Data linkage protocols using a statistical linkage key

The Australian Institute of Health and Welfare is Australia's national health and welfare statistics and information agency. The Institute's mission is *better health and wellbeing for Australians through better health and welfare statistics and information*.

Please note that as with all statistical reports there is the potential for minor revisions of data in its life. Please refer to the online version at <www.aihw.gov.au>.

DATA LINKAGE SERIES
Number 1

Data linkage protocols using a statistical linkage key

Rosemary Karmel

November 2005

Australian Institute of Health and Welfare
Canberra

AIHW cat. no. CSI 1

A complete list of the Institute's publications is available from the Business Promotion and Media 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>.

ISSN 1833-1238

ISBN 1 74024 511 3

Suggested citation

AIHW: Karmel R 2005. Data linkage protocols using a statistical linkage key. AIHW cat. no. CSI 1. Canberra: AIHW (Data Linkage Series no. 1).

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

Community Services Integration and Linkage Unit
Australian Institute of Health and Welfare
GPO Box 570
Canberra ACT 2601
Phone: (02) 6244 1000

Published by Australian Institute of Health and Welfare

Printed by Pirion

Contents

Tables and figures

Acronyms

ACAT	Aged Care Assessment Team
ACCMIS	Aged and Community Care Management Information System
AIHW	Australian Institute of Health and Welfare
C3C2-SLK	Statistical linkage key based on first three consonants of family name, first two consonants of given name, date of birth and sex
CACP	Community Aged Care Package
DoHA	Department of Health and Ageing
HACC	Home and Community Care
MDS	Minimum data set
RACS	Residential aged care service
SAAP	Supported Accommodation Assistance Program
SLK	Statistical linkage key
SLK-581	Statistical linkage key based on five letters of name (5), date of birth (8) and sex (1) (see Section 1.2)

Symbols

. .	when used in a table, means not applicable
—	when used in a table, means nil or rounded to zero (including null cells)

Acknowledgments

This report was prepared as part of a project funded jointly by the Australian Institute of Health and Welfare and the Department of Health and Ageing. The report was authored by Rosemary Karmel of the Community Services Integration and Linkage Unit within the Institute. Important contributions were also made by a number of staff (past and present) of the Institute's Ageing and Aged Care Unit, including Evon Bowler, Peter Braun, Anne Jenkins and Heather Logie. Useful comments on various drafts of the report were provided by Diane Gibson, Ruel Abello and Ann Peut of the Australian Institute of Health and Welfare, and by David Martin, Tom Goff, Gilian Lee, D'Arcy Jackson and John Patroni of the Department of Health and Ageing.

Summary

Three main programs provide care services to older people in Australia: the Home and Community Care (HACC) program, Community Aged Care Packages (CACPs) and residential aged care services (RACS). With the development of the quarterly collection for the Home and Community Care program, it has become possible to develop a picture of transitions between these programs by linking the various data collected on the three programs.

Complete demographic data—such as name, address and date of birth information—are not available for all three programs. However, data linkage is possible using a statistical linkage key (named the SLK-581) which is either directly collected or derivable from the data available for the three programs. Previous analysis had shown that the likelihood of common SLK-581 linkage keys for different individuals in large aged care data sets is very low (AIHW: Ryan T et al. 1999), indicating that this key could be useful for linking across data sets.

This report examines the quality of the data available for undertaking statistical data linkage between the programs and describes the protocols followed to ensure that the privacy of individuals is not compromised. Practices that allow consistent linkage procedures to be used over time and across data sets are also outlined. A second report, *Transitions Between Aged Care Services* (AIHW: Karmel 2005), examines the validity of the links established via SLK-581 and analyses the resulting linked data.

HACC MDS data quality

Data on services received by people using the HACC program comes from the quarterly HACC national minimum data set (MDS). A statistical linkage key (SLK-581) is included in this collection. In the HACC data it is assumed that the linkage key uniquely identifies an individual. Consequently, for the HACC data it has not been possible to investigate either the extent of common linkage keys for more than one client or the extent to which individuals may have more than one linkage key due to name variations.

Incomplete information affects between 3% and 4% of the linkage keys in the quarterly HACC MDS, with the main cause being poor date of birth information (identified by 1 January birth dates). This raises the question of using additional data to identify links for records with poor linkage key information. Use of postcode of the client's usual residence has been identified as providing high discriminatory power when some of the linkage key information is missing or incomplete. However, even when postcode data are available, valid date of birth information is required before considering matching for records with some poor linkage key data. Information relating to cultural diversity does not generally provide additional discriminating power, even when much of the linkage key information is valid.

Although the above issues affect any linkage undertaken with the HACC MDS, the greatest cause of missed links between programs is likely to be the non-participation of some agencies in the HACC data collection which results in some clients being absent altogether from the data sets. For the two quarters used in this study (the September and December quarters, 2002), 78% and 86% of HACC agencies participated in the data collection; however, as many non-participating agencies have very small numbers of clients, the percentages of clients included in the collections are expected to be higher.

CACP and RACS data quality

For CACP and RACS, administrative by-product data is collected and stored on the Department of Health and Ageing's Aged and Community Care Management Information System (ACCMIS). Complete name, date of birth and sex data are held on ACCMIS for both CACP and RACS clients, and these can be used to construct the SLK-581 linkage key corresponding to the key directly collected for the HACC MDS. The quality of linkage key data on ACCMIS is generally high. The main causes of errors are the use of 1 January birth dates as default dates when full date of birth is not known and multiple representation of clients in the data. For clients with admissions between 1 July 2000 and 30 June 2003, just under 0.5% of RACS clients had reported 1 January birth dates, and 0.9% of CACP clients had such birth dates. Multiple representation on the ACCMIS database is less common, affecting fewer than 0.3% of client records relating to admissions over the 3 years. For CACP clients, very few cases of multiple representation were identified—0.05% of client records were affected.

Privacy protection protocols

The key features of the protocol developed to protect clients' privacy when undertaking the data linkage for this project are the separation of personal identifying information from service information, and the absence of any record identifiers which would allow linkage back to the source data. To achieve this, several underlying principles were adopted when developing the protocol used for this study. These relate to data handling procedures and so do not necessarily require that the people undertaking the linkage are different from those doing the subsequent analysis. The following linkage principles were used:

- Data linkage is not to be carried out directly between original complete data sets.
- Data linkage is to be undertaken using data sets that contain only the data required for establishing and validating links.
- Links between data sets are to be recorded using *project-specific* unique record identifiers so that links identified for a particular project cannot be used to establish links between other data sets using a chain of links ('consequential' linking).

- Analysis files are not to contain identifying data (such as name and address, or the record number from the original data set).
- Intermediate data sets and the project-specific record identifiers are to be deleted following development of the final linked analysis data sets.

Linkage protocols

The purpose of linkage protocols is to ensure that appropriate methods are used when undertaking the linking and that a linkage analysis is reproducible. Two options are described for undertaking linkage, depending on the additional data available to enhance the linkage. Briefly, these are:

- *Option 1 – Basic linking*, excluding all records with linkage keys incorporating missing or poor information from the data set for linking.
- *Option 2 – Enhanced linking*, retaining, as far as possible, all valid data in the linkage key, and using this in conjunction with other information to establish links with other data sets.

When linking data for aged care programs, the individual data sets should be analysed to identify the extent of poor quality linkage keys and multiple representation of clients. Where comparable client postcode data are available, enhanced linkage which uses postcode to augment linkage keys with some missing data can be used. Among the aged care programs considered in this report, due to differences in data collection methods such enhanced linkage is appropriate only when linking between the HACC and CACP programs. Using basic linkage between HACC and CACP, rather than enhanced linkage, marginally reduces the HACC records available for linking. In the final linked data set, differences between variables common to the two source data sets should be resolved using pre-determined rules.

1 Introduction

There are several programs which provide care services to older people in Australia. Information on the movement of people between these programs would help policy analysts and researchers to understand when and why people move between these services, and would provide insight into the interplay of the various programs in the Australian aged care system.

If sufficient data were available on two or more programs, data linkage could be used to link the program data sets and thus allow examination of relationships and movements between programs. In such a scenario, only statistical linkage would be required as 'the individual unit…is important only in terms of its contribution to the pattern of use of the client group overall' and 'the identity of the individual unit is unimportant for "statistical" linkage (whereas the identity of the unit is critical for "administrative" linkage purposes)' (NCSIMG 2004:5). Such statistical linkage between three aged care programs is the subject of this report.

Often statistical linkage between data sets is based on full name and other demographic data, and the data is linked using probabilistic methods based on the similarity of the demographic data in records in the data sets being linked (see Box 1; NCSIMG 2004:10–11). However, complete name data are not essential for data linkage if sufficient data are available to distinguish between individuals with high probability. In some data sets, a statistical linkage key—'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' (NCSIMG 2004:12)—is available which can be used to distinguish between individuals. The statistical linkage key can then be used to link records either deterministically or probabilistically between data sets.

Box 1: Statistical linkage methods

Deterministic matching *links records using a fixed set of variables and involves exact one-to-one character matching of these variables. When linking records, only those with exactly the same data for the set of linkage variables are considered to match.*

Probabilistic matching *uses mathematical algorithms to determine the likelihood or probability that two or more records from the same, or different, data sets represent the same person or entity. When comparing two records, each variable is compared and assigned a score based on how well it matches. Matching on a rare characteristic is given a higher score than matching on a common one. The final score for the comparison is the sum of the scores for the individual variables. The decision on whether two records match is based on the total match score; the higher the score the more likely it is that the records match. Cut-off scores are used to distinguish between matches and non-matches. Often some clerical review of matches is undertaken for those comparisons with match scores around the cut-off. Variations in reported data, for example in names or dates of birth, can be allowed for in probabilistic matching.*

Three key programs which provide care services to older Australians are the Home and Community Care (HACC) program, Community Aged Care Packages (CACPs) and residential aged care services (RACS). For many years administrative by-product data have been collected on the clients of the RACS and CACP programs and stored in the Department of Health and Ageing's Aged and Community Care Management Information System (ACCMIS). However, it was not until the implementation of the quarterly national minimum data set (MDS) collection for the Home and Community Care program in 2001 that sufficient data became available to support data linkage between the data sets for the three programs. Different data are available for the three programs and because, unlike the two other programs, the HACC data does not contain full name information, full name-based probabilistic linkage cannot be used. However, the HACC MDS contains data for a statistical linkage key which can also be derived for the other programs. Therefore, in this study, a statistical linkage key (based on parts of name, date of birth and sex) has been used to distinguish between clients. Deterministic matching is then used to link records across data sets; that is, only records in different data sets with exactly the same linkage key are linked. The derivation of linked data sets through use of this key make it possible to identify the movement of clients between services, and to develop a picture of transitions between the main community and residential aged care programs.

The aged care programs included in this study and the statistical linkage key used to link the data sets (SLK-581) are described below. Chapters 2 and 3 discuss analyses undertaken to establish the utility of this linkage key in distinguishing between clients and the extent of non-unique linkage keys within the quarterly HACC MDS collections and within the CACP and RACS data. Other factors which might also affect the utility of SLK-581 for linking these data sets, such as data quality and rules for distinguishing between multiple keys for the same person and identical keys for different people, are also discussed. It has not been possible, however, to examine whether the utility of SLK-581 is changing over time, either as a result of the increasing numbers of older people or as name patterns change (for example, due to the ageing of Asian immigrants who often have short names).

Linking data from different sources using a statistical linkage key needs to be subject to clear protocols, both to ensure that the privacy of individuals is not compromised, and to ensure that consistent linkage procedures are used over time and across data sets. Data handling protocols that protect the privacy of individuals when linking data sets are outlined in Chapter 4, and practices to ensure consistency of linkage across data sets and over time are discussed in Chapter 5. While the analysis carried out in the course of establishing linkage protocols and practices (and discussed in this report) does not involve linking data sets, ethics approval to undertake the linkage and subsequent analysis was obtained from the Australian Institute of Health and Welfare's Ethics Committee before the project began. Analysis of quarterly movements between programs using linked data derived using the protocols presented in this report are described in a second report *Transitions Between Aged Care Services* (AIHW: Karmel R 2005).

1.1 The aged care services

The aged care services included in this study are the Home and Community Care program, Community Aged Care Packages and residential aged care services.

The bulk of home- and community-based services for older people are provided under the auspices of the HACC program. The HACC target population is people of all ages requiring assistance due to disability and/or frailty (and their carers). The aim of the program is to enhance the independence of people in these groups and avoid their premature or inappropriate admission to long-term residential care. The program includes home nursing services, delivered meals, home help and home maintenance services, transport and shopping assistance, allied health services, home- and centre-based respite care, and advice and assistance of various kinds. HACC also provides brokered or coordinated care for some clients through community options or linkages projects. During 2002–03, at least 661,000 clients received services through Home and Community Care; of these, just over three-quarters were aged 65 or more (DoHA 2003b).

Community Aged Care Packages provide support services for older people with complex needs living at home who would otherwise be eligible for admission to 'low-level' residential care. They provide a range of home-based services, excluding home nursing assistance (which may, however, be provided through HACC), with care being coordinated by the package provider. To receive a package, an Aged Care Assessment Team (ACAT) recommendation specifically for a CACP is required. On 30 June 2003, there were 26,573 people in receipt of a Community Aged Care Package, and during 2002–03 there were 14,719 admissions to a package (AIHW 2004b).

Residential aged care services provide accommodation and support for older people who can no longer live at home. To enter residential care, people must have the appropriate recommendation from an Aged Care Assessment Team. Two levels of care are available: low-level care and high-level care. Short-term respite care services are also available. During 2002–03 there were 51,200 admissions into permanent residential aged care and 45,445 into respite care, and on 30 June 2003 142,846 people were in permanent residential aged care and 2,549 were in residential respite care (AIHW 2004e).

1.2 The statistical linkage key (SLK-581)

The first step in a data linkage process using a statistical linkage key (SLK) is to construct a consistent linkage key in all relevant data sets. The linkage key to be used when linking aged care data collections was first proposed during the development of the HACC minimum data set. Analysis at that time showed that the likelihood of common linkage keys for different individuals in large aged care data sets is very low, with 0.6% out of the 440,000 people in the test aged care data set having a non-unique value for the proposed key (AIHW: Ryan et al. 1999:75–79).

The HACC statistical linkage key (HACC SLK) for a person is the concatenation of the 2nd, 3rd and 5th letters of the family name, the 2nd and 3rd letters of the given name, date of birth as a character string of the form *ddmmyyyy*, followed by the character '1' for male and '2' for female. Non-alphabetic letters in names are excluded (for example, hyphens and apostrophes), and where a name contains insufficient letters, the character '2' is used as a place marker for absent key letters. The character '9' is used for any other missing data so that the linkage key always has a length of 14 characters.

Although the HACC MDS does not contain the full name of clients, the letters required for the HACC SLK are reported, and the key is constructed for the MDS using this information in conjunction with date of birth and sex. It is also possible to construct the same linkage key for residential aged care service and Community Aged Care Package clients from the ACCMIS data as this database contains full name, date of birth and sex.

Other data sets which contain the HACC SLK include the censuses for Community Aged Care Packages, Extended Aged Care at Home and Day Therapy Centres (see AIHW 2004a, AIHW 2004c and AIHW 2004d), and the Commonwealth State/Territory Disability Agreement National Minimum Data Set (see AIHW 2003). The linkage key is also being considered for inclusion in a number of other collections. Because of its increasing use in other data collections, the HACC statistical linkage key is referred to as SLK-581 in the remainder of this paper to emphasise its broader use (where the '5' represents the five letters of name, the '8' represents the eight digits of date of birth, and the '1' represents the single character for sex).

Before any analysis of linked data is undertaken, it is first necessary to establish whether the SLK-581 provides a sufficiently accurate linkage key to generate a linked database useful for investigating movement between aged care services. For aged care programs, in the absence of a universal client identifier, an accurate statistical linkage key (or client identifier) would create a unique reproducible key for each client thereby allowing use of various programs to be linked appropriately. A unique linkage key is one where there is only one instance of the specific linkage key in a particular data set. A non-unique key is one where more than one record on the database contains the same linkage key. The lower the proportion of non-unique linkage keys in the population of interest, the greater the likelihood of appropriate linking, leading to greater validity in the linkage process.

The success of a linkage key in accurately linking data for individuals across data sets requires that the populations to which the linkage key is being applied are sufficiently diverse with respect to the characteristics contributing to the linkage key so that it can distinguish between individuals with a high degree of accuracy. For many analytical purposes, particularly when investigating patterns and trends among large groups, 100% accuracy is not required. Recalling that the three components of the linkage key are date of birth, sex and a combination of letters taken from given and family names, the ideal population for the application of the SLK-581 linkage key would be characterised by independent uniform distributions in

all three of these components. With respect to aged care clients, there is some divergence from this situation: there is a preponderance of female clients, and a general tendency to be dealing with older age groups which leads to clustering in the date of birth component. An associated factor is that given names go in and out of fashion, so that there is likely to be clustering around certain names for clients with close years of birth. The final combined distribution of all the linkage key elements determines the efficiency of the SLK-581 in distinguishing between individuals in the aged care data.

While different people may legitimately have the same linkage key, it is also possible for an individual to have more than one linkage key, either within the same or across different data collections; that is, the linkage key is not reproduced exactly on all occasions of data collection. This happens when different names or dates of birth are recorded at different times. When an individual has more than one linkage key it is likely that some or all of their information will not be linked. For example, an individual appearing as Joseph Smith in residential care may be identified as Joe Smyth in a HACC collection leading to different SLK-581 keys in the two data collections. Consequently the records for this individual would not be linked across these data sets using SLK-581. Similarly, records for people who change their name (for example, due to marriage or divorce) may not be linked either within or across data sets.

The extent to which multiple keys are derived for individuals cannot be quantified without cross-checking with other identifying data. The occurrence of multiple keys for individuals in hospital and death records was examined in an analysis of the Western Australian linked health database. In that study it was found that 2.1% of the 205,000 people identified in the records for 1 year using extensive name-based linkage processes (which allowed for changes in names and other demographic information) would have had more than one SLK-581 based on the demographic data reported at different times (NCSIMG 2004:62). Considering 3 years' data, 4.3% of the 470,000 people identified in the records would have had multiple SLK-581 keys over that period. These results cannot be generalised to other data collections as the method of collecting the linkage key data can also affect the prevalence of multiple keys. However, they do show that a proportion of people provide different identifying information on different occasions, with this being more likely as the time period increases. Similar analysis for the aged care data collections is not currently available, and so it is not possible to undertake comprehensive investigations at this stage. Nevertheless, it has been possible to examine this issue to some extent for RACS and CACP clients.

2 SLK-581 within the HACC MDS

The prevalence of non-unique and poor quality SLK-581 linkage keys has been investigated using data for two quarters of the 2002–03 HACC MDS: the July–September quarter containing 386,299 records, and the October–December quarter containing 405,404 records.

2.1 Non-unique keys

When looking at the HACC MDS, only client-based data sets have been examined; that is, where there is nominally one record per client. To obtain this data, SLK-581 is *assumed* by the HACC Data Repository to identify HACC clients uniquely, and service data from contributing agencies is collapsed across SLK-581s ensuring that the data contains only one amalgamated service record for each distinct SLK-581 linkage key. Consequently, there are only unique linkage keys in the HACC client-level data sets.

The above treatment of the HACC MDS collection began with data for the 2002 April quarter. Prior to this, clients with the same linkage key reported by the same HACC agency were assumed to be different people; in all other cases the linkage key was assumed to identify the client. Using this earlier rule, the proportions of records with unique linkage keys in the HACC MDS were 98.7% and 98.3% for the July–September and October–December quarter collections in 2001, respectively. These results suggest that the prevalence of non-unique linkage keys for different people is low and is unlikely to affect many analyses. However, as can be seen from the above example, the rules used to identify different people with the same linkage key affects the prevalence of non-unique keys within data sets, and care must be taken to choose appropriate rules for differentiating between identical, or coincident, linkage keys for different people and multiple, or replicated, keys for the same person.

2.2 Data quality

The quality of the data used to derive the linkage key also affects the utility of the key when linking both within and across data sets. Errors or missing data in any of the constituent elements will affect the quality of any links identified using the key.

Date of birth

Date of birth is an integral part of the SLK-581 linkage key. The guidelines to the HACC MDS state that a client's date of birth should be recorded as accurately as possible (DoHA 2004:17). In some cases, however, exact birth date may be unknown, and in this situation the guidelines advise using 1 January along with a known year of birth or a year estimated from the age of the client. The HACC MDS does not identify which dates of birth contain estimated information and which are complete.

Table 2.1: Frequency counts for dates of birth in SLK-581 linkage keys, HACC MDS

	July–September 2002			October–December 2002		
	1 January		Mean number per date in the decade[a] (Other 1 January dates)	1 January		Mean number per date in the decade[a] (Other 1 January dates)
Year	Number of SLKs	Per cent all SLKs		Number of SLKs	Per cent all SLKs	
Missing or unknown						
1900	1,198	0.31	4.54 (18[b])	1,727	0.43	4.67 (15[b])
1901	345	0.09	4.54 (18[b])	340	0.08	4.67 (15[b])
Total	*1,543*	*0.40*	. .	*2,067*	*0.51*	. .
Start of decade dates						
1910	124	0.03	26.94 (173)	112	0.03	27.41 (167)
1920	1,485	0.38	38.00 (241)	1,672	0.41	40.28 (247)
1930	747	0.19	16.83 (137)	817	0.20	18.15 (146)
1940	321	0.08	7.53 (69)	405	0.10	8.10 (84)
1950	532	0.14	4.96 (47)	550	0.14	5.15 (52)
1960	249	0.06	3.50 (38)	272	0.07	3.64 (48)
1970	95	0.02	2.32 (18)	127	0.03	2.35 (21)
1980	23	0.01	1.93 (9)	32	0.01	1.96 (11)
1990	25	0.01	1.84 (9)	29	0.01	1.84 (11)
2000	115	0.03	2.63 (151)	129	0.03	2.63 (158)
Total	*3,716*	*0.96*	. .	*4,145*	*1.02*	. .
Total missing /unknown/start of decade	*5,259*	*1.36*	. .	*6,212*	*1.53*	. .
Other 1 January dates	7,115	1.84	. .	7,517	1.85	. .
All 1 January dates	**12,374**	**3.20**	**. .**	**13,729**	**3.39**	**. .**
All SLKs	**386,299**	**100.0**	**11.64**	**405,404**	**100.0**	**12.13**

(a) Average based on birth dates which occur in the data set, excluding 1st of the decade.

(b) Excludes 1 January 1900 and 1901.

Note: A small number (under 10 per quarter) of SLKs contained birth dates earlier than 1 January 1900. These have been excluded from the analysis of missing, unknown and 01/01 birth date information.

Source: AIHW analysis of HACC MDS.

For HACC data, use of default dates results in unusually high frequencies of 1 January birth dates, with 39 of the 40 birth dates most frequently recorded in the July–September 2002 quarter involving a 1 January date. All of the 40 birth dates most frequently recorded in the October–December 2002 quarter involved a 1 January date. In total, 1 January birth dates affected over 3% of SLK-581 linkage keys in the two HACC quarterly collections considered (Table 2.1).

While the HACC MDS guidelines recommend recording estimated year of birth when date of birth is not known, some agencies use either 1 January 1900 or 1 January 1901 as default birth dates when the date is either unknown or missing from their records. For the September and December 2002 HACC quarterly collections, 1 January 1900 and 1901 dates were present in around 0.5% of all SLKs recorded. Analysis of reported dates also suggests that for around half of the 1 January birth dates very rough estimates for year of birth are being provided, with clustering especially evident at the start of each decade (1910, 1920, …). For both quarters examined, the largest cluster of dates was for 1 January 1920, with over of 1,400 SLK-581 keys including this birth date, compared with an average of around 40 per birth date in the 1920's. Overall, start of decade birth dates (1910 and on) were used in 1% of SLK-581 linkage keys for both periods, and other 1 January dates (excluding 1900 and 1901) in a little more than 1.8% of SLKs for both the September and December quarters, respectively. In general, these other 1 January dates were between four and 10 times more common than other dates in their decade, except for 1 January 2001 and 2002 birth dates which were about 60 times more common suggesting possible data entry errors for many of these cases.

Table 2.2: Number of HACC agencies contributing data per SLK-581 linkage key, HACC MDS

	July–September 2002		October–December 2002	
Error type	**Number**	**Mean number of agencies per SLK-581**	**Number**	**Mean number of agencies per SLK-581**
No error	372,744	1.22	390,395	1.23
1 January 1900 date of birth	1,198	1.02	1,727	1.03
1 January 1901 date of birth	345	1.01	340	1.01
Other 1st of decade date of birth[(a)]	3,717	1.03	4,147	1.04
Other 1 January date of birth	7,115	1.06	7,519	1.06
With missing name/sex[(b)]	1,958	1.02	2,146	1.01
Total	**386,299**	**1.21**	**405,404**	**1.22**

(a) Includes a very small number of cases with 01/01 birth date before 1900 (under 3 for both quarters).

(b) Includes cases with date of birth also missing: 778 in the September quarter, and 870 in the December quarter.

Source: AIHW analysis of HACC MDS.

Inaccuracies in variables contributing to the linkage key can result in either under-identification of individuals, through the amalgamation of service data for different clients, or over-enumeration of people (double counting) as records for clients are not

amalgamated due to different linkage key information being reported by different HACC agencies. If poor data were leading to under-identification, we would expect keys based on inaccurate information to have more HACC agencies contributing service data to the corresponding amalgamated records. However, on average, SLK-581 linkage keys with no apparent inaccuracies had more agencies contributing service data (averaging around 1.22) than keys containing possibly estimated dates of birth (averaging less than 1.06) (Table 2.2). Furthermore, none of the SLKs reported by a very high number of agencies (eight or more) had 1 January birth dates. These findings suggest that a poor link key date is more likely to lead to some double counting rather than to under-identification of individuals.

Name and sex

Like date of birth, letters of name and sex are essential to the SLK-581 key. Overall, at around 0.5%, missing name and sex information occurred slightly more often than 1 January 1900 and 1901 dates of birth (Table 2.3). Sex was slightly more likely to be missing than given name—both missing for between 0.24% and 0.28% of linkage key—while letters of family name were rarely missing (in only 0.01% of SLKs).

Table 2.3: Community services linkage keys using missing name, sex or date of birth data, HACC MDS

	July–September 2002		October–December 2002	
Missing information	**Number**	**Per cent**	**Number**	**Per cent**
Letters of given name	937	0.24	997	0.25
Letters of family name	22	0.01	38	0.01
Sex	1,012	0.26	1,125	0.28
Letters of name(s) or sex	*1,958*	*0.51*	*2,146*	*0.53*
Date of birth unknown or missing (1 January 1900 or 1 January 1901)	1,543	0.40	2,067	0.51
Letters of name(s), sex, or date of birth missing or unknown	**3,152**	**0.82**	**3,805**	**0.94**
Letters of name(s) or sex missing, or date of birth missing, unknown or 01/01/19N0	**6,623**	**1.71**	**7,673**	**1.89**
Letters of name(s) or sex missing, or date of birth missing, unknown or 01/01/19Nn	**13,555**	**3.51**	**15,009**	**3.70**
All SLKs	**386,299**	**100.0**	**405,404**	**100.0**

Note: A small number (under 10 per quarter) of SLKs contained birth dates earlier than 1 January 1900. These have been excluded from the analysis of missing, unknown and 01/01 birth date information.

Source: AIHW analysis of HACC MDS.

Poor quality data on name and sex is not generally coincident with poor date of birth information. Overall, 0.8% of SLK-581 keys for the September quarter and 0.9% of SLKs for the December quarter had either missing or unknown date of birth and/or

some missing name or sex information, with an additional 0.9% having 1st of the decade birth dates. If all 1 January dates are considered inaccurate, almost 4% of SLK-581 linkage keys in the quarterly HACC collections contain some poor quality information (that is, for name, date of birth or sex).

As with erroneous date of birth information, poor name and/or sex data are more likely to lead to double counting of clients than to under-identification. On average, SLK-581 keys which included missing name and/or sex had relatively few agencies contributing service data, averaging under 1.02, compared with an average of around 1.22 for keys based on known information (Table 2.2).

Agency non-participation

Incorrect link key data affects the identification of individuals, and the validity of particular links within and between data sets. The complete absence of data on particular HACC clients is another source of error. For the September and December 2002 quarters, 78% and 86% of HACC agencies, respectively, participated in the data collection (DoHA 2003b:4). Clients who only used non-participating agencies were therefore not represented in the MDS for those quarters. The proportion of HACC clients not included in the collection is not known, although it is expected to be lower than the agency non-participation rate because of the use of multiple agencies by individuals and because—according to advice from the Department of Health and Ageing—larger agencies (in terms of funding) are more likely to participate in the collection than smaller agencies.

Under-coverage of clients in the HACC MDS due to agency non-participation means that it is not possible to identify all movements between HACC and other aged care programs. Consequently, some CACP and RACS clients that have received HACC services will not be identified by linking the data sets because their use of HACC is not recorded in the HACC MDS. As a result, the amount of movement between HACC and other services will be underestimated.

Linking when the linkage key contains missing data

All three components of the linkage key are important when linking records. For example, Table 2.4 indicates that date of birth (within sex) on its own does not differentiate between individuals using HACC services. Also, the five letters of name used in the SLK-581 key by themselves are not very good identifiers of clients: for the July–September 2002 quarter, the 386,299 reported complete SLKs contained 124,732 distinct letters of name combinations (ignoring sex), and only 70,006 of these occurred only once. Consequently, just 18% of all linkage keys had unique letters of name. In addition 90 combinations occurred at least 100 times, with four combinations reported more than 300 times. Similar results were found for the following quarter, with again only 18% of all linkage keys having unique letters of name. Furthermore, even taking sex into account, the number of different combinations of letters of name and sex within the HACC MDS is under 40% of the number of unique SLK-581s.

Table 2.4: Different combinations of SLK-581 linkage key components combined with additional information, HACC MDS (as per cent of total SLKs)[(a)]

	Additional variables used[(b)]								
Key data available	None	Client state	Client post-code	Country of birth	Main language spoken	Indig-enous status	A|C|D|E tog.	C|D|E tog.	All
July–September 2002		**A**	**B**	**C**	**D**	**E**	**F**	**G**	**H**
None (number)	. .	9	2,620	202	99	6	4,882	2,242	45,290
None (per cent)	. .	—	0.7	0.1	—	—	1.3	0.6	12.2
Letters of family name	1.6	6.3	66.0	9.7	6.3	3.6	26.3	16.5	79.2
Letters of names and sex	39.5	64.6	96.1	54.1	46.8	47.6	78.8	61.3	97.9
Letters of names, year of birth and sex	88.0	95.2	98.3	92.2	90.1	90.5	97.5	93.7	99.1
Date of birth and sex	15.2	44.1	96.8	38.3	27.1	24.7	70.1	47.8	98.4
October–December 2002									
None (number)	. .	9	2,638	208	99	6	4,981	2,279	47,326
None (per cent)	. .	—	0.7	0.1	—	—	1.3	0.6	12.1
Letters of family name	1.6	6.1	65.4	9.5	6.1	3.5	26.1	16.3	78.9
Letters of names and sex	38.8	63.7	96.0	53.5	46.1	46.9	78.4	60.9	97.9
Letters of names, year of birth and sex	87.7	95.0	98.2	91.9	89.7	90.2	97.4	93.5	99.1
Date of birth and sex	14.6	42.8	96.7	37.7	26.5	24.1	69.5	47.4	98.4

(a) Table was derived using cases with all contributing SLK-581 information known. All 1 January birth dates were considered to be estimated and so were excluded. For July–September 2002 quarter there were 372,645 distinct SLKs based on reliable data; for October–December 2002 quarter there were 390,395 such SLKs.

(b) Cases with missing data for additional variables have been included. For the September quarter 2002, postcode was missing for 2.9% of cases with a complete SLK-581, country of birth for 7.1%, main language spoken in the home for 6.2% and Indigenous status for 11.2%. For the December quarter, postcode was missing for 2.8% of cases with a complete SLK-581, country of birth for 7.3%, main language spoken in the home for 6.4% and Indigenous status for 11.4%.

Source: AIHW analysis of HACC MDS.

There are two options for dealing with poor link key data:

- *Option 1 – Basic linking*: Exclude all 'clients' with SLK-581 linkage keys incorporating missing or poor information from the data set for linking.
- *Option 2 – Enhanced linking*: Retain all valid data in the linkage key, and where possible use this in conjunction with other information to establish links with other data sets. These additional variables must of course be available and comparable on both data sets.

The first of these two approaches is the easiest, requiring the fewest decisions and little effort. However, some movements between services will not be identified because some clients with missing data will not be included in the data set for linking. Dropping those cases with unreliable SLKs (including all those with 1 January dates of birth) would result in the exclusion of up to 4% of SLK-581 linkage

keys in the quarterly HACC MDS collections. While this is not desirable, its effect is likely to be less than that of agency non-participation. In addition, some of these clients may also have accurate linkage keys in the data sets, thus ameliorating the effect.

The second approach involves using available valid link data in conjunction with additional information to establish initial links. Further information could then be used to distinguish between links to non-unique keys. Variables which could be used in this process include state of client's residence, postcode of client's residence, country of birth, main language spoken at home and Indigenous status. However, in many cases, the last three variables are poor discriminators because of the concentration of a large majority of people in a single category. Also, differences between definitions of the variables in different collections will affect their utility (see Chapter 5). The additional discriminatory value of a range of variables can be broadly gauged by combining the different elements of the SLK-581 with other variables and examining the resulting number of different, or distinct, keys using these new combinations compared with the original number of SLKs (Table 2.4).

By themselves, the variables available to provide additional linkage data—state of residence, postcode of residence, country of birth, main language spoken at home and Indigenous status—do not provide sufficient information to discriminate between individuals. Even taken altogether these variables result in only 45,300 different combinations for the 372,645 distinct SLK-581 linkage keys based on reliable data in the July–September 2002 HACC collection—a 12% distinct key rate (see column H of Table 2.4). Adding letters of family name—which is rarely missing—to this set of variables increases the distinct key rate to around 79%. Either date of birth and sex or letters of name and sex are required in conjunction with these variables to bring the distinct key rate above 95%, suggesting that linkage should not be attempted with other data sets if either date of birth or all letters of both names (and sex) are not known.

Of the variables considered, postcode of residence provides the greatest additional discrimination—primarily by virtue of their relatively large number—and adding postcode to known linkage key data greatly increases the number of distinct keys. If either date of birth or name data are missing, adding postcode to the known SLK-581 data increases the distinct key rate to over 96% from just 15% if only name data are missing, from under 40% if only date of birth is missing, and from 88% if only day and month of date of birth are missing. Interestingly, despite the smaller number of categories, adding state of client residence has a greater impact on the number of distinct combinations than adding either country of birth or main language spoken at home.

In general, because most postcodes lie within a particular state, state of residence contains little additional information once postcode has been included, and Table 2.4 suggests that country of birth may help to distinguish between non-unique keys more effectively than main language spoken at home or Indigenous status. However, using the variables relating to cultural diversity could lead to some bias in the linkage process because links are more likely to happen by chance in larger

population groups, that is, in the Australian-born, English-speaking, non-Indigenous population.

The results in Table 2.4 also indicate that, even when combined and added to available SLK-581 data, the non-regional variables considered cannot distinguish effectively between clients. Even when only day and month of birth are dropped from the linkage key and replaced by country of birth, main language spoken at home and Indigenous status, the number of distinct linkage keys drops by 6% when compared with the number in the original set of complete SLK-581 linkage keys (see column G). If full date of birth is replaced by the three cultural diversity variables, the number of distinct linkage keys drops by nearly 40%; the corresponding fall if letters of name are replaced is just over 50%. The situation improves if state of residence can also be included in the adjusted linkage key (column F). However, even in this case, if either full date of birth or letters of name are missing, the number of distinct linkage keys is more than 20% below the number of unique keys using complete SLK-581 data.

The above results suggest that linkage using SLKs incorporating missing data should not be considered if comparable postcode data are not available in the two data sets being linked. In addition, some incomplete linkage keys may contain insufficient valid data so that even when combined with postcode, the resulting key is not sufficiently accurate to undertake linkage. As an extreme example, if only letters of family name are available adding postcode still leads to only to two-thirds the number of different keys when compared with the full SLK-581 linkage key (column B of Table 2.4).

The question then arises as to which parts of the key need to be available before considering undertaking linkage using these components in conjunction with postcode. To address this issue, the effect of adding postcode to a wide range of reduced keys is examined in Table 2.5. From this it can be seen that, in data sets of around 400,000, using an SLK-581 type linkage key that excludes sex results in very little loss of information. Consequently, combining postcode with the reduced key increases the accuracy of the key only marginally; a linkage key consisting of the SLK-581 letters of name, date of birth and postcode would result in 0.5% of complete keys having non-unique adjusted linkage keys. Overall, if full date of birth is known and at least the letters of either the family or given name are known, fewer than 3.5% of complete keys have non-unique adjusted linkage keys (based on known SLK-581 data along with postcode). If either full date of birth or full letters of name information is known, and some other data apart from sex is known, then under 5% of complete keys have non-unique adjusted linkage keys. Note, however, that if only decade of birth is known then sex must also be known along with the letters of name to keep the per cent of non-unique keys down to around 5%.

Table 2.5: Different combinations of SLKs excluding selected elements, with and without postcode, HACC MDS

Reduced key[b]	July–September 2002				October–December 2002			
	Prevalence of poor data	Different combinations[a]		Non-unique keys with postcode	Prevalence of poor data	Different combinations[a]		Non-unique keys with postcode
		Key only	With postcode			Key only	With postcode	
	Number	Per cent of total SLKs[a]			Number	Per cent of total SLKs[a]		
Valid date of birth with some name data								
S3G2|dob|__	609	99.7	99.8	0.5	688	99.7	99.7	0.5
__G2|dob|sex	7	84.4	99.5	1.1	10	83.8	99.4	1.1
__G2|dob|__	—	79.8	99.2	1.6	—	93.3	99.2	1.7
S3__|dob|sex	559	97.3	98.8	2.4	576	97.2	98.8	2.4
S3__|dob|__	5	95.8	98.4	3.2	2	95.7	98.4	3.3
Valid name data with poor date of birth, or valid date of birth and sex only								
S3G2|yob|sex	6,928	88.0	98.2	3.5	7,334	87.7	98.2	3.6
S3G2|yob|__	42	85.1	97.9	4.1	39	84.6	97.9	4.2
S3G2|decade|sex	3,470	63.1	97.4	5.3	3,866	62.4	97.3	5.4
____|dob|sex	—	15.2	97.2	5.5	—	14.6	97.1	5.7
S3G2|decade|__	51	57.1	96.9	6.2	56	56.3	96.8	6.3
S3G2|__|sex	1,203	39.5	96.3	7.4	1,666	38.8	96.2	7.5
Incomplete name and date of birth data, or complete name or complete date of birth data only								
S3__|yob|sex	132	35.4	96.2	7.6	134	34.6	96.1	7.8
____|dob|__	—	8.9	96.1	7.7	—	8.5	96.0	8.0
S3G2|__|__	297	32.7	95.6	8.8	328	32.1	95.5	9.1
__G2|yob|sex	5	6.6	90.2	19.7	6	6.4	89.8	20.3
Other missing	251	. .	. .	. .	307	. .	. .	. .
Total	**13,559**	**372,740**	**362,144**	**. .**	**15,012**	**390,392**	**379,573**	**. .**

(a) Based on cases with complete SLK-581 data, and valid postcode when combining postcode with the key.

(b) S3 = three letters of family name as in the SLK-581; G2 = two letters of given name as in the SLK-581; dob = date of birth; yob = year of birth; decade = decade of birth.

Note: Dates of birth prior to 1892 have been assumed to be erroneous, and have been set to missing (4 and 3 cases in the September and December quarters, respectively).

Source: AIHW analysis of HACC MDS.

When linking between data sets, links between non-unique keys necessitates deciding which link to choose. This problem is exacerbated if there are non-unique keys in both data sets. When using adjusted keys, postcode data has already been used to establish the links, therefore other information must be used to distinguish between links. While in some cases using other demographic data—such as the cultural diversity information discussed above—might allow appropriate links to be identified, in many cases they will not because of high concentrations in one or two categories. Therefore, random choice may be required when deciding which link to use. Furthermore, if several adjusted linkage keys are derived to allow for differing missing components, the adjusted keys must be derived for all clients with sufficient

information in the other data set—not just for those with similarly missing information—as there is no reason to think that only clients with a particular SLK component missing in one data set will have the same component missing in the other data set. Consequently, using adjusted linkage keys will result in multiple comparisons between records in the two data sets. This, in conjunction with the lesser accuracy of the adjusted linkage keys, will lead to a greater chance of inaccurate links.

Taken altogether, these findings suggest the following strategy if enhanced linking is to be used:

- Client postcode should be used in conjunction with the valid SLK data to identify links for cases when some SLK-581 information is missing.
- At least complete date of birth data and some name data need to be available before linking should be attempted using postcode-adjusted linkage keys. If only sex is missing, the reduced key (that is, the key based on valid data only) can be used on its own without the addition of postcode. Within quarterly HACC data, this approach will result in 3% or fewer clients having non-unique adjusted keys.
- Random selection can be used to choose between links involving non-unique keys.
- If client postcode data are not available, enhanced linking should not be attempted.

Using the above strategy, some 1,200 HACC records per quarter with incomplete SLK-581 data would be considered for postcode-enhanced linking (Table 2.5). However, since missing date of birth information is the main cause of incomplete linkage keys, the majority of records with missing linkage key data (around 90% of poor quality keys per quarter) would not be included.

The above discussion also suggests a method for identifying clients who appear more than once on the HACC MDS due to missing information in one of their reported SLKs. By comparing the postcode-enhanced reduced keys for records with missing linkage key data with those for records with a complete linkage key, likely multiple representation of clients on the MDS can be identified. Using this approach, for the two quarters being examined around 350 (or over one-quarter) reduced keys considered to contain sufficient data for linking are highly likely to have been for clients who also had a complete linkage key (Table 2.6). Among records with reduced keys with insufficient data for linking but with either complete name or date of birth information, it is estimated that up to 9% are likely to have been for clients who also had a complete SLK-581. Note that these figures slightly overestimate the number of clients with more than one linkage key because of coincidence among the various keys, especially among those with insufficient data for efficient matching.

Looking at particular reduced keys, over one-quarter of those with either only missing sex information or missing given name were associated with a record with a complete SLK-581. Also, perhaps as many as one-fifth of keys with only missing date of birth information were for clients who also had a complete key. Keys with poor date of birth data, as opposed to missing information, were less likely to be

associated with a record with a complete SLK-581 key than other reduced keys. For example, only 6% of linkage keys with missing day and month of birth were likely to have been associated with a complete SLK-581. This suggests that it may be difficult to obtain a complete date of birth for many of these clients.

Table 2.6: Poor quality keys possibly associated with complete SLK-581 linkage keys in the HACC MDS, using postcode as discriminating variable

	July–September 2002				October–December 2002			
Valid SLK-581 data[a]	**Prevalence**	**With valid postcode**	**Associated with complete keys[b]**		**Prevalence**	**With valid postcode**	**Associated with complete keys[b]**	
Sufficient for linking		**Number**		**%[c]**		**Number**		**%[c]**
S3G2|dob|__[d]	609	560	188	30.9	688	651	209	30.4
__G2|dob|sex	7	7	1	14.3	10	9	1	10.0
__G2|dob|__	—	—	—	—	—	—	—	—
S3__|dob|sex	559	536	167	29.9	576	551	154	26.7
S3__|dob|__	5	1	0	0.0	2	1	0	0.0
Total	*1,180*	*1,104*	*356*	*30.2*	*1,276*	*1,212*	*364*	*28.5*
Sufficient only for identifying likely associated keys								
S3G2|yob|sex	6,928	6,667	397	5.7	7,334	7,023	426	5.8
S3G2|yob|__	42	41	0	0.0	39	38	1	2.6
S3G2|decade|sex	3,470	3,280	269	7.8	3,866	3,640	331	8.6
____|dob|sex	—	—	—	—	—	—	—	—
S3G2|decade|__	51	48	3	5.9	56	53	2	3.6
S3G2|__|sex	1,203	1,176	287	23.9	1,666	1,635	377	22.6
Total	*11,694*	*11,212*	*956*	*8.2*	*12,961*	*12,389*	*1,137*	*8.8*
Insufficient information								
Other	*685*	*676*			*775*	*750*		
Total	**13,559**	**12,992**	**1,312**	**9.7**	**15,012**	**14351**	**1,501**	**10.0**

(a) S3 = three letters of family name as in the SLK-581; G2 = two letters of given name as in the SLK-581; dob = date of birth.

(b) Adjusted linkage key is the same for a linkage key based on full data and one based on incomplete data.

(c) As a per cent of cases both with and without valid postcode data.

(d) Adjusted key does not include postcode as key is sufficiently accurate without it (see Table 2.5).

Source: AIHW analysis of HACC MDS.

The above analysis indicates that using enhanced rather than basic SLK-581 linking will increase the number of HACC records available for linking by between 800 and 900 records per quarter, or by just over 0.2% (excluding those associated with complete SLKs). Furthermore, 3.3% of records—relating to about 11,500 clients in the September 2002 quarter and 12,600 clients in the December 2002 quarter with poor quality linkage key data—would still be excluded from the linking.

2.3 Summary

Incomplete data affects between 3% and 4% of SLK-581 linkage keys in the quarterly HACC MDS. For some studies, use of other variables to enhance the linkage may overcome some of this problem. In particular, postcode of the client's usual residence has been identified as providing high discriminatory power when some of the linkage key information is missing or incomplete. However, even when postcode data are available, valid date of birth information is required before considering matching for records with some poor linkage key data. Information relating to cultural diversity is generally not sufficient to discriminate between clients, even when much of the linkage key information is valid.

Using the data available, it has not been possible to investigate the extent of non-unique keys in the HACC data, nor has it been possible to look at the effect of variations in name. However, these problems are expected to be greater for HACC than for RACS and CACP data sets, both because of the greater number of clients involved and because of the less co-ordinated nature of the care provided by HACC (and hence the greater opportunity for variation in the reported linkage key components).

Although these issues affect any linkage undertaken with the HACC MDS, the greatest cause of missed links is likely to be the non-participation of some agencies in the HACC data collection which results in some clients being absent altogether from the data sets.

3 SLK-581 within ACCMIS

As stated before, the ACCMIS database contains the Department of Health and Ageing's administrative data on residential aged care and Community Aged Care Packages. On ACCMIS, individual clients are identified via name and other demographic data, and given a distinct client number (Client ID). All admissions into a residential care service or commencements on a package are recorded. Client numbers are assigned separately for residential aged care and Community Aged Care Packages, so that clients that have used both CACPs and residential aged care have two client numbers on ACCMIS.

The data extracted from the ACCMIS database for the current study included all admissions over the 3-year period beginning on 1 July 2000 relating to permanent and respite residential care and to Community Aged Care Packages. Assuming that there is a one-to-one relationship between an ACCMIS Client ID and a client, 324,444 residential aged care admissions relating to 191,413 clients were extracted from the database for analysis. For CACPs, data was extracted for 44,602 admissions involving 42,298 clients.

For residential aged care and Community Aged Care Packages, the SLK-581 for a client can be derived from the personal identifiers held on ACCMIS. In this chapter, the effectiveness of the SLK-581 in identifying individual clients within these two programs is examined.

3.1 Data quality

When a non-unique SLK-581 is generated in the ACCMIS data set it could be for a number of reasons.

1. Due to the composition of the linkage key, two or more people with the same name components and the same date of birth and sex might independently generate the same linkage key. For example, Barbara Butler and Maria Vuttesque, both born on the 2 January 1923 both have the linkage key 'UTEAR020119232'; that is, two different people have coincident linkage keys.
2. A single person has been allocated more than one Client ID within the data set so that these records independently generate the same linkage key; that is, the records with the linkage keys are replicates relating to the same person.
3. Unclean or missing data (for example, misspelt names, pseudonyms and use of default values for unknown/missing date of birth) affects a component of the linkage key. In this case, the non-unique linkage keys may or may not refer to the same person.

Table 3.1 demonstrates some of these issues.

Table 3.1: Problems associated with construction of unique SLK-581 linkage keys

Client ID	Given name	Surname	Date of birth	Sex	SLK-581	Issue
SPARC00001	BENJAMIN	GREGORY	09/12/1930	M	REOEN091219301	Client has more than one client ID
SPARC00002	BENJAMIN	GREGORY	09/12/1930	M	REOEN091219301	
SPARC00003	BARBARA	BUTLER	15/08/1923	F	UTEAR150819232	Two clients have same sex, components of name and date of birth
SPARC00004	MARIA	VUTTESQUE	15/08/1923	F	UTEAR150819232	
SPARC00005	JIMMY	BLACK	01/01/1920	M	LAKIM010119201	Default date of birth used
SPARC00006	MALVERN	GREY	01/01/1920	M	RE2AL010119201	Default date of birth used
SPARC00007	JOHN	SMITH	20/05/1922	M	MIHOH200519221	Two clients have a same name and date of birth
SPARC00008	JOHN	SMITH	20/05/1922	M	MIHOH200519221	
SPARC00009	LAVINIA	WALTERS	12/02/1916	F	ALEAV120219162	Pseudonyms
SPARC00010	WINNY	WALTERS	12/02/1916	F	ALEIN120219162	
SPARC00011	ZU	LU	06/06/1937	M	U22U2060619371	Short name
SPARC00012	XXXX	XXXXX	22/11/1907	M	XXXXX221119071	Missing name

Note: Table uses fictitious clients.

For the 3 years under consideration, the construction of the SLK-581 linkage key using ACCMIS data yielded 190,921 client records with unique linkage keys for RACS and 42,264 client records with unique linkage keys for CACP, so that overall there were 492 Client IDs in the RACS data and 34 in the CACP data with non-unique linkage keys (Table 3.2). Consequently, 99.73% of aged care residents (as identified by the ACCMIS Client ID) had unique keys and 99.92% of CACP recipients had unique keys.

Table 3.2: Unique SLK-581 linkage keys for RACS and CACP clients, 1 July 2000 – 30 June 2003

	RACS data		CACP data	
	Number	Per cent	Number	Per cent
Unique SLK-581	190,921	99.74	42,264	99.92
Non-unique SLK-581	492	0.26	34	0.08
Unique Client IDs	**191,413**	**100.0**	**42,298**	**100.0**

Source: AIHW analysis of ACCMIS database.

These results indicate that SLK-581 is very good at identifying individual clients in residential aged care and Community Aged Care Packages. The percentage of data lost would be small if records for clients with non-unique linkage keys were either eliminated or combined; when looking at 3 years of admissions, data for less than 0.3% of clients for RACS and 0.08% for CACP would be affected in any linkage analysis. For shorter periods, even fewer clients would have their admissions either dropped or incorrectly combined. Removing records with non-unique keys from the database altogether leads to a (very slight) undercount of clients and introduces the possibility of missing a link between data sets. On the other hand, using the SLK-581 linkage key as the sole way of identifying clients would lead to data from a small

number of clients being inappropriately amalgamated. This would also lead to some—albeit very small—bias.

As outlined above, non-unique linkage keys can result from an individual having two Client IDs. If, however, for some reason a person has more than one Client ID with different personal information recorded against their various Client IDs then the client could have several distinct SLK-581 keys. In this case, the multiple occurrence of a client on the database would go undetected using SLK-581.

As in the HACC MDS, the quality of the data contributing to the SLK-581 affects its efficiency in distinguishing between clients. If coincident and replicated keys can be identified through appropriate data cleaning processes, it may be possible to keep valid (coincident key) records in the linkage data set while amalgamating across replicated keys for the same client.

Date of birth

As with the HACC data, frequency of dates of birth will influence the uniqueness of the resulting linkage key. In particular, any default values being used for date of birth can skew the data to those dates, reducing the effectiveness of the linkage key, both due to the occurrence of coincident SLK-581 linkage keys for different people and due to inaccuracies in the linkage key itself.

In the ACCMIS extracts used for this study, there were no missing values for date of birth in the RACS or CACP data sets. There was, however, evidence that 1 January dates are sometimes recorded when exact date of birth is not known (Table 3.3). In particular, 1 January 1920 and 1 January 1930 were reported over four times more frequently than other birth dates in their respective decades for both RACS and CACP clients. However, the prevalence of such 1 January dates was much less than that observed in the HACC MDS, where, for example, 1 January 1920 dates were 40 times more common than other dates in the 1920s. The most commonly reported birth date was 1 January 1920 (occurring 65 times), and, overall, 1 January dates of birth were recorded for 0.5% of RACS Client IDs, with fewer than one-eighth of these being for the first of a decade. This compares with 3% of linkage keys in the HACC MDS involving 1 January birth dates, with around half of these relating either to 1 January 1900 or 1901 or to first of decade birth dates. Four other 1 January birth dates featured in the 40 most common RACS birth dates, but none of these related to the first of a decade.

For CACP there is a slightly larger effect, with 1 January birth dates reported for 0.9% of Client IDs, and one in six of these relating to first of decade birth dates. In addition, 1 January birth dates for 1925, 1930 and 1920 were the three most common birth dates recorded for CACP recipients.

Apart from the problem of individuals receiving different SLKs in different data sets due to inconsistencies in the linkage key data, the main concern about the prevalence of default birth date values is that they could be the source of non-unique SLK-581 linkage keys. However, in the data examined, these 1 January dates of birth had no effect on the occurrence of non-unique keys for ACCMIS data because among the 34

CACP records and 492 RACS records with non-unique SLKs, none had a 1 January date of birth. Summarising, while there is some evidence that 1 January is being used when date of birth is not known precisely, the prevalence of such dates is not large and suggests that date of birth is generally well recorded in the ACCMIS database.

Table 3.3: Frequency counts for dates of birth for RACS and CACP clients, 1 July 2000 – 30 June 2003

	RACS			CACP		
	1 January			1 January		
Year	Number of Client IDs	Per cent of all Client IDs	Mean number per date in the decade[(a)]	Number of Client IDs	Per cent of all Client IDs	Mean number per date in the decade[(a)]
Turn of century dates						
1900	3	0.002	6.38	2	0.005	1.71
1901	6	0.003	6.38	1	0.002	1.71
Total	*9*	*0.005*	. .	*3*	*0.007*	. .
Start of decade dates						
1910	24	0.013	24.76	8	0.019	4.99
1920	65	0.034	16.48	18	0.043	4.33
1930	22	0.011	4.06	21	0.050	1.85
1940	8	0.004	1.73	8	0.019	1.26
1950	3	0.002	1.24	4	0.009	1.09
1960	—	—	1.06	1	0.002	1.00
1970	—	—	1.03	—	—	1.00
1980	—	—	1.00	—	—	1.00
1990	—	—	1.00	—	—	0.00
Total	*122*	*0.064*	. .	*60*	*0.142*	. .
Other 1 January dates	**866**	**0.452**	. .	**308**	**0.728**	. .
All 1 January dates	**997**	**0.521**	. .	**371**	**0.877**	. .
All Client IDs	**191,413**	**100.00**	**10.44**	**42,298**	**100.00**	**3.34**

(a) Average based on birthdays which occur in the data set.

Notes

1. For CACP Client IDs there were 20 birth dates before 1900, including 1 before 1890 (in the 1850s).
2. For RACS Client IDs there were 241 birth dates before 1901, including 1 before 1890 (in the 1880s).
3. There were 190,921 client records with unique SLK-581 linkage keys for RACS and 42,264 for CACP.

Source: AIHW analysis of ACCMIS database.

Sex

There were no records with missing sex in the RACS records on ACCMIS. Of the 51 CACP client records in ACCMIS with missing sex, the only record with any associated admissions related to a 1993 admission and a 1994 discharge. Consequently, missing sex is generally not an issue for RACS or CACP data.

Letters of name

The SLK-581 for ACCMIS data can be affected by the presence of invalid name information and by the use of pseudonyms when recording client data on the database. These two issues are discussed below.

Missing name data

On ACCMIS, missing—or crossed out—names appear to be indicated using X's. Names were crossed out in a very small number of cases over the period of interest: for only one RACS Client ID and three CACP Client IDs. In addition, one CACP client had a missing given name indicated by '.'. For these cases the letters of the missing names used in the SLK-581 should be changed to '9' before linking is considered.

Invalid name data

An anomaly was identified when examining records with non-unique SLKs; it appeared that some clients have non-name information attached to the value in the family name field. Looking at all the cases with non-unique SLKs, and using truncated and compressed ACCMIS name data, non-name strings discovered included USETHEOTHER, USEID, DONOTUSE, USE, USEOTHER, DONTUSE, PLSUSE, USEOTHERONE, SAMEAS, USERESI, NOTTHISONE, USEBUILDING, SEE, and NONONPO.

Returning to the original ACCMIS data, further investigation revealed that these strings were actually the truncated textual components of more descriptive instructions that had been included in the family name field at some point of the data entry or cleaning process. Moreover, in some cases the non-name string provided information that could be used to identify a number of Client IDs that were for the same individual. In others, including some strings that appeared to provide pseudonyms, the tags were more random and less identifiable. These records contained values such as HENRY (BILL) or LAVINIA_(WIN) in the name field. This additional information is always going to be unpredictable and difficult to screen systematically and yet could influence the construction of the linkage key.

Table 3.4: Estimated prevalence of ACCMIS family name fields with additional information, 1 January 2000 – 31 December 2002

	Sample 1		Sample 2	
	Number	Per cent	Number	Per cent
Instruction tag present (e.g., DO NOT USE)	4	0.04	5	0.05
Identified as pseudonym	5	0.05	6	0.06
Total with additional information present	*9*	*0.09*	*11*	*0.11*
Total records in sample	**10,000**	**100.0**	**10,000**	**100.0**

Source: AIHW analysis of ACCMIS database.

An attempt was made to quantify how often such additional information is included in the name data on ACCMIS. Two separate random samples of 10,000 client records with an admission in the 3-year period from 1 January 2000 were extracted from the original ACCMIS data set. The two samples were found to have a very low number of records with additional information in the family name field—both around 0.1% (Table 3.4).

Special characters in name fields

In earlier ACCMIS records, additional instructions in name fields for data entry personnel were often associated with special characters such as an asterisk or an underscore, or referred to a specific Client ID number as the replicate. The utility of finding numbers and special characters in name fields as a means of identifying multiple Client IDs for individuals was therefore investigated.

While instructions referring explicitly to other Client ID numbers were identified in earlier entries, no records for people whose first admission was between 1 July 2000 and 30 June 2003 contained numbers; that is, none referred to specific Client IDs. In the RACS data, the family name field for 2,497 Client IDs contained non-alphabetic non-numeric characters—predominately hyphens and apostrophes. Of these, 18 were associated with additional instructions: 17 were variations of 'DO NOT USE' and one included the phrase 'USE BUILDING'. For the given name field, 229 Client IDs contained special characters, but again these generally resulted from hyphenated names. For 25 records the special characters were associated with inclusion of a nickname; however, none of these contained instructions, and none involved given names with fewer than three characters so that the formation of the SLK-581 linkage key was not affected.

For Community Aged Care Package recipients, the family name data for 540 Client IDs included special characters. As for RACS clients, these cases predominately related to valid cases where names included hyphens or apostrophes (531, or 98%). Among given names, there were 138 which included special characters; almost one-third of these (41) were valid, involving hyphenated names. In nearly all other cases the special characters were associated with either nicknames or titles; in one case a missing name was indicated by a '.'. Again, none of the given names with nicknames were less than three letters long, so that the formation of the SLK-581 linkage key was unaffected by their presence.

These results indicate that identifying non-alphabetic characters in name fields does not assist greatly in finding those clients with multiple Client IDs on ACCMIS.

Instruction tags

From looking at the range of instruction tags in the samples, instruction tags commonly indicate that a particular client has more than one Client ID on the database. Consequently, the occurrence of specific character strings which had been identified in the samples as components of instructions was investigated.

Using automated searching for aged care residents admitted between 1 July 2000 and 30 June 2003, no names (given, family, or middle names) were identified which

contained instructions with the strings 'SEE', 'THIS', 'ONE', 'RESI', 'LATEST'—these had been used in earlier records. However, records for 105 Client IDs (or 0.05%) contained variations of 'DO NOT USE' (103 of the tags were in the family name, and one each was in a given and middle name). With the exception of one record where the given name was 'DO NOT USE', none of these instructions were at the start of the name field (although this did occur in earlier records). Among Client IDs for CACP recipients, instruction tags were found in only two names—both being 'DO NOT USE' appended to the family name.

In terms of obtaining an accurate data set for linking, it would be desirable if the instruction tags could be used to eliminate multiple Client IDs from the data set prior to linking. Since all the instructions tags identified to date advise that the Client ID containing the tag should not be used, flagging those Client IDs with instruction tags could assist in identifying clients that appear more than once in the data. However, the tags by themselves do not assist in finding the partner Client ID implied by the instruction.

It should be noted that the tags do not seem to be drawn from a specified list, and the people entering data on the database can generate new ones every day. Therefore, automated identification of tagged Client IDs may not be 100% effective. It is also possible that the additional information in the name fields could interfere in the appropriate construction of the SLK-581, particularly for short names. Consequently, the SLK-581 linkage key may also not always identify related Client IDs. For these reasons more than one approach is needed to remove multiple Client IDs for individuals. The identification of individuals with more than one Client ID is discussed further in the next section.

3.2 Differentiating between coincident and replicated SLK-581 keys

As discussed above, non-unique SLKs may be either for the same person, as in the case where one client has more than one Client ID, or for different people, as in the case where two clients have the same date of birth, sex and SLK name components. Different approaches should be taken when dealing with these two types of duplicates.

If common linkage keys are constructed for the same person with two Client IDs then the preferred treatment would be to collapse all data relating to both Client IDs against the one SLK-581. This treatment would then necessitate making a decision about which Client ID's demographic information should be used; for example, the data relating to the most recently used Client ID could be used. If identical linkage keys are constructed for different people, the preferred treatment would be to retain all occurrences of coincident keys in the pre-linkage data set. If, during linkage, a matching SLK-581 was found in the second data set, additional identifying information present on the two linked data sets could then be used to make a decision about which is the matching record, or the matching record could be chosen randomly.

From the above it can be seen that before non-unique keys can be treated appropriately, we must first decide whether records with a particular SLK-581 relate to the same or different clients. There are two possible strategies for identifying duplicate type: either manual inspection can be used, or decision making can be automated using programmed deterministic rules. Manual inspection is very time consuming, and therefore costly, so the second approach is preferred if appropriate rules can be developed.

Since, by definition, non-unique keys cannot be distinguished using the SLK-581, variables other than elements used in the SLK-581 key must be used to determine whether the keys are for the same or different clients. One obvious candidate for distinguishing between clients with common SLKs is the remainder of the name information. A manual scan of the records for the RACS and CACP programs suggested that name is significantly different for two records with the same SLK-581 for around 30% of records with a non-unique SLK-581 (Table 3.5). This would suggest that in about 70% of cases, the repeated SLK-581 is the result of a single client having more than one Client ID, as is the case for SPARC00001 and SPARC00002 in the examples given in Table 3.1.

Table 3.5: Manual assessment of duplicate SLK-581 linkage keys as 'same' or 'different' clients, duplicate SLK-581 keys for 1 July 2000 – 30 June 2003

	RACS		CACP	
Manual assessment	**Number**	**Per cent**	**Number**	**Per cent**
Records with non-unique SLK referring to a **different** client (coincident keys)	164	33	12	35
Records with non-unique SLK referring to a **same** client (replicated keys)	328	67	22	65
Total records with non-unique SLK-581	492	100	34	100

Notes

1. The assessment of 'same' or 'different' client has been made with reference to a number of variables in the data. Duplicates are assumed to have been generated for two records relating to the **same** person unless certain conditions are met:
   (a) Name is significantly different (TURNER and BURKE are assessed to be different but ROSEANNAH and ROSANNA are not).
   (b) Periods of care overlap.
   (c) Periods of care indicate admission after death; for example, two periods of care 13/12/02–7/03/03 and 9/03/03–17/03/03 for Client IDs with the same SLK-581 are assumed to relate different people if the first episode ended in death.
2. An assessment of same name is made with an appropriate degree of flexibility to include spelling errors and some abbreviation. For example, Ken Williams and Kenneth Williams are assumed to be the same name and Anya Mikhaela and Anya Mikhala are assumed to be the same name. Additional information including state of client, country of birth, marital status, admission and discharge dates as well as reason for discharge were used to validate a 'same' or 'different' decision. For example, it was assumed that a client could not be accessing permanent and respite care in different states at the same time, or discharge for the reason of 'death' more than once.

Source: AIHW analysis of ACCMIS database.

In most cases where several Client IDs have common SLK-581 keys, a manual assessment of 'different' was made when the two names associated with the shared SLK-581 had the same letter of name components used in the construction of the SLK-581 (2nd, 3rd and 5th letter of family name and 2nd and 3rd letter of given name) but were otherwise quite different. In these cases, common keys were

generally the result of the high prevalence of certain letters in these positions in name.

The prevalence of certain letters in various positions in name has previously been explored in relation to the Supported Accommodation and Assistance Program (SAAP) linkage key which uses different letters of name components to the SLK-581 (AIHW: Karmel 2000 (unpublished)). Analysis of the prevalence of certain letters in critical positions that make up the SAAP linkage key was undertaken with reference to names in the National Death Index. This work demonstrated that the letters A, E and O together accounted for 57% and 52% of second letters in given and family names on the National Death Index, respectively (reproduced in Table A.1). The uniqueness of characters in the SLK-581 is necessarily influenced by this clustering of second letters around three vowels.

Information other than name could also be used to identify multiple Client IDs for particular clients. Such information could include client postcode (or state), country of birth, preferred language and Indigenous status. Of these, postcode has the greatest spread across categories (see discussion in Section 2.2), and so could be useful in many situations. However, it will not be useful when people have changed their location between admissions, when postcodes are reported differently on different occasions, or if postcode data are missing. Over the time period under study, postcode was missing for 0.14% of RACS Client IDs (273 cases), and for 1.2% of CACP Client IDs (526 cases).

Using client postcode

Among the 492 RACS Client IDs with non-unique SLK-581 linkage keys, 186 (38%) also had the same postcode as their double(s). All of these had been identified in the manual assessment as relating to people with more than one Client ID. For CACP Client IDs, 18 out 34 (53%) with non-unique SLKs also had the same postcode as their double(s), and again all of these were identified manually as relating to the same client.

From this it can be seen that postcode can be used to identify a proportion of the non-unique SLK-581 link keys as relating to multiply-recorded clients; that is, they were replicated SLKs. However, comparisons with the manual assessment indicate that some of the remaining non-unique keys also relate to people with more than one Client ID. This is particularly true among RACS clients, where using postcode did not identify 43% of non-unique SLKs manually assessed as relating to the 'same' people, compared with 18% of CACP recipients. The difference between the results for RACS and CACP clients most likely arises from the way that postcode is recorded—as the place where the Aged Care Assessment Team can contact the client (before 2003)—and the fact that Community Aged Care Packages are delivered to people still living in their homes (see Section 5.3 for further discussion of client postcode on ACCMIS).

Using other name data

A secondary linkage key was tested for its ability to identify individual clients with common SLK-581 keys within ACCMIS. This key, termed the C3C2-SLK, is constructed in a similar way to SLK-581 but uses different letters, selecting the first three consonants of the family name and first two consonants of the given name. That is, excluding A, E, I, O, U and Y from the names, C3C2-SLK is the concatenation of the first three consonants of family name, the first two consonants of given name, date of birth and gender coded as '1' for male and '2' for female. A different form of construction was chosen, rather than absolute place of letters in names, to try to get around some misspelling of names. C3C2-SLK resulted in slightly more non-unique linkage keys than the SLK-581 for residential aged care clients (517 Client IDs), and slightly fewer for CACP clients (30 Client IDs) (Table 3.6).

Table 3.6: Uniqueness of C3C2-SLK for RACS and CACP clients, 1 July 2000 – 30 June 2003

	RACS data		CACP data	
	No. of records	Per cent	No. of records	Per cent
Unique C3C2-SLK	190,896	99.73	42,268	99.93
Non-unique C3C2-SLK	517	0.27	30	0.07
Unique Client IDs	**191,413**	**100**	**42,298**	**100**

Source: AIHW analysis of ACCMIS database.

As expected due to the differences in their construction, not all non-unique SLK-581 keys relate to Client IDs with non-unique C3C2 keys: 167 out of 492 (34%) Client IDs in RACS with non-unique SLK-581 keys had unique C3C2 keys; 18 out of 34 (53%) Client IDs in CACP were in a similar position (Table 3.7). On the other hand, of the 517 Client IDs with non-unique C3C2 keys in the RACS data, 192 (37%) had unique SLK-581 linkage keys. The corresponding proportion for CACP Client IDs was 47% (14 out of 30).

Table 3.7: Comparing uniqueness of SLK-581 and C3C2-SLK for RACS and CACP clients, 1 July 2000 – 30 June 2003 (number)

		C3C2-SLK		
		Unique	Non-unique	Total
RACS				
SLK-581	Unique	190,729	192	190,921
	Non-unique	167	325	492
	Total	**190,896**	**517**	**191,413**
CACP				
SLK-581	Unique	42,250	14	42,264
	Non-unique	18	16	34
	Total	**42,268**	**30**	**42,298**

Source: AIHW analysis of ACCMIS database.

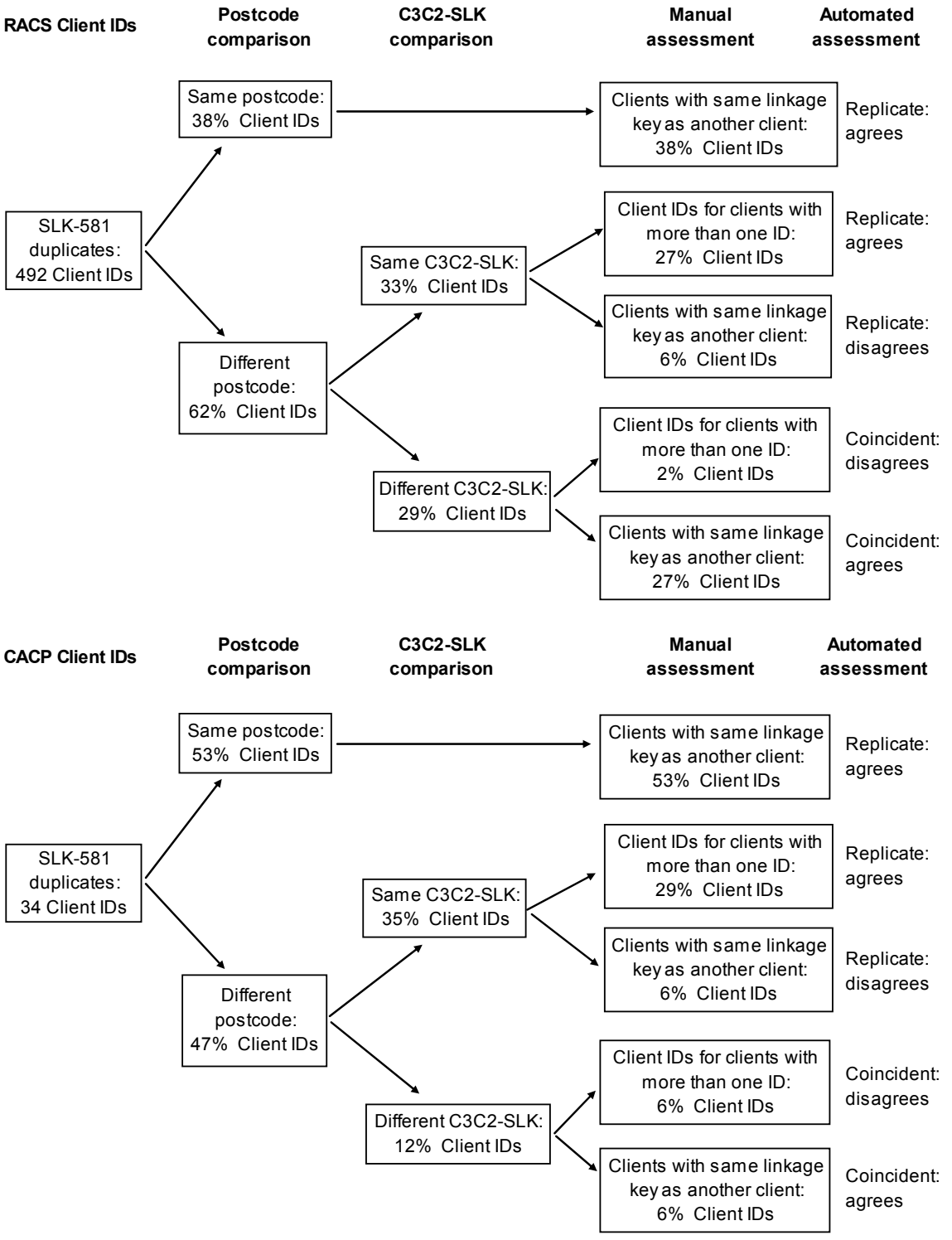
From the above, it can be seen that, in the RACS data, 66% (325/492) of Client IDs generating a non-unique SLK-581 also generated a non-unique C3C2-SLK. The corresponding figure for the smaller CACP data set was 47% (16/34). Whether a unique C3C2 key but non-unique SLK-581 can be used to identify different clients with the same SLK-581, and whether a non-unique key in both keys can be used to identify clients with multiple Client IDs, can be examined by comparing the dual link key results with the original manual assessment. The procedure could be further refined by using the finding that repeated SLK-581 keys which also have the same postcode were always manually assessed as belonging to the same person.

Using these results, a possible automated assessment algorithm is then as follows:

- Identical SLK-581 keys for different Client IDs with identical client postcodes are classified as replicates; that is, they relate to the same person;
- Identical SLK-581 keys for different Client IDs with different client postcodes but with identical C3C2-SLKs are classified as replicates; that is, they relate to the same person;
- Identical SLK-581 keys for different Client IDs with different postcodes *and* different C3C2-SLKs are classified as coincident SLKs; that is, they relate to different people.

For the RACS records with repeated SLK-581 linkage keys, there was 92% agreement (451 out of 492) between the manual assessment and the above combined automated SLK-581/postcode/C3C2-SLK assessment when distinguishing between cases where one client has more than one Client ID and those where different people have the same SLK-581 (Figure 3.1 and Table 3.8). The level of correct allocation was higher for replicates than coincident SLK-581 keys—with 97% and 81% correct allocation, respectively. Consequently, coincident keys are more likely to be identified as replicates than the other way around, leading to a slight undercount in the total number of clients. In addition, 38 of the 349 records identified as replicates had instruction tags (equivalent to 'DO NOT USE') appended to the family name. Similar results were found for the much smaller number of non-unique keys seen in the CACP data where there was 88% agreement between the manual and automated assessment methods. One of the 22 identified replicates for CACP Client IDs had an instruction tag.

As Figure 3.1 shows, postcode is more effective in distinguishing between replicates and coincident linkage keys for CACP clients than for RACS clients. This reflects that, prior to January 2003, the address recorded on ACCMIS for a client was that used by Aged Care Assessment Teams to contact the client about an assessment. As Community Aged Care Packages are for people still living in the community, the contact address for a CACP client is highly likely to be their home address; on the other hand, many RACS clients are assessed in hospital, so a home address is less likely (see Chapter 5 for further discussion).



Source: Table 3.8.

Figure 3.1: Comparing manual assessment of Client IDs with non-unique SLK-581 linkage keys with combined SLK-581, postcode and C3C2-SLK assessment

Table 3.8: Comparing manual assessment with assessment combining SLK-581, postcode and C3C2-SLK, for RACS and CACP clients, 1 July 2000 – 30 June 2003

		Automated assessment for Client IDs with same SLK-581			
			Different postcode		
		[a]Same postcode: replicate	[a]Same C3C2-SLK: replicate	[b]Different C3C2-SLK: coincident	Total
RACS			**Number**		
Manual assessment	Replicate[a]	186	132	10	328
	Coincident[b]	—	31	133	164
	Total	**186**	**163**	**143**	**492**
CACP					
Manual assessment	Replicate[a]	18	2	2	22
	Coincident[b]	—	2	10	12
	Total	**18**	**4**	**12**	**34**
RACS			**Per cent**		
Manual assessment	Replicate[a]	37.8	26.8	2.0	66.7
	Coincident[b]	—	6.3	27.0	33.3
	Total	**37.8**	**33.1**	**29.1**	**100.0**
CACP					
Manual assessment	Replicate[a]	52.9	5.9	5.9	64.7
	Coincident[b]	—	5.9	29.4	35.3
	Total	**52.9**	**11.8**	**35.3**	**100.0**

(a) Replicate: multiple SLK-581s (and therefore Client IDs) for a particular client.

(b) Coincident: identical SLK-581s for different clients.

Note: See notes to Table 3.5 on manual assessment.

Source: AIHW analysis of ACCMIS database.

The above results suggest that by using postcode and C3C2-SLK as secondary linkage variables, non-unique SLK-581 keys derived from the ACCMIS can be allocated as either relating to the same (replicated SLK-581 keys) or different (coincident SLK-581 keys) people with around 90% accuracy. Given the very small number of duplicates involved relative to the size of the data set, this provides a cost-effective way of distinguishing between the two. Any remaining misidentification will have an extremely marginal effect on analyses; in the RACS data for the 3 years examined, only 41 non-unique SLK-581 linkage keys would be wrongly identified as coincident or replicated keys, out of records for 191,413 Client IDs. For CACP clients, only four SLK-581 keys out of 42,298 would be misidentified. As in general the number of non-unique linkage keys increases with the number of people in a data set, for larger data sets the effects would be greater, and so a more detailed approach involving additional demographic data could be required to distinguish between coincident and replicated keys.

3.3 Other multiple representation on ACCMIS

Multiple representation of individuals on the database can lead to multiple SLKs being generated for some clients. In some cases the data for the SLK-581 linkage key will not change, leading to identical SLKs (as discussed above); in other cases, keys from the multiple Client IDs will differ and so the connection between related Client IDs will not be identified by the SLK-581 linkage key. The SLK-581 is just one way of identifying clients included more than once on the ACCMIS database. An indication of the extent of other multiple Client IDs is given below using three approaches.

Spelling variations

Differences in spelling names can lead to people with multiple Client IDs not being identified via the SLK-581. The preceding analysis suggests that using an alternative name-based linkage key in conjunction with postcode could identify other multiple Client IDs. For RACS data, a total of 224 Client IDs had a non-unique C3C2-SLK/ postcode combination. Of these, 62 were Client IDs with a unique SLK-581 (that is, they had not been identified as potential replicates using the SLK-581), including nine which had been identified by ACCMIS staff as they had a 'DO NOT USE' tag (or similar) appended to the family name. On manual inspection, four of the common C3C2-SLK/postcode combinations appeared to be for different people, with the remaining 58 having admissions consistent with multiple Client IDs being assigned for particular individuals. Four Client IDs for CACP recipients with a unique SLK-581 had a non-unique C3C2-SLK/postcode combination; all of these were identified manually as replicates.

Nicknames

One source of multiple Client IDs that may not be identified using either the SLK-581 or C3C2 keys is the use of different names (nicknames or pseudonyms) when recording people on the data set. This issue is investigated below for given names for which there are existing sets of alternative versions. The use of different family names by individuals is not considered as it is not possible to identify changes in name due to marriage or reversion to earlier names.

Case studies

A person may have more than one identification number where a second client record has been created using a pseudonym or nickname. For example, the given name for a client may be recorded as Molly for one Client ID and as Mary in a second; similarly Beth could replace Elizabeth and Jack replace John. The occurrence of multiple Client IDs involving nicknames was investigated using the National Death Index's standard list of pseudonyms. Initially, six pseudonym groups were investigated to gauge the prevalence of duplicates among very common given names. The groups chosen were those relating to John, Francis, Henry, Mary, Margaret and Elizabeth. The names included in the pseudonym groups are listed in

Appendix Table A.2. Among both RACS and CACP clients, 'Margaret' and 'Mary' were the largest pseudonym groups—both at around 6,700 occurrences for RACS clients and 1,500 for CACP recipients over the 3 years in the study. 'John' was the next most common group, with nearly 5,400 and 1,000 RACS and CACP Client IDs, respectively.

Table 3.9: Name variations in names, selected pseudonym name groups, for RACS and CACP, 1 July 2000 – 30 June 2003

Pseudonym group	Number of name variations	Number of Client IDs with given name in pseudonym group	
		RACS	CACP
John	26	5,365	1,034
Francis	18	2,217	458
Henry	19	1,905	374
Mary	13	6,693	1,507
Margaret	34	6,742	1,556
Elizabeth	44	3,608	861
Total	**154**	**26,530**	**5,790**

Note: Pseudonym groups use the National Death Index standard list of pseudonyms (see Table A.2).

Source: AIHW analysis of ACCMIS database.

A linkage key (Linkey3) based on the appropriate given name pseudonym group, the first four letters of the family name, date of birth and sex was created and the resulting non-unique keys were reviewed manually. The manual assessment of whether identical Linkey3s related to the same person considered demographic data including country of birth, Indigenous status, marital status, sex and the client's state of residence. However, the most critical factor when ruling out a match was overlapping dates of service provision and death as the reason for discharge prior to a later admission. At the same time, adjoining service periods weighted heavily in concluding that different Client IDs did in fact belong to the same person.

Among the 26,530 RACS Client IDs associated with the six names included in the study, just 78 (or 0.3%) had a non-unique Linkey3, resulting in 39 duplicate pairs (Table 3.10). Manual assessment indicated that 25 of these pairs (or 64%) involved multiple Client IDs for an individual, with the remainder being coincident Linkey3s for different people. In addition, the majority of duplicate pairs (34 out of 39) had either a common SLK-581 or C3C2-SLK. Of the 10 Client IDs (five pairs) not previously identified as possibly relating to the same client as another Client ID, Client IDs for two pairs were manually assessed as relating to the same client (including one Client ID with an instruction tag) and the remaining three were not.

For CACP clients, there were four non-unique Linkey3s (or 0.07%) out of the 5,790 Client IDs associated with the six chosen name groups. For one of these two pairs, the two Client IDs were manually assessed as relating to the same person; for the other the Client IDs were assessed as relating to different people. On its own, the SLK-581 distinguished correctly between the coincident and replicated Linkey3s, while both pairs had common C3C2-SLKs.

Table 3.10: Multiple Client IDs for individuals in selected pseudonym name groups, RACS and CACP clients, 1 July 2000 – 30 June 2003

Pseudonym group (Link name)	Duplicate pairs using Linkey3	Manual assessment		SLK-581		C3C2-SLK	
				Non-unique	Unique	Non-unique	Unique
RACS				Number of Linkey3 duplicate pairs			
John	6	Replicate[(a)]	5	3	2	3	2
		Coincident[(b)]	1	1	0	1	0
Francis	4	Replicate[(a)]	3	2	1	2	1
		Coincident[(b)]	1	1	0	1	0
Henry	3	Replicate[(a)]	1	1	0	1	0
		Coincident[(b)]	2	1	1	1	1
Mary	9	Replicate[(a)]	7	6	1	6	1
		Coincident[(b)]	2	2	0	2	0
Margaret	11	Replicate[(a)]	5	3	2	3	2
		Coincident[(b)]	6	3	3	3	3
Elizabeth	6	Replicate[(a)]	4	3	1	3	1
		Coincident[(b)]	2	0	2	0	2
Total	**39**	**Replicate[(a)]**	**25**	**18**	**7**	**23**	**2**
		Coincident[(b)]	**14**	**8**	**6**	**11**	**3**
CACP							
Henry	1	Replicate[(a)]	1	1	0	1	0
Elizabeth	1	Coincident[(b)]	1	0	1	1	0
Total	**2**	**Replicate[(a)]**	**1**	**1**	**0**	**1**	**0**
		Coincident[(b)]	**1**	**0**	**1**	**1**	**0**

(a) Multiple Linkey3s (and therefore Client IDs) for a particular client.

(b) Identical Linkey3s for different clients.

Source: AIHW analysis of ACCMIS database.

All names

The overall effect of the use of pseudonyms in the given name can be examined by deriving Linkey3 for all CACP and RACS Client IDs. However, for around 25% of Client IDs on ACCMIS the associated given name is not included on the National Death Index standard list of pseudonyms. For these clients, the reported given name was assigned as the standard name, and Linkey3 was derived as the first four letters of the family name combined with the standard given name, date of birth and sex.

Overall, a non-unique Linkey3 was derived for 379 RACS Client IDs and 14 CACP Client IDs. Among the 379 RACS Client IDs, only 24 (6%) had both a unique SLK-581 and C3C2-SLK; that is, only 6% had not previously been considered as a potential replicated record. Using manual assessment, 16 of these 24 related to clients with multiple Client IDs, including two with 'DO NOT USE' tags. Only three of the eight manually identified pairs of replicates had the same postcode as well. None of the 14

CACP Client IDs with a non-unique Linkey3 had both a unique SLK-581 and a unique C3C2-SLK.

The above findings indicate the majority of multiple Client IDs associated with nicknames or pseudonyms are identified using SLK-581 and/or C3C2-SLK in conjunction with postcode. Consequently, using Linkey3 adds only marginally to the identification of clients with multiple Client IDs.

Other names with instruction tags

For the period of interest, 105 RACS Client IDs with instruction tags on name data were identified (see Table 3.11). Forty-nine of these (along with the preferred Client ID) were found using SLK-581, postcode and C3C2-SLK. Another two (with their partner) were identified using Linkey3, leaving 54 cases where a client has more than one Client ID, but the related IDs have not been identified. For CACP Client IDs, only two names had tags, of which one was identified using SLK-581; the remaining tagged Client ID was not found using any of the three linkage keys considered.

Overall, the SLK-581 linkage key identifies a proportion of the clients with multiple Client IDs. Using C3C2-SLK in conjunction with postcode, some of the other multiply-occurring clients can be identified. However, even after using both these linkage keys there are still some duplicates on the database, as indicated by instruction tags in the name data. Compared with the number of clients, the number of unidentified clients remaining with multiple Client IDs is estimated to be very small: after allowing for those found using SLK-581 and the C3C2 linkage key, instruction tags indicate that 56 clients with multiple Client IDs remained unidentified, out of a total of 191,413 Client IDs in the period of interest (or 0.03%). For CACP clients, only two instruction tags were found, one of which was identified using the three linkage keys in conjunction with postcode.

Table 3.11: Instruction tags in name data, RACS and CACP clients, 1 July 2000 – 30 June 2003

	Instruction tag in name data		
	No	Yes	Total
RACS			
Unique SLK-581, C3C2-SLK and Linkey3	190,651	54	190,705
Identified through SLK-581 and/or C3C2-SLK			
Replicate based on SLK-581/postcode	158	28	186
Replicate, with both SLK-581 and C3C2-SLK not unique	153	10	163
SLK-581 not unique, C3C2-SLK unique	141	2	143
Replicate based on C3C2-SLK/postcode	53	9	62
C3C2-SLK not unique, SLK-581 unique	130	—	130
Total	*635*	*49*	*684*
Others identified through Linkey3			
Linkey3/postcode not unique	6	—	6
Other Linkey3 not unique	16	2	18
Total	*22*	*2*	*24*
Total	**191,308**	**105**	**191,413**
CACP			
Unique SLK-581 and C3C2-SLK	42,249	1	42,250
Replicate based on SLK-581/postcode	17	1	18
Other (SLK-581 not unique, or C3C2-SLK not unique, or Linkey3 not unique)	30	—	30
Total	**42,296**	**2**	**42,298**

Source: AIHW analysis of ACCMIS database.

3.4 Hidden clients

While some clients may be given multiple Client IDs, it is also possible for different people to be assigned erroneously to the same Client ID. Using the data on ACCMIS it is not possible to gauge the full extent of this phenomenon. However, whether clients are sometimes incorrectly assigned to an existing Client ID can be examined by looking at the prevalence of multiple deaths relating to a single Client ID, and the existence of overlapping periods of care.

Within the 3-year period 1 July 2000 and 30 June 2003, 70,286 residential aged care admissions had 'death' recorded as the reason for discharge (out of 244,602 completed periods of care). Some of these were reported against the same Client ID. Overall 649 RACS Client IDs—or 0.3% of all Client IDs—had 'death' recorded as the reason for discharge on more than one occasion. For the vast majority of these, two deaths were recorded (for 594, or 92%). However, four deaths were recorded against each of 11 Client IDs. Examining the multiple 'deaths' more closely, it was found that in all but three cases there was a re-admission (almost always to the same service) within a day of the recorded death, suggesting that although deaths had been

reported, the earlier 'deaths' had not in fact occurred and therefore multiple use of a Client ID was not involved.

In addition to the occurrence of multiple reported deaths, 10 RACS Client IDs had overlapping periods of care. Inspection showed that in two of these cases the overlap was valid, involving a same-day stay in a residential aged care service on the same day as the start of a longer stay in another service. A further six involved overlapping stays in the same agency, suggesting that errors in recorded dates could have been the cause rather than multiple use of a Client ID. However, for two Client IDs the overlapping care periods were in different agencies, and the periods of care were still incomplete by 30 June 2003. While these cases do not include those where the later user of the Client ID did not have periods of care overlapping those for another client, overall these results suggest that multiple use of RACS Client IDs rarely happens, especially when clients have overlapping periods of care.

For Community Aged Care Packages, 5,049 out of 26,933 periods of care that finished during the period of interest reportedly ended in death. Two deaths were reported against just three CACP Client IDs (or 0.1% of all Client IDs). In one case, re-admission to a CACP happened within a day of the first 'death'. In the other two cases, longer gaps were involved. However, neither of these clients had commonly occurring names, suggesting that misidentification was unlikely. Moreover, there were no incidents of overlapping periods of care within a Client ID for CACP records.

The above analysis suggests that there are very few cases (less than a handful) where events for several people are recorded against the same ACCMIS Client ID. In nearly all cases identified as possibly involving hidden clients, reporting practice or error seems to underlie apparently conflicting events, rather than the multiple use of a Client ID.

3.5 Summary

The quality of data on ACCMIS used to construct the SLK-581 linkage key is generally high. The main causes of errors when considering linkage are the use of 1 January birth dates and the occurrence of multiple Client IDs for individuals resulting in replicated SLKs. For the 3 years examined, just under 1,000 RACS Client IDs (0.5%) had reported 1 January birth dates, and 371 (0.9%) CACP Client IDs had such birth dates. Multiple representation on the ACCMIS database is less common, affecting fewer than an estimated 500 RACS Client IDs out of the 191,308 (0.3%) used for admissions between 1 July 2000 and 30 June 2003. For CACP clients commencing over the same period, fewer than 20 cases of multiple representation were identified among 42,298 Client IDs (or 0.05%).

4 Privacy protection protocols

The protection of the privacy of individuals is of key concern when linking between data sets. The security of data held by the Australian Institute of Health and Welfare has been of the highest importance since the Institute was established. The Explanatory Memorandum which accompanied the introduction of the *Australian Institute of Health Act 1987* stated: 'An important aspect of the Bill is the provision to protect the confidentiality of personal information given to the Institute. Any publications based on the work of the Institute may not identify an individual (including a deceased person...)' (Parliament of the Commonwealth of Australia House of Representatives 1987). The Act's confidentiality provisions are contained in Section 29 which prohibits disclosure or communication of information held under the Act, even to a court of law. Furthermore, Section 29(4)(e) of the Act includes the requirement not to disclose even the source of the information or 'the whereabouts, existence or non-existence of a document concerning a person'. Newly appointed Institute staff, including those employed on a short-term basis and staff of collaborating units, are required to sign an Undertaking of Confidentiality as soon as they start work.

While the Act governing the Institute includes strong measures for ensuring data security, protocols for handling confidential data within the Australian Institute of Health and Welfare further protect the privacy of individuals. In particular, staff only have access to those data sets required for their work. In addition, analyses involving the linkage of data for individuals must be approved by the Institute's Ethics Committee, and the Institute will not permit its data to be linked for administrative or regulatory purposes. The AIHW web site contains more detailed information on the strategies in place to ensure the confidentiality of its data (see <http://www.aihw.gov.au/privacy_of_data.html>).

In terms of data-handling, one of the prime ways to preserve privacy is to ensure that identifying information (such as name, date of birth and address) is not transferred from one data set to another. The protocol presented below for linking aged care data sets has been developed to ensure that in the course of linking two data sets held within the Institute, such identifying information is not transferred between the two. The privacy protocol was developed with reference to the framework suggested by the National Community Services Information Management Group for statistical data linkage in community services data collections (NCSIMG 2004). The principles underlying the protocol are that:

- Data linkage is not carried out directly between original complete data sets.
- Data linkage is undertaken using data sets that contain only the data required for establishing and validating links.
- Links between data sets are recorded using *project-specific* unique record identifiers so that links identified for a particular project cannot be used to

establish links between other data sets using a chain of links ('consequential' linking).

- Analysis files do not contain identifying data (such as name and address, or the record number from the original data set).
- Intermediate data sets and the project-specific record identifiers are deleted following development of the final linked analysis data sets.

The above principles relate to data-handling procedures and so do not necessarily require that the people undertaking the linkage are different from those doing the subsequent analysis. The key features of this protocol are the separation of personal identifying information from service information, and the absence of any record identifiers which would allow linkage back to the source data.

4.1 The privacy protocols

The privacy protocols to be used in the current project involving linking aged care services data sets within the AIHW involve nine steps. These are described below, and illustrated in Figure 5.1 using HACC and RACS data as an example.

Step 1. Obtain approval for the linkage project from the Institute's Ethics Committee.

Step 2. Derive the separate data sets for the programs to be included in the linkage analysis, adding a unique project record identifier. Data contained on these unlinked *source files* include:

- a project- and data set- specific unique record identifier
- data to be used to establish links (i.e. *the data linkage items*)
- data items to be used to clarify or check links (i.e. *linkage check items*)
- data items to be used in the final analysis.

Full name and detailed address data are not included on any of these source files.

Step 3. For each data set separately, derive a *match file* which contains:

- the data set specific unique project record identifier
- the data linkage items: SLK-581 linkage key and postcode
- additional data items to be used when choosing between non-unique links across data sets. Such data items will be limited to characteristics such as area of residence, country of birth, preferred language, and Indigenous status, and the C3C2-SLK when linking between CACP and RACS.

Step 4. Undertake linkage of the match files, and produce a *link file* showing the links using the project record identifiers. For a particular pair of match files, the link file contains only three data items:

- the unique project record identifier from the first match file

- the unique project record identifier from the second match file
- a unique link number for each record in the link file.

No other data are contained on the link file.

Step 5. Following linkage, delete the match files.

Step 6. Derive an *analysis file* from each of the source files by dropping the SLK-581 linkage key and retaining only data items to be used in the analysis. The analysis files will therefore contain:

- the data set specific unique project record identifier
- data items (including demographic items) to be used in the final analysis.

Potentially identifying data such as linkage keys, full name and detailed address data are not included on any of the analysis files.

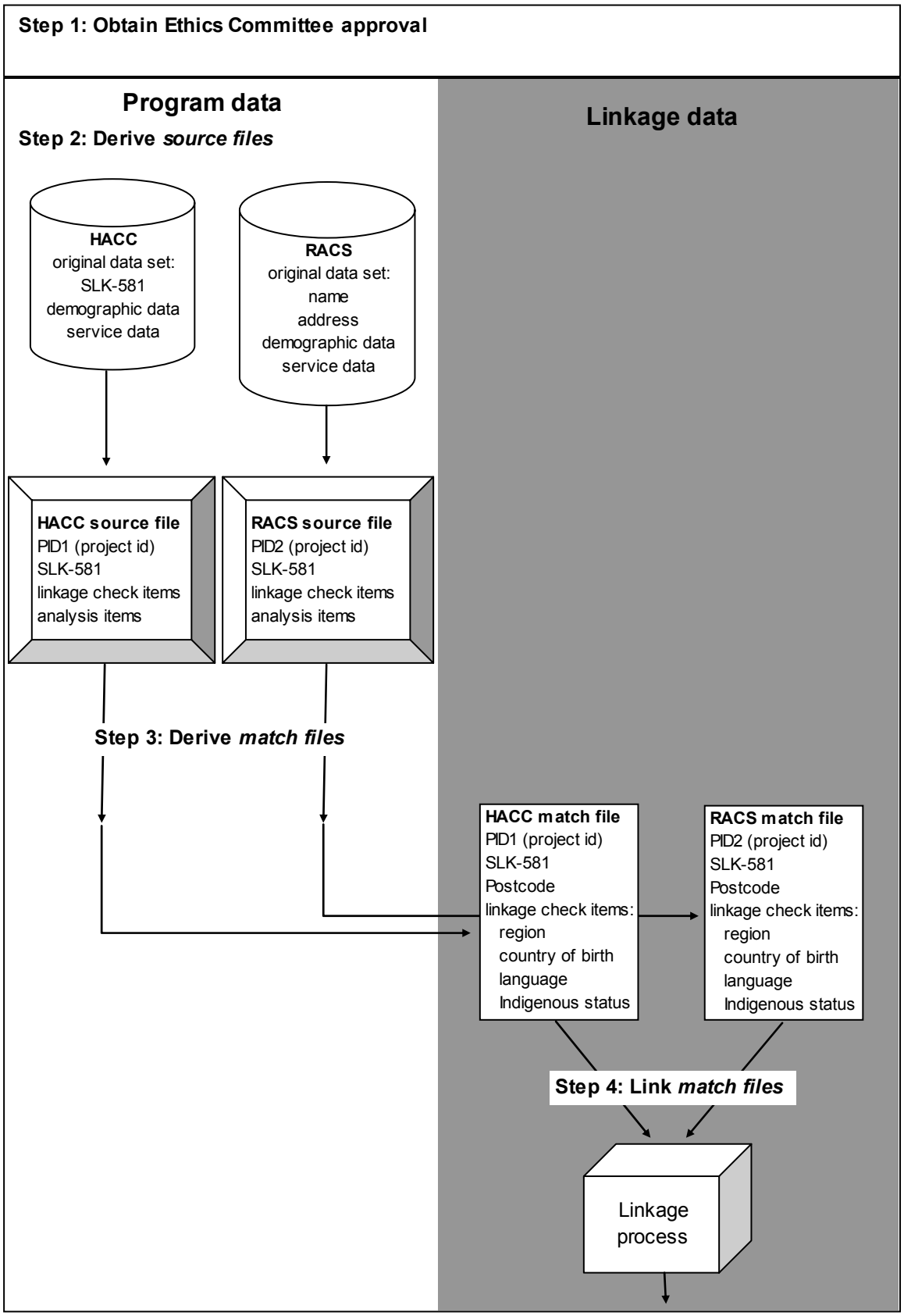
Step 7. Using the unique project identifiers, add the link number to the individual analysis files. The analysis files will therefore now contain:

- the link number of each record
- data items (including demographic items) to be used in the final analysis.

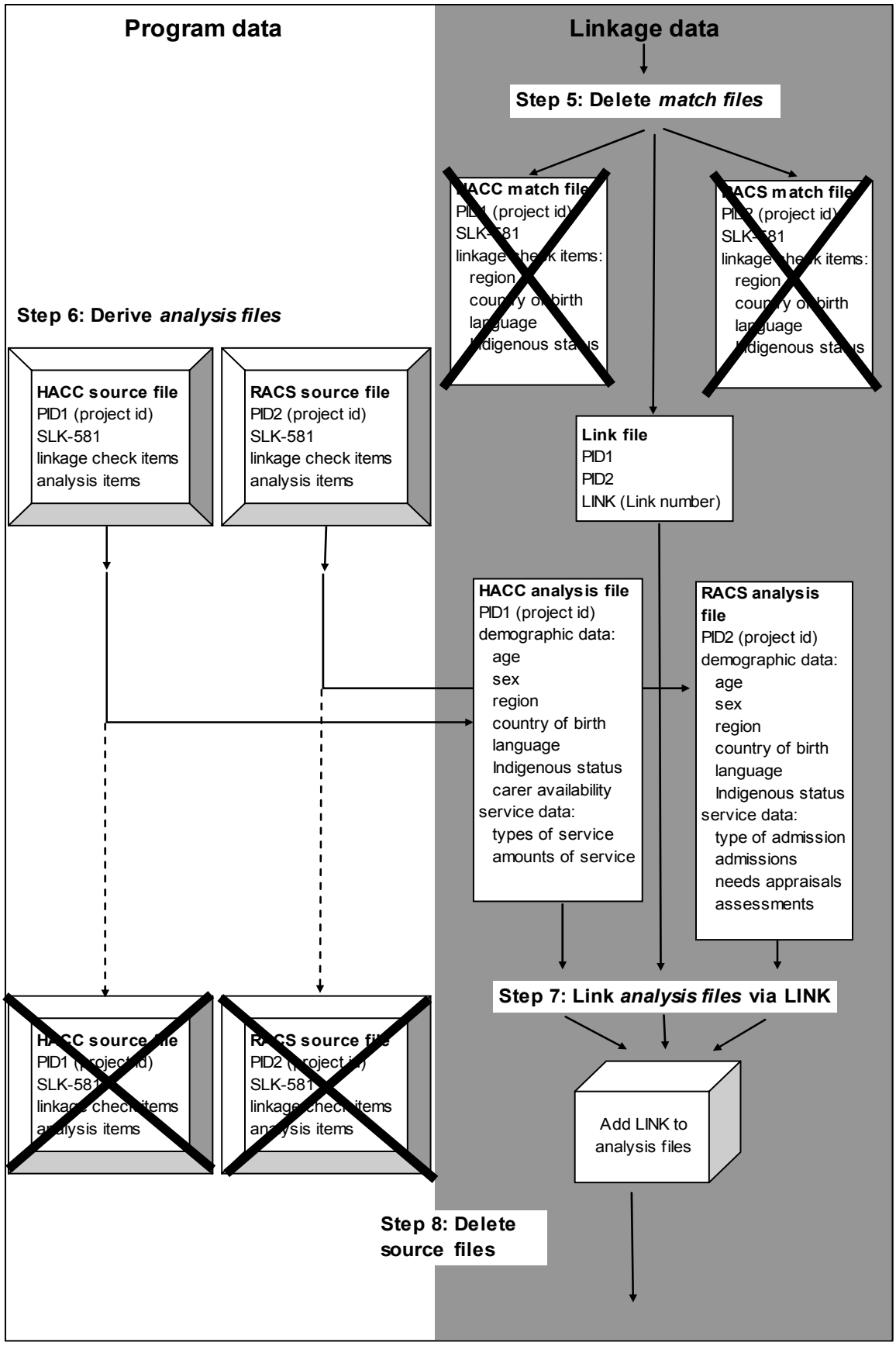
Step 8. Delete the source files so that there can be no linkage back to the original data sets.

Step 9. Link the analysis files using the *link number*.

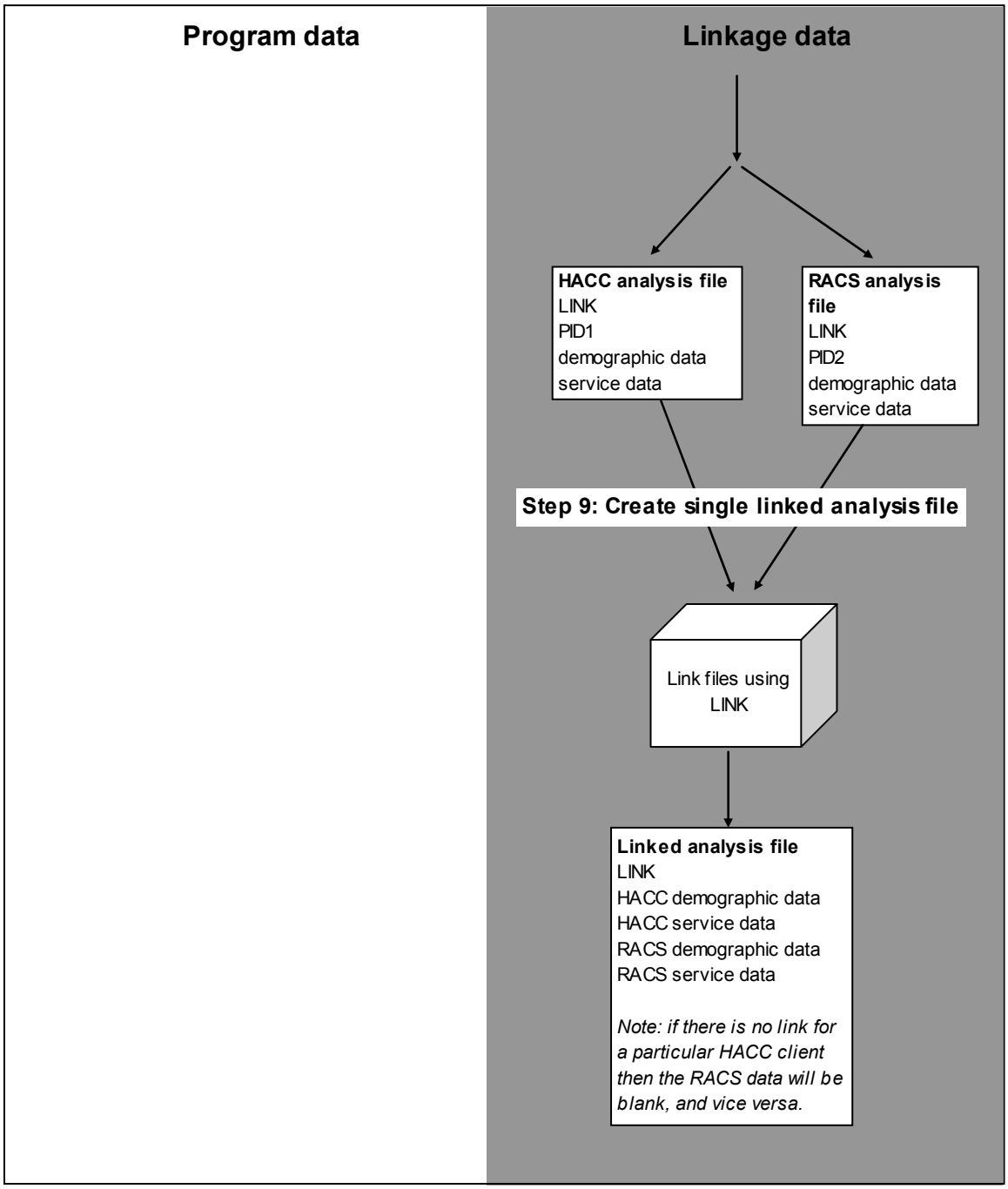
Figure 4.1: Data linkage protocol: an example linking data from the HACC and RACS programs



(continued)


Program data
Linkage data
Step 5: Delete match files
HACC match file
PID1 (project id)
SLK-581
linkage check items:
region
country of birth
language
Indigenous status
RACS match file
PID2 (project id)
SLK-581
linkage check items:
region
country of birth
language
Indigenous status
Step 6: Derive analysis files
HACC source file
PID1 (project id)
SLK-581
linkage check items
analysis items
RACS source file
PID2 (project id)
SLK-581
linkage check items
analysis items
Link file
PID1
PID2
LINK (Link number)
HACC analysis file
PID1 (project id)
demographic data:
age
sex
region
country of birth
language
Indigenous status
carer availability
service data:
types of service
amounts of service
RACS analysis file
PID2 (project id)
demographic data:
age
sex
region
country of birth
language
Indigenous status
service data:
type of admission
admissions
needs appraisals
assessments
Step 7: Link analysis files via LINK
HACC source file
PID1 (project id)
SLK-581
linkage check items
analysis items
RACS source file
PID2 (project id)
SLK-581
linkage check items
analysis items
Add LINK to analysis files
Step 8: Delete source files

(continued)


Program data
Linkage data
HACC analysis file
LINK
PID1
demographic data
service data
RACS analysis file
LINK
PID2
demographic data
service data
Step 9: Create single linked analysis file
Link files using LINK
Linked analysis file
LINK
HACC demographic data
HACC service data
RACS demographic data
RACS service data
Note: if there is no link for a particular HACC client then the RACS data will be blank, and vice versa.

5 Linkage protocols

The purpose of linkage protocols is to ensure that appropriate methods are used when undertaking the linking and that a linkage analysis is reproducible. In Chapter 2, two options were put forward for undertaking linkage, depending on the additional data available to enhance the linkage.

- *Option 1 – Basic linking:* Exclude all records with SLK-581 linkage keys incorporating missing or poor information from the data set for linking.
- *Option 2 – Enhanced linking:* As far as possible, retain all valid data in the linkage key, and use this in conjunction with other information to establish links with other data sets.

In the protocols described below, if non-unique links occur then the final link to be used is chosen randomly. Given the rare occurrence of non-unique SLKs in the aged care data sets, this approach will affect the final linked data set only marginally and avoids any possible bias introduced by using highly clustered cultural diversity items to distinguish between links. Moreover, it provides additional privacy protection by introducing an additional—albeit small—random element into the matching process.

Note that if the data used to establish links can change for a client over time, then the linkage should be re-done as data are updated, or as different periods are examined. In the current study, information used in the linkage and subject to change (as opposed to being reported erroneously) includes name data and client region.

5.1 Preparing the HACC MDS for linkage

The findings in Chapter 2 suggest the following strategies for the two options put forward for undertaking linkage with the HACC MDS, depending on the availability of additional data to enhance the linkage. However, before preparing the data specifically for basic or enhanced linkage, a number of steps need to be performed to identify cases with poor quality linkage key information.

Initial data cleaning

HACC data is prepared for linkage using the following four steps.

1. Identify all SLK-581 linkage keys derived using poor quality data (N_1 out of N SLKs), including:
    - birth dates more than 110 years before the period of interest
    - all 1 January birth dates
    - any missing name data (first and/or family name)

- missing sex.

Calculate the proportion or SLKs affected by each of these problems, separately and combined.

2. Replace missing or invalid client postcodes by '9999'. Note that valid postcodes are between 200 and 8000.
3. Where at least complete date of birth data and either letters of the family name or given name are available, use postcode data to derive adjusted linkage keys incorporating all known information in the incomplete SLK-581 key. Note, however, that if only sex is missing from the key the adjusted key need not include postcode. Identify those cases (N_2) highly likely to be associated with a complete SLK-581 in the same collection. If service data is being analysed, the service information for the N_2 incomplete SLKs can be combined with that for the associated complete SLK-581. If more than one incomplete SLK-581 is associated with a complete SLK-581, then all records are assumed to be for the same client. In the unlikely case of links to more than one complete SLK-581, randomly choose one as the associated SLK. This is only necessary if service data are to be combined.
4. In other cases with incomplete linkage keys, where at least either complete date of birth or letters of name data are available, use postcode data to derive adjusted linkage keys. For reporting purposes, identify those cases (N_2') possibly associated with a complete SLK-581 in the same collection.

Following this initial examination of the linkage key information, the data are then prepared for the specific linkage option being used.

Option 1: Basic linking using SLK-581 only

This option should be used when data on client postcode (or similar small area) are either not available or not comparable on the data sets that are being linked. The steps taken to prepare the data, and to obtain important pre-linkage information on the quality of the data, are as follows:

1. Identify variables common to both data sets that could be used to validate the linkage; for example, broad geographic variables and cultural diversity information.
2. Exclude all N_1 SLKs derived using poor quality data from subsequent linking.
3. Using results from the initial data cleaning, estimate the number ($N_1 - N_2$) of HACC clients with inadequate SLK-581 keys for linkage across data sets (as per Table 2.5), and the number (N_2') of these that could possibly be associated with a complete linkage key. The number of HACC records being linked is then ($N - N_1 + N_2$), relating to between ($N - N_1 + N_2$) and ($N - N_1 + N_2 + N_2'$) clients. These estimates should be reported to indicate the coverage of the linkage process, along with the agency non-participation rate.
4. For the ($N - N_1$) records with complete SLK-581 linkage keys, link data sets using SLK-581.

5. In the rare case of non-unique links being identified, randomly choose the link to be retained. Report the number of times such random choice is used.

Option 2: Enhanced linking using SLK-581 with additional data

This option can be used when data on client postcode is available and comparable for the data sets that are being linked. If comparable postcode (or similar) data are not available in both data sets then Option 1 should be adopted.The steps taken to prepare the data, and to obtain important pre-linkage information on the quality of the data, are as follows:

1. Identify variables common to both data sets that can be used to validate the linkage; for example, broad geographic variables and cultural diversity information.
2. Exclude cases (N_3) where at least complete date of birth data in conjunction with either letters of the family name or given name are not available, or the postcode data is invalid where sufficient linkage key data are available. This limits the adjusted keys to those where the distinct key rate is least 97% or more—when combined with postcode—among data sets with up to 400,000 distinct SLK-581 link keys (see Table 2.5). Report the number of cases excluded by this process as part of the analysis to indicate the coverage of the linkage process, along with the agency non-participation rate.
3. Excluding the N_2 cases identified in the data cleaning process as highly likely to be associated with a record with a complete SLK-581, retain all other records ($N - N_2 - N_3$) for linkage across data sets, and estimate the number ($N_1 - N_2 - N_3$) of HACC clients using additional information when linking across data sets. Report these estimates as part of the analysis to give a measure of the coverage of the linkage process.
4. For the ($N - N_1$) records with complete SLK-581 linkage keys, link data sets using SLK-581.
5. Derive adjusted linkage keys using known SLK-581 components in conjunction with client postcode, and use these linkage keys to link the ($N_1 - N_2 - N_3$) clients with incomplete SLKs with similarly adjusted keys—for all clients not previously linked—in the other data set. Note that several adjusted keys will need to be derived for all records on both data sets included in the linkage to allow for different missing information in either set. Five reduced keys are used, being applied in order of linking efficiency (see the top five keys in Table 2.5), with linked records being removed from the process at each stage to avoid non-unique links. As when identifying associated linkage keys, if only sex is missing from the key it is not necessary to include postcode for this reduced key.
6. In the rare case of non-unique links being identified, randomly choose the link to be retained. Report the number of times such random choice is used.

5.2 Preparing ACCMIS data for linkage

The analysis in Chapter 3 on data quality suggests the following approach for preparing RACS and CACP data for linkage.

Initial data cleaning

Again, before preparing the data specifically for basic or enhanced linkage, a number of steps needs to be undertaken to identify cases with poor quality linkage key information. This stage is more complex than that for the HACC MDS because of the more detailed name information available.

1. Identify all SLK-581 linkage keys derived using possibly poor quality data (N_1 out of N SLKs) including:
   - birth dates more than 110 years before the period of interest
   - all 1 January birth dates
   - any missing name data (given and/or family name). In the very few cases for which it is relevant, replace missing (or X-ed out) name information by '99999'
   - missing sex. In the very few cases for which it is relevant, replace missing information by '9'.
2. Replace missing or invalid client postcodes by '9999'. Note that valid postcodes are between 200 and 8000.
3. In cases with incomplete date of birth (excluding the very few with missing name or sex) use postcode data to derive adjusted linkage keys. For reporting purposes, identify those cases (N_2') possibly associated with a complete SLK-581 in the same collection. Note that as date of birth is overwhelmingly the cause of possibly-incomplete keys on ACCMIS, it is not necessary to try to identify associated keys for cases where name and/or sex data are missing.
4. Identify Client IDs with tags in the name data.
5. Identify non-unique SLK-581 linkage keys (N_4), and distinguish between replicates (N_{4r}) and coincident keys (N_{4c}) using SLK-581/postcode/C3C2-SLK. Replicated SLK-581 keys for a single client are identified by identical postcode *and/or* identical C3C2-SLK information. Coincident SLK-581 keys for different clients are identified by non-identical postcode *and* non-identical C3C2-SLK.
6. Identify other clients with multiple representation on the data set using C3C2-SLK/postcode.
7. For replicates with a tagged partner (identified using either key), retain the Client ID without the tag. Otherwise, retain the most recently used Client ID. All events related to the excluded Client ID should be treated as belonging to the retained Client ID.
8. Retain all other tagged Client IDs (under 60 cases in the current study), noting the number involved (N_5). This number should be reported as part of the write up of

the analysis. These cases are not dropped because this would result in loss of admission data.

Option 1: Basic linking using SLK-581 only

The steps taken to prepare the data, and to obtain important pre-linkage information on the quality of the data, are as follows:

1. Identify variables common to both data sets that can be used to validate the linkage; for example, broad geographic variables and cultural diversity information.
2. Out of the total of N linkage keys, exclude all N_1 SLK-581 keys derived using poor quality data from subsequent linking. Note that this includes any 1 January birth dates considered to be of poor quality in either data set being linked. For example, if linking to HACC data poor birth dates include all 1 January birth dates, but if linking to RACS or CACP data only first of the decade and birth dates earlier than 110 years before the period under analysis are considered to be of poor quality.
3. Using results from the initial data cleaning, report the numbers of Client IDs N_1, N_2', N_{4r}, N_{4c}, N_5. The number of Client IDs being linked is then $(N - N_{4r} - N_1)$, relating to between $(N - N_{4r} - N_5 - N_1)$ and $(N - N_{4r} - N_5 - N_1 + N_2')$ clients.
4. For the $(N - N_{4r} - N_1)$ Client IDs with complete SLK-581 linkage keys, link data sets using the SLK-581.
5. Noting that the above steps result in N_{4c} non-unique SLKs on the prepared data set, in the rare case of non-unique links being identified, randomly choose the link to be retained. Report the number of times such random choice is used.

Option 2: Enhanced linking using SLK-581 with additional data

This option can be used when data on client postcode is available and comparable for the data sets that are being linked. If comparable postcode (or similar) data are not available in both data sets then Option 1 should be adopted. Note that for RACS and CACP data, poor date of birth data is the overwhelming cause of incomplete SLK-581 linkage keys, with data for name and/or sex very rarely missing. Consequently, enhanced linking solely using postcode cannot be used to identify links between these two programs for cases with poor linkage key data. Furthermore, given the extremely small number of RACS and CACP cases on ACCMIS with missing name or sex data (fewer than five over 3 years of admissions) no data preparation additional to that carried out for basic linking is considered necessary when using enhanced linking between these and other data sets.

The steps taken to prepare the data, and to obtain important pre-linkage information on the quality of the data, are as follows, with steps 1 to 4 being identical to steps 1 to 4 in the basic linking option:

1. Identify variables common to both data sets that can be used to validate the linkage; for example, broad geographic variables and cultural diversity information.
2. Out of the total of N linkage keys, exclude all N_1 SLK-581 keys derived using poor quality data from subsequent linking. Note that this includes any 1 January birth dates considered to be of poor quality in either data set being linked. For example, if linking to HACC data poor birth dates include all 1 January birth dates, but if linking to RACS or CACP data only first of the decade and birth dates earlier than 110 years before the period under analysis are considered to be of poor quality.
3. Using results from the initial data cleaning, report the numbers of Client IDs N_1, N_2', N_{4r}, N_{4c}, N_5. The number of Client IDs being linked is then $(N - N_{4r} - N_1)$, relating to between $(N - N_{4r} - N_5 - N_1)$ and $(N - N_{4r} - N_5 - N_1 + N_2')$ clients.
4. For the $(N - N_{4r} - N_1)$ Client IDs with complete SLK-581 linkage keys, link data sets using the SLK-581.
5. Derive adjusted linkage keys using known SLK-581 components in conjunction with client postcode. These are used to link to records with reduced linkage key data in the second data set. Note that several adjusted keys will need to be derived to allow for different types of missing information in the second data set. Five reduced keys are used, being applied in order of linking efficiency (see the top five keys in Table 2.5), with linked records being removed from the process at each stage to avoid non-unique links.
6. Noting that the above steps result in N_{4c} non-unique SLKs on the prepared data set, in the rare case of non-unique links being identified, randomly choose the link to be retained. Report the number of times such random choice is used.

5.3 Linking between programs

Two options have been proposed when linking data from different programs: basic linking using the SLK-581 only (Option 1), and enhanced linking using SLK-581 in conjunction with postcode (or equivalent) (Option 2). The choice of option for a particular project will depend on a number of factors, including the time available to carry out the linkage, and the additional geographic data available for inclusion in any adjusted linkage keys. In addition, a range of comparable data items in the two data sets are required to allow validation of the linked data set once links have been established.

Most of the variables considered in Chapter 2 for refining the linkage—state of usual residence, country of birth, language spoken and Indigenous status—could also be used to validate the resulting linked data set. However, as when using postcode for linking, their utility depends on the comparability of these items across the HACC MDS, CACP and RACS data sets. The comparability of data items across the various data sets is discussed below.

Data comparability

While the postcode numbers for a particular area are rarely changed, there is a significant difference in the way client postcode is recorded on the HACC MDS and ACCMIS, especially prior to the introduction in January 2003 of a new assessment form for use by Aged Care Assessment Teams (ACATs) (DoHA 2003a). The client postcode recorded on the HACC MDS refers to the postcode in which the client lives (AIHW 1998:59, DoHA 2004:34). However, for ACCMIS data this is not necessarily the case. Prior to January 2003, the client postcode recorded by the ACAT—and therefore included on ACCMIS—related to the address at which a client could be contacted by the Assessment Team (DHAC 1999); this latter could have been the address of a relative, hospital or a residential aged care service. From January 2003, using the new form, ACATs have reported the address 'where the client usually lives'.

Since recipients of CACPs are still living in the community, it is highly likely that even before 2003 in most cases the contact postcode was also that for their usual residence.[1] This is also likely to be the case for the majority of people using residential respite care, although 25% of assessments prior to admission into residential respite care take place in a hospital (Table 5.1). However, for people moving into permanent residential aged care more than half of the associated assessments take place in hospital, suggesting that in many cases the contact address and address of usual residence may be different. These results suggest that client postcode for pre-2003 RACS clients (especially once they are moving permanently into residential aged care) may not be comparable with that recorded for clients of the community-based programs of HACC and CACP.

Table 5.1: Place of assessment of last ACAT assessment prior to first admission into residential aged care during 1 July 2001 – 30 June 2002, by type of first admission

Place of assessment	Respite	Permanent	Total
		Number	
Aged care facility	1,270	2,086	3,356
At home	20,359	11,057	31,416
Hospital	7,528	18,951	26,479
Other	1,479	1,531	3,010
Total	**30,636**	**33,625**	**64,261**
		Per cent	
Aged care facility	4.1	6.2	5.2
At home	66.5	32.9	48.9
Hospital	24.6	56.4	41.2
Other	4.8	4.6	4.7
Total	**100.0**	**100.0**	**100.0**

Note: First admissions in the year resulting from transfer within care type are not included: 7,176 permanent to permanent and respite to respite transfers have been excluded.

Source: AIHW analysis of ACCMIS database.

[1] Data on place of assessment for CACP recipients is not available on ACCMIS.

A further issue which complicates the use of ACCMIS postcode data is that only one is retained on the database for each RACS or CACP client. That is, no history is retained for this data item so that if a person changes 'ACAT' address—for example, becomes a permanent aged care resident—the client postcode also changes. As a consequence, if postcode is to be included in the linkage process, different snapshots (or refreshes) of the database need to be accessed to pick up the client postcode relevant at different times.

Table 5.2: Comparison of data items on HACC MDS and ACCMIS

Data item	HACC MDS	ACCMIS
Postcode (or state)	Postcode (or state) for person's residence	– For assessments prior to January 2003: postcode (or state) of address where can usually be contacted. – For assessments post January 2003: postcode (or state) for where client usually lives.
Country of birth	Uses SACC	– Prior to AIHW December 2003 refresh[(a)]: ASCCSS for RACS, different code set for CACP – From (and including) AIHW December 2003 refresh[(a)]: SACC
Language	Main language spoken at home, using an adaptation of ASCL	– For assessments prior to January 2003: preferred language (unknown classification) – For assessments post 1 January 2003: language spoken at home other than English, using ASCL ('not stated' has different code than on HACC MDS)
Indigenous status	Only Aboriginal, only Torres Strait Islander, both Aboriginal and Torres Strait Islander, or neither Aboriginal nor Torres Strait Islander	– For assessments prior to January 2003: yes/no to either Aboriginal or Torres Strait Islander – For assessments post 1 January 2003: 'Yes' only Aboriginal, or 'Yes' only Torres Strait Islander, or 'Yes' both, or 'No' neither.

(a) Refresh (or database snapshot) as received by AIHW.

Notes

1. SACC = Standard Australian Classification of Countries
2. ASCCSS = Australian Standard Classification of Countries for Social Statistics
3. ASCL = Australian Standard Classification of Languages

Sources: AIHW 1998; Aged Care Application and Approval (DHAC form 2624 (9908)); Aged Care Client Record (DoHA form 3020 (0302)).

Turning to other data items that could be used for linkage validation, the data collected and classifications used in the HACC MDS and on ACCMIS are reasonably comparable for country of birth, language spoken and Indigenous status, with comparability increasing since January 2003 with the introduction of the new assessment form (Table 5.2). Prior to this, preferred language was recorded on ACCMIS, rather than main language spoken at home as recorded in the HACC MDS, and there were some differences in the method used to collect Indigenous status. In addition, quite different code sets were used for country of birth, but in 2003 the country of birth codes used on the ACCMIS database were standardised to the Standard Australian Classification of Countries (SACC)—as used in the HACC MDS—with earlier data being recoded. While there are still minor differences in some of the codes used, concordance between the code sets used on the two databases is high.

Choice of linkage type

The choice between basic and enhanced linkage is dictated to a large extent by the availability of comparable postcode information on the two data sets being linked.

Linking between HACC and CACP

It is likely that the client postcode recorded for CACP recipients on ACCMIS refers to the client's usual residence, even before 2003 when the recorded client referred to a contact address. Therefore, when linking data from the HACC and CACP programs, the enhanced linkage (Option 2) can be used, using client postcode to augment the SLK-581. If time is limited, the basic option could be used with only a small fall in the coverage of the linkage.

Linking between HACC and RACS

Before 1 January 2003

Before the introduction of the new ACAT form in 2003, a RACS client's usual residence and contact address could have been different in a significant number of cases, with around one-quarter of RACS respite admissions and one-half of permanent admissions involving an in-hospital assessment. Consequently, client postcode as recorded on the HACC MDS and on ACCMIS for RACS clients may not be comparable in many cases, particularly for people entering permanent residential aged care. Consequently, because using enhanced linking (Option 2) adds relatively few cases to the number of HACC clients records being considered for linkage, enhanced linking is not recommended for linking between HACC use and either permanent or respite admissions into RACS before 2003.Therefore, basic SLK-581 linkage (Option 1) should be used when linking HACC and RACS data prior to 2003.

After 1 January 2003

After the introduction of the new ACAT form, the HACC and RACS client postcode data are compatible, and so Option 2 can be used, with client postcode enhancing the

SLK-581. However, if resources are limited, the basic option could be used with only a small fall in the coverage of the linkage.

Linking between CACP and RACS

As discussed previously, few RACS and CACP clients have missing data for the SLK-581 linkage key, and in nearly all these cases there is insufficient data to consider using enhanced matching. Therefore, simple linkage (Option 1) should be used when linking between these two programs

5.4 Choosing values for common data items

When linking two data sets, both may contain information relating to a number of characteristics which are not expected to change over time. For example, both the HACC MDS and ACCMIS contain information on country of birth, language spoken at home and Indigenous status. While in most cases the information recorded against these items will be the same in both data sets, in some cases there may be conflicting information. In this situation a decision must be made as to which data should be used in any analyses.

The following rules for choosing between differing information on two data sets were drawn up after considering the source of the information. In the HACC MDS the demographic variables recorded at the most recent HACC assessment are retained on the data set, irrespective of whether or not they are missing or whether there is conflicting information from assessments from other agencies. Also, although ACAT assessments are valid for 12 months, on ACCMIS there is no requirement for the demographic details of a client to be updated using more recent ACAT assessment information. Therefore, information on the system for the date of the most recent assessment may not reflect the currency of the data.

Rules for choosing values for common data items

1. Missing data should never be selected over a valid response.
2. A client identified as Indigenous on *either* data set should be recorded as Indigenous on the linked data set. Note that the distinction between Aboriginal and Torres Strait Islander is rarely used in analysis, but if required and there is a conflict then rules 4 and 5 should be used.
3. A client identified as non-English speaking, or as from a mainly non-English speaking country, on *either* data set should be recorded as such on the linked data. Note that if there is a conflict between two non-English speaking responses, then rules 4 and 5 should be used.
4. Given the more formal assessment arrangements for RACS and CACP clients compared with HACC clients, and the greater contact involved in these programs, the values recorded for CACP and RACS are to be preferred over those recorded in the HACC MDS.

5. Given the greater contact involved in residential aged care compared with Community Aged Care Packages, the values recorded for RACS are to be preferred over those recorded for CACP.

5.5 Summary

When linking data for aged care programs, the individual data sets should be analysed to identify the extent of poor quality linkage keys and multiple representation of clients. Where comparable client postcode data are available, enhanced linkage which uses postcode to augment linkage keys with some missing data can be used. Among the aged care programs considered in this report, such enhanced linkage is appropriate only when linking between the HACC and CACP programs; basic linkage using only SLK-581 should be used when linking between HACC and RACS and between CACP and RACS. Using basic linkage between HACC and CACP, rather than enhanced linkage, marginally reduces the HACC records available for linking. In the final linked data set, differences between variables common to the two source data sets should be resolved using pre-determined rules.

Appendix tables

Table A.1: Frequency of letters in names

Letter position		Letter	%
	Given name		
First	More than 10%	J	11.0
	Closest under 10%	M	8.7
	Least common	X, Q	<0.1
	L	L	5.0
Second	More than 10%	A	21.5
		E	16.9
		O	18.3
	Closest under 10%	I	9.5
	Least common	X, Q	<0.1
	L	L	7.7
Third	More than 10%	N	10.2
		R	15.5
	Closest under 10%	L	9.4
	Least common	Q	<0.1
	L	L	9.4
	Name shorter than 3 letters		0.1
Fourth	More than 10%	E	11.8
		N	13.8
	Closest under 10%	A	8.4
	Least common	Q	0.1
	L	L	8.1
	Name shorter than 3 letters		3.0
	Family name		
First	More than 10%	M	10.3
	Closest under 10%	B	9.5
	Least common	X	<0.1
	L	L	4.3
Second	More than 10%	A	22.5
		E	12.1
		I	10.5
		O	17.1
	Closest under 10%	R	7.8
	Least common	Q, X	<0.1
	L	L	3.8
Third	More than 10%	L	11.2
		R	12.6
	Closest under 10%	A	8.2
	Least common	Q	0.1
	L	L	11.2
Last	More than 10%	E	10.4
		N	20.3
		S	14.4
	Closest under 10%	Y	8.8
	Least common	Q	<0.1
	L	L	5.0
Second last	More than 10%	E	19.0
		O	13.4
	Closest under 10%	A	9.4
	Least common	Q, X	<0.1
	L	L	8.7
Third last	More than 10%	A	10.4
		E	10.2
	Closest under 10%	L	8.9
	Least common	Q	<0.1
	L	L	8.9
	Name shorter than 3 letters		0.1

Note: Table uses National Death Index data for people born between 1913 and 1987 (inclusive), and died between 1980 and January 2000.

Source: Reproduced from AIHW: Karmel 2000 (unpublished):Table A.1.

Table A.2: Name variations for selected pseudonym name groups, National Death Index standard list of pseudonyms

Pseudonym name group	Name variations		
John	Giouanna	Jaques	Juan
	Giovanna	John	Sean
	Giovanni	Johnathan	Shane
	Giovannin	Johnathon	Shaun
	Ivan	Johnnie	Shawn
	Jack	Johnny	Shayne
	Jacko	Jon	Yannis
	Jackson	Jonathan	Yvan
	Jacques	Jonathon	
Francis	Cisco	Francesca	Franco
	Fan	Francesco	Frank
	Fannie	Francie	Frankie
	Fanny	Francine	Franklin
	Fran	Francis	Frans
	Frances	Francisco	Franz
Henry	Enrico	Halley	Hendrick
	Enrique	Hank	Henrey
	Enriquez	Har	Henrick
	Hal	Harold	Henrique
	Halbert	Harrie	Henry
	Halden	Harrold	
	Haley	Harry	
Mary	Maire	Maria	Mollie
	Mame	Marie	Molly
	Mamie	Mary	Moyra
	Mare	Marya	
	Maree	Moira	
Margaret	Greta	Margery	Marjory
	Gretal	Margie	Meaghan
	Gretchen	Margo	Meg
	Madge	Margorie	Megan
	Maggie	Margot	Miriam
	Maisie	Margret	Peg
	Marg	Marguerita	Peggie
	Margaret	Marguerite	Peggy
	Margareta	Margurita	Reta
	Margarita	Marj	Rita
	Marge	Marjorey	
	Margerita	Marjorie	
Elizabeth	Babette	Elisabetta	Lisabeth
	Bes	Elise	Lisbeth
	Bess	Elissa	Lise
	Bessie	Eliz	Lisette
	Bessy	Eliza	Lissa
	Bet	Elizabeth	Liz
	Beth	Ella	Liza
	Betsey	Elle	Lizabeth
	Betsy	Elli	Lizbeth
	Bett	Ellie	Lizette
	Bettie	Els	Lizzie
	Bettina	Elspeth	Lizzy
	Betty	Lecia	Lyssa
	Elisa	Libby	Panagiota
	Elisabeth	Lisa	

Source: National Death Index.

References

AIHW (Australian Institute of Health and Welfare) 1998. HACC data dictionary version 1.0. Canberra: Commonwealth Department of Health and Family Services.

AIHW 2003. Australia's National Disability Services Data Collection: redeveloping the Commonwealth State/Territory Disability Agreement National Minimum Data Set (CSTDA NMDS). Cat. no. DIS 30. Canberra: AIHW.

AIHW 2004a. Community Aged Care Packages Census 2002. Cat. no. AGE 35. Canberra: AIHW (Aged Care Statistics Series no. 17).

AIHW 2004b. Community Aged Care Packages in Australia 2002–03: a statistical overview. Cat. no. AGE 39. Canberra: AIHW (Aged Care Statistics Series no. 19).

AIHW 2004c. Day Therapy Centres Census 2002. Cat. no. AGE 34. Canberra: AIHW (Aged Care Statistics Series no. 16).

AIHW 2004d. Extended Aged Care at Home Census 2002: a report on the results of the census conducted in May 2002. Cat. no. AGE 33. Canberra: AIHW (Aged Care Statistics Series no. 15).

AIHW 2004e. Residential aged care in Australia 2002–03: a statistical overview. Cat. no. AGE 38. Canberra: AIHW (Aged Care Statistics Series no. 18).

AIHW: Karmel R 2000 (unpublished). Duplicates in the SAAP linkage key. Canberra: AIHW.

AIHW: Karmel R 2005. Transitions between aged care services. Cat. no. CSI 2. Canberra: AIHW (Data Linkage Series no. 2).

AIHW: Ryan T, Holmes B & Gibson D 1999. A national minimum data set for Home and Community Care. Cat. no. AGE 13. Canberra: AIHW.

DHAC (Department of Health and Aged Care) 1999. Aged care application and approval. Canberra: Department of Health and Aged Care.

DoHA (Department of Health and Ageing) 2003a. Aged care client record 3020(0302). Canberra: Department of Health and Ageing.

DoHA 2003b. Home and Community Care Program Minimum Data Set, 2002–03 annual bulletin. Canberra: Department of Health and Ageing. Viewed 19 March 2004, <www.hacc.health.gov.au/mds/statindx.htm>.

DoHA 2004. Guidelines to the HACC MDS: a companion to the HACC data dictionary V 1.0 (May 1998). Canberra: DoHA. Viewed 24 June 2004 <www.hacc.health.gov.au/pub_mds_gdd.htm>.

NCSIMG (National Community Services Information Management Group) 2004. Statistical data linkage in community services data collections: a report prepared by the Statistical Linkage Key Working Group. Canberra: AIHW.

Parliament of the Commonwealth of Australia House of Representatives 1987. Australian Institute of Health Bill 1987, Explanatory Memorandum. pp. 11264/87, Cat. no. 87 4080 2.