkage key represents the source population and on the extent of duplication (that is, multiple keys for one individual as well as multiple individuals sharing the same key) (AIHW 2000b). Arbitrary decisions have then been made that a certain level of duplication is acceptable for planning and research purposes. A more thorough measure of the effectiveness should be gained by seeking to answer the question of whether the analysis of data linked by different SLKs (for example, 'direct collection' versus 'constructed' SLKs, deterministic versus probabilistic linkage methods) leads the researcher to reach significantly different results and conclusions.

Dr John Bass is currently investigating this problem in collaboration with Professor D'Arcy Holman and members of the Data Linkage Unit in Perth (a collaborative project between the Health Department of Western Australia and the Department of Public Health at the University of Western Australia). Preliminary results from this study have been made available for this paper, and a technical report on the effectiveness of the SLKs, including details of the methods used in this study and the results of the first two analyses, is attached at Appendix B. An abbreviated version of this appendix follows below.

#### 5.3.1 Scope of the comparative study

The study used a data set containing seven years of hospital and death records (1993–1999) of individuals older than 19 years from Western Australia

(2,844,030 hospital unit records). HACC and SAAP SLKs were created for all of these records, and deterministic linkages based on these keys were performed to link records within the hospital data as well as to a copy of the Western Australia death register to which the HACC and SAAP SLKs had been added. The data also contain an arbitrary project identifier which has been generated by the Data Linkage Unit, and which has been assigned to the results of the probabilistic linkage of full demographic data (all names, sex, date of birth, address, country of birth and Indigenous status) that was undertaken. This Western Australia personal identifier (WA PID) has been improved by linkage to other data sets such as the State electoral roll that provides historical information on name and address changes. Significant effort has also been put into validation of the links.

While not perfect, the WA PID and the associated demographic data are an excellent standard for assessing the comparative effect of the SLKs. Apart from the extensive resources that have gone into linking the WA information, the data sets include all of the typical problems found in administrative data.

### 5.3.2 Summary of results

Initial expectations of the group undertaking the Western Australian study were that analyses of data linked by SLKs would not vary greatly in terms of accuracy, but that they would be less precise (that is, have greater variance). If this turned out to be true, then data linked by SLKs would be expected to produce valid results with the finer details sometimes obscured by broader confidence limits. In statistical terms, it was expected that average values would not differ significantly but that there would be a significantly larger variance.

The first analysis, making use only of hospital data, looked at the total number of days in hospital per patient, a statistic commonly used in economic analyses of health and community services data. The second analysis, making use of death data as well as hospital data, looked at the relative risk of death within the cohort of hospital patients.

The results of both analyses, contrary to expectations, showed significant differences in average values (that is, variation in accuracy) and virtually constant variance (that is, no significant changes in precision). These findings mean that analyses of de-identified data linked by means of the existing HACC and SAAP SLKs may carry a significant level of inaccuracy.

These differences may be significant, but are they large enough to make an impact in practical applications? Each case has to be judged on its own merits, but the current analyses at least indicate the extent to which results may be inaccurate. An example is mentioned here, with full details at Appendix B.

This example covers the relative risk of death for patients of Indigenous Australian descent as compared to other patients. Using data linked according to the WA PID, Indigenous Australian patients had a relative risk of 2.3 compared to non-Indigenous patients. The equivalent figures for data linked by the HACC and SAAP keys were 1.2 and 1.5 respectively. In other words, these patients had a 130% increase in their risk of death according to the WA PID linkage, but only a 20% and 50% increase respectively according to the HACC and SAAP keys. These differences are significant at the 95% confidence level.

### 5.3.3 Conclusions

There is a range of linkage keys which can be used to 'join up' data across collections for statistical and research purposes in the community services sector. These results illustrate the need to consider the effects of using the different linkage methods before undertaking any planning or research projects dependent on linked data. The results indicate that the analyses of data linked by the 'direct collection' HACC or SAAP keys (which have been developed primarily to collect and analyse data for the HACC and SAAP program) may lead to greater inaccuracies than those analyses based on a 'full demographic' linkage key, which provides a better linkage of data across programs through employing probabilistic linkage methodology. These effects are more pronounced where longitudinal data or small client groups are being analysed.

Variation in data quality between different demographic groups may result in marked differences after linkage by different methods. As demonstrated, the estimation of relative risk of death of Indigenous Australians varies from 20% (HACC key), 50% (SAAP key) or 130% (WA PID) greater, compared to the non-Indigenous population.

Comparisons of analyses on data linked by different SLKs may be particularly doubtful if the two SLKs are affecting the analyses in opposite directions. For instance, the HACC key produces an estimate of average days in hospital for all patients that is 6% less than that produced by the WA PID. By contrast, the SAAP key produces an estimate that is 10% greater than that produced by the WA PID. If the corresponding estimates produced by the HACC and SAAP keys are compared, that of the SAAP data is 17% greater compared to the HACC data. Comparisons between two linked data sets based on different SLKs should therefore be treated with extra caution.

Every distinct analysis needs a separate decision as to whether a particular linkage method is sufficiently accurate and precise. It is clear that some linkage/analysis combinations lead to results that are of questionable quality.

The causes of these marked differences are still being investigated. What these results do show is that the use of different linkage methods can lead to significantly varied (and unexpected) results. If SLKs are to be used for linkage, then the quality of that linkage in respect of any analysis should be routinely and thoroughly investigated.

Ideally, linkage should be performed using probabilistic methods based on as much demographic data as possible (rather than utilising existing 'direct collection' keys generated primarily for data collection purposes, and linked using 'deterministic' methods) to increase the reliability of the data available for analysis.

#### Summary of key points

- *The community services sector is a distinct environment from the health sector, with specific sensitivities affecting the way linkage for statistical and research purposes are implemented.*
- *However, many of the issues around implementing linkage methodologies across both sectors are similar.*
- *Statistical linkage keys (SLKs) are currently used in four community services data collections, namely the:*
  - *Home and Community Care (HACC) Minimum Data Set;*

- *Supported Accommodation Assistance Program (SAAP) National Data Collection;*
- *Commonwealth and State Disability Agreement (CSDA) Minimum Data Set collection; and the*
- *Reconnect program (previously the Youth Homelessness Pilot Project).*

- *Protocols have been developed for these linkage processes covering issues relating to client consent, the purpose and usage of linked data, role of the data repository, access and sharing of data between agencies and data security issues (that is, use of encryption).*
- *Previous measures of the effectiveness of SLKs tend to focus on how well the target population is represented by the key (that is, levels of participation or consent to linkage) or measures of the accuracy of the key (that is, in relation to the number of duplicate keys created).*
- *A comparative study looking into the effects on analysis of using different linkage keys, including existing 'direct collection' SLKs and a 'full demographic' SLK, has been undertaken. The results show that the type of key used can significantly affect the results obtained through the analysis of linked data, especially where either longitudinal data are used, or small client groups are subject to analysis.*
- *Every statistical linkage proposal therefore needs to consider whether linkage using a particular key is sufficiently accurate and precise. Some linkage/analysis combinations lead to results that may be inaccurate.*

# 6 Privacy and legal considerations

While the implementation of a linkage process for statistical analysis and research purposes is of great potential benefit and is technically feasible, the linkage of data across jurisdictions raises significant privacy and legislative issues that need to be carefully addressed. Protocols must be developed to address these privacy issues and to provide a workable framework to support future statistical linkage projects using community services sector data sets.

The following section outlines the key privacy and legislative issues that require careful consideration by each agency considering involvement with a statistical linkage key project. These issues have been identified by the SLKWG and considered in the development of the suggested protocol identified in Section 7.3.

While complex and sometimes complicated, the many privacy and legislative issues raised here can be addressed successfully by agencies considering implementing linkage key methodologies for statistical and research purposes. The key issue for each agency will be in accepting responsibility for ensuring that the proposed linkage project(s) are implemented using the most complete and comprehensive safeguards and protocols to minimise, to the greatest extent possible, the chances that relevant privacy and legislative frameworks are contravened.

It has not been possible for the SLKWG to provide the NCSIMG (and the community services sector) with definitive advice on the appropriateness of statistical linkage projects for all community services sector data collections. Each project needs to be considered on its merits and with regard to its particular circumstances by the participating agencies. The suggested protocol aims to assist agencies to ensure that relevant privacy and legislative issues are addressed. Responsibility, however, rests with each agency to ensure that record linkage is undertaken in a manner consistent with existing legislation, including privacy legislation

## 6.1 HACC linkage key experience

The SLKWG initially investigated the way in which the HACC Minimum Data Set project team had addressed the legislative and privacy issues around implementing the HACC linkage key.

Prior to the introduction of the HACC linkage key, the project team sought information on the applicable privacy legislation governing the use of a linkage key. The key findings were:

- *the Information Privacy Principles (IPPs) contained in the* Privacy Act 1988 *should be used as a framework for considering the privacy issues relating to record linkage in the HACC program; and*
- *Responsibility rests with each agency to ensure that record linkage is undertaken in a manner consistent with existing legislation, including privacy legislation.*

The HACC linkage key project team then considered a range of options to maximise the privacy and confidentiality safeguards associated with the introduction and use of a linkage key. These options included:

- *using contracts between the data providers and users of the data specifying their respective roles and responsibilities in relation to the use of the HACC linkage key;*
- *the development of adequate protocols for the data linkage process and encryption process to govern the protection, confidentiality and use of HACC client data;*
- *the recognition that HACC service providers needed to be adequately informed as to the purpose, scope and process of the HACC linkage key for statistical data linkage purposes;*
- *the need to make clear decisions as to the level of client information that would be released to relevant jurisdictions for data analysis (that is, level of aggregation of data); and*
- *alterations to the names of data fields (and aggregations of data items) before the data were released to allay fears that unit record files could be used by relevant jurisdictions to identify individual HACC clients.*

The project team also recommended that the linkage process be undertaken by an independent, trusted third party who would be subject to stringent, ethical guidelines and privacy safeguards consistent with the Information Privacy Principles contained in the *Privacy Act 1988*. Clients of HACC are informed by the service provider how their data will be used and their consent is sought prior to the collection of data. While the precise wording of the consent statement differs between States/Territories and individual agencies, all clients are generally advised that their information will be sent to third party data repositories and will be used by specific agencies for statistical and planning purposes.

Similarly, when clients are admitted to the HACC program, the service provider issues a confidentiality statement prior to collecting the client details that specifies the service provider may release non-identifiable information about HACC clients to the Department of Health and Ageing and to the National Data Repository. The specified use of the non-identified information is for statistical purposes, and to gain information about HACC services and their consumers.

### 6.2 Role of the Office of the Federal Privacy Commissioner

The Office of the Federal Privacy Commissioner (OFPC) has a pivotal role in the promotion and ongoing development of privacy safeguards in Australia. The purpose of the OFPC is to promote an Australian culture that respects privacy while having due regard for other important social interests, such as the free flow of information and the need for government and business to operate efficiently.

The OFPC carries out a range of functions, including:

- *providing advice to individuals on their rights under the Privacy Act and related legislation;*

- *providing general advice about the Privacy Act and privacy issues (eg. OFPC 2001a, 2001b, 2001c);*
- *promoting best practice in privacy standards;*
- *providing advice (in response to written requests from Federal and ACT government agencies and private organisations) on how to comply with the Privacy Act and related legislation;*
- *examining proposed legislation for its privacy implications; and*
- *undertaking regulatory and compliance functions under the legislation, including handling complaints, conducting investigations and audits.*

For more information on the scope of the Privacy Commissioner's functions, refer to Section 27 of the Privacy Act 1988.

The SLKWG originally intended to approach the OFPC to seek a determination on how a statistical linkage key process fitted within the framework of the Privacy Act 1988. However it became clear that, with a process as broad as statistical linkage, and in the absence of specific detail of particular data linkage projects, it would not be possible to obtain a general, unequivocal position on the appropriateness of statistical linkage key methodologies within existing privacy legislation.

### 6.3 Legal pro forma response from SLKWG members

The review of the HACC Linkage Key project team's approach and detailed consideration of the role of the OFPC demonstrated to the SLK WG two main points:

- *it would be difficult to gain general advice on the appropriateness or suitability of Statistical Linkage Key methodologies under existing legislative frameworks (eg.* Privacy Act 1988*) without specific detail on the proposed linkage and parties involved: and*
- *each agency would ultimately be responsible for identifying and complying with relevant legislation, including the* Privacy Act 1988.

Due to this better appreciation of the role of the OFPC, the SLKWG determined that it would be necessary to identify for each agency the legislative frameworks and agency-specific privacy protocols which may impact on the adoption of a statistical linkage key process. This task was clearly beyond the scope of the SLKWG for all community services sector agencies. However, as a start the SLKWG asked each participating member to seek advice from its respective legal section on how any existing legislation or protocols would affect that agency in participating in a statistical linkage key project.

To assist agencies in this process, a pro forma of four specific questions to be answered was developed by the SLKWG. These questions were:

- *What legislation do you work under and how might that legislation affect the development and use of a statistical linkage key by your agency?*
- *Are there any privacy protocols, practices or policies specific to your agency which might affect the development or use of a statistical linkage key?*

- *Would it be possible for your agency to provide to a third party (for example, a data repository) data on an individual which:*

– *does not allow the individual to be identified (that is, the data has been de-identified);*

– *has an SLK added; and*

– *is governed by a set of privacy protocols developed by the SLKWG for data linkage?*

*If not, what might need to be done to allow this to happen?*

- *Does your agency currently link data with an external agency? If so, please provide a brief description and identify what legislation/policies/protocols and/or practices underlie this arrangement.*

In response to these pro forma questions, each agency found it extremely difficult to obtain clear guidance from its respective legal section. While the identification of the relevant legislation and/or protocols was possible, it was less simple to determine from these whether or not linkage of data *for statistical purposes* was permitted. The legal advice from each agency was also limited by the need for lawyers to have specific details of the proposed linkage, including details as to which data items, from which collections, held by which agencies would be linked.

The SLKWG concluded from the legal pro forma exercise that the legal consideration of whether or not a statistical linkage methodology would be allowable under existing legislation would have to be considered on a case-by-case basis, with regards to the:

- *agencies involved in the linkage;*
- *relevant legislation and/or protocols; and*
- *specific data items and data collections being linked.*

While the legal pro forma exercise was limited in its ability to provide the SLKWG with a clear view as to the full scope of legislative constraints on implementing an SLK process across the community services sector, it did raise a number of significant questions relating both to privacy and legislative issues. These issues were considered by the SLKWG and are outlined below.

## 6.4 Legislative issues

The legislative issues identified by the SLKWG which require consideration by agencies considering implementing statistical linkage projects across the community services sector are outlined below for consideration.

### 6.4.1 Legislative privacy protection

The SLKWG identified three main levels of legislation relating to the protection of clients' privacy which need to be considered in the implementation of statistical linkage projects in the community services sector. A list of the relevant legislation identified by the SLKWG in addressing this question is provided at Appendix D.

The first level relates to Commonwealth legislation encapsulated in the *Privacy Act 1988*, covering the data handling practices of both Commonwealth agencies and, after 21 December 2001, private sector organisations.

The second level relates to specific State or Territory laws which contain provisions affecting the treatment of data and providing protection for the rights to privacy of individuals providing information. Most of these State and Territory laws follow in the spirit of the *Privacy Act 1988* and the associated Privacy Principles. However, there are particular differences in their application and coverage in each State and Territory, and in how the State/Territory legislation articulates with the relevant Commonwealth legislation.

For example, New South Wales became the first State or Territory to pass a comprehensive State personal data privacy law, with the Privacy and Personal Information Protection Act 1998. However, this generic legislation was intended to govern only public sector activities, and has received criticism (for example, O'Connor 1999) due to the agencies exempted from compliance such as law enforcement agencies. In contrast, Victoria and the ACT have passed specific legislation governing the management of personal health information (ie. the Victorian Data Protection Bill 1999 and Health Records Act 2002, and the ACT Health Records (Privacy and Access) Act 1997) whether the information is held in the private or public sector. Both the Victorian and ACT legislation specifically relates to health record data, providing health-specific protection to individuals including a general right of consumer access to their health records. Both the New South Wales and Northern Territory parliaments are intending shortly to introduce legislation similar in intent to the Victorian and ACT laws (eg. the draft Health Records and Information Privacy Act in NSW).

The third level of legislation identified by the SLKWG relates to the agency-specific legislation under which both Commonwealth and State/Territory Government agencies operate.

The conceptual separation of these three levels of legislation reflects the boundaries of jurisdictions and agencies. In practical terms, they all are relevant and will differentially influence the appropriateness or legality of a statistical linkage methodology depending on the parties proposing the linkage and the data items/collections being linked.

These legislative boundaries are also dynamic. Due to the relatively recent and rapid expansion of information technology in the delivery of community and health services, and the enhanced capacity this provides organisations to use data in new ways, privacy legislation is continually evolving to maintain relevance and adequate protection for clients.

Because the application of legislation depends on the specific characteristics of the statistical linkage proposal (that is, Commonwealth or State agencies involved, data collections involved) and because of the dynamic nature of the legislation, the protocol proposed by the SLKWG to cover statistical linkage requires adequate consideration to be given to the need for privacy legislation to be assessed on a case-by-case basis.

### 6.4.2 Commonwealth *Privacy Act 1988*

The Commonwealth *Privacy Act 1988* and the associated Privacy Principles provide the cornerstone for privacy legislation in Australia. Much of the State and Territory privacy legislation is modelled on this Act. Therefore, the SLK WG considered this legislation in detail in relation to its regulation of Commonwealth and ACT Government agencies participating in a statistical linkage key methodology project, and has considered the implications of this for linkage projects in other States and Territories (cognisant that legislation other than the Privacy Act 1988 will apply)."

The *Privacy Act 1988* and the associated Privacy Principles specify for Commonwealth (and ACT) government agencies and the private sector (including private sector health service providers) the manner in which personal information must be collected, used and disclosed.

'Personal information' has a specific meaning under the Act, namely:

> *'personal information' means information or an opinion (including information or an opinion forming part of a database)…about an individual whose identity is apparent, or can reasonably be ascertained, from the information or opinion.*

### 6.4.3 Use of personal information to participate in a statistical linkage project

All Commonwealth agencies hold personal information on their clients. The use of this information (for example, in the construction of a statistical linkage key) by the agency is therefore governed by its own legislation and the *Privacy Act 1988*. Therefore, before considering whether or not statistical linkage can take place between agencies, each agency has to determine whether (under the *Privacy Act 1988*) it can use personal information already held (or being collected) to construct a statistical linkage key for de-identified research. Information Privacy Principles Two and Ten are relevant to this question.

Information Privacy Principle Two (IPP2) relates to the need for the data collection agency to explain to any person providing personal information why the information is being gathered (i.e. for what purpose the information is required). The various purposes for which an individual's information might be used should be made clear at the point of collection (or as soon as practicable after collection). The individual should be able to form a reasonable expectation as to what their information will be used for by the agency collecting that information.

Information Privacy Principle Ten (IPP10) refers to the requirement for information to be used only for the purpose for which it was collected. It should be clear that IPP10 must be consistent with the purposes stated in IPP2 that were provided when the information was originally collected from the client. IPP10 has a number of exceptions where personal information can be used for other purposes of which the client may not be aware, including the use of personal information in the protection of public revenue (IPP10(d)). It is therefore important for each Commonwealth agency to initially establish the uses and purposes for which the information was initially gathered from clients.

As an example, a range of information is gathered directly from clients by Aged Care Assessment Teams for the Commonwealth Department of Health and Ageing in assessing client entry to the aged care system. This information is recorded on an

'Aged Care Application and Approval' assessment form (No. 2624). There are three main purposes identified to clients as to why this information is gathered, namely for:

*(a) the provision of aged care, or other community, health or social services, to the person;*

*(b) assessing the needs of the person for aged care, or other community, health or social services; or*

*(c) reporting on, and conducting research into, the level of need for, and access to, aged care and other community, health or social services.*

Under the *Privacy Act 1988* and the IPPs, the use of an individual's personal information gathered in this form by the Department to construct a statistical linkage key for research purposes would have to be consistent with the research uses referred to in provision (c) above.

At the point of data collection, most Commonwealth agencies inform clients that the information they provide may be used by the agency (and other nominated bodies) for specified research purposes. Commonwealth agencies considering using an SLK methodology must therefore ensure that clients understand that the data collected may be used for specified research purposes, including (in some cases) statistical linkage projects. As long as this requirement is met, then the relevant Commonwealth or ACT Government agency appears (under the *Privacy Act 1988)* to be permitted to use the personal information collected to participate in a statistical linkage research project.

### 6.4.4 Does the *Privacy Act 1988* relate to statistical linkage projects?

A further consideration for Commonwealth agencies is whether the *Privacy Act 1988* and the IPPs actually relate to a constructed SLK linked to de-identified data.

The construction of an SLK usually involves the use of personal information by a specific agency. As specified in Section 3.3.2, the SLK should still be considered to be an identifier. This means that where an SLK is attached to de-identified data, the SLKWG considers this to be 'personal information' as defined under the *Privacy Act 1988*.

However, once an SLK has been encrypted it is no longer identifiable. Therefore, when an encrypted SLK is attached to de-identified data, the SLKWG considers it is no longer identifiable 'personal information' under the *Privacy Act 1988* (as long as the identity of an individual can not be 'reasonably ascertained' from the de-identified data).

The identity of an individual could theoretically be 'reasonably ascertained' from de-identified data if a particular combination of de-identified data items allows a person with particular characteristics to be identified. For example, de-identified data relating to a specific disability or illness (for example, lower limb amputee) coupled with small area locality data (for example, suburb) and specific age or date of birth information (for example, 45 years, or date of birth 1/1/1956) may allow a data custodian to identify from a data collection that the record belongs to a specific individual.

Controlling the level of aggregation of specific de-identified data items such as age (into age ranges) and locality (into ABS Statistical Locality Areas) would

address this issue to ensure de-identified data items do not allow individuals to be reasonably identified.

Therefore, the legislation that covers the use of an encrypted SLK attached to de-identified and aggregated data appears to be either the agency-specific legislation (for example, the *Aged Care Act 1997*) or any agency-specific privacy and/or information handling protocols.

It is unlikely that agency-specific legislation would refer to whether or not the agency is permitted to use or exchange encrypted, de-identified data between agencies. However, there may be some (agency-specific) privacy or information handling protocol which may be relevant to this process. In the absence of any legislative guidance from agency-specific legislation, an agency could proceed with the linkage process by adopting a statistical linkage protocol as suggested in Section 7.3. Where agency-specific privacy or information protocols exist, the agency must ensure that the adoption of a linkage protocol does not breach these existing agreements. Some examples of these protocols are now discussed.

### 6.4.5 Agency-specific protocols

A further protection offered by some agencies to their clients in the treatment of personal information is afforded through agency-specific information handling practices and protocols. Such protocols, on the other hand, may impose a duty of care or other consideration that compels an agency to override some aspect of privacy.

The SLKWG identified through the legal pro forma exercise some agency-specific protocols which had been developed by agencies to provide clients with an assurance that their personal information would be handled sensitively and with due regard for their rights to privacy. These protocols or procedures are in addition to the existing legislative protection afforded through specific privacy legislation or through agency-specific legislation.

For example, the data collection and dissemination activities of the Australian Institute of Health and Welfare (AIHW) are subject to ethical assessment from the AIHW Ethics Committee. The AIHW Ethics Committee is constituted under the *Australian Institute of Health and Welfare Act 1987* and has the power to release identifiable health and community services data for research and statistical purposes under conditions specified by the AIHW Ethics Committee. The Committee has consistently not allowed AIHW data to be used for client management purposes.

The South Australian Department of Human Services (DHS) has also developed a Code of Fair Information Practice that applies to DHS, funded service providers and anyone who has access to personal information in the South Australian public health, housing and community welfare sectors. The standards set in this code are based on the National Privacy Principles contained in the *Privacy Act 1988*.

Clearly, these agency-specific agreements or protocols need to be considered carefully by community services agencies in terms of their effect on the implementation of a statistical linkage research project by a particular agency. For example, an agency-specific protocol or policy (such as a 'duty of care') may potentially override the proposed protocol developed in Section 7.3 governing the use of statistically linked data.

This may compel the relevant agency (once it is participating in a linkage project) to pursue any potential instance of wrongdoing identified from statistically linked data.

However, the steps and safeguards outlined in the proposed protocol in Section 7.3 (that is, aggregation, encryption and replacement of the SLK with a project identification number) should ensure that the relevant agency, even if compelled to administratively follow up the matter by an agency-specific protocol, could not use the statistically linked data to do so. While the statistically linked data may identify the scale of a potential problem, the administrative pursuit of this problem would have to be undertaken by the agency using different means and as a separate process to the statistically linked data.

Where a conflict is identified by an agency (prior to commencing a linkage project) between their agency-specific protocols/policies and the proposed statistical linkage protocol identified in Section 7.3, the agency must consider very carefully whether it can take part in linkage projects and, if unresolvable, avoid participating in the project altogether.

### 6.4.6 *Privacy Amendment (Private Sector) Act 2000*

The *Privacy Amendment (Private Sector) Act 2000* was passed by Federal Parliament in December 2000. This legislation amended the *Privacy Act 1988* (which had only covered Commonwealth and ACT public sector agencies), extending its cover to most private sector organisations around Australia, including non-government organisations (for example, community sector service providers).

The legislation introduced in 2000 for private sector organisations differs in both nature and intent to that which remains applicable to the Commonwealth public sector. The amendments are known as 'light touch' legislation, proposing generic, high level principles rather than detailed legislation which applies to the whole of the private sector. These privacy principles for the private sector are known as the ten National Privacy Principles (NPPs). In addition to these NPPs, specific codes can be developed by specific industries to provide increased protection for specified consumer groups.

In 1999, at the request of the Federal Attorney-General, the Privacy Commissioner consulted widely on whether the National Principles covering the private sector should be varied to provide appropriate protection for personal health information under the new legislation. From these consultations, the Privacy Commissioner concluded that the National Principles, with relatively few modifications, *and supported by appropriate guidelines*, would be able to form the basis of an appropriate framework for personal health information.

### 6.4.7 Health privacy guidelines

The *Guidelines on privacy in the Private Health Sector* (to support the Principles) were released by the OFPC in November 2001. The guidelines provide advice on how private sector organisations can ensure the protection of health information, as well as providing examples on how the legislation operates in specific circumstances (OFPC 2001b).

The definitions of 'health information' and 'health service provider' used in the Guidelines are very broad. For example, the definition of 'health service provider' includes many community sector organisations, such as:

- *private aged care facilities;*
- *other health and allied health professionals in private practice (including psychologists);*
- *phone counselling services or drug and alcohol services; and*
- *Indigenous community controlled health organisations.*

The Guidelines provide important information about the development and implementation of nationally consistent guidelines and standards for access, storage and use of personal health information that are relevant to a statistical linkage key project.

For example, NPP2 relates to the use and disclosure of personal information about an individual. NPP2.1 allows personal information to be used or disclosed for:

- *the primary purpose for which the information was originally collected; or*
- *a related secondary purpose that is within the individual's reasonable expectations (or, where the information involved is sensitive information, this must be a directly-related secondary purpose that is within the individual's reasonable expectations); or*
- *another secondary purpose where the individual has given their consent, or where another of the limited exceptions to NPP 2 applies.*

The Guidelines also discuss the use of de-identified data (as opposed to identified health information) by private organisations for research purposes. The OFPC has indicated that organisations conducting research should use de-identified data where possible. If and when a client makes a complaint, an organisation must be able to justify why de-identified data was not used.

The *Guidelines on Privacy in the Private Health Sector* and *Information Sheet 9-2001 Handling Health Information for Research and Management* provide more information. Both are available at www.privacy.gov.au.

As the SLKWG believes community services data collections form a continuum with health sector collections. As these Guidelines now impact on service providers from within both the health and community services sectors, it is reasonable to view the Guidelines as an appropriate model for the community services sector (for both government and non-government organisations).

### 6.4.8 Data-matching guidelines

The Office of the Privacy Commissioner's *The use of Data Matching in Commonwealth Administration—Guidelines* (OFPC 1998) may also provide some guidance to community service agencies considering participation in a statistical linkage process.. While the Guidelines refer to linkage for administrative or client management

purposes (rather than statistical or research purposes, as considered by this report) they still provide guidance on issues common to all linkage exercises. These issues include:

- *identifying the costs and benefits of the linkage activity;*
- *engaging public awareness of the linkage activity;*
- *informed consent by participants in linkage projects;*
- *specifying a protocol to define and guide the linkage activity; and*
- *handling and disposing of linked records.*

The Guidelines apply to linkage activity as an administrative, public revenue protection (that is, anti-fraud) or law enforcement tool. As identified in Section 2.1, this is not the context within which the current report is framed. However, the Guidelines have been considered by the SLKWG in the development of the protocol outlined in Section 7.3.

## 6.5 Privacy issues

A recent report from the United States General Accounting Office (2001) identifies three interrelated issues to be considered in discussing privacy-related concerns for record linkage projects:

- *personal privacy — related to the individual's status and rights;*
- *confidentiality — status accorded to information and control over its disclosure; and*
- *security — issues relating to safeguards placed on the data, such as encryption.*

The major privacy issues identified by the SLKWG that require consideration by agencies considering implementing statistical linkage projects across the community services sector are outlined below against this framework.

### 6.5.1 Personal privacy issues

The individual client of a community services sector agency has rights prescribed under relevant legislation protecting their information and its use by the agency. While the purposes and uses of the individual's information by the agency must be in line with the reasons for which the information was initially collected [2], many clients may not be aware that 'research' uses include statistical linkage projects. Where the linkage project involves the sharing of information across jurisdictions, the individual may feel that they had a right to provide consent to the agency for this use of their information.

The client's sensitivity to the issue of whether or not consent has been given will necessarily be increased where the nature of the data is perceived to be 'sensitive'. Personal information gathered by some community sector agencies relating to the safety or security of an individual (for example, SAAP data) or to client characteristics such as disability or criminal history will obviously generate much greater sensitivity on the part of the client to the use and linkage of this information. While linkage for statistical

[2] With some exceptions (see Section 6.4.4).

purposes makes use of de-identified data, many clients may feel that this use is one that they should have been made specifically aware of at the time the information was gathered.

The SLKWG concluded that any proposed linkage project involving 'sensitive' data be subject to greater rigor in determining under what purposes the information was initially collected from clients. This may involve community services agencies reviewing all procedures and forms used at the point of data collection and ensuring that, where necessary, the use of information for de-identified linkage research is made clear to clients. Clients should also have the opportunity (as currently exists in the HACC and SAAP data collections) of opting not to allow their personal information to be used by agencies for statistical linkage research purposes. A further protection (which will be discussed in the protocol presented in Section 7.3) includes the involvement of client or consumer representatives on the steering committees of data linkage projects, to ensure particular sensitivities are appropriately handled.

### 6.5.2 Control of information and data

The issue of community sector agencies sharing data for linkage research across jurisdictional boundaries also requires special consideration.

Data custodians from each agency are legislatively and ethically responsible for the distribution and use of information in their care. Many agencies are extremely wary about any record linkage exercise involving data in their care because of justifiable fears that, once control is relinquished, unauthorised copies may proliferate and/or the data may be used for improper or illegal purposes. This concern leads to caution in providing access to identified data, either within or between organisations.

The SLKWG recognises that the planning of linkage projects needs to take into account these concerns of data custodians relating to data control. In developing the proposed protocol outlined in Section 7.3, the SLKWG has identified a number of steps to ensure the data custodians retain appropriate levels of control over the use of data in their care.

These steps include:

- *selection of an independent agency to do the data linkage. Staff members in the agency with access to the data would be named and sign individual confidentiality agreements;*
- *selection of an independent agency as data repository (or data management team)—this may be the same agency as that which does the linkage, although the individual staff members involved would be from separate teams;*
- *a steering committee including representatives of data custodians for each data set involved;*
- *appropriate scrutiny and approval of the research from the relevant ethics committee (as identified by the steering committee);*

- *individual agreements for each and every research project using the linked data, with each data custodian (via the steering committee) having full power of veto over any proposed project that uses their data;*
- *supply of linked data for research projects only to the named researchers from the data repository, who conduct the analysis on behalf of the steering committee members (see Section 7.4.2 for a discussion of this approach);*
- *the named analysts from the data repository being required to sign individual confidentiality agreements, and be under strict conditions covering access to the data, the supply of copies to other parties, and deletion of the data at the conclusion of the project; and*
- *the steering committee receiving from the data repository only the agreed analyses from the linked data (that is, they do not receive the linked, de-identified and aggregated data file).*

### 6.5.3 Security issues

The physical and virtual security of the information in the care of a data custodian before, during and after a linkage process is also a major issue in relation to the privacy protections offered to community sector clients.

The protocol identified in Section 7.3 provides some guidelines on the safe storage and handling of information during a statistical linkage process. Other protections identified by the SLKWG include conducting the linkage on a non-networked (stand-alone) computer. The guidelines and protections around the handling of data during the linkage process are also governed by relevant agency-specific protocols and privacy protection laws.

#### 6.5.3.1 Encryption algorithms

There are specific scrambling techniques which can be used to encode the SLK attached to the de-identified and aggregated data to protect its confidentiality during transmission between agencies. Encryption algorithms are well developed and have been widely used to preserve the confidentiality of data transmission, especially in the financial sector to preserve the confidential transmission of financial details and information.

The options for using encryption to preserve the confidentiality of data during transmission will be discussed here in terms of the SLK only, rather than encrypting the entire data stream. The SLKWG has recommended that, for the best possible linkage using probabilistic methods, the SLK would contain full demographic data and be attached to suitably aggregated, service experience information. As such, it should be sufficient to encrypt the SLK only, with the remaining data stream not encrypted.

Where agencies consider that it may be necessary to encrypt both the SLK and the attached, de-identified and aggregated data stream, it would be necessary to use a reversible encryption algorithm, for reasons discussed in detail below. The second option outlined below (that is, a non-reversible encryption algorithm) would not be suitable.

As indicated, encryption algorithms usually fall into one of two main categories: reversible and non-reversible. Reversible encryption algorithms are shared between the contributing agency and the receiving agency. The reversible algorithm depends on an agreed decryption key being used by the receiving agency to decrypt the data received. This decryption key may or may not be known by the contributing agency, and must at all times be kept secret and secure by the receiving agency. Leaking of the decryption key used to deconstruct the algorithm would obviously mean that the confidentiality of the encryption process would be compromised, with the data in danger of being decrypted by unauthorised individuals. On receipt of the data from the contributing agency, the receiving agency would use the decryption key to 'unlock' the SLK (composed of full demographic data) and then use the full demographic data to link records across collections using probabilistic linkage methods.

The second option would involve a non-reversible encryption algorithm being held by the contributing agency only. The contributing agency would encrypt the SLK using an algorithm which could not be decrypted, add the encrypted SLK to the de-identified and aggregated data and provide this to the receiving agency. On receipt of the data, the receiving agency would only be able to link the records received using deterministic (character to character matching) linkage methodology. Use of the non-reversible encryption algorithm by the contributing agency would necessarily be limited only to the SLK, with the remainder of the data stream (that is, the de-identified and aggregated service experience data) never being able to be encrypted.

Non-reversible encryption algorithms permit agencies to freely exchange the encryption algorithm and encrypted SLK without risk of identifying personal client details.

The Department of Health Services in South Australia has recently investigated (with the ABS) the use of a non-reversible encryption algorithm for use with Gambling Rehabilitation Fund data (collected by the Break Even services for gambling counselling). The encryption algorithm works by summing the ASCII representation of names, gender and date of birth, dividing by 256 and converting the remainder to hexadecimal representation. The algorithm provides 2 to the power 36 unique codes, and has an error rate of a fraction of a per cent, and is therefore nearly unique. It is irreversible since anyone intercepting the data cannot tell how many lots of '256' were in the original sum, or even how many letters there were in the first or last names entered.

The choice of which form of encryption to use will need to be made by each relevant agency (or steering committee) participating in the linkage project. In some cases, it may be deemed appropriate by the relevant steering committee to use a full range of client personal demographic data to construct the SLK (for example, entire name, birth details), then encrypt this using a non-reversible algorithm and link the data using deterministic methods.

However, reversible encryption algorithms are generally more commonly understood and allow the use of full demographic data by the receiving agency (that is, the linkage agency) to perform the linkage using probabilistic linkage methods. Thus, the draft protocol described in Section 7.3 proposes use of a reversible encryption algorithm. The SLKWG recommends that some form of encryption be used in any linkage process, to ensure that the maximum security is afforded to data items in transmission and, through the use of a reversible algorithm, that the conditions supporting the best

possible linkage are provided.

### 6.5.4 Consumer consultation

One of the most important issues considered by the SLKWG in relation to privacy involves the need for any future statistical linkage methodology to be undertaken with the greatest possible degree of transparency and openness. This would involve the active involvement or representation of client or consumer groups in the development and implementation of statistical linkage projects as a prerequisite to any project going forward.

Until recently, data linkage projects were generally regarded as sensitive due to legitimate public concern about the use of personal data for such purposes. Data linkage projects (and especially 'data matching' projects) can attract a high level of public concern and awareness. This legitimate concern can quickly escalate into an ill-informed public debate of the issues when the legitimate and positive goals of the research are misrepresented by the media as an Orwellian invasion of each individual's personal privacy. It makes an excellent 'bad news' story.

It is therefore critical for all data linkage projects, and statistical linkage in particular, to:

- *identify clearly the benefits to be gained from the linkage;*
- *engage relevant consumer/stakeholder groups in dialogue during the project; and*
- *demonstrate clearly the protections afforded to any individual's private information through the relevant memorandum of understanding and statistical linkage protocols.*

The SLKWG has considered the argument that it may be difficult to identify the appropriate consumer representation required, and that the technical or statistical aspects of potential linkage projects may not require consumer representation. The SLKWG is also conscious that the addition of this representation may slow down or limit legitimate research. However, the SLKWG considers that it should be possible to engage these groups either indirectly (through representation on properly constituted ethics committees) or more directly through participation or representation on the relevant project steering committee for the linkage project. The potential benefits to the specific linkage project of this approach far outweigh expected difficulties in the identification and successful engagement of these groups.

An example of a valuable project that neglected a number of these factors recently occurred in Canada. The Longitudinal Labour Force File (LLFF) was a Canadian databank made up from several contributing public agencies linked together for policy development and research purposes. Information came from employment insurance data, from the Customs and Revenue Agency, and from social assistance data. The LLFF was used to help design and improve programs such as assistance for Canadians seeking employment, the doubling of parental leave under employment insurance, and the employment insurance family supplement. Analysts could look at the impacts of a wide range of policy options, taking into account such factors as business

cycles, changes in provincial/territorial policies, regional differences, gender, age, income and education.

A number of factors combined to lead to the rapid dismantling of the LLFF. Highly emotive and negative media reporting of the potential capacities of the LLFF stirred up intense and vocal public opposition to the project. By the time the media had run with the story, it was too late for the LLFF to argue for the many potential benefits to consumers of its existence. Insufficient or limited consultation with consumer groups in the development of the LLFF combined with the media overplay of the story, and the wider public (which remained unclear as to what the LLFF was for) became justifiably upset. The Privacy Commissioner of Canada then expressed concerns about the approach to the management of policy analysis and research information and data in May 2000 and soon afterwards the LLFF was dismantled. The LLFF is slowly being rebuilt in a more transparent and publicly acceptable manner.

**Summary of key points**

- *The privacy and legal considerations in implementing a statistical linkage project are varied. However, they can be addressed successfully by agencies that are considering implementing linkage key methodologies for statistical and research purposes.*
- *Each agency has to accept responsibility for ensuring that the proposed linkage is implemented using the most complete and comprehensive safeguards and protocols to minimise, to the greatest extent possible, the chances that relevant privacy and legislative frameworks are contravened.*
- *Generic legal advice cannot be given regarding the status of statistical linkage projects. Legal advice needs to be sought on a case-by-case basis, within the structures identified in Section 7.3 (if appropriate).*
- *Each agency (coordinated by the relevant project-specific steering committee) needs to review the conditions and proposed uses under which information to be used in the linkage has been gathered.*
- *At the point of collection, clients should be able to form a reasonable expectation that their data will be used in linkage projects by certain agencies.*
- *Commonwealth agencies may participate in an SLK project only where IPP2 and IPP10 have been met.*
- *Within the steering committee, participating agencies need to agree on a standard level of aggregation of specific data items (for example, age bands, locality areas) to ensure de-identified data items will not allow a reasonable possibility that individuals can be identified.*
- *Agencies considering participating in SLK processes will need to review their jurisdiction and agency-specific legislation and/or existing protocols governing the use of data for research purposes.*
- *Health privacy (2001) and data-matching (1998) guidelines have been developed by the Office of the Federal Privacy Commissioner, and provide assistance to community services sector agencies in the consideration of the issues relating to linkage activity.*
- *Personal privacy issues relate mainly to the issues of consent, client awareness of uses of the information provided and the the type of data being collected.*
- *Confidentiality issues have been identified that relate to the responsible control and use of data in the care of the data custodian.*
- *Security issues have also been identified relating to the secure transmission and storage of data, including encryption and safe handling of data.*
- *Adequate consumer representation and consultation (that is, the 'transparency' of the project) is essential to the success of data linkage projects.*
- *The education and involvement of consumers and/or consumer groups in the uses, benefits and safeguards of linkage projects are extremely important.*
- *The positive benefits of statistical linkage projects need to be properly publicised and emphasised with relevant consumer representative groups.*

# 7 Draft protocol for statistical linkage key research

This section outlines a proposed protocol for conducting statistical linkage research within the community services sector. This protocol is intended to guide the process decisions of community services agencies[3] considering participating in a linkage project, to ensure the linking of data collections across agencies occurs with the greatest possible protections for both the privacy of clients' information and the security and access/use of their information for legitimate policy and research purposes.

In describing the process and likely operation of this protocol, a range of assumptions and key issues around different aspects of the protocol will also be raised. A brief discussion of these assumptions and issues is presented in Section 7.5, after the broad process of the protocol has been described.

## 7.1 Scope of the protocol

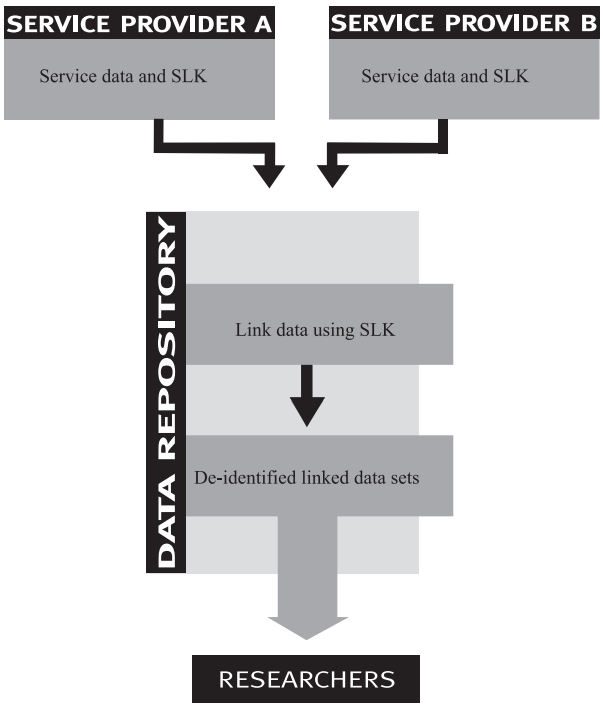
The main task of the SLKWG has been to develop an appropriate protocol to address the privacy and legislative issues raised in the consideration of statistical linkage projects for the sector.

The SLKWG considered that the draft protocol(s) should cover (as a minimum) the:

- *process through which SLK proposals are developed/endorsed, including ethical clearance;*
- *community service sector client consultation processes;*
- *data collection process and procedures;*
- *defined roles of key stakeholders (that is, community services agencies, third party data repository, relevant ethics committees and/or linked data analysts);*
- *construction of the SLK;*
- *type/level of information to be attached to the SLK;*
- *encryption procedures required/used;*
- *secure channels for the transmission of information/data;*
- *responsibility for storage and defined uses of the linked data (especially in relation to longitudinal linked data sets);*
- *agreement between agencies (and responsibility for) the destruction of linked data after analysis; and*
- *process for dissemination of linked data (or analyses of linked data) to contributing agencies.*

[3] The protocol has been developed principally from the perspective of statistical linkage between government administrative agencies, at both the State and Commonwealth levels. While the process would be similar for non-government community sector agencies, the protocol might need to be modified to reflect differing and more relevant structures applying in this area.

**Figure 1: Example of the HACC MDS linkage process**




### 7.2 Example of an existing statistical linkage process

The HACC MDS can be used to demonstrate the main steps involved in a currently operational statistical linkage research process in the community services sector. The process of exchanging HACC data for linkage between service providers and the national data repository has formed the basis from which the protocol outlined in Section 7.3 has been developed. This process is broadly represented in Figure 1.

The service providers 'A' and 'B' identified in this process represent HACC service providers collecting the individual's personal identified information and generating the SLK. The de-identified service information is then transmitted with the SLK to an independent, third party data repository. The data repository then links the data received from each service provider using the HACC linkage key, and provides the de-identified, linked data to specified agencies (for example, Department of Health and Ageing, relevant State Government agencies) for research and program planning purposes.

### 7.3 Proposed statistical linkage research protocol

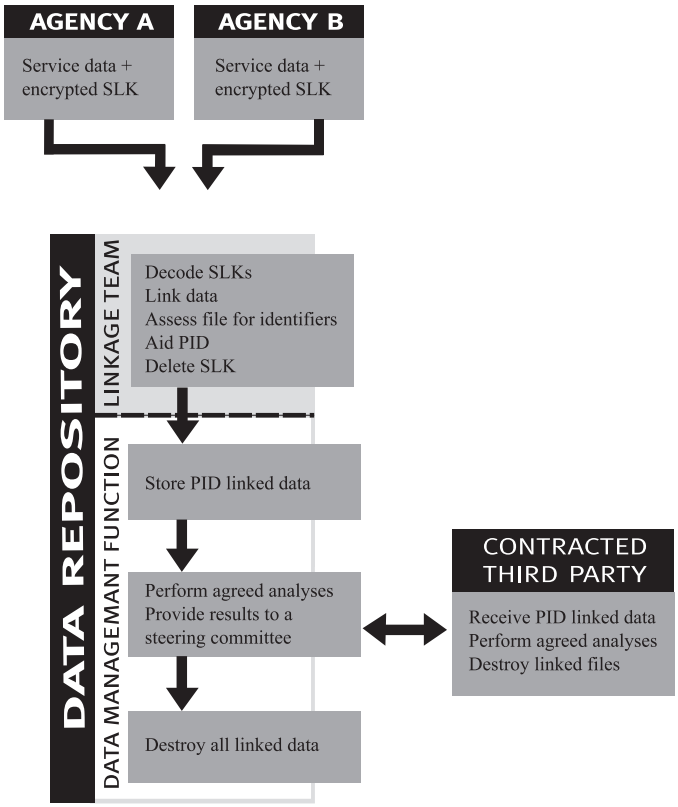
As described in Section 7.2, the process outlined under which statistical linkage takes place for the HACC MDS forms the basis from which the following protocol has been developed.

The protocol has been developed around three main stages of the life of a statistical linkage project, namely the:

- *pre-linkage phase;*
- *statistical linkage of data; and*
- *post-linkage phase (research applications).*

The basic process around the second phase of the protocol (statistical linkage of data) is represented in Figure 2. This representation distinguishes between the activities of the linkage team (who perform the linkage of files) and the data management team (who are responsible for the storage, access and possible analysis of the linked data). The linkage function and the data function may exist within the same organisation (that is, the data repository) or they may be separate entities or organisations.

**Figure 2: Proposed statistical linkage protocol**



### 7.3.1 Pre-linkage phase

The following steps are recommended by the SLKWG as necessary for the pre-linkage phase of a statistical linkage protocol.

1. *Relevant jurisdictions/agencies identify and agree to a set of defined research questions that address policy issues and require linkage of de-identified data at the unit record level to be adequately addressed. The policy questions and benefits to be gained through the statistical linkage research need to be clearly justified.*
2. *A steering committee involving representatives from each of the participating jurisdictions/agencies is formed. Representation on the steering committee should be at a sufficiently senior level within each organisation to provide delegation for the release of data to a third party. Relevant community sector client representatives may also be identified at this stage and may be invited to participate in the steering committee process for the project.*
3. *The steering committee identifies a relevant ethics committee or committees (within each agency, or externally) who will be responsible for assessing the linkage proposal and approving the implementation of the linkage process. If appropriate, consumer representation or involvement in the assessment of the proposal may be sought. The specific arrangements or requirements of the ethics committee are identified and guide the approach of the steering committee throughout the pre-linkage phase.*
4. *The steering committee agrees on the data collections and data items relevant to the research. Only data collections or items directly relevant to the research question should be included in the linkage proposal.*
5. *The steering committee identifies and agrees on a linkage agency and data management agency (collectively referred to as the 'data repository'). The linkage and data management agencies may be different entities, or they may be from the same organisation. The data repository works under the direction of the steering committee. The criteria for selection of the data repository would include the capacity for secure storage and transmission of information.*
6. *Each agency identifies the purpose(s) under which the data proposed for linkage was initially collected. An assessment is then made by each agency against relevant legislation and agency-specific protocols as to whether the agency can use the data for research purposes (in this case, using linkage for statistical and research purposes). Legal advice is sought as appropriate.*
7. *The steering committee develops a research proposal to be presented to the relevant ethics committee(s) identified in (3) above, identifying the issues raised above and seeking approval to proceed with the statistical linkage research.*

8. *An agreement (that is, memorandum of understanding) is developed by the steering committee and signed by the jurisdictions/agencies and linkage agency/data repository. A draft memorandum of understanding is provided at Appendix E for reference. The memorandum would include as a minimum the following issues:*
   - *policy purpose/benefits of linkage;*
   - *relevant data custodians;*
   - *relevant data collections and items;*
   - *specified treatments to the data to preserve confidentiality (that is, encryption processes, aggregation of data items);*
   - *specified research personnel from each agency;*
   - *specified personnel from the linkage agency/data repository;*
   - *responsibilities of the relevant agencies;*
   - *confidentiality agreements from linkage agency/data repository staff;*
   - *proposed analyses and applied uses of the linked data;*
   - *legal considerations identified in (6) above; and*
   - *rights of each agency to publish research results.*

### 7.3.2 Statistical linkage of data

These steps are recommended as a minimum by the SLKWG as necessary in the second phase of the proposed statistical linkage protocol.

9. *The data custodians from each jurisdiction prepare a data file for linkage containing:*
   - *as much full demographic data (that is, the SLK) as is common for clients across agencies. This will be used by the linkage agency to link the data files across agencies using probabilistic linkage methods. The SLK data used for the linkage would be encrypted using an agreed (reversible) algorithm; and*
   - *specified relevant service experience data, suitably aggregated and de-identified.*
10. *Coordinated by the steering committee, these files are transmitted from each agency to the linkage team/agency of the data repository.*
11. *The linkage team then decrypts the SLK data and links the data received from each agency using probabilistic linkage methods. The linkage agency then:*
    - *strips away the SLK data used to merge the data files from the merged data set, and re-enumerates it with an arbitrary project identification (PID) number (that is, each record is now identified by numbers 1 to 'n'); and*
    - *destroys the original extract data sets.*

12. *The linkage team then assesses the linked data file for any single person cells and/or very low cell sizes which may not have been picked up in the earlier aggregations, and advises the steering committee where the linked data may divulge information on individuals or small groups of potentially identifiable individuals. The steering committee will then advise the linkage team of how best to treat these cases, through further aggregation, random alteration, deletion or other appropriate methods.*
13. *The linkage team then transfers the PID-encoded service experience data to the data management team. The linkage agency/team destroys all data used to create the linked data file once the data management team has confirmed its receipt.*

The replacement of the decrypted SLK data in the merged data file (step 11) with an arbitrary PID number is extremely important. This step ensures the SLK data (that is, the full demographic data) do not remain linked to the individual's service experience data, and also ensures the linked data file can never be used by the data management team (or anyone else) to merge with other data sets to re-identify participants in the linked data file. The destruction of the SLK data used to perform the linkage (step 11, dot point two) by the linkage team also prevents all further use of the decrypted SLK and service data.

It is important to note that the SLK data may contain some variables which are required for further analysis (for example, gender, date of birth). These variables may be provided to the linkage team only as a part of the SLK data (that is, not attached in the de-identified, aggregated service experience data). Therefore, prior to the destruction of the SLK data, the linkage agency may be required to maintain some analytical elements of the SLK, or derive and aggregate others (such as changing date of birth into an age value, and converting this into an agreed age range) for future analysis.

### 7.3.3 Post-linkage phase (research applications)

These steps are considered by the SLKWG as necessary in the research application phase of the proposed statistical linkage protocol.

14. *The data management team is responsible for the secure storage and use of the linked data file (consisting of the PID-encoded service experience data). The data management team then undertakes the pre-defined and agreed analyses as specified by the steering committee on behalf of the participating agencies identified in the memorandum of understanding.*
15. *Where the data management team does not have the skills or capacity to undertake these analyses, the data management team then provides the linked data file to a third party contracted by the steering committee to perform this analysis. On completion of these analyses, the data management team then receives the completed analyses and provides them to the steering committee for distribution to each agency.*
16. *The agencies participating in the linkage project then assess the agreed analyses and prepare reports for publication based on the data. Approval for publication is provided through the steering committee, in line with the memorandum of understanding agreement specified in Step 8.*

17. *Researchers from each agency may request further information or analyses (through the steering committee and relevant ethics committee) not identified in the original the memorandum of understanding. If approved, the data management team extracts the linked data file, replaces the original PID with another arbitrary PID (to ensure the additional information cannot be linked to the linked data previously supplied) and performs the secondary analysis (or provides it to the contracted third party if used).*
18. *The data management team stores the linked, de-identified data for a period of time as specified by the steering committee. After the expiry of this time, the linked data is destroyed.*

It may also be necessary to require both the linkage team (at step 13) and the data management team (at step 18) to sign statutory declarations to the effect that the data has been destroyed as required by the steering committee.

## 7.4 Assumptions and issues relating to the proposed protocol

A number of assumptions and key issues have been considered by the SLKWG in the development of this protocol. These are outlined below and are intended to promote discussion between agencies considering participation in a statistical linkage process, to ensure that the proposed protocol above is understood in terms of its strengths and limitations.

### 7.4.1 Focus of the proposed protocol

As indicated in the introduction to this section, this protocol has been developed largely to support the statistical linkage of data sets held by Commonwealth and/or State/Territory administrative agencies. Thus, some of the activities referred to during the protocol may not be appropriate or directly applicable for non-government agencies interested in conducting statistical linkage research in the community sector (for example, university researchers, independent statistical agencies, service providers, regional authorities). To be of greater relevance to the parties involved, the proposed protocol will therefore need to be modified in some instances where a statistical linkage is proposed that does not involve a Commonwealth or State/Territory Government jurisdiction to be of greater relevance to the parties involved. The proposed protocol will be important to assist identification of comparable safeguards and structures within a particular environment to provide the same or greater protection to the linkage process.

### 7.4.2 Type of statistical linkage projects

The protocol has been developed to describe the required steps involved in a one-off linkage project between a number of agencies. Such a linkage project could involve the use of data held or collected over a number of years by agencies, as demonstrated in the section comparing the effectiveness of linkage keys (Section 5.3) where seven years worth of data was employed for the analysis.

However, the current protocol has not been designed to cover linkage projects which create a longitudinal, linked data collection, stored permanently and added to incrementally by agencies over time. Neither does the protocol cover any form of staged, multi-program linkage project, where an initial linkage between two programs

is supplemented at a later date by adding in a third program, and later again a fourth program. The protocol also is not intended to cover linkage between population survey data and administrative by-product data sets.

If a decision was made in the future to set up a data linkage project using either longitudinal, staged multi-program or population survey data sets, then the current protocol would need to be strengthened especially with regard to the enhanced roles of the data repository and the project steering committee (or another relevant body) in managing and maintaining the data. The creation of such forms of linked data collections to inform policy analysis and statistical research is technically feasible, and the current protocol would provide the foundation from which specialised protocols could be developed to provide adequate protections for these developments.

### 7.4.3 Proposed encryption algorithm

As described in Section 6.5.3.1, the two main forms of encryption that could be employed in any potential protocol are reversible or non-reversible. To allow the linkage agency to use full demographic data with probabilistic linkage methods, the SLKWG has based the protocol around use of a reversible encryption algorithm applied to the SLK only. Extension of this algorithm to de-identified, aggregated data could easily be achieved if this were felt necessary by the relevant project steering committee. Agencies may also consider it more appropriate in their particular circumstances to apply a non-reversible encryption algorithm to the SLK only, and to link the data using deterministic methods, although this process is not described in the current protocol.

### 7.4.4 Return of linked data to source agencies

The SLKWG has considered carefully the issues around return of the linked data files (with an arbitrary project identification number) to source agencies for analysis.

The first option involves not returning the linked data to source agencies after linkage. Under this scenario, the data repository (or another contracted third party) would perform a range of agreed analyses on the linked data on behalf of the source agencies, with the results of these analyses provided to the relevant source agencies.

The main advantages of not returning the linked data to source agencies include the following:

- *All analyses of the data are specified by the relevant steering committee (and cleared through the relevant ethics committee) prior to the linkage taking place.*
- *It guarantees that source agencies can never attempt to identify participants in the linked data file by matching with administrative data collections.*
- *Use of the data is strictly controlled by the steering committee through its contractual arrangement with the data repository and/or contracted third party.*

The disadvantages of not returning the linked data to source agencies include the following:

- *Some analyses may only become apparent/relevant after inspection of the linked data, and so may be missed.*
- *The data repository may not have the capacity, skills or experience to undertake the required analyses. In such a case, the steering committee would have to obtain, through a separate procurement process, the skills and services required through a contracted third party.*

A second option considered by the SLKWG involves the data repository *returning* the PID-encoded, de-identified and aggregated data file to nominated researchers from within the source agencies to conduct the required (and pre-specified) analyses.

The main advantages of returning the linked data to source agencies include the following:

- *All analyses of the data are still specified by the relevant steering committee (and cleared through the relevant ethics committee) prior to the linkage taking place.*
- *Nominated researchers from the source agencies are free to inspect the linked data and decide if further analyses are required, which would then be cleared through the relevant steering committee.*
- *Source agencies generally have a very good understanding of their own data needs and policy imperatives, which guides the correct interpretation of the analyses.*
- *In many cases, source agencies are capable of performing the required analyses without the expense of contracting out this function.*

The main disadvantages of returning the linked data to source agencies include the following.

- *The steering committee and consumer are asked to 'trust' that the source agency will only use the data for the prescribed purposes/analyses (that is, the steering committee has less direct control over any unauthorised use of the linked data).*
- *The 'bamboo curtain' distinction between the nominated researchers with access to the linked data, and the data custodians from the same source agency, may not be perceived by consumers to offer strong enough privacy protections.*

In consideration of these two options, the SLKWG has framed the protocol around the *non-return* of linked data to source agencies. As with the encryption issue, agencies considering participation in a statistical linkage process may deem that the return of the linked data to the source agencies is a manageable risk and provides more benefits than the non-return option. The protocol would therefore need to be amended slightly to reflect this change. The decision regarding the return or non-return of the data to the source agency therefore also has to be made on a case-by-case

basis with due regard for the agencies participating in the linkage, the sensitivity of the data, the skills required for the analysis of the linked data and the degree of separation that exists between different areas within the one agency.

### 7.4.5 Type of 'data custodian'

As discussed above, the SLKWG and the protocol has made a clear distinction between the roles of the 'data custodian' (that is, the original custodian of the administrative data) and the 'analysts' of the linked data for research and policy uses. The protocol indicates that these roles are distinct and exclusive, with the analysis and use of the linked data solely undertaken by the data repository (or a contracted third party) on behalf of the data custodians, who receive only the agreed analyses.

This distinction (discussed in greater detail in the next section) is critically important where the 'data custodian' also has an administrative responsibility towards the data. This distinction is important from the perspective of government agencies, from which the protocol has been developed. However, there are many agencies which act as a data custodian without the administrative interest or responsibility attached (for example, AIHW, universities, contracted private sector data repositories). The distinction made in the protocol between 'data custodians' and 'analysts' may be less critical where the agency has no administrative responsibilities. The protocol would therefore need to be amended to reflect this circumstance, specifying under what conditions a 'non-administrative' data custodian would be permitted to access the de-identified, aggregated, PID-encoded linked data.

### 7.4.6 Staff training and development needs

While the technical implementation of statistical linkage software is not in itself difficult, it is important to note that personnel with the proper skills, experience and knowledge of data linkage processes, privacy sensitivities and protocols are relatively uncommon. In the evaluation of candidates for the data repository role agencies should consider fully the candidates' skills and experience in linkage of data, its management and analysis. This should ensure that the linkage projects are conducted using appropriate methodologies, with due regard to the possible privacy sensitivities and with an understanding of the effect these techniques have on the analysis of linked data."

As implied in Section 7.4.2, if the linked data were to be returned to the contributing agencies for analysis from the data repository, each agency would need to consider the specialised training and development requirements of its staff to ensure that the statistically linked data is managed and analysed correctly.

### 7.4.7 What is the 'best' SLK to use?

The consideration of which is the 'best' SLK to use by agencies participating in a statistical linkage process is a critical decision for the steering committee. The choice of the SLK data items will be determined by the data items that are both available and common to the agencies participating in the project. The choice of the SLK will also be governed by the choice of linkage methodology (that is, probabilistic or deterministic), the level of accuracy required by the analyses proposed and consideration of how well the privacy of clients is protected (and perceived to be protected) by the SLK.

While a degree of error can be tolerated in any statistical linkage process, there is no necessity for the linkage to be either artificially flawed or overly imperfect. The better the statistical linkage of data records, the more likely it is that any analysis will produce valid and useful results.

As a general principle, the amount of identifying information used in linkage should be kept to a minimum. The ideal is a compromise between the minimal amount of information and a linkage of sufficient quality to produce results of appropriate accuracy and precision when analysing de-identified linked data files.

If the source data from each agency are properly de-identified (and appropriately aggregated), and the SLK is encrypted, the risk of an individual client's information being 'identifiable' is extremely low. As the encrypted SLKs and de-identified information are not exchanged between agencies, but provided directly to the linkage team, the SLK from each agency should contain as much identifying information as possible to allow the 'best' possible linkage to occur.

Existing linkage keys in the community services sector such as the HACC and SAAP linkage keys are all 'direct collection' keys, employed primarily as data collection tools and not designed to facilitate linkage across programs or data collections. As would be expected, deterministic linkages based on these keys appear subject to greater errors or inaccuracies than linkages using probabilistic methods employing full demographic data. To achieve the best possible statistical linkage of data for research and public policy purposes, community sector agencies therefore need to consider carefully the benefits of using 'full demographic' linkage keys, constructed from the maximum amount of identifiable demographic information available and common to each participating agency.

As an example, three years ago the board of Silver Chain, the largest home care agency in Western Australia, agreed to release their data into the Western Australia Linked Health Data system. The object was to enable better planning and research by being able to track clients through the acute hospital system and death records. Caution about privacy concerns meant that Silver Chain demographic data made available for linkage was restricted to an SLK made up of sex, date of birth, postcode, first initial and a six-character phonetic version of the surname. Concern about the quality of linkage obtained using this key, together with confidence in the security of the linkage protocol, have led to a recent decision to repeat the linkage using full demographic data including full names and addresses.

Within the last two months the Disability Services Commission of Western Australia has obtained ethical approval to link their data into the Western Australia Linked Health Data system. The actual linkage should begin by early 2002.

The SLKWG recommends that agencies consider using an SLK based on full demographic data to construct a project-specific SLK based on all details common across agencies. Use of a different SLK for each linkage project also assists in addressing privacy concerns in that the different SLKs for different projects will not be able to be linked in the future across agencies or projects.

**Summary of key points**

- *The proposed protocol has been developed around three main stages of the life of a statistical linkage project, namely the:*
  - *pre-linkage phase;*
  - *statistical linkage of data; and*
  - *post-linkage phase (research applications).*
- *Key steps relating to the pre-linkage phase relate to the formation of an appropriate steering committee for the project, the identification of relevant ethics committee(s) to approve the proposal and the development of a memorandum of understanding specifying the roles and responsibilities of all parties.*
- *A steering committee oversees each linkage project and coordinates the cross-agency interaction and interaction with relevant ethics committees.*
- *Every data custodian can (via the steering committee) veto any proposed research project.*
- *Key steps relating to the statistical linkage of data phase relate to the roles of the linkage team (in decrypting, aggregating and encoding the data with an arbitrary PID) and the data management team (in storing, accessing and analysing the linked data).*
- *The actual linkage is performed by the independent linkage team, distinct from the data management team and function (although both functions may be performed by the same organisation).*
- *The linkage team and the data management team are named individually under the memorandum of understanding and each of the members should be bound by a separate confidentiality agreement.*
- *Use of an arbitrary project identification number by the linkage team ensures that no-one will be able to use the de-identified, linked data to try and re- identify clients in the linked file.*
- *Key steps relating to the post-linkage phase involve the transmission of agreed analyses to the steering committee, managing any further analyses and destroying the data at a time specified by the steering committee.*
- *Any subsequent request for further linked analyses by agencies would be approved through the steering committee.*
- *A separate PID would be attached to any subsequent linked data to ensure two linked files could not be merged against each other, or against existing data collections.*
- *The proposed protocol has been based on a number of assumptions regarding the agencies participating in the linkage, the type of linkage and encryption algorithms used, the use of the data repository to perform the analyses, staff skills and choice of linkage methodology.*
- *Agencies considering a linkage project therefore may need to have some scope to alter this protocol after consideration of these issues to suit their particular needs.*

# 8 Recommendations

The recommendations of the SLKWG to the NCSIMG based on the findings of this report are outlined below under the following five headings:

- *Framework for data linkage;*
- *Statistical linkage methods;*
- *Privacy and legal considerations;*
- *Engagement with the community; and*
- *Coordination with the health sector.*

## 8.1 Framework for data linkage

**Recommendation 1:**
*The NCSIMG endorses the use of statistical linkage methodologies for research, planning and policy analysis.*

**Recommendation 2:**
*The NCSIMG endorse the principle that data collections produced by linkage for statistical and research purposes should not be used subsequently for administrative or client management purposes*

## 8.2 Statistical linkage methods

**Recommendation 3:**
*The NCSIMG acknowledges the need for a statistical data linkage protocol and:*

*(a) notes the proposed draft protocol outlined in Section 7.3 provides a proposed framework for statistical linkage projects in the community services sector and is intended to guide the development of SLK projects, rather than to prescribe a set methodology and process for undertaking such projects;*

*(b) requests the jurisdictions represented to assess the impact of the proposed protocol and report to enable their finalisation at the next NCSIMG meeting in early 2002;*

*(c) refers the protocol and the report to the National Community Services Data Committee for its consideration.*

**Recommendation 4:**

*The NCSIMG notes that in some instances the use of a third party data repository in community services sector statistical linkage projects may be desirable (for example, for cross-jurisdictional statistical data linkages) and their use should be formally considered by each statistical linkage project.*

**Recommendation 5:**

*The NCSIMG recognises that the linkage of data is context-specific, and there is no one preferred method for statistical data linkage. Where possible, the use of full demographic data is appropriate for statistical linkage, but this does not preclude the use of more limited linkage methods.*

**Recommendation 6:**

*The NCSIMG recognises that security of data in transmission between agencies and any third party data repository is essential, and that the encryption of an SLK provides one option to ensure this security.*

### 8.3 Privacy and legal considerations

**Recommendation 7:**

*The NCSIMG recognises that the privacy, client consultation and legal implications of each statistical linkage project will have to be identified, assessed and resolved on a case-by-case basis by the relevant steering committee (and ethics committee) involved in each project.*

**Recommendation 8:**

*The NCSIMG recommend to member agencies considering participating in statistical linkage projects that they review the purposes under which clients contribute data to their agency. The review should seek to ensure that the potential use of information for research and planning purposes (based on statistical linkage) is made clear to clients.*

**Recommendation 9:**

*The NCSIMG recommend to agencies currently using SLK methodologies that the privacy and legal implications are considered in the light of the issues raised in this report*

### 8.4 Engagement with the community

**Recommendation 10:**

*The NCSIMG endorses the involvement of relevant community sector consumer representatives in the development and implementation of statistical linkage projects. The appropriate level of involvement will be determined by the relevant steering committee, and mechanisms (such as a memorandum of understanding) built into each project's work program.*

**Recommendation 11:**

*The NCSIMG acknowledges that the participation and education of both community services sector agencies and consumers is important to the successful implementation of statistical linkage in the sector.*

### 8.5 Coordination with the health sector

**Recommendation 12:**

*The NCSIMG acknowledges that the issues in implementing statistical linkage projects for research purposes in the community services sector are in many cases the same as those being considered by the health sector.*

**Recommendation 13:**

*The NCSIMG considers that the further development of statistical linkage methodology for the community services sector should occur in close consultation with similar developments in the health sector.*

**Recommendation 14:**

*The NCSIMG seeks to cooperate with the health sector (possibly through the NHIMG) where relevant infrastructure (such as ethics committees, data repositories) or expertise can be shared, to facilitate efficient and appropriate linkage implementation across both sectors.*

# References


Anderson P 2000. CSDA linkage key trial, in *Disability support services provided under the Commonwealth/State Disability Agreement: national data*. Canberra:AIHW.

Armstrong B & Kricker A 1999. Record linkage — a vision renewed. *Australian and New Zealand Journal of Public Health*, 23, 5, 451–452.

Australian Bureau of Statistics (ABS) 2001. *Community services catalogue 1999– 2000*, cat. no. 8696.0. Canberra:ABS.

Australian Institute of Health and Welfare (AIHW) 1999a. *SAAP National Data Collection annual report 1997–98 Australia*, SAAP NDCA Report Series 3, AIHW cat. no. HOU 24 Canberra:AIHW.

____ 1999b. *National Community Services Information Development Plan*, Standing Committee of Community Services and Income Security Administrators, AIHW cat. no. AUS 14. Canberra:AIHW.

____ 2000a. *SAAP National Data Collection annual report 1998–99 Australia*, SAAP NDCA Report Series 4, AIHW cat. no. HOU 38. Canberra:AIHW.

____ 2000b. *The use of linkage key for statistical work in community services: background paper for the Statistical Linkage Project of the National Community Services Information Management Group*. Unpublished paper, AIHW.

Brameld K, Holman CDJ, Thomas M & Bass J 1999. Use of state data bank to measure incidence and prevalence of a chronic disease: end stage renal failure. *American Journal of Kidney Diseases*, 34, 6, 1033–1039.

Department of Health and Aging (DoHA) 1999. *Health information and data protection issues*. Unpublished paper, DoHA.

Holman CDJ, Bass J, Rouse I & Hobbs M 1999. Population-based linkage of health records in Western Australia: development of a health services research linked database. *Australian and New Zealand Journal of Public Health*, 23, 5, 453–459.

Holman CDJ, Wisniewski Z, Semmens J, Rouse I & Bass J 1999. Mortality and prostate cancer risk in 19,598 men after surgery for benign prostatic hyperplasia. *British Journal of Urology*, 84, 37–42.

Holman CDJ, Wisniewski Z, Semmens J, Rouse I & Bass J 2000. Population-based outcomes after 28,246 in-hospital vasectomies and 1,902 vasovasostomies in Western Australia. *British Journal of Urology*, 86, 1043–1049.

Karmel R 2000. *Duplicates in the SAAP linkage key.* Unpublished report, AIHW.

Kelman CW, Bass AJ & Holman CDJ 2001. A 'best practice' for creation of linked health data drawn from two or more organisations. Under submission to the *Australian and New Zealand Journal of Public Health* (2001).

Marshall R 2001. *Business rules for the use of unique patient identifiers in statistical collections*. Draft paper prepared for the National Health Information Management Group (NHIMG).

National Health Information Management Advisory Council (NHIMAC) 2000. *A national health information standards plan for Australia: setting the standards*. Canberra:AGPS/ NHIMAC.

O'Connor K 1999. *Protection of personal information by law in Australia: 11 years after the passage of the federal Privacy Act*. Paper to the Records Management Association Conference, Darwin.

Office of the Federal Privacy Commissioner (OFPC) 1994. *Plain English guidelines to information privacy principles 1–3: advice to agencies about collecting personal information.* Canberra:OFPC.

____ 1998. Guidelines for the use of data matching in Commonwealth administration. Canberra:OFPC.

____ 2001a. Guidelines approved under section 95A of the Privacy Act 1998: December 2001. Canberra, OFPC.

____ 2001b. Guidelines on privacy in the Private Health Sector: November 2001. Canberra, OFPC.

____ 2001c. Guidelines to the National Privacy Principles : September 2001. Canberra, OFPC.

Ryan T, Holmes B & Gibson D 1999. *A national minimum data set for Home and Community Care*. Canberra:AIHW.

United States General Accounting Office 2001. *Record linkage and privacy: issues in creating new federal research and statistical information,* cat. no. GAO-01-126SP.

# APPENDIXES

**APPENDIX A**

## Statistical linkage key Working Group membership

**The following individuals were members of the Statistical Linkage Key Working Group (SLKWG):**

| | |
|---|---|
| Mr Andrew Stuart (Chair) | Commonwealth Department of Health and Ageing |
| Dr Ching Choi | Australian Institute of Health and Welfare |
| Mr James Jordan | Commonwealth Department of Family and Community Services |
| Mr Leon Pietsch | Australian Bureau of Statistics |
| Mr Paul Basso | Department of Human Services (South Australia) |
| Mr Allan Dernee | Department of Ageing, Disability, and Home Care (New South Wales) |

**The following individuals also contributed to the work of the SLKWG:**

| | |
|---|---|
| Mr D'Arcy Jackson | Commonwealth Department of Health and Ageing |
| Mr Mark Thomann | Commonwealth Department of Health and Ageing |
| Ms Trish Ryan | Australian Institute of Health and Welfare |
| Ms Margaret Fisher | Australian Institute of Health and Welfare |
| Mr Geoff Neideck | Commonwealth Department of Family and Community Services |
| Mr James Kemp | Commonwealth Department of Family and Community Services |
| Mr John Fulop | Commonwealth Department of Family and Community Services |
| Mr Proshanta Dey | Department of Ageing, Disability, and Home Care (New South Wales) |

In addition, the Commonwealth Department of Health and Ageing (on behalf of the SLKWG) contracted Dr John Bass to provide expert advice and analysis on data linkage issues, especially in relation to the measure of the effectiveness of statistical linkage keys.

APPENDIX B

# Measure of the effectiveness of statistical linkage keys

Previous measures of the effectiveness of SLKs in use in the community services sector have tended to focus on two measures of completeness and accuracy (for example, AIHW 2000b).

The first of these measures concerns the availability of data for the construction of an SLK. Client refusals to allow details to be used for linkage purposes, as well as incomplete/missing data items attached to a client's record, reduce the number of links that can be made. The missing data may be biased compared to the overall client population. Some demographic groups may have an increased aversion to allowing the use of their data, and the quality of data may also vary according to socioeconomic or demographic factors. The proportion of clients for which data are unavailable and the extent of selection bias amongst those clients are measures of the representativeness of an SLK. Most measures of effectiveness have examined the proportions of clients for which data are unavailable, with little information on whether these clients are representative of the whole population.

The second measure relates to the proportion of incorrect linkage keys being generated from the source data. These errors fall into two main types:

1. *errors in the source information leading to the generation of multiple keys for one individual, such as, when a surname is misspelt ('Smith' / 'Smythe') or when there is a name change (as often occurs at marriage or divorce); and*
2. *multiple clients sharing similar identifying information leading to the construction of a single linkage key.*

Errors of the second type will be more prevalent in linkage keys containing less information (that is, they are more likely with the SAAP key than the HACC key). As a measure of the effectiveness of linkage keys, these two errors are often added together as an overall 'mismatch' or 'duplication' rate.

### Existing effectiveness measures of HACC, SAAP and CSDA SLKs

The quality of the HACC linkage key has been tested in terms of duplication rates using three sets of data: the Commonwealth Aged Care database, Silver Chain (a large HACC service provider in Western Australia) and the National Death Index. The testing found a key duplicate rate of between 0.6% and 1% against these collections, which was considered to be acceptable for statistical research purposes (Ryan, Holmes & Gibson 1999).

Two SAAP collections made in 1998–1999 and 1999–2000 reported 25% and 21% client refusals with a further 3.5% and 2.5% missing due to insufficient data. Estimates of duplication rates ranged from 3.3% to 5% (AIHW 2000). These estimates were within a level of accuracy acceptable to the SAAP Data and Research Advisory Committee.

A further test of the SAAP mismatch rate has been conducted by the AIHW (Karmel 2000). This involved testing the SAAP linkage key against a model based on synthetic populations of unique individuals that approximate the year of birth distribution of the SAAP population. These synthetic populations were constructed using data from the National Death Index. The mismatch (duplicate) rate was estimated to be about 3.3% over all year of birth groups. The mismatch rate also increased with the number of people within a particular year of birth, and was higher among younger SAAP clients than older clients. The test also shows that the mismatch rate is expected to be higher if data for more than one year are linked.

The CSDA-linked records from 1999 showed that about 3% of records were of insufficient quality to construct a linkage key. This was an improvement from the levels of invalid data in the 1998 test, which ranged from between 3.7% to 6% (Anderson 2000).

For the Reconnect program (using the HACC linkage key) a consent rate of 80% was achieved as at November 2000. Approximately 3% of the client group provided insufficient information for the construction of a linkage key. There has not been a test on mismatches using the linkage key, so information is not currently available on this aspect of the quality of the linked data. It is hoped that this linkage key will be used in the future to give an indication of multiple use and repeated use of services within the Reconnect program and perhaps to link to the SAAP data collection. Consultation with the community services sector will be undertaken if such linkage is to occur. Plans for this work have not yet been developed.

The results of these broad measurements of the completeness and accuracy of SLK methodology have generally been taken to indicate that these keys are adequate for statistical research purposes.

### Current measures of the effectiveness of SLKs

As outlined above, existing measures of SLKs have usually focused on how well the linkage key represents the source population and on the extent of duplication that is, multiple keys for one individual as well as multiple individuals sharing the same key. It is a far more difficult task to ascertain whether the analysis of data linked by deterministic matching of SLKs leads to significantly different conclusions than would be obtained through analysis of 'real' linked data.

Dr John Bass is currently investigating this problem in collaboration with Professor D'Arcy Holman and members of the Data Linkage Unit in Perth (a collaborative project between the Health Department of Western Australia and the Department of Public Health at the University of Western Australia). Some preliminary results from the study have been made available for this paper.

A data set has been constructed containing seven years of hospital and death records (1993–1999) of individuals older than 19 years from Western Australia (2,844,030 hospital unit records). HACC and SAAP SLKs were created for all of these records, and deterministic linkages based on these keys were performed to link records within the hospital data as well as to a copy of the death register to which the HACC and SAAP SLKs had been added. The data also contain the project identifier (WA PID) created by the Data Linkage Unit, based on probabilistic linkage of full demographic data (all names, sex, date of birth, address, country of birth and Indigenous status). This WA PID has been improved by linkage to other data sets such as the State electoral roll that provides historical information on name and address changes. Significant effort has

also been put into validation of the links (Holman et al. 1999).

While not perfect, the WA PID and the associated demographic data are an excellent standard for assessing the comparative effect of the SLKs. Apart from the extensive resources that have gone into linking the Western Australia information, the data sets involved include the typical problems found in administrative data. The demographic information for an individual is often inconsistent, with varied dates of birth, names, addresses, race and (surprisingly) sex.

The files being used for analysis outside the Data Linkage Unit have had all identifying variables (including the SLKs) encrypted to ensure full protection of privacy. The files were obtained by a standard application to the Data Linkage Unit for de-identified linked data, a process which includes obtaining the signatures of the custodians of all data sets involved as well as that of the General Manager of the Health Information Centre at the Health Department of Western Australia.

The primary aim of the study is to compare the results of typical analyses of linked data from the same set of hospital and death records linked by means of the HACC and SAAP SLKs as well as the WA PID. The effects of increasing the time period over which data are collected, Indigenous status (a group where linkage is usually difficult and liable to an increased error rate) and sample size are all being examined.

### Duplication rates for HACC and SAAP SLKs

Duplication rates of the HACC and SAAP keys in the Western Australian study are summarised in Table 2. For each key, the first row shows the percentage frequency of multiple HACC keys for one individual (that is, one WA PID) while the second row shows the percentage frequency of more than one individual sharing one HACC key. The third row shows the ratio of these two percentages while the fourth row shows their sum.

**Table 2: Duplication rates of HACC and SAAP keys compared to WA PID**

| Duplication rate (%) | 1 year 1993 | 2 years 1993–1994 | 3 years 1993–1995 | 5 years 1993–1997 | 7 years 1993–1999 |
|---|---|---|---|---|---|
| HACC keys/ WA PID | 2.1 | 3.3 | 4.3 | 5.7 | 6.7 |
| WA PIDs/ HACC key | 0.02 | 0.04 | 0.06 | 0.10 | 0.17 |
| Ratio | 105 | 83 | 72 | 57 | 39 |
| **Total** | **2.1** | **3.3** | **4.4** | **5.8** | **6.9** |
| SAAP keys/ WA PID | 1.4 | 2.2 | 3.0 | 4.1 | 4.9 |
| WA PIDs/ SAAP key | 4.6 | 7.6 | 9.8 | 13.0 | 15.4 |
| Ratio | 0.3 | 0.3 | 0.3 | 0.3 | 0.3 |
| **Total** | **6.0** | **9.8** | **12.8** | **17.1** | **20.3** |
| Approximate number of WA PIDs | 205,000 | 350,000 | 470,000 | 650,000 | 785,000 |

Table 2 shows that the rate of multiple HACC keys per individual PID increases steadily from 2.1 to 6.7% over periods of one to seven years. The rate of multiple WA PIDs per HACC key is very low, ranging from 0.02 to 0.17%. The ratio of the duplication types provides a measure of the prevalence of type 1 errors (multiple keys per individual) to type 2 errors (multiple individuals per key). For the HACC key this ratio ranges from 105 over one year to 39 over seven years.

The SAAP key displays a markedly different picture, with the ratio of the duplication types constant at 0.3. The rate of multiple SAAP keys per individual ranges from 1.4 to 5% (slightly lower than that for the HACC key), while the rate of multiple individual WA PIDs per SAAP key is much higher, ranging from 5 to 15%. This is to be expected because the SAAP key contains less information than the HACC key, increasing the chances of more than one individual having the same key.

These results show that the HACC and SAAP keys both produce inaccurate linkages compared to that resulting from the WA PID. The pattern and extent of these biases is different in the HACC and the SAAP keys, and the question arises as to whether analyses of different data sets linked by these two keys might produce different results.

### Comparisons of analyses based on data linked on HACC and SAAP keys

Initial expectations of the group undertaking the Western Australian study were that analyses of data linked by SLKs would not vary greatly in terms of accuracy, but that they would be less precise (that is, have greater variance). If this turned out to be true, then data linked by SLKs would be expected to produce valid results with the finer details sometimes obscured by broader confidence limits. In statistical terms, it was expected that average values would not differ significantly but that there would be a significantly larger variance.

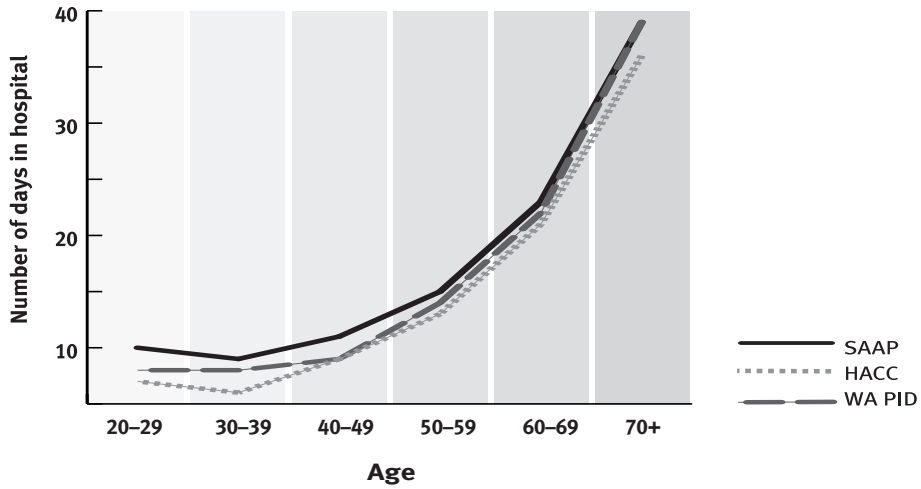
Results from the two analyses completed at the current time are presented here. The first, making use only of hospital data, looks at the total number of days in hospital per patient, a statistic commonly used in economic analyses of health and community services data. The second analysis, making use of death data as well as hospital data, looks at relative risk of death within the cohort of hospital patients.

### Number of days in hospital

Figure 3 is a graph showing the number of days in hospital per patient by age group according to data linked by the HACC and SAAP keys and the WA PID.

It is quite clear that data linked with the HACC key under-estimate the number of days in hospital relative to the WA PID data. Data linked with the SAAP key consistently over-estimate the number of days in hospital, except for the oldest age group where the SAAP and WA PID data are virtually identical. These differences are significant at the 95% confidence level (in most cases, at the 99% confidence level) except for the SAAP/WA PID data in the oldest age group. In the age groups under 60 years of age the HACC results are closer to the WA PID data than are the SAAP results, but this is reversed in people of 60 years and older.

**Figure 3: Number of days in hospital by age group according to data linked by HACC and SAAP keys and the WA PID**




The number of unique HACC keys in these hospital data is higher than the number of unique SAAP keys. It follows that the average number of days in hospital per 'individual' will be lower in data linked by the HACC key that in data linked by the SAAP key.

These differences may be significant, but are they large enough to make an impact in practical applications? Table 3 shows the average number of days in hospital by age group (together with the 95% confidence limits) for the WA PID, HACC and SAAP linkages.

**Table 3: Average number of days in hospital by age group (with 95% confidence limits)**

| Age group | WA PID | | HACC | | SAAP | |
|---|---|---|---|---|---|---|
| 20–29 | 8.8 | (8.7 – 8.9) | 8.2 | (8.1 – 8.2) | 9.9 | (9.8 – 10.0) |
| 30–39 | 8.3 | (8.2 – 8.4) | 7.9 | (7.8 – 8.0) | 9.5 | (9.3 – 9.6) |
| 40–49 | 9.3 | (9.2 – 9.5) | 8.9 | (8.8 – 9.0) | 10.5 | (10.3 – 10.6) |
| 50–59 | 13.2 | (13.0 – 13.4) | 12.5 | (12.3 – 12.7) | 14.2 | (14.0 – 14.5) |
| 60–69 | 21.7 | (21.4 – 22.0) | 20.2 | (19.9 – 20.5) | 23.0 | (22.6 – 23.3) |
| 70+ | 39.0 | (38.5 – 39.4) | 35.8 | (35.4 – 36.2) | 39.8 | (39.4 – 40.2) |
| All ages | 14.6 | (14.5 – 14.6) | 13.7 | (13.6 – 13.8) | 16.0 | (15.9 – 16.1) |

Table 3 shows that, in the example of the 70+ years age group, the average number of days in hospital per patient according to the WA PID is 39.0 compared with 35.8 days for data linked by the HACC key and 39.8 days for data linked by the SAAP key. The 95% confidence limits for the average number of days according to the WA PID range from 38.5 to 39.4. This means that we can be 95% certain that the true value of the average (estimated at 39.0) occurs in this range.

Table 4 displays the percentage difference between the average number of days according to the WA PID and HACC keys, the WA PID and SAAP keys, and the HACC and SAAP keys, also indicating which comparisons are significantly different at the 95% confidence level.

**Table 4: Percentage difference of days in hospital by age group**

| Age group | WA PID > HACC | WA PID > SAAP | HACC > SAAP |
|---|---|---|---|
| 20–29 | –6.9 * | 13.5 * | 21.8 * |
| 30–39 | –5.3 * | 13.7 * | 20.1 * |
| 40–49 | –4.7 * | 11.9 * | 17.4 * |
| 50–59 | –5.6 * | 7.9 * | 14.3 * |
| 60–69 | –6.8 * | 5.8 * | 13.6 * |
| 70+ | –8.2 * | 2.1 | 11.2 * |
| All ages | –6.1 * | 10.1 * | 17.3 * |

(* = 95% significant)

The HACC data average 6% less than the WA PID data with no consistent pattern except for a small rise in the oldest age group. The SAAP data are on average 10% greater than the WA PID data, with a clear pattern of larger differences in the younger age groups (over 13%) falling to 2% in the oldest age group. The only comparison not significant at the 95% confidence level was that between the WA PID and the SAAP data in the oldest age group.

Initial expectations that the different linkage keys would not produce significantly different results in terms of accuracy were clearly wrong.

What about the expectation that precision would be decreased in data linked by the SLKs? Table 5 shows the standard errors of the average values in Table 3.

**Table 5: Standard errors of average values in Table 3**

| Age group | WA PID | HACC | SAAP |
|---|---|---|---|
| 20–29 | 0.05 | 0.05 | 0.06 |
| 30–39 | 0.06 | 0.06 | 0.07 |
| 40–49 | 0.08 | 0.07 | 0.09 |
| 50–59 | 0.11 | 0.10 | 0.12 |
| 60–69 | 0.15 | 0.14 | 0.16 |
| 70+ | 0.22 | 0.20 | 0.22 |
| All ages | 0.04 | 0.04 | 0.05 |

The standard errors of the average values do not vary greatly or in a consistent pattern. The HACC averages are generally slightly more precise than the WA PID averages, with the SAAP linkage showing a slightly larger variance. If the data in Table 5 are normalised to remove the effect of differences in the average values, then the WA PID and HACC standard errors are virtually identical with the SAAP data displaying a consistent small (and not significant) increase.

The initial expectations were therefore wrong on both counts — this analysis shows significant differences between the three different linkages in the average values (that is, variation in accuracy) with virtually constant standard errors (that is, consistent precision) in these values. Analyses of three de-identified linked data sets based on the HACC or SAAP keys or the WA PID led to significantly different results in each case.

## Indigenous status

Linkage of data from persons of Indigenous Australian descent is often more difficult compared to linkage of other cultural groups, with frequent name changes and relatively poor recording of dates of birth and other demographic details. Tables 6 through 8 show the results of an analysis of the number of days in hospital per patient by Indigenous status rather than by age group.

**Table 6: Average number of days in hospital by Indigenous status (with 95% confidence limits)**

| Indigenous status | WA PID | | HACC | | SAAP | |
|---|---|---|---|---|---|---|
| Not Indigenous | 14.2 | (14.1 – 14.3) | 13.4 | (13.3 – 13.5) | 15.7 | (15.6 – 15.8) |
| Indigenous | 27.7 | (26.7 – 28.7) | 22.1 | (21.3 – 22.8) | 26.3 | (25.5 – 27.0) |
| Total | 14.6 | (14.5 – 14.6) | 13.7 | (13.6 – 13.8) | 16.0 | (15.9 – 16.1) |

The results in Table 6 show that the estimates of number of days in hospital per Indigenous patient covered a wide range from just under 28 (WA PID) through about 26 (SAAP) to just over 22 (HACC). The significance and extent of these differences are summarised in Table 7.

**Table 7: Percentage difference of days in hospital by Indigenous status**

| Indigenous status | WA PID > HACC | WA PID > SAAP | HACC > SAAP |
|---|---|---|---|
| Non-Indigenous | –5.7 * | 10.6 * | 17.2 * |
| Indigenous | –20.4 * | –5.3 | 18.9 * |
| Total | –6.1 * | 10.1 * | 17.3 * |

(* = 95% significant)

**Table 8: Standard errors of average values in Table 5**

| Indigenous status | WA PID | HACC | SAAP |
|---|---|---|---|
| Non-Indigenous | 0.05 | 0.05 | 0.06 |
| Indigenous | 0.06 | 0.06 | 0.07 |
| Total | 0.08 | 0.07 | 0.09 |

Tables 7 and 8 show a similar pattern in the analysis for Indigenous status as that shown by the analysis for age groups, with significant differences between the average values (except for the WA PID/SAAP figures for Indigenous patients) and virtually constant precision.

The extent of the differences in average values is sufficient to raise serious concerns about the validity of some of these linkages. For instance, the estimate of the number of days in hospital for Indigenous patients is 20% lower for the HACC linkage than for the WA PID linkage.

### Relative risk of death

The quality of the death data linkages was investigated by performing a Cox regression for the WA PID, HACC and SAAP linked data sets to show the relative risk of death by age group, sex and Indigenous status. Details of this analysis are provided in Figure 4 and Tables 9, 10 and 11.

As far as age groups are concerned, the HACC and SAAP keys display consistently lower estimates of the relative risk of death compared to the WA PID linkage. Differences between the HACC and WA PID linkages are less than 5% except for the 70+ age group where the HACC linkage has a difference of just over 9%. The variances of the relative risk estimates for the different age groups are relatively high and the differences are not significant except for the SAAP and WA PID linkages in the two oldest age groups (60–69 and 70+ years).

Estimates of the relative risk of death for males are remarkably similar for all three linkages, and there are certainly no significant differences.

**Figure 4: Relative risk of death by age group for data linked by HACC and SAAP keys and the WA PID**

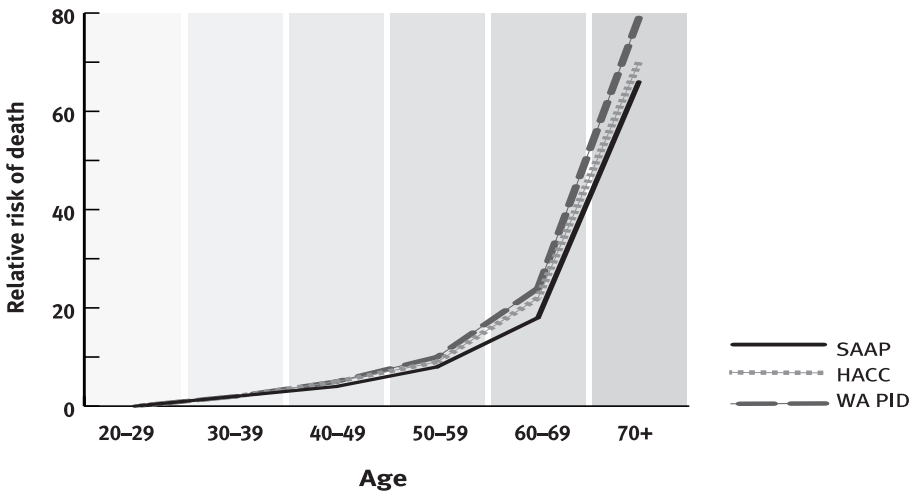


**Table 9:** **Relative risk of death by age group compared to 20–29 year olds; males compared to females; and Indigenous patients compared to non-Indigenous patients (with 95% confidence limits)**

| Age group | WA PID | | HACC | | SAAP | |
|---|---|---|---|---|---|---|
| 30–39 | 1.7 | (1.6 – 1.8) | 1.7 | (1.6 – 1.8) | 1.6 | (1.5 – 1.7) |
| 40–49 | 3.8 | (3.5 – 4.1) | 3.8 | (3.5 – 4.1) | 3.4 | (3.2 – 3.7) |
| 50–59 | 9.2 | (8.6 – 9.8) | 8.9 | (8.3 – 9.5) | 8.1 | (7.5 – 8.7) |
| 60–69 | 23.2 | (21.8 – 24.7) | 22.2 | (20.8 – 23.7) | 20.1 | (18.8 – 21.5) |
| 70+ | 75.7 | (71.2 – 80.4) | 68.8 | (64.5 – 73.3) | 65.5 | (61.4 – 69.9) |
| Sex | | | | | | |
| Male | 1.5 | (1.5 – 1.5) | 1.5 | (1.5 – 1.5) | 1.5 | (1.5 – 1.5) |
| **Indigenous status** | | | | | | |
| Indigenous | 2.30 | (2.2 – 2.4) | 1.2 | (1.1 – 1.3) | 1.5 | (1.4 – 1.5) |

**Table 10:** **Percentage difference of relative risk of death**

| Age group | WA PID > HACC | WA PID > SAAP | HACC > SAAP |
|---|---|---|---|
| 30–39 | 0.0 | –7.1 | –7.1 |
| 40–49 | –0.8 | –10.5 | –8.8 |
| 50–59 | –3.1 | –12.0 | –9.2 |
| 60–69 | –4.4 | –13.2 * | –9.2 |
| 70+ | –9.1 | –13.4 * | –4.7 |
| **Sex** | | | |
| Male | 0.7 | 0.7 | 0.0 |
| **Indigenous status** | | | |
| Indigenous | –47.4 * | –36.5 * | 20.7 * |

(* = 95% significant)

For patients of Indigenous descent the figures are markedly different, ranging from 2.3 for the WA PID linkage through 1.5 for the SAAP key to 1.2 for the HACC key. These relative risk estimates are all significantly different from each other. This is emphasised when one considers that, according to the HACC key, Indigenous patients are 20% more likely to die than non-Indigenous patients but, according to the WA PID, this figure is increased to 130%.

Table 11 shows a marked increase in standard errors with increase in age group. This reflects the sharp increase in risk of death among older patients. For the WA PID linkage, the relative risk of death increases by a factor of 45 from the 30–39 age group to the 70+ age group, while the standard error increases by a factor of 35. Taking the increase in risk into account, there is therefore only a small increase in variance among the values for the older patients.

Table 11: Standard errors of average values in Table 8

| Age group | WA PID | HACC | SAAP |
|---|---|---|---|
| 30–39 | 0.1 | 0.1 | 0.1 |
| 40–49 | 0.1 | 0.1 | 0.1 |
| 50–59 | 0.3 | 0.3 | 0.3 |
| 60–69 | 0.7 | 0.7 | 0.7 |
| 70+ | 2.3 | 2.3 | 2.3 |
| Sex | | | |
| Males | 0.01 | 0.01 | 0.01 |
| **Indigenous status** | | | |
| Indigenous | 0.06 | 0.04 | 0.04 |

## Conclusions

These results illustrate the need to consider the effects of using different linkage methods before undertaking any planning or research projects dependent on de-identified linked data. While the measures of effectiveness relating to duplication rates could easily lead to the conclusion that the HACC key provides a better linkage variable than the SAAP key, an analysis of bed use in elderly patients might well be more accurate using data linked with the SAAP key.

Variation in data quality between different demographic groups may result in marked differences after linkage by different methods. The estimation of the relative risk of death in Indigenous compared to non-Indigenous patients is 20% greater in data linked by the HACC key, compared to 50% greater for the SAAP key and 130% greater for the WA PID data.

Comparisons of analyses on data linked by different SLKs may be particularly doubtful if the two SLKs are affecting the analyses in opposite directions. For instance, Table 3 shows that, for all patients, the HACC key produces an estimate of average days in hospital that is 6% less than that produced by the WA PID. By contrast, the SAAP key produces an estimate that is 10% greater than that produced by the WA PID. If the corresponding estimates produced by the HACC and SAAP keys are compared, that of the SAAP data is 17% greater compared to the HACC data. Comparisons between two linked data sets based on different SLKs should be regarded with extra caution.

Decisions as to whether a particular linkage method is sufficiently accurate and precise need to be made separately for every distinct analysis. It is clear that some linkage/analysis combinations lead to results that are, at the very least, of dubious quality.

The causes of these marked differences are still being investigated. What these results do show is that the use of different linkage methods can lead to significantly varied (and unexpected) results. If SLKs are to be used for linkage, then the quality of that linkage in respect of any analysis should be routinely and thoroughly investigated. Ideally, linkage should be performed using probabilistic methods with as much demographic data as possible.

APPENDIX C

# Western Australian Diabetes Linkage Project protocol

## Model for cross-jurisdictional data linkage

### Proposed approach

The process involves two separate stages. The first stage is a memorandum of understanding between participating agencies to share data for an agreed purpose and to prepare a linkage key file (using probabilistic methods). The second stage involves the production of linked, de-identified data files for an undefined number of separate (approved) research projects. Each project will be covered by its own agreement, the data for the project being supplied directly to the researchers by the various data custodians.

For each research project, a unique set of project IDs will be generated by the custodian of the linkage keys and will provide the only way of combining the data files into a single linked de-identified file.

This two-stage process will ensure that data custodians have full control over the distribution and usage of their data, as each project will need to be well defined and then individually approved before proceeding. No research will be undertaken without the written approval of every data custodian supplying data to the project.

Linked data files will be provided only to the individually identified researchers doing the analysis for each project, and will be destroyed when the analyses are complete. A different ID will be used in each project, thus making it extremely difficult to merge the linked data for two projects (such a process is, in any case, specifically prohibited).

### Linkage key file

The linkage key file will be produced by a small technical team specialising in data matching, including personnel from several of the participating institutions. All people involved in the actual linkage and therefore requiring access to the data used in the linkage process will sign confidentiality agreements and be named on a list provided to the steering committee. Any changes to this list will be reported in writing to this committee. No other personnel will be allowed access to the files used in this process, as they will contain private and confidential information. The work will be done on an isolated computer, and all personal demographic data will be destroyed as soon as the linkage is complete. Transfer of these data files will be done only via tape, diskette or CD-ROM personally carried by those personnel taking part in the data matching. The linkage personnel will not be permitted to take any part in the analysis of the linked data, or to have any communication about these data with the researchers.

### Linked de-identified data

The linkage key file will contain no actual data but will provide coded keys to the data sets involved. Every custodian will supply the approved records from their databases, together with a project ID number, directly to the nominated researchers for that project. These researchers will also sign confidentiality agreements. They will link the data together using the project ID, and will be the only people granted access to the de-identified linked information. They will be specifically forbidden to disseminate copies of the data files, and will be required to destroy these files on completion of the analysis.

### Ethics approvals

Ethics approvals from the researcher's institution as well as the confidentiality or ethics committees of each of the participating institutions are mandatory.

APPENDIX D

# Related legislation on health and privacy

### Commonwealth health-related legislation

*Australian Institute of Health and Welfare Act 1987*
*National Health Act 1953*
*Medicare Levy Act 1986*
*National Health and Medical Research Council Act 1992*

### Commonwealth and State privacy legislation

**Commonwealth**
*Privacy Act 1988*
*Privacy Amendment (Private Sector) Act 2000*
*Data Matching Program Assistance and Tax Act 1990*

**State and Territory**
*Health Records (Privacy and Access) Act 1997* (ACT)
*Health Rights Commission Act 1991* (ACT)
*Freedom of Information Act 1992* (Qld)
*Privacy and Personal Information Protection Act 1998* (NSW)
*Health Administration Regulation 2000* (NSW)
*Freedom of Information Act 1999* (Vic)
*Data Protection Bill 1999* (Vic)
*Information Privacy Act 2000* (Vic)
*Health Records Act 2000* (Vic)
*Health Commission Act 1976* (SA)
*Housing Trust Act 1995* (SA)
*Community Housing Authority Act 1991* (SA)

### Agency-specific legislation

*Social Security Administration Act 1999* — FaCS
*Child Care Act 1972* — FaCS
*Disability Services Act 1986* — FaCS
*Home and Community Care Act 1985* — DoHA
*Aged Care Act 1997* — DoHA
*Public Sector Management Act 1995* — DHS
*Family and Community Services Act 1997* — DHS
*Public and Environmental Health Act 1987* — DHS
*Transplant and Anatomy Act 1983* — DHS
*Mental Health Act 1993* — DHS

**APPENDIX E**

# Draft linkage documentation

## Version 1: Data repository

**MEMORANDUM OF UNDERSTANDING**

(reference number **XXXXX**)

**between**

**AAAAA**

**and**

**BBBBB**

**concerning**

***one-line description of overall project***

THIS MEMORANDUM OF UNDERSTANDING
(reference number **XXXXX**) is made between:

***AAAAA*** *representing* ***XXXXX*** *(in this memorandum of understanding called* ***AAAAA****);*

**AND**

***BBBBB*** *representing* ***XXXXX*** *(in this memorandum of understanding called* ***BBBBB****).*

The parties have determined that they wish to cooperate to enable the completion of a project to link specific person – level data from data sets ***list the data sets and the purpose of the project.***

## OVERVIEW

***A brief overview of the memorandum, with emphasis on the protocol to be observed and stressing the attention paid to the protection of privacy and confidentiality. A sample follows:***

This memorandum covers the linkage and extraction of AAAAA data relating to clients living in [State]. The data will be supplied to nominated analysts as de-identified linked files for use in planning and research on [Program — for example, aged and community care services]. The period of interest covers [time period 1] through [time period 2] and information will be included on individuals living in the State and registered as [program] clients.

The fundamental protocol aims to:

- maximise the conservation of individual privacy;
- minimise access to identified data;
- allow data custodians full control over the dissemination and use of de-identified data files;
- provide linked data files only to named analysts involved in specific approved projects;
- provide analysts with no more than the minimal data required for their analyses; and
- ensure that all copies of named data and all linked data files are destroyed immediately after use.

The process involves two separate stages. The first stage is a memorandum of understanding to share data for an agreed purpose and to prepare a linkage key file (using probabilistic methods) and a master copy of a de-identified linked data file to be stored in a safe repository.

The second stage includes a defined number of separate research projects.

Each separate research project will be covered by its own agreement, the data for the project being supplied by the data repository to analysts to conduct research on behalf of the steering committee. An agreement pro forma, to be completed by the researchers and sent to the steering committee by way of application for data, is in Attachment 2. For each research project, a unique set of project identification numbers (PID) will be generated by the repository custodian and attached to the copy of the master de-identified linked data file provided to the analysts.

This two-stage process will ensure that data custodians have full control over the distribution and usage of their data, as each project will need to be well defined and then individually approved before proceeding. No research will be undertaken without the written approval of every data custodian supplying data to the project. Linked data files will be provided only to the individually identified analysts from the data repository doing the analysis for each project, and will be destroyed when the analyses are complete. A different PID will be used in each project, thus making it extremely difficult to merge the linked data for two projects (such a process is, in any case, specifically prohibited).

The linkage key file will be produced by a small technical team specialising in data linkage. All people involved in the actual linkage and therefore requiring access to the data used in the linkage process will sign confidentiality agreements and be named on a list provided to the steering committee. Any changes to this list will be reported in

writing to this committee. No other personnel will be allowed access to the files used in this process, as they will contain private and confidential information. The work will be done on an isolated computer, and all personal demographic data (with the exception of those items required to derive variables for further analysis) will be destroyed as soon as the linkage is complete. Transfer of these data files will be done only via tape, diskette or CD-ROM personally carried by those personnel taking part in the data linkage. The linkage personnel will not be permitted to take any part in the analysis of the linked data, or to have any communication about these data with the analysts.

The repository will supply linked data directly to the nominated analysts (from the data repository, or to a contracted third party) for each project. These analysts will also sign confidentiality agreements as required by all parties. They will be the only people granted access to the de-identified linked information. They will be specifically forbidden to disseminate copies of the data files, and will be required to destroy these files on completion of the analysis.

The arbitrary reference number (XXXXX) enables agreements for individual research projects to refer to this over-arching first-stage document.

### INTERPRETATION

In this memorandum of understanding unless the contrary intention appears:

- 'MOU' means this memorandum of understanding signed by the parties and includes any schedules or attachments hereto.
- A reference to this MOU or another instrument includes any variation or replacement of them.
- The singular includes the plural and vice versa.
- The masculine includes the feminine and neuter; the feminine includes the masculine and neuter; the neuter includes the masculine and feminine.
- The word 'person' includes a firm, an unincorporated association or any authority.
- A reference to a person includes a reference to the person's executors, administrators, successors, substitutes (including, without limitation, a person taking by novation) and assigns.
- An agreement, representation or warranty on the part of or in favour of two or more persons binds, or is for the benefit of them, jointly and severally.
- A reference to any thing (including, without limitation, any amount) is a reference to the whole of any part of it and a reference to a group of persons is a reference to any one or more of them.
- A reference to all clauses, exhibits, annexures or schedules shall, unless otherwise provided, be the clauses, exhibits, annexures or schedules of or to this MOU.
- Headings have been inserted for ease of reference only and shall not be regarded as forming any part of the context of this MOU.
- A reference to a statute shall include all statutes amending, replacing or consolidating the statute referred to.
- 'Intellectual property' includes all copyright, all rights in relation to inventions (including patents), registered and unregistered trademarks (including service marks), registered designs, confidential information and circuit layouts, and all other rights resulting from intellectual activity associated with the design, development, delivery or findings of this project, including applications or rights to apply for registration of any of these rights.

- 'Personal information' means information or an opinion (including information or an opinion forming part of a database), whether true or not, and whether recorded in a material form or not, about an individual whose identity is apparent, or can reasonably be ascertained from the information or opinion (refer to s.6(1) of the *Privacy Act 1988*).
- 'Management committee' means the committee established under clause 3.1 of this MOU.
- 'Steering committee' means the committee established under clause 11.1 of this MOU.
- ***Add any further definitions that may be necessary.***

## OPERATION

This MOU and its operations shall be managed by a steering committee, consisting of:

– The **Secretary** of **AAAAA**, or nominee.
– The **Chief Executive Officer** of **BBBBB**, or nominee.
– ***Add further members as agreed.***

The **Secretary of XXXXX**, or nominee, will chair the steering committee which will meet if and as required, by teleconference or by face-to-face meeting. Parties are to meet their own costs.

The establishment of the steering committee is described at clause 11.

The work protocol to be followed is described in Attachment 1.

Data will be provided in accordance with the requirements of the agreed work protocol.

The provision of the data shall be subject to the requirements of ***list here the appropriate acts, guidelines and ethics committees.***

All results of research or analysis undertaken through approved projects will be placed in the public domain as soon as is feasible according to specification in the Agreement for each project, with each party acknowledged as the source of the data.

## OWNERSHIP AND USE OF MATERIAL

Analyses and resulting manuscripts based on each de-identified linked data set will be made available to the steering committee for comment within 30 working days. Any comments will be forwarded to the analysts for their consideration. Analysts should make every effort to take all comments into account and respond to any comments offered but not accepted.

The title to and intellectual property rights in all materials generated in relation to this MOU shall vest jointly upon its creation in each party to the MOU.

Steering committee members are required to disclose all intended publications and reports of results to the steering committee. The parties may decide to further disseminate material received, with full acknowledgment of the source, including through their own publications and other output.

## CONFIDENTIALITY

Each party acknowledges that any data provided under this MOU are subject to the confidentiality provisions of the Act under which they were collected.

Officers handling identified data for the purposes of this MOU will be required to sign confidentiality agreements and/or unilateral deed polls as required by each of the Parties.

The parties recognise that in accepting identified demographic data from AGENCY A, AGENCY B, AGENCY C etc., each party becomes subject to the provisions of:

- ***List here the appropriate acts and other documents that are appropriate and relevant. The following are some typical examples:***
- ***Section 95 of the* Privacy Act 1988*;***
- ***The Information Privacy Principles in section 14 of the* Privacy Act 1988*;***
- ***Section 135A(4) of the National Health Act 1953;***
- ***The Privacy Commissioner notes, May 1997;***
- ***The National Statement on Ethical conduct in Research involving Humans — National Health and Medical Research Council, 1999.***

Each party will carefully observe these provisions and the conditions of the undertakings signed prior to receipt of data.

The steering committee will maintain appropriate administrative records of all data supplied to the linkage group and of all de-identified linked data supplied to researchers, as well as the undertakings signed by those people authorised to have access to the data.

On completion of the record linkage described in the work protocol, the de-identified linkage key file will be created and all files used by matching staff during the record linkage will be permanently and irretrievably destroyed. The master copy of the de-identified linked data file will then be created.

Agency C will act as custodian of the data repository, providing secure long-term storage for the linkage key file and the de-identified linked data file.

When a specific approved research program is complete, the analysts are to inform the steering committee in writing, attesting to the permanent and irretrievable destruction of the data set used in that project.

Each party will, at all reasonable times, give to the other parties, or to any person authorised in writing by the parties, permission to inspect the arrangements for storage and security of any data relating to the project. Researchers will be required to allow similar inspections of arrangements for the storage and security of their de-identified linked data.

## DISPUTES

Where there is a conflict between the parties over any matter related to issues covered by this MOU, parties will seek to resolve the issue through the steering committee.

Should the parties fail to resolve a conflict, the matter shall be referred for resolution to the [relevant arbitrator].

### ENTIRE MEMORANDUM OF UNDERSTANDING AND VARIATION

This MOU along with any attachments is to be extended by the addition of specific agreements for each research project proposed. This overarching MOU constitutes the entire agreement between the parties and supersedes all communications, negotiations, arrangements and agreements, either written or oral, between the Parties with respect to the matter hereof, except where otherwise required in law.

Each research project will require the signature of all parties on a detailed specification in writing as described in the project schedule (see Attachment 2).

No variation or extension to this MOU shall be binding upon any party unless in writing and signed by all parties.

### TERM

This MOU shall commence when signed by all parties and shall remain in force and effect until otherwise agreed by all parties, or unless terminated in accordance with the terms hereof. At termination of the MOU all parties will destroy any de-identified linked data sets in their custodianship, and the Australian Institute of Health and Welfare will archive a copy of the linkage key file before destroying all other copies.

This MOU may be reviewed upon written request by any party.

### TERMINATION

A party shall have the right to request termination of this MOU by written notice to the steering committee if any other party or parties fails to comply with any of the terms and conditions of this MOU.

Any party may terminate this MOU by giving the other parties three months notice in writing to terminate.

Upon termination pursuant to clauses 9.1 or 9.2 all materials and data relating to the project shall be destroyed by all parties, with the steering committee determining the timing and manner whereby any approved projects are to be terminated.

Any termination under clauses 9.1 or 9.2 by any party shall result in the termination of the MOU for all other parties.

### ASSIGNMENT

Each party may not assign or otherwise deal with their rights under this MOU without the prior written consent of the other parties, which consent may be given on such terms or conditions as the other parties think fit.

### STEERING COMMITTEE

The project shall be guided by a steering committee consisting of:
(a) a nominee from **AAAAA**; and
(b) a nominee from **BBBBB**.
***Add further members to this committee as agreed by the parties to the memorandum.***

A nominee of XXXX will chair the steering committee which will meet if and as required, by teleconference or by face-to-face meeting. Parties are to meet their own costs.

## NOTICES

Any notice or other communication required or given under this MOU shall be deemed to have been given and received at the time of delivery or electronic receipt or within seven days of posting if it is sent by pre-paid post.

The address for service of a notice or invoice shall be as follows:

***AAAAA***

| | |
|---|---|
| *Address:* | ***xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx*** |
| | ***xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx*** |
| | ***xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx*** |
| *Postal address:* | ***xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx*** |
| | ***xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx*** |
| | ***xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx*** |
| *Attention:* | ***xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx*** |
| *Facsimile number:* | ***xxxxxxxxxxx*** |

***BBBBB***

| | |
|---|---|
| *Address:* | ***xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx*** |
| | ***xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx*** |
| | ***xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx*** |
| *Postal address:* | ***xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx*** |
| | ***xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx*** |
| | ***xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx*** |
| *Attention:* | ***xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx*** |
| *Facsimile number:* | ***xxxxxxxxxxx*** |

## INSURANCE

The parties shall have and maintain in force all normal and appropriate insurance including public liability insurance relating to the performance of their responsibilities under this MOU.

## GOVERNING LAW

The interpretation and construction of this MOU shall be governed and determined in accordance with the law of the Australian Capital Territory and the parties submit to the non-exclusive jurisdiction of the courts of that Territory.

## ETHICAL CONSIDERATIONS

The overall project must be approved by ***list the appropriate ethics committees***. Such approvals may be sought before or after the signing of this MOU, but must be completed before any linkage work is undertaken.

## INDEMNITY

Each of the parties hereby indemnifies and shall keep indemnified each of the other parties, their respective officers, servants and agents from and against all actions, claims, demands, proceedings, costs and expenses made, sustained, brought or prosecuted in any manner based upon, occasioned by or attributable to any injury or any person (including death), or loss or damage to property which may arise from or be a consequence of any unlawful or negligent act or omission of that Party or its officers, employees, servants or agents, in carrying out any of the obligations imposed by this MOU.

Each party must not represent itself, and must ensure that its employees do not represent themselves, as being employees or agents of the other parties.

Each party shall not by virtue of this MOU be, or for any purpose be deemed to be, an employee or agent of another party.

### RESPONSIBILITIES OF AAAAA

***List responsibilities for each agency as agreed by the parties. Examples are shown below:***

***AAAAA** will identify records of the **XXXX** client population for the period from **XXXX** through **XXXX** inclusive, and will prepare files of demographic data for these individuals to be used in linkage to data supplied by the other parties.*

***AAAAA** will encrypt the full demographic data from **XXXX** data collection to be used for linkage.*

***AAAAA** will extract and supply de-identified, aggregated service data to the data repository (linkage team) to be merged into a single de-identified linked data file.*

***AAAAA** will take the lead in collaborating with the other parties to select a suitable agency as custodian of the repository. This agency will link together the demographic files and provide secure storage and access to the de-identified linked data set for the nominated analysts.*

### RESPONSIBILITIES OF BBBBB

**BBBBB** will identify records of the **XXXX** client population for the period from **XXXX** through **XXXX** inclusive, and will prepare files of demographic data for these individuals to be used in linkage to data supplied by the other parties.

**BBBBB** will encrypt the full demographic data from **XXXX** data collection to be used for linkage.

**BBBBB** will extract and supply de-identified, aggregated service data to the data repository (linkage team) to be merged into a single de-identified linked data file.

**Executed as a memorandum of understanding:**

SIGNED for and on behalf of **Agency A**

_____Date_____

**Xxxx X Xxxxxxxx, Chief Executive Officer, Agency A**

SIGNED for and on behalf of **Agency B**

_____Date_____

**Xxxx X Xxxxxxxx, Chief Executive Officer, Agency B**

## Attachment 1 — Work protocol

### Background

*A brief description of the background to the project as a whole.*

### Objectives

*A brief description of the objectives of the project.*

### Scope of study

*List the populations and data sets on which the study will be based.*

### Work plan

Production of linkage key file and master de-identified linked data file

Demographic data from **XXXXX** to be supplied by **AAAAA**

***List fields to be included in this file, e.g.:***

Identification number (encrypted)

| | | |
|---|---|---|
| Surname | First given name | Second given name |
| Gender | Date of birth | |
| Country of birth | Indigenous status | |
| Address | Suburb/town | Postcode |
| First date of contact | Last date of contact | |

Demographic data from **XXXXX** to be supplied by **BBBBB**

***List fields to be included in this file e.g.:***

Identification number (encrypted)

| | | |
|---|---|---|
| Surname | First given name | Second given name |
| Gender | Date of birth | |
| Country of birth | Indigenous status | |
| Address | Suburb/town | Postcode |
| First date of contact | Last date of contact | |

Date of death (if known to be deceased)

### Linkage process

**AAAAA** and **BBBBB** will deliver their demographic files to **Agency C**, where the primary linkage will be done. The linkage will result in the creation of a demographic file containing [add characteristics of linked file here–for example, linked pairs of hospital patients and aged and community care clients].

Delete all the demographic files used to create the linked file.

### Merging of service data

The linkage team will merge together the service data, making use of the information contained in the linkage file, to create the master copy of the de-identified linked data file.

### Extraction of de-identified linked data files

The linkage team will prepare extracts from the master de-identified linked data file to be sent to the nominated analysts (that is, either from the data repository or from a contracted third party). A unique project identification number (PID) will be generated and attached to an extract from the master file containing only the fields approved for the particular research project. The data file will be sent directly to the analysts named in the project.

### Analysis

The analysts will be the only people to have access to the combined linked files, and will be under an obligation not to disseminate copies of the files, or to allow any other personnel to have access to the files. When the analyses are complete the data files will be destroyed.

### Summary of entire process

| | | |
|---|---|---|
| **Implementation** | Data custodians | Agree to work together on project. |
| | | Select agency to perform linkage and act as data repository. |
| | | Draw up memorandum of understanding. |
| | | Obtain ethical approvals where necessary. |
| | | Sign memorandum of understanding. |
| | | Create steering committee. |
| **Linkage** | Data custodians | Prepare demographic files containing at least one record for each individual. Each record will contain encrypted identifying demographic variable(s) and de-identified service experience data (may also be encrypted). |
| | | Supply files to linkage/repository agency. |
| | Linkage agency | Supply list of linkage personnel to steering committee. |
| | | Linkage personnel sign confidentiality forms and send them to steering committee. |
| | | Link files by means of the identifying variables. |
| | | Delete all identifying variables to create 'linkage key file'. |
| **Master copy of linked data file** | Data management team | Supply list of personnel who will have access to the data to steering committee. |
| | | Personnel with access to data sign confidentiality forms and send them to steering committee. |
| | | Use master copy of de-identified linked data file for specified analyses. |
| **Supply of analyses** | Steering committee researchers | Receive from steering committee completed analyses. |
| | | Provide comments/requests for further analyses through steering committee within 30 days. |
| | Steering committee | Request data repository to provide subsequent analyses/incorporate comments. |
| | Linkage team | Add arbitrary PID to copy of linked data file. |
| | | Extract copy of master de-identified linked data file containing only additional variables approved for the further analyses. |
| | | Supply extracted data file to data management team. |
| | Data management team | Perform analysis and prepare reports for steering committee. |
| | | Send copies of all output to steering committee. |
| | | Delete data as soon as practicable, notifying steering committee. |

## Attachment 2 — Pro forma for agreement on research project

### Agreement for research project

(Project reference number **XXXXX/xx**)

using de-identified linked data available as a result of the memorandum of understanding ***put name of project here***

(reference number **XXXXX**)

**between**

**Agency A**

**and**

**Agency B**

**to approve a research project led by**

***enter chief investigator's name and (in parentheses) institution***

### INTERPRETATION

In this agreement, unless the contrary intention appears, all interpretation is as defined in the over-arching memorandum of understanding (reference number **XXXXX**).

### OVERVIEW OF RESEARCH

***enter a brief overview of the research project (no more than two pages)***

### PROPOSED OUTPUT / PUBLICATIONS

***for example, internal report, planning paper, academic journal (name potential journals)***

### DATA TO BE USED IN ANALYSES

***list all required fields***

### CONDITIONS FOR DATA ACCESS

Personnel listed under paragraph 6 of this agreement are the only people to be allowed access to these data files. These individuals will be required to sign confidentiality agreements as supplied by the steering committee. Any changes to this list must be notified in writing to the chairperson of the steering committee.

By accepting de-identified unit record data from **Agency A** and **Agency B**, these individuals and their host institution become subject to the provisions of:

***list all appropriate acts, guidelines, etc.***

The data files must be kept on secure computer systems requiring encrypted password entry.

No part of these data will be copied or made available (in any format) to any other individuals or institutions.

These data will not be linked or merged with any other data sets, including data sets generated under a separate agreement covered by the same memorandum of understanding as that for this research project.

These data will only be used for the analyses and output defined in sections 1 and 2.

No attempt will be made to identify any individual whose data are in these files.

Publications and other output will not contain tables or other information that might allow readers to identify any individual whose data have been used in this project.

All analyses and resulting manuscripts from this project will be made available (via the chairperson of the steering committee) to all the parties to the memorandum of understanding giving them opportunity to comment (within 30 working days). The researchers have the right to publish results that suitably address any qualifications or comments by the parties.

All reports and publications resulting from this project must acknowledge **Agency A** and **Agency B** as providers of the data.

If the analysis is not concluded within twelve months from provision of the data, a progress report and request for continued access must be provided to the steering committee.

At the conclusion of the analyses all copies of the data files will be destroyed with written notification to this effect to the chairperson of the steering committee. If necessary, new copies of the data files can be supplied by the data custodians subject to the approval of the steering committee.

The steering committee may request a progress report on this project at any time.

## RESEARCH PROJECT PERSONNEL

By signing this agreement, the following personnel agree to observe the terms and conditions listed in this agreement and in the over-arching memorandum of understanding. These are the only individuals permitted to have access to the linked de-identified data files extracted for the purposes of this research project. Any changes to this list must be notified in writing to the chairperson of the steering committee. This notification must include full details (as under) as well as the dated signature(s) of any personnel added to this list. Signed copies of all confidentiality undertakings and/or unilateral deed polls must be supplied for these extra personnel who will have to comply with the terms of the agreement in the same manner as the personnel listed below.

List chief investigator first, then all others in the following format:

**Chief investigator:**

Title and name:_____

Address:_____

_____

Telephone number:_____

Fax number: _____

Email:_____

Signed: _____Date: _____

**Other investigators:**

Title and name:_____

Address:_____

_____

_____

Telephone number:_____

Fax number: _____

Email:_____

Signed:_____Date: _____

**Executed as a agreement:**

SIGNED for and on behalf of **Agency A**

_____Date_____

**Xxxx X Xxxxxxxx, Chief Executive Officer, Agency A**

SIGNED for and on behalf of **Agency B**

_____Date_____

**Xxxx X Xxxxxxxx, Chief Executive Officer, Agency B**

### Attachment 3 — Notifications of approval from ethics committees

*Include copies of the approval documents from each ethics committee.*

# Abbreviations

| | |
|---|---|
| ABS | Australian Bureau of Statistics |
| ACT | Australian Capital Territory |
| AIHW | Australian Institute of Health and Welfare |
| ASCII | American Standard Code for Information Interchange (a standard computer code) |
| CSDA | Commonwealth/State Disability Agreement |
| CSDA MDS | Commonwealth/State Disability Agreement Minimum Data Set |
| CSMAC | Community Services Ministers' Advisory Council |
| DoHA | Department of Health and Ageing (Commonwealth) |
| FaCS | Department of Family and Community Services (Commonwealth) |
| HACC | Home and Community Care |
| HACC MDS | Home and Community Care Minimum Data Set |
| HDWA | Health Department of Western Australia |
| HIC | Health Insurance Commission |
| IPP | Information Privacy Principles |
| IVF | in-vitro fertilisation |
| LLFF | (Canadian) Longitudinal Labour Force File |
| MCHRDB | Maternal and Child Health Research Data Base (Western Australia) |
| MOU | memorandum of understanding |
| NCSIMG | National Community Services Information Management Group |
| NPP | National Principles (for the fair handling of Personal Information) |
| NSW | New South Wales |
| OFPC | Office of the Federal Privacy Commissioner |
| PID | Project Identification Number |
| Qld | Queensland |
| SA | South Australia |
| SAAP | Supported Accommodation Assistance Program |
| SIDS | sudden infant death syndrome |
| SLK | statistical linkage key |
| SLKWG | Statistical Linkage Key Working Group |
| Vic | Victoria |
| WA PID | Western Australian personal identifier number |

# Notes

# Notes