Australian Government
Australian Institute of Health and Welfare

Better information and statistics
for better health and wellbeing

DATA LINKAGE SERIES
Number 11

Comparing an SLK-based and a name-based data linkage strategy

An investigation into the PIAC linkage

February 2011

Australian Institute of Health and Welfare
Canberra

Cat. no. CSI 11

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

This publication is part of the Australian Institute of Health and Welfare's Data linkage series. A complete list of the Institute's publications is available from the Institute's website <www.aihw.gov.au>.

ISSN 1833-1238

ISBN 978-1-74249-124-0

Suggested citation

Australian Institute of Health and Welfare 2011. Comparing an SLK-based and a name-based data linkage strategy: an investigation into the PIAC linkage. Data linkage series no. 11. Cat. no. CSI 11. Canberra: AIHW.

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

Data Linkage Unit

Australian Institute of Health and Welfare

GPO Box 570

Canberra ACT 2601

Phone: (02) 6244 1000

Email: phil.anderson@aihw.gov.au

Published by the Australian Institute of Health and Welfare

Please note that there is the potential for minor revisions of data in this report. Please check the online version at <www.aihw.gov.au> for any amendments.

Contents

Acknowledgments

The authors of this report were Andrew Powierski, Rosemary Karmel and Phil Anderson of the Data Linkage Unit at the Australian Institute of Health and Welfare (AIHW).

This report uses data from the National Health and Medical Research Council-funded Pathways in Aged Care (PIAC) study. Diane Gibson (University of Canberra) and Ann Peut (Ageing and Aged Care Unit, AIHW) designed the PIAC study. Stephen Duckett (University of Queensland) provided advice on research design, particularly in relation to maximising policy relevance. Yvonne Wells (La Trobe University) provided advice on the interpretation and use of the Aged Care Assessment Program (ACAP) National Minimum Data Set (NMDS). Rosemary Karmel was the principal developer of the linkage strategy and undertook the data linkage. Phil Anderson provided statistical advice on developing the linkage strategy. Andrew Powierski undertook the name-based data linkage.

The authors thank the Australian Department of Health and Ageing for permission to use data from the Aged and Community Care Management Information System and AIHW for permission to use the National Death Index for this study.

The report was funded by the AIHW.

Abbreviations

ACCMIS	Aged and Community Care Management Information System
ACAP	Aged Care Assessment Program
AIHW	Australian Institute of Health and Welfare
CACP	Community aged care packages
DoHA	Department of Health and Ageing
DOB	Date of birth
DOD	Date of death
EACH	Extended Aged Care at Home
EACHD	Extended Aged Care at Home Dementia
FMR	False match rate
NDI	National Death Index
NMDS	National Minimum Data Set
PIAC	Pathways in Aged Care
PPV	Positive predictive value
RAC	residential aged care
SLK	Statistical Linkage Key
TCP	Transition Care Program
YOB	Year of birth

Summary

In Australia, many community service program data collections developed over the last decade, including several for aged care programs, contain a common statistical linkage key (SLK-581) to enable derivation of client level data and to determine which clients use a number of programs. A direct comparison between an SLK-based linkage strategy and a name-based linkage strategy was carried out to gauge the quality of the SLK-based linkage. The scope of the comparison was limited to linkage for a single stage of the Pathways in Aged Care (PIAC) study which linked Aged and Community Care Management Information System (ACCMIS) data with National Death Index (NDI) data using an SLK-based linkage strategy. The purpose of this report is to examine the accuracy of the SLK-based strategy and utility of the resulting matched data, using the name-based strategy as the reference standard.

Methods

Both the ACCMIS and NDI data sets have full name information available as well as the data required for the keys used in the PIAC SLK-based data linkage. In the PIAC linkage, a detailed stepwise deterministic record linkage algorithm was developed to link data sets. The strategy used a general person identifier (SLK-581) in conjunction with additional data items (e.g. region and date of death). Measures of likely match accuracy were used to select match keys and ensure match quality.

Both a name-based linkage strategy and the PIAC SLK-based linkage strategies were applied to link the data sets. The name-based strategy was probabilistic and involved running a series of passes allowing for variation in name and demographic data and using clerical review to identify matches. Matches made under the two strategies were directly compared using the name-based strategy as the reference standard. The name-based strategy made 172,776 links and the SLK-based strategy made 170,928 links.

Results

Overall, the study confirms that the utility of the SLK-based linkage strategy (which used SLK-581 in conjunction with other common data items) is comparable to that of the name-based linkage strategy. More specifically, the study showed that:

- the SLK-based strategy was highly effective in identifying matches, with a positive predictive value (PPV) of 99.7% and a sensitivity of 98.5%.
- the name-based strategy was not infallible. A very small number of name-based matches were identified as false. Also, detailed comparisons showed that one-third of the small number of the matches made only by the SLK-based strategy, were identified as true matches after close clerical review.
- some minor improvements could be made to the stepwise SLK-based linkage process.
- the SLK-based linkage strategy resulted in linked data that largely reflected the name-based linkage strategy in terms of the distributions across key variables.

Furthermore, the use of the detailed stepwise SLK-based linkage process, which utilises additional common data items, was justified when compared with using a single-step SLK-581 linkage, identifying an extra 10% of all name-based links (sensitivity of 98.5% compared with 88.4%, respectively).

1. Introduction

Data linkage is a powerful tool both for identifying multiple appearances of individuals within a dataset and for integrating client information across datasets. It is increasingly being used to identify the use of a range of community services by individuals in order to obtain a person-based view of program use.

To enable the derivation of client-level data, many of the community service program data collections contain a common statistical linkage key (SLK) based on the concatenation of selected letters of name, date of birth (DOB) and sex. This is known as SLK-581. Its purpose is specifically to enable data linkage for statistical and research purposes, while protecting client privacy. SLK-581 consists of:

- the 2nd, 3rd and 5th letters of the family name.
- the 2nd and 3rd letters of the given name.
- DOB.
- Sex.

In 2005, a research team centred at the Australian Institute of Health and Welfare (AIHW) successfully applied for a National Health and Medical Council Strategic Award to undertake the Pathways in Aged Care (PIAC) cohort study. The purpose of the project was to link service use data for aged care programs to enable analysis of pathways through the aged care services to death. As not all of the program data sets contained name, but all contained the data required for a common SLK (SLK-581) this was used to link data for the PIAC cohort study (Karmel et al. 2010).

The linkage strategy for the PIAC project involved stepwise deterministic matching using SLK-581 in conjunction with other available data to allow for variation in reported SLK-581. Results from the SLK-based linkage suggested that the matches were of a high quality. However, without a comparison to a reference standard the accuracy of the linkage can not be directly gauged.

Name-based strategies are generally considered more accurate than SLK-based strategies. Therefore in 2010, a direct comparison between the PIAC SLK-based linkage strategy and a name-based linkage strategy was carried out to measure any differences between the two approaches, and to determine if the SLK-based linkage strategy leads to any biases in the match data set compared with name-based linkage. It is these comparative analyses that are the subject of this report. The scope of the comparison was limited to a single stage of the PIAC linkage (Stage 2 out of 7 steps) which involved linking Aged and Community Care Management Information System (ACCMIS) data with National Death Index (NDI) data. Stage 2 was used as both the ACCMIS and NDI data sets have full-name and other demographic information suitable for the SLK-based and name-based strategies.

This report compares the match results obtained using the name-based and SLK-based linkage strategies. In particular, it examines:

- the accuracy of the SLK-based strategy, using the name-based strategy as the reference standard
- ways to refine the original SLK linkage strategy
- factors causing contested, missed and false links made by the SLK linkage (when compared with the name-based linkage strategy)

- the utility of the SLK-matched data in analyses by comparing results derived using the name-based and SLK-based matched data.

1.1 Data

The ACCMIS and NDI data linked as part of the PIAC project were used in the comparative study. The PIAC project included people who used an ACCMIS program from 1 July 2002 to 30 June 2006 and deaths from 1 July 2003 to 31 December 2006. The data items used for linkage were based on:

- First name and surname of client.
- DOB.
- Sex.
- State of usual residence.
- Postcode of usual residence.
- Date of death (DOD).

This section introduces the two data sets used in the two linkage methods.

ACCMIS data

The Department of Health and Ageing's (DoHA) administrative aged care data is maintained on ACCMIS. This database contains information on client's use of the Residential Aged Care (RAC) program and the community aged care package programs, including Community Aged Care Packages (CACP), Extended Aged Care at Home (EACH) packages, Extended Aged Care at Home Dementia (EACHD) packages and the Transition Care Program (TCP).

The ACCMIS system is person-based. Identification of new and continuing clients in the two program groups is carried out by DoHA staff. Individual clients are identified via name and other demographic data, and given a distinct client identification number (client ID). Occasionally repeat program use by a client is not identified, resulting in multiple client IDs for an individual within a program (AIHW: Karmel 2005). ACCMIS identifies two types of clients:

1. RAC+: those using RAC or any of the small EACH, EACHD or TCP programs
2. CACP: those using CACP.

A person will have two client IDs if they used both a RAC+ program and a CACP program (cross-program use). These data were merged into a single record to represent a client in Stage 1 of the PIAC linkage process.

After accounting for cross-program use and client duplication, the estimated number of people that used ACCMIS programs between 1 July 2002 and 30 June 2006 was 415,057.

National Death Index data

The NDI is a database, housed at the AIHW, which contains records of all deaths occurring in Australia since 1980. The data are obtained from the Registrars of Births, Deaths and

Marriages in each state and territory. For each deceased person name and basic demographic data are stored on the NDI.

A person who dies is assigned a record with a unique mortality identifier. As a person can only die once, it would be expected that they should only have one record on the NDI. However, a person may have multiple records under the same mortality identifier when new or revised data on their death is added to the NDI. Therefore the data cleaning process included identifying and removing these duplicates. However, duplicates with substantially different data were retained to improve the matching process.

After removing the duplicates with similar data, the NDI data consisted of 470,121 records for people who died between 1 July 2003 and 31 December 2006.

For a full description of the structure of the two data sets and the scope and quality of the data used in the two linkage methods see Appendix 1.

1.2 Report structure

In covering the various phases of the comparison, the report is structured as follows. First, the two linkage methods being compared (name-based linkage strategy and the SLK-based linkage strategy) are described in detail in Section 2. Section 3 contains a comparison of the matches resulting from the SLK-based linkage and the name-based linkage strategies. In particular, there are investigations into the contested, missed and false links identified in the comparison. This is done primarily to identify ways of improving the SLK-based linkage strategy. In addition, direct and refined estimates of the sensitivity and positive predictive value (PPV) are reported. Finally, Section 4 shows the analytical effect of any differences in the linked data sets using a number of variables of interest.

2. Linkage Strategies

The purpose of the data matching was to link a single ACCMIS client record to a single NDI death record for the same person, that is create a 1 to 1 match. Two different types of data linkage methods were used in the comparison study: the PIAC SLK-based linkage strategy which used stepwise deterministic record linkage, and a name-based linkage strategy using probabilistic record linkage. Both of these strategies, and their results, are discussed below.

2.1 Types of data linkage

Data linkage is a powerful tool for identifying multiple appearances of individuals within a data set and for integrating client information across data sets. As the information recorded for an individual may vary from data set to data set — due to either differences in reporting (e.g. in first name) or errors — a robust linkage process should allow for some discrepancy in characteristics (Karmel et al. 2010). There are two main types of data linkage:

- Deterministic record linkage: The linkage of records is based on exact agreement of the linkage variables. Simple (one-step) deterministic record linkage cannot allow for variation in reporting. For example, linking two data sets based solely on the exact agreement of the SLK-581 only. However, deterministic linkage can be constructed to allow for variation in linkage elements: "An intricate deterministic algorithm can be as successful — or more successful — than probabilistic algorithms in identifying valid links" (Campbell 2005).
- Probabilistic record linkage: This allows for variation in reported characteristics by deriving a measure of similarity across variables used to identify matches, called the match weight. This is then used to decide whether a particular pair-wise comparison between records on two data sets is accepted (high weight) or rejected (low weight) as a match. Clerical review of possible record matches is often used to decide both the total weight above which record pairs are acceptable as a match and to determine whether matches with weights near this boundary should be considered to be valid.

2.2 SLK-based linkage

The strategy

SLK-based linkage was undertaken using stepwise deterministic match passes in conjunction with an algorithm for identifying suitable deterministic match keys and the order in which they should be used. The deterministic match keys were based on components of the SLK-581, postcode of usual residence and DOD. This approach was chosen because it does not rely on clerical review using full-name and/or address data — information which is not available on many of the data sets in the PIAC project. The method had four distinct steps.

Step 1: Defining the structure of the match keys

The match keys used for linkage were defined in terms of the SLK-581 (divided into five components), postcode and DOD. Different match keys were obtained by using different combinations of those components. More specifically, the components included for matching were:

- s3: the 2nd, 3rd and 5th letters of the family name.
- g2: the 2nd and 3rd letters of the given name.
- day and month of birth
- year of birth
- sex
- region (using state, postcode or first two digits of postcode)

In addition, two versions of keys using the above components were considered:

- those using DOD.
- those not using DOD.

Step 2: Incorporating cross-program use

As discussed in Section 1, a person in the ACCMIS system may have two client records. This may result in differing information for a person: RAC+ client information and CACP client information. Two versions of the SLK-581 were used to establish links: one version using the SLK-581 as reported for RAC+ clients and one version using the SLK-581 as reported for CACP clients. This allowed for name variation in reported data. Furthermore, two postcodes were used for people who had been in permanent RAC. To reduce the likelihood of false matches this additional information was only used for matching when using keys which used either all components of SLK-581, or when using DOD in conjunction with all SLK-581 information except for one component.

Step 3: Identification of keys to use in matching

There are many combinations of the components specified in Step 1 that could be used to define match keys. To ensure that any employed match keys were based on combinations which both discriminated well between individuals and would not introduce too many false matches, a key was identified as suitable for matching using three criteria:

1. *Discriminating power:* 97.5% of clients within each dataset had to have a unique value for the match key.
2. *Likelihood of introducing false matches:* The estimated theoretical false match rate (FMR) for links established using the match key could be no more than 0.5%.
3. *Trade-off between additional true and additional false matches:* The estimated theoretical trade-off between additional true and additional false matches made with the key given matches already identified had to be at least 2 to 1.

The first of these criteria limits the testing of suitability to those keys that distinguish between individuals with reasonably high probability, while the second and third criteria ensure that an employed key adds few false matches given any matches which have been made in earlier passes. Any combination of SLK-581 elements, region of usual residence and DOD that met all of the above criteria was used for matching the two data sets. Overall, 71

keys were identified for use in the linkage between ACCMIS and NDI. For a full description of the criteria used to select these keys see Karmel et al. 2010.

Step 4: Stepwise matching using selected match keys

Using the selected match keys, stepwise linkage was then carried out, with order of use determined by the discriminating power of the keys (going from high to low). Variation in match key elements identified when combining RAC+ and CACP clients was also incorporated into match steps where relevant (see Step 2). All links identified by the selected match keys were accepted as valid, with the exception of duplicate matches. In this case, a duplicate was selected at random.

Results

Overall, the SLK-based linkage strategy made a total of 170,928 links (Table 2.1). Three match keys made up just over three-quarters (77%) of all links:

- 44% of links matched on all linkage variables (SLK-581, postcode and DOD).
- 18% of links matched on all linkage variables, except postcode.
- 14% of links matched on all linkage variables, except DOD (remembering that this information was not available for all ACCMIS records).

More specifically, the match rates of the individual linkage components in linked records were as follows (Table 2.2):

- Name data (s3 and g2) matched in 166,759 cases (97.6%).
- DOB matched entirely in 157,381 cases (92.1%).
- Sex matched in 170,306 cases (99.6%).
- Postcode matched in 121,410 cases (71.0%).
- DOD matched in 120,892 cases (70.7%).

When available, DOD was an important matching variable. Approximately 133,000 linked records (78%) had DOD information available for linkage, that is, the person was reported as dying while using an ACCMIS program and had a DOD before 1 July 2006. Of those records, 91% (120,892, Table 2.1) matched exactly on the DOD; that is, 9 per cent of linked records with a DOD on both data sets had a different DOD on the ACCMIS and NDI data sets. Of the linked records that did not match on DOD, closer inspection of these showed that 92.1% of dates were less than 8 days different. Also, these records were more likely to be for community care (CACP) clients for whom DOD is less likely to be reliably reported by the service provider (17% CACP only compared with 10% of the ACCMIS data set).

The distribution of links across match elements was very similar for those that matched with or without the DOD. For example, of those that matched on DOD, 63% of links also matched on SLK-581 and postcode of usual residence, with a similar result for those matching records with different DOD. Similarly, of those that did not have DOD available, 64% matched on SLK-581 and postcode. The main difference was the higher proportion of matches (3.7%) made using 'other keys' and DOD, compared with under 0.5% when DOD did not match or was not available.

Table 2.1: Distribution of links using the SLK-based linkage strategy

DOD before 1 July 2006	DOD match status	Elements matched on	Number	Per cent	Overall per cent
Yes[(a)]	**Matched on DOD**	s3g2|dob|sex|pc	75,923	62.8	44.4
		s3g2|dob|sex|_	30,913	25.6	18.1
		s3g2|dob|_|pc	258	0.2	0.2
		s3g2|_|sex|pc	7,292	6.0	4.3
		_g2|dob|sex|pc	460	0.4	0.3
		s3_|dob|sex|pc	1,603	1.3	0.9
		Other keys	4,443	3.7	2.6
		Total	*120,892*	*100.0*	*70.7*
	Did not match on DOD	s3g2|dob|sex|pc	7,681	63.2	4.5
		s3g2|dob|sex|_	3,448	28.4	2.0
		s3g2|dob|_|pc	25	0.2	0.0
		s3g2|_|sex|pc	749	6.2	0.4
		_g2|dob|sex|pc	3	0.0	0.0
		s3_|dob|sex|pc	209	1.7	0.1
		Other keys	34	0.3	0.0
		Total	*12,149*	*100.0*	*7.1*
No	**Not applicable**	s3g2|dob|sex|pc	24,114	63.6	14.1
		s3g2|dob|sex|_	10,949	28.9	6.4
		s3g2|dob|_|pc	82	0.2	0.0
		s3g2|_|sex|pc	2,083	5.5	1.2
		_g2|dob|sex|pc	5	0.0	0.0
		s3_|dob|sex|pc	544	1.4	0.3
		Other keys	110	0.3	0.1
		Total	*37,887*	*100.0*	*22.2*
		Total number of links	**170,928**	**..**	**100.0**

(a) A person only had a DOD on the ACCMIS data if they died while in use of an ACCMIS program and had a DOD before 01 July 2006 otherwise they were recorded on the extract used for matching as not having a DOD.

Table 2.2: Distribution of links using the SLK-based linkage by individual linkage elements

DOD before 1 July 2006	DOD match status	Elements matched on	Number	Per cent	Overall per cent
Yes[(a)]	**Matched on DOD**	s3|g2	117,511	97.2	68.7
		s3|_	2,646	2.2	1.5
		_|g2	690	0.6	0.4
		_|_	45	0.0	0.0
		Full DOB	110,230	91.2	64.5
		Sex	120,465	99.7	70.5
		Postcode	85,892	71.1	50.3
		Total	*120,892*	..	*70.7*
	Did not match on DOD	s3|g2	11,929	98.2	7.0
		s3|_	215	1.8	0.1
		_|g2	5	0.0	0.0
		_|_	0	0.0	0.0
		Full DOB	11,389	93.7	6.7
		Sex	12,102	99.6	7.1
		Postcode	8,674	71.4	5.1
		1 day DOD difference	8,735	71.9	5.1
		2 - 7 days DOD difference	2,456	20.2	1.4
		8+ days DOD difference	958	7.9	0.6
		Total	*12,149*	..	*7.1*
No	**Not applicable**	s3|g2	37,319	98.5	21.8
		s3|_	558	1.5	0.3
		_|g2	10	0.0	0.0
		_|_	0	0.0	0.0
		Full DOB	35,762	94.4	20.9
		Sex	37,739	99.6	22.1
		Postcode	26,844	70.9	15.7
		Total	*37,887*	..	*22.2*
		Total number of links	**170,928**	**..**	**100.0**

(a) A person only had a DOD on the ACCMIS data if they died during the use of an ACCMIS program and had a DOD before 1 July 2006 otherwise they were recorded on the extract for matching as not having a DOD.

2.3 Name-based linkage strategy

A name-based linkage strategy was applied to examine the accuracy of the SLK-based linkage strategy. This was possible as both the ACCMIS and NDI data had full name information.

The strategy

The name-based linkage strategy used probabilistic matching. It involved running a series of passes allowing for variation in full name information and demographic data. Each pass consisted of matching on selected variables. Within each pass, a weight was calculated for each pair-wise match based on the similarity of match variables (high weight for very similar or exact data, low weight for quite different data). These weights were used during clerical review to identify matches where there was variation in reported match data. Finally, all accepted match pairs were output into a data set of accepted links. Unmatched records were then returned for matching in the next pass. Figure 2.1 provides a visual representation of the strategy applied to perform the name-based linkage. A more detailed explanation of the four steps used by the process follows.

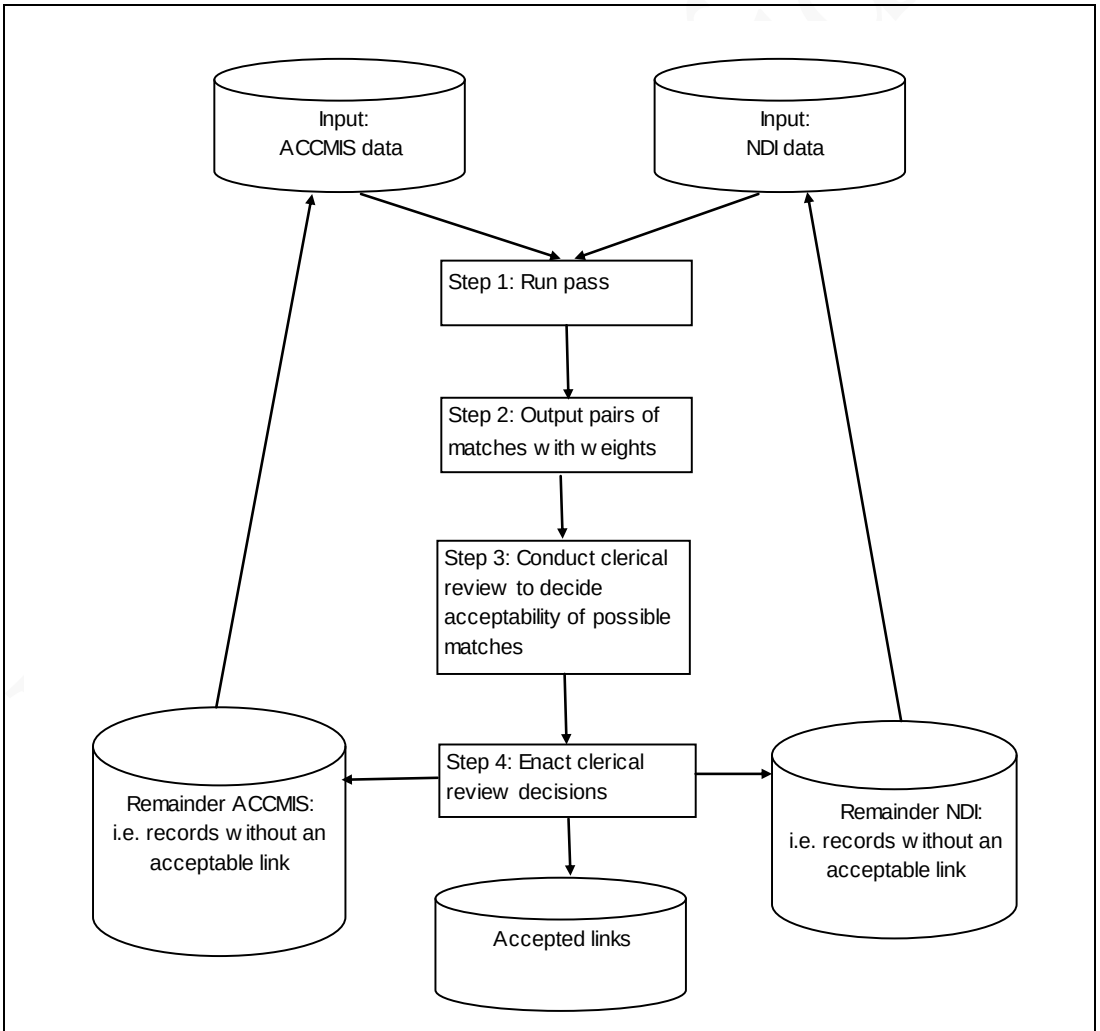


Figure 2.1: Flow diagram of the name-based linkage strategy

Step 1: Run pass

A pass matched the ACCMIS and NDI data sets based on particular blocking variables and match variables. A blocking variable is one that must match on both data sets before a pair of records can be considered for matching. For example, in Pass 2 (Table 2.3) only records that match exactly on the blocking variables surname, DOB and sex are considered any further in the pass. The match variables are used in the calculation of the weight (discussed in Step 2) to identify matches within a block. In the Pass 2 example, there is only one match variable (first name of client). In this case, first names are compared for all possible match pairs within a surname, sex and DOB block. An overview of the passes used in the name-based linkage strategy is provided in Table 2.3 .

Table 2.3 : Overview of passes run in the name-based linkage process

Pass	Blocking variables	Match variables
1	First name, surname, DOB and sex	Nil
2	Surname, DOB and sex	First name
3	First name, DOB and sex	Surname
4	DOB	First name, surname and sex
5	Day and year of birth	First name, surname, month of birth and sex
6	Month and year of birth	First name, surname, day of birth and sex
7	Day and month of birth	First name, surname, year of birth and sex
8	a-code[(a)] and year	First name, surname and sex
9	s-code[(b)] and year	First name, surname and sex
10	Year of birth	First name, surname, day and month of birth and sex
11	Day of birth	First name, surname, month and year of birth and sex
12	Month of birth	First name, surname, day and year of birth and sex
13	Nil	First name, surname, DOB and sex

(a) a-code checks for dates of birth that are expressed in the American form (for example, 01/30 as compared to 30/01).

(b) s-code represents another type of coding error in the DOB. That is, the second digit of the day has been swapped with the second digit of the month (for example, 28/04 on one record and 24/08 on another).

In addition to the 13 passes listed in Table 2.3 there were a further four sub-passes run within each pass. This was done to improve the efficiency of the data linkage process by reducing the number of match pairs into smaller segments for clerical review. The four sub-passes were based on the difference between the NDI DOD and the ACCMIS DOD (or last seen ACCMIS date):

(a) 0 days difference;

(b) 1 day difference;

(c) 2-7 days difference;

(d) More than 8 days difference (this includes those that did not have DOD on the ACCMIS extract).

Overall, 52 passes were run in the name-based linkage process.

Step 2: Output pairs of matches with weights

After running a pass we obtained all possible match pairs within a block and their weights. As stated earlier, the name-based linkage was a probabilistic linkage method. This involves the comparison of data fields on two data sets to obtain evidence on whether two records belong to the same person. This evidence is summarised in the form of a cumulative weight.

In the name-based linkage strategy each link was assigned a cumulative weight based on the names, sex and birth year.

- Name: The main contribution to the weight was based on the names. Two factors were used in determining a name weight:
    1. The first of these was based on the frequency of the name. For example, a match of 'John' to 'John' receives far less weight (6.5) than a match of 'Zybygniew' to 'Zybygniew' (20.0) because the high frequency of 'John' as a name means that a match on this name is much more likely to occur by chance.
    2. The second allowed for name similarities. The Jaro algorithm with improvements from McLaughlin and Winkler is used to determine how "close" the two names are (Jaro 1989; Porter & Winkler 1997). Names that were very similar received almost the same weight that would have been earne d had they been the same. For example, a match of 'John' to 'Jon' receives a weight of 6.1 as compared with an exact match of 'John' to 'John' which receives a weight of 6.5.
- Sex: As sex is not very discriminating in terms of identifying individuals, the weight for sex was small, +1 when the same and -1 when different, respectively.
- DOB: In passes where the year of birth was allowed to be different, a weight penalty of -1 applied for each year that the pair's DOB disagreed.

The above three match variable weights for a specific match pair were summed to form a cumulative weight to help decide whether a particular pair-wise comparison between records would be accepted (high weight) or rejected (low weight).

Step 3: Clerical review

Clerical review is the name given to the process that involves examining possible match pairs manually and deciding whether to accept or reject the match. Clerical review was conducted after each pass in the name-based linkage strategy. For link pairs for which it was

not clear if a link should be accepted or rejected a 'possible' was assigned to the match in the event that there may be a better match in a later pass.

In general, for clerical review, after each pass, the match pairs in the results file (i.e. a data set with all possible match pair for a pass) are ordered by descending weight (i.e. from highest quality matches to lowest quality). Therefore the majority of the matches that are accepted are in the top region of the file, and the majority of matches that are rejected are in the bottom region. There is also a region between these two where good and poor matches are mixed together.

The clerical review process for the current study involved manually examining the weights, names, sex, DOB, postcode and DOD to decide whether the link should be accepted, rejected or assigned as a possible link. Records that are a part of a possible link-pair are retained for matching in future passes in order to see if a better link exists. After running all 52 passes, any remaining possible match-pairs were either accepted or rejected.

Step 4: Enact clerical review decisions

After each pass was run, the clerical review decisions were assigned to the input ACCMIS and NDI data sets for running future passes. Depending on the clerical review decision made different actions were taken:

- A match pair that was accepted was added to a data set of accepted links. Furthermore, as the match pair had been linked and accepted we did not want to attempt to link these records in future passes. Therefore they were dropped from the input NDI and ACCMIS data sets for the next pass.
- If it was decided that the match should be rejected or assigned as a possible match, for later consideration, the records were retained on the input ACCMIS and NDI data sets for matching in future passes.

Results

Overall, the name-based linkage strategy made a total of 172,776 links (Table 2.4). This was 1,848 more links than made by the SLK-based linkage process. It was of no surprise that the name-based linkage process identified more links than the SLK-based process as the former had more name-based information for identification of links and probabilistic matching allowed for unidentified variation. Three match types made up almost three-quarters (72.1%) of all links:

- 42% of links matched on all linkage variables (first name, surname, DOB, postcode and DOD).
- 17% of links matched on all linkage variables, except postcode.
- 13% of links matched on all linkage variables, except DOD.

The distribution of these three match types was very similar to that obtained using the SLK-based linkage process.

Of the 172,776 name-based links (Table 2.5):

- first name and surname matched in 159,225 cases (92.2% compared with 97.6% of links matching on s3 and g2 under the SLK-based linkage).
- DOB matched in 157,887 cases (91.4% compared with 92.1% under the SLK-based linkage).

- sex matched in 172,170 cases (99.6%, same as under the SLK-based linkage).
- postcode matched in 121,592 cases (70.4% compared with 71.0% under the SLK-based linkage).
- when available, DOD matched in 120,745 cases (90.4% of matches with DOD available, compared with 90.9% under the SLK-based linkage).

The distribution of links across match types was very similar for those that matched on DOD, did not match on DOD although a date was available and did not have any DOD information to match by (Table 2.4). However, there was a higher percentage of links made without matching exactly on name data in the name-based linkage as opposed to the SLK-based linkage (compare Table 2.2 and Table 2.5). This result is to be expected as the name-based linkage can allow for more name variation than the SLK-based linkage.

Table 2.4: Distribution of links using the name-based linkage strategy

DOD before 1 July 2006	DOD matching	Elements matched on	Number	Per cent	Overall per cent
Yes[(a)]	**Matched on DOD**	surname|firstname|dob|sex|pc	72,259	59.8	41.8
		surname|firstname|dob|sex|_	29,376	24.3	17.0
		surname|firstname|dob|_|pc	238	0.2	0.1
		surname|firstname|_|sex|pc	6,701	5.5	3.9
		_|firstname|dob|sex|pc	1,360	1.1	0.8
		surname|_|dob|sex|pc	4,243	3.5	2.5
		other keys	6,568	5.4	3.8
		Total	*120,745*	*100.0*	*69.9*
	Did not match on DOD	surname|firstname|dob|sex|pc	7,255	56.8	4.2
		surname|firstname|dob|sex|_	3,215	25.2	1.9
		surname|firstname|dob|_|pc	22	0.2	0.0
		surname|firstname|_|sex|pc	731	5.7	0.4
		_|firstname|dob|sex|pc	176	1.4	0.1
		surname|_|dob|sex|pc	511	4.0	0.3
		other keys	852	6.7	0.5
		Total	*12,762*	*100.0*	*7.4*
No	**Not applicable**	surname|firstname|dob|sex|pc	22,903	58.3	13.3
		surname|firstname|dob|sex|_	10,268	26.1	5.9
		surname|firstname|dob|_|pc	69	0.2	0.0
		surname|firstname|_|sex|pc	1,937	4.9	1.1
		_|firstname|dob|sex|pc	467	1.2	0.3
		surname|_|dob|sex|pc	1,437	3.7	0.8
		other keys	2,188	5.6	1.3
		Total	*39,269*	*100.0*	*22.7*
		Total number of links	**172,776**	**..**	**100.0**

(a) A person only had a DOD on the ACCMIS data if they died during the use of an ACCMIS program and had a DOD before 1 July 2006 otherwise they were recorded on the extract for matching as not having a DOD.

Table 2.5: Distribution of links using the name-based linkage strategy by individual linkage elements

DOD before 1 July 2006	DOD matching	Elements matched on	Number	Per cent	Overall per cent
Yes[(a)]	**Matched on DOD**	surname|firstname	111,526	92.4	64.5
		surname|_	6,861	5.7	4.0
		_|firstname	2,098	1.7	1.2
		_|_	260	0.2	0.2
		Total	*120,745*	*100.0*	*69.9*
		Full DOB	110,219	91.3	63.8
		Sex	120,321	99.7	69.6
		Postcode	85,692	71.0	49.6
		Total	*120,745*	*..*	*69.9*
	Did not match on DOD	surname|firstname	11,589	90.8	6.7
		surname|_	848	6.6	0.5
		_|firstname	283	2.2	0.2
		_|_	42	0.3	0.0
		Total	*12,762*	*100.0*	*7.4*
		Full DOB	11,548	90.5	6.7
		Sex	12,718	99.7	7.4
		Postcode	8,817	69.1	5.1
		1 day DOD difference	9,219	72.2	5.3
		2 – 7 days DOD difference	2,585	20.3	1.5
		8+ days DOD difference	958	7.5	0.6
		Total	*12, 762*	*..*	*7.4*
No	**Not applicable**	surname|firstname	36,110	92.0	20.9
		surname|_	2,320	5.9	1.3
		_|firstname	759	1.9	0.4
		_|_	80	0.2	0.0
		Total	*39,269*	*100.0*	*22.7*
		Full DOB	36,120	92.0	20.9
		Sex	39,131	99.7	22.6
		Postcode	27,083	69.0	15.7
		Total	*39,269*	*..*	*22.7*
		Total number of links	**172,776**		**100.0**

(a) A person only had a DOD on the ACCMIS data if they died during the use of an ACCMIS program and had a DOD before 01 July 2006 otherwise they were recorded as not having a DOD.

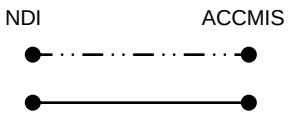
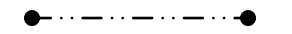
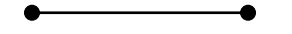
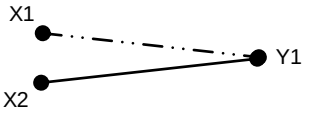
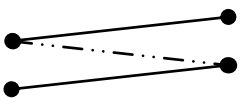
3. A comparison of two linkage strategies

The primary objective of this report is to compare the name-based and SLK-based linkage strategies. This section presents:

- a look at the type of link comparisons possible when comparing two linkage strategies
- results on the link comparisons obtained in this comparison study
- an investigation into the contested, false and missed links identified in the comparison
- estimates of the sensitivity and the PPV of the SLK linkage strategy.

3.1 Type of link comparisons

When comparing one linkage strategy with a reference linkage strategy we can easily identify whether a record has been linked and whether it should have been linked. However, in the current context it is also important to establish whether a record in one data set matches with the correct record in the other data set. That is, as well as knowing whether or not a record should have been linked, we need to know when two strategies result in a record in one data set being linked to the same or a different record in the second data set. Figure 3.1 illustrates the type of link comparisons possible when comparing the SLK-based and name-based links, namely: identical links, SLK-based link only, name-based link only, mixed links and multiple mixed links.

Type of link comparison	Description	Outcome
A – identical link	Same link under SLK-based and name-based linkage	
B – SLK-based link only	Link using SLK-based linkage only	
C – name-based link only	Link using name-based linkage only	
D – mixed links	NDI record links to different ACCMIS records under name-based and SLK-based linkage	
E – mixed links	ACCMIS record links to different NDI records under name-based and SLK-based linkage	
F – multiple mixed links	Combination of D or E links	

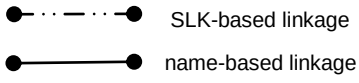


Figure 3.1: Types of link comparisons possible when comparing SLK-based and name-based links

Perspective

There are three types of perspectives that are used in this analysis to classify links:

- NDI perspective (One-way links to ACCMIS): describes links looking from the NDI record to the ACCMIS record.
- ACCMIS perspective (One-way links to NDI): describes links looking from the ACCMIS record to the NDI record.
- Two-way ACCMIS–NDI links: combines the links made looking from the NDI and ACCMIS perspectives.

Link type E in Figure 3.1 can be used to illustrate these perspectives. Link E is a mixed link that consists of two links: an SLK-based link in which a record on the NDI (X1) matches to a record on ACCMIS (Y1) and a name-based link in which a record on the NDI (X2) matches to a record on ACCMIS (Y1). The classification of these two links is dependent upon the perspective one is looking from. For example:

- from the NDI perspective, there are two links: an SLK-based link (like link type B) and a name-based link (like link type C).
- from the ACCMIS perspective, there is a single mixed link: that is, an ACCMIS record (Y1) links to different NDI records under SLK-based linkage (X1) and name-based linkage (X2).
- from the two-way ACCMIS–NDI links there is a single mixed link (link type E).

3.2 Comparing links

In Section 2 we saw that the two linkage strategies resulted in different numbers of matches:

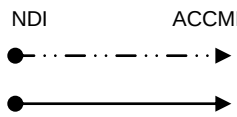
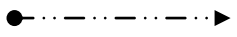
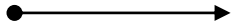
- the name-based linkage strategy gave 172,776 links
- the SLK-based linkage strategy gave 170,928 links.

Overall, the two linkage strategies resulted in 173,577 distinct link pairs.

NDI perspective

Each of the distinct link pairs was categorised based on looking from each NDI record to the corresponding ACCMIS record; that is, looking at ACCMIS links from the NDI perspective. These links are shown in Figure 3.2: and are classified as identical links (A), SLK-based links only (B), name-based links only (C) and mixed links (D).

Overall, 170,127 (98.0% of all distinct links) NDI records matched to the same ACCMIS record under the name-based and SLK-based linkage strategies. Under the name-based linkage there was a total of 2,498 (1.4%) matched NDI records that were not linked under the SLK-based linkage process (type C links). In contrast, there were a total of 650 links made under the SLK-based linkage process not made by the name-based process (type B links). Finally, the remainder of matches were cases in which NDI records linked to different ACCMIS records under name-based and SLK-based linkage (302, type D links).

Type of link comparison	Outcome	Description	SLK-based links	Name-based links	All links
A		Links to the same ACCMIS record under name-based and SLK-based linkage	170,127	170,127	170,127
B		Links using SLK-based linkage only	650	..	650
C		Links using name-based linkage only	..	2,498	2,498
D		NDI record links to different ACCMIS records under name-based and SLK-based linkage	151	151	302
		Total number of links	170,928	172,776	173,577

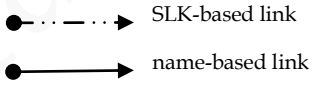
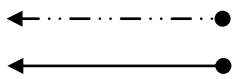
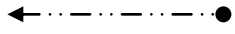


Figure 3.2: Links from NDI perspective: classifying SLK-based links and name-based links

ACCMIS perspective

In this case, each of the distinct link pairs were categorised based on looking from each ACCMIS record to the corresponding NDI record; that is, looking at links to the NDI from the ACCMIS perspective. These types of links are shown in Figure 3.3 and can again be classified as identical links (A), SLK-based links only (B), name-based links only (C) and mixed links (D).

As before, the total number of links to the same NDI record under name-based and SLK-based linkage was 170,127 (98.0%). Under the name-based linkage there were 2,608 (1.5%) ACCMIS records that linked to NDI records but which were not linked under the SLK-based linkage process (type C links). There were 760 links made under the SLK-based linkage process not made by the name-based process (type B links). Finally, the remainder of matches were cases in which ACCMIS records linked to different NDI records under name-based and SLK-based linkage (82, type E links).

Type of link comparison	Outcome	Description	SLK-based links	Name-based links	All links
A	NDI ACCMIS 	Links to the same NDI record under name-based and SLK-based linkage	170,127	170,127	170,127
B		Links using SLK-based linkage only	760	..	760
C		Links using name-based linkage only	..	2,608	2,608
E		ACCMIS record links to different NDI record under name-based and SLK-based linkage	41	41	82
		Total number of links	170,928	172,776	173,577

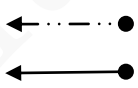
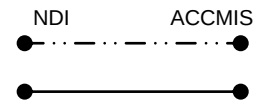
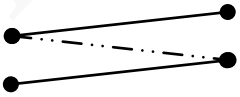


Figure 3.3: Links from ACCMIS perspective: classifying SLK-based links and name-based links

Two-way ACCMIS–NDI link comparison

After combining the results from the two one-way perspectives we obtain a two-way classification for each link. For example, looking at the SLK-based links, from the one-way ACCMIS perspective there were 760 SLK-based links only (type B link, Figure 3.3). These links are represented in the two-way comparison in Figure 3.4 by 613 type B SLK-based links only plus 147 SLK-based type D links. The two-way results are of primary interest to our comparison study as it is here we identify our false links and missed links for further analysis. As seen in Figure 3.2: and Figure 3.3, there were 170,127 links made under both the name-based linkage and the SLK-based linkage (type A link, Figure 3.4). The total number of links made by the SLK-based linkage process was only 613 (type B link), and 2,460 links were made by the name-based linkage process only (type C link). The remainder of links were mixed links or multiple mixed links, also known as contested links. Link types B to F are scrutinised in the following sections in order to identify the causes of these differences and to assess if there are any areas of improvement that can be made to the SLK-based linkage process.

Type of link comparison	Outcome	Description	SLK-based links	Name-based links	All links
A		Same link under name-based and SLK-based linkage	170,127	170,127	170,127
B		Links using SLK-based linkage only	613	..	613
C		Links using name-based linkage only	..	2,460	2,460
D		NDI record links to different ACCMIS record under name-based and SLK-based linkage	147	148	295
E		ACCMIS record links to difference NDI record under name-based and SLK-based linkage	37	38	75
F		Example multiple mixed links	4	3	7
		Total number of links	170,928	172,776	173,577

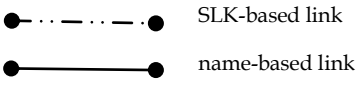


Figure 3.4: The two-way link comparison: classifying SLK-based links and name-based links

3.3 Analysis of SLK-based contested, missed and false links

When linking records four outcomes are possible: a true link, a true non-link, a false link (false positive) and a missed link (false negative). In the current analysis, the name-based linkage provided the reference standard, and so the status of the SLK-based links (that is, whether a link was a true link, a true non-link, a false link or a missed link) was determined by comparing the SLK-based links with the name-based links (Figure 3.5).

	Name-based links	**Name-based non-links**
Linked by SLK-based linkage (SLK links)	SLK true links	SLK false links
Not linked by SLK-based linkage (SLK non-links)	SLK missed links	SLK true non-links

Figure 3.5: Classification of SLK-based links when compared with name-based links

The reasons for the SLK-based linkage strategy missing some of the links identified by name-based linkage (SLK missed links), for making links that were not identified by name-based linkage (SLK false links), and for getting different links (contested links) are of interest for two reasons:

1. analysis may identify ways to improve the SLK-based linkage strategy
2. any patterns in the missed and false links could indicate whether there are likely to be biases in the SLK-based linked data set.

A range of investigations into these issues were carried out, and the results are summarised below.

Contested (mixed) links

Contested, or mixed, links result when two strategies result in a record in one data set being linked to different records in the second data set. For example, an ACCMIS record links to different NDI records under name-based and SLK-based linkage.

ACCMIS perspective

There were 37 link pairs where an ACCMIS record linked to a different NDI record under name-based and SLK-based linkage (type E link from Figure 3.1). Clerical review was conducted on these mixed links and the resulting decisions are presented in Figure 3.6. The data used for clerical review included first name, surname, sex, DOB, DOD and postcode to assist in determining if each link should be accepted or rejected.

After clerical review, the most common outcome was that the name-based link was accepted and the SLK-based link was rejected. This occurred for 25 (67%, Figure 3.6) pairs. For 8 link pairs the name-based link was rejected and the SLK-based link accepted. This provides evidence that the name-based linkage process is itself not 100% accurate. Furthermore, there were 3 pairs in which both the SLK-based links and the name-based links were true — these related to duplicates on the NDI. Finally, there was a single pair in which both the name-based link and the SLK-based link were rejected showing again that clerical review is a subjective process.

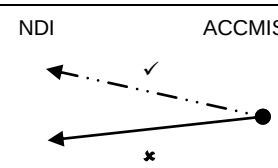
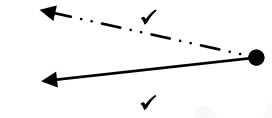
Type of mixed link	ACCMIS record links to different NDI records under name-based and SLK linkage	After clerical review	Number of pairs
A		SLK-based link – yes name-based link – no	8
B		SLK-based link – no name-based link – yes	25
C		SLK-based link – no name-based link – no	1
D		SLK-based link – yes name-based link – yes	3
		Total number of links	37

Figure 3.6: Clerical review decisions for ACCMIS records linking to different NDI records under name-based and SLK-based linkage

NDI perspective

There were 147 cases where an NDI record linked to a different ACCMIS record under our two linkage processes (type D link in Figure 3.1). Clerical review was conducted on each of these 147 mixed links and the results are presented in Figure 3.7.

The analysis of mixed links from the NDI perspective identified data quality issues. As discussed in Section 1, creating the ACCMIS data for linkage involved identifying cross-program use and removing duplicate client records to consolidate multiple records for a client into one record using an SLK linkage strategy. However, variation in name information, DOB and postcode data meant that this process was imperfect. This became apparent as for 142 of the 147 cases in which the NDI record linked to different ACCMIS

records clerical review indicated that both of the links were most likely correct (type D link, Figure 3.7). Closer inspection of the ACCMIS records revealed that the alternative links were generally duplicates for people who had used both RAC+ and CACP programs but whose data were not able to be combined in the earlier SLK-based linkage process for the PIAC project. For example, the name-based linkage may have linked the NDI record to the RAC+ client while the SLK-based linkage matched the record to the CACP client, with there being slightly different SLK-581 information between the two cases.

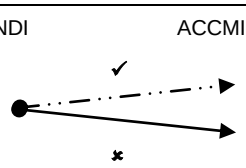
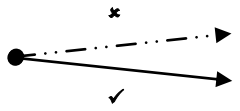
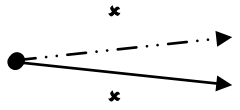
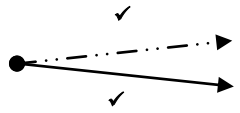
Type of mixed link	NDI record links to different ACCMIS records under name-based and SLK linkage	After clerical review	Number of pairs
A		SLK-based link – yes name-based link – no	0
B		SLK-based link – no name-based link – yes	5
C		SLK-based link – no name-based link – no	0
D		SLK-based link – yes name-based link – yes	142
		Total number of links	147

Figure 3.7: Clerical review decisions for NDI records linking to different ACCMIS records under name-based and SLK-based linkage

SLK missed links

Overall, there were 2,460 links made via the name-based linkage but not by the SLK-based strategy (SLK missed links, Figure 3.4). Name-based only links were examined to determine the reason(s) that matches were missed by the SLK-based strategy, and so to identify ways to improve the SLK-based linkage strategy by identifying additional keys that could have been used in the original SLK-based linkage process without introducing too many false matches.

Identifying additional keys

The match status (matched/different) of all components used in the SLK-based linkage strategy was established for SLK missed links; that is, whether the records in the missed links matched on name components, DOB components, sex, postcode or DOD. Combinations of elements that were common to many missed links were used to identify potential keys that could be used to improve the SLK-based linkage strategy.

As discussed in Section 2, the SLK-based linkage process used three criteria to determine, theoretically, if a key discriminated well between individuals and would not introduce too many false links.

1. *Discriminating power:* 97.5% of clients within each data set had to have a unique value for the match key.
2. *Likelihood of introducing false matches:* The estimated theoretical false match rate (FMR) for links established using the match key could be no more than 0.5%.
3. *Trade-off between additional true and additional false matches:* The estimated theoretical trade-off between additional true and additional false matches made with the key given matches already identified had to be at least 2 to 1.

Keys that met criteria 1 (discriminating power) were investigated as potential keys that could be added to the SLK-based linkage process. To check whether a key identified in this way could be used without adding too many false links, the possible additional keys were applied to unlinked ACCMIS and NDI records (unlinked under the SLK-based strategy) and the results compared with the name-based linkage.

Generally, the keys that picked-up name-based links only (SLK missed links) were well spread with most only adding a few matches. Overall, from the comparison around 13 keys were identified as possible keys for inclusion, and these were investigated further (Table 3.1 and Table 3.2).

Possible additional keys to use

Six keys could be used without adding too many false matches (Table 3.1). These divided into two groups:

- keys that split day and month for DOB and DOD. This was not done in the original SLK linkage process. Using these keys would have added 246 matches (including 217 good, Table 3.1), or
- keys that allowed for surname spelling differences (also, not done in the SLK linkage process). The name based component of the SLK includes the 2nd, 3rd and 5th letters of the family name (s3). There were a number of links made under the name-based linkage only that matched given name (g2) and had minor spelling differences in the surname. Therefore it was decided to investigate whether we should split up the surname; that is, rather than match on all three elements of the surname, allow only two out of three letters to match. Allowing for this variation would have added 145 matches, few of which were false (142 good, Table 3.1).

Introducing the additional keys presented in Table 3.1 would have added a total of 391 extra links to the SLK-based linkage process, of which 359 were true matches (92%).

Table 3.1: Possible additional keys to use to improve the SLK-based linkage strategy

Possible additional keys	Number of links made in the name-based approach	Number of links made by including the key in the SLK-based approach	False matches[(a)]	Trade-off true: false
Split day and month of birth / death				
s3g2|dm_ob|s|pc2|_mYOD (key new_1)	63	64	1	63.0
s3g2|_mYOB|s|pc2|_mYOD (key new_2)	83	84	1	83.0
s3g2|_mYOB|s|_|_mYOD (key new_3)	71	98	27	2.6
Total	**217**	**246**	**29**	
Surname spelling differences				
s2g2|dmYOB|s|pc (s2 = 2,3) (key new_4)	121	123	2	60.5
s2g2|dmYOB|s|pc (s2 = 2,5) (key new_5)	11	11	0	undefined
s2g2|dmYOB|s|pc (s2 = 3,5) (key new_6)	10	11	1	10.0
Total	**142**	**145**	**3**	

(a) The name-based linkage strategy was used as the reference standard.

Keys that would incorporate too many false links

The remaining seven potential keys would have added too many false links (Table 3.2). These included three keys that were dropped in the original SLK-based linkage process (key new_11 for low trade-off, key new_12 for high FMR and key new_13 considered as too coarse to be considered at all). Overall, using these keys would have added 8,600 false matches to get 970 (possibly) good matches (Table 3.2). The exclusion of key new_11, key new_12 and key new_13 is confirmed as being appropriate.

Two potential keys (key new_7 and key new_9) met criteria 1 and 2 for inclusion in the matching. However, in practice the keys had trade-offs between 1 and 2 and so failed on this criterion. Together they would have added 273 good matches and 178 false matches (Table 3.2). Their inclusion would have had little impact on the overall quality of the matching.

Table 3.2: Additional keys that would introduce too many false links

Additional keys that would introduce too many false links	Number of links made in the name-based approach	Number of links made by including the key in the SLK-based approach	False matches[(a)]	Trade-off true: false
s3g2|_mYOB|s|pc2 (key new_7)	198	329	131	1.5
s3g2|_mYOB|s (key new_8)	200	3,563	3,363	0.1
s3g2|d_YOB|s|pc2 (key new_9)	75	122	47	1.6
s3g2|d_YOB|s (key new_10)	99	1,189	1,090	0.1
_g2|dmYOB|s|pc (key new_11)	70	139	69	1.0
s3g2|dm_ob|s|pc2 (key new_12)	102	286	184	0.6
s3g2|dm_ob|s (key new_13)	152	3,892	3,740	0.0
Total	**967**	**9618**	**8,651**	**0.1**

(a) The name-based linkage strategy was used as the reference standard.

Data quality – missing postcode on the NDI

For the 2,460 SLK missed links, data quality was a contributing factor. The NDI does not contain postcodes for all states and territories; in particular, postcode is generally missing for deaths in Tasmania and Western Australia, and also in some cases for the Australian Capital Territory. However, other address information is available, and, while it could not be used in the SLK-based strategy, it was available for the name-based clerical review.

Table 3.3: SLK missed links by state and territory

State and territory	Number of SLK missed links	Per cent of SLK missed links	Per cent of all name-based links [(a)]
New South Wales	708	28.8	35.4
Victoria	611	24.8	24.7
Queensland	213	8.7	17.6
Western Australia	509	20.7	8.1
South Australia	200	8.1	10.0
Tasmania	122	5.0	2.8
Australian Capital Territory	49	2.0	1.1
Northern Territory	48	2.0	0.3
Total	**2,460**	**100.0**	**172,776**

(a) The name-based linkage strategy was the reference standard in which all links are assumed to be correct.

Of the 2,460 missed SLK links, 758 (30.8%) had no postcode information on the NDI. In addition, Western Australia, Tasmania and the Australian Capital Territory all accounted for a higher per cent of SLK missed links, compared with all links made in the name-based linkage. For example, of the SLK missed links, over 20% were from Western Australia compared with only 8% of all name-based links coming from Western Australia (Table 3.3). Furthermore, 5% of SLK missed links were from Tasmania relative to 3% of all name-based links. These results lend support to there being a small regional bias in the SLK-based links.

SLK false links

Using name-based matches as the reference standard, SLK only matches represent false matches made by the strategy. Excluding mixed matches, there were 613 SLK only matches.

During the investigation of the contested links it was established that a small proportion of name-based links were false links. Similarly, this investigation considers that some SLK only links may in fact be true. Clerical review was again used to examine these links.

After clerical review of a sample of 180 SLK only matches approximately 35% (standard error of 3 percentage points) of links from the sample were most likely true links. This equates to an estimated 215 out of 613 SLK only matches and shows that not all SLK only links were 'false links'. The main reason why these links were not picked-up by the name-based linkage process was name inconsistencies between the ACCMIS and NDI data sets.

Excluding the estimated 35% true links leaves an estimated 400 SLK only links that were false. These false links can be attributed to cases in which people had the same SLK name information, but entirely different names. This is possible since the SLK extracts the second, third and fifth letter of the surname and the second and third letter of the first name, only.

For example, 'Coral Lindsay' and 'Dorothy Windsor' have the same SLK information (s3 = 'INS' and g2 = 'OR'). The sample of 180 SLK only matches was examined to identify the rate at which this event happened. Of our sample, 18% (standard error of 5 percentage points) had the same SLK name information for a link but very different names. For the remainder of SLK only links, there were differences in various subsets of match variables but no consistent pattern explaining why these links were not picked-up by the name-based linkage strategy.

3.4 Overall link quality

There are two key measures commonly used when comparing matches. Using terminology originating in epidemiology and clinical trials these are:

- Sensitivity: the percentage of all true links that are identified by the SLK linkage strategy
  = SLK true links/(SLK true links + SLK missed links)
  = SLK true links/all true links
- Positive predictive value (PPV): the percentage of SLK links that are true links
  = SLK true links/(SLK true links + SLK false links)
  = SLK true links/SLK links

At the outset of this analysis the name-based linkage strategy was defined as the reference standard to identify true and false SLK matches and contested matches. Upon closer inspection of the contested links it became clear that a small number of the links made using the name-based linkage process were in fact false links and the strategy missed some true matches. In particular, of the 37 pairs where an ACCMIS record linked to different NDI records under name-based and SLK-based linkage the name-based link was wrong in eight cases and the SLK-based linkage was correct. Furthermore, after inspecting the SLK only links we estimated that approximately 35% of the 613 SLK only links were in fact true links.

Therefore two sets of estimates for sensitivity and PPV are presented below: one that treats the name-based linkage as a direct reference standard and a second set of refined estimates that utilises all information and re-classifies links as applicable, based on the investigations into contested, 'missed' and 'false' links.

Direct estimates

Overall, using the name-based linkage strategy as our reference, 172,776 SLK-based links were true links (Table 3.4). This includes 170,127 links made by both linkage strategies. The direct estimate of the sensitivity for the SLK-based linkage strategy using the name-based linkage as a reference is therefore 98.5% (Table 3.4); that is, 98.5% of all name-based links were identified using the SLK-based linkage strategy. The corresponding direct estimate for the PPV is 99.5%; that is, 99.5% of all SLK links were true. The high estimates for PPV and sensitivity for the SLK linkage strategy lend support to the utility of the SLK-based linkage strategy.

Table 3.4: Direct estimates of the PPV and sensitivity of the SLK-based linkage strategy, using name-based linkage as the reference standard

Match strategy	True links (A)	Additional links (B)	Missed links (C)	Total links (D = A + B)	PPV (A/D)	Sensitivity (A/F)
	Number				Per cent	
Name-based linkage	172,776 **(F)**		..	..	..	
SLK-based linkage	170,127	801[(a)]	2,649	170,928	99.5	98.5
SLK-581 linkage[(b)]	152,783	245	19,993	153,028	99.8	88.4

(a) Traditionally, the Additional links (B) would be defined as false links. However, it has been shown that some of these links are in fact true, the count of additional links includes contested links.

(b) Single-step deterministic matching using SLK-581.

Results for matches made with a single-step deterministic linkage using SLK-581 are presented in Table 3.4; that is, linking records that match on the SLK-581 only. This is of interest as SLK-581 is collected across a number of community service data sets specifically for statistical data linkage. A total of 152,783 links were made using SLK-581 (Table 3.4). The PPV for the SLK-581 linkage was slightly higher when compared with the stepwise SLK-based linkage process (99.8% compared with 99.5%). This result is of no surprise as the stepwise SLK-based linkage allows for variation in the SLK-581 and other demographic variables and hence is more likely to introduce false matches. However, allowing for this variation enables the identification of a significant number of valid links. Accordingly, the sensitivity of the SLK-based linkage process was 10% more than the SLK-581 linkage (98.5% as compared with 88.4%, respectively). That is, by allowing for variation in the linkage process the stepwise SLK-based method obtained an extra 10% of all name-based links.

Refined estimates

After re-classifying links based on the clerical review of contested, 'missed' and 'false' SLK links, there was a total of 173,139 true links between ACCMIS and NDI data (F, Table 3.5). This includes 35% (215) of the SLK only links and 156 contested links identified as true links; that is an extra 371 SLK-based links identified as true links (Table 3.5). Ten name-based links identified as false have also been excluded. After refining the estimate, the PPV rose from 99.5% (direct estimate) to 99.7% (refined estimate). That is, after incorporating the clerical review decisions, 99.7% of all SLK links are identified as true. The sensitivity based on the refined estimate increased by 0.01% when compared with the direct estimate (98.5%); however, the two sensitivity estimates are the same to one decimal place.

Table 3.5: Refined estimates of the PPV and sensitivity of the SLK-based linkage strategy

Match strategy	Same links (A)	Additional links		True links (D= A + B)	Missed links	Total links (E = A + B + C)	PPV (D/E)	Sensitivity (D/F)
				Number			Per cent	
		True (B)	False (C)					
Combined	..	..	..	**(F)**173,139	..	..	..	..
Name-based linkage	170,127	2,639	10	172,766	215	172,776	100.0	99.8
SLK-based linkage	170,127	371	430	170,498	2,460	170,928	99.7	98.5

3.5 Summary

The study involved comparing 172,776 name-based links with 170,928 SLK based links. Overall, there were 173,577 distinct link pairs. The comparison involved using the name-based linkage strategy as the reference standard.

There were:

- 170,127 (98.0% of all distinct links pairs) identical links made using the name-based linkage strategy and the SLK-based linkage strategy.
- 613 (0.4%) SLK only links.
    - After performing clerical review on a sample of the false links, it was estimated that 35% of these links were most likely true; that is, SLK only links are not false links in all cases.
    - It was estimated that approximately 18% of false links had the same SLK name information for a link but very different names.
- 2,460 (1.4%) name-based only links.
    - The investigation into the missed links allowed identification of possible additional keys that could be used in the SLK linkage process to improve the SLK linkage results.
    - Six additional keys were identified that would improve the SLK-based linkage process. These used finer decomposition of linkage data than that used for the PIAC linkage. In total, these six keys would have added a further 391 SLK links (359 true links or 15% of missed links). Other potential keys would have added too many false matches.
    - Missing region data contributed to missed links in the SLK-based linkage process. The NDI does not contain postcodes for all states. Of the 2,460 name-based only links, 758 (31%) had no postcode information.

An examination into the contested links provided evidence indicating that because of the subjective nature of the name-based linkage process, it is not 100% accurate. Allowing for this, the SLK-based strategy had a PPV of 99.7% and a sensitivity of 98.5% (compared with 99.5% and 98.5%, respectively, when using the name-based strategy directly as the reference standard).

The use of the stepwise SLK-based linkage process was justified when compared with using a single step SLK-581 linkage, with the former method identifying an extra 10% of the name-based links.

4. Analysis comparisons

In general, when undertaking analyses of linked data there is no reference standard for the links and so it is not possible to identify the false matches. As a result, all identified matches, both true and false, are used. Consequently, knowledge of biases in the complete match set is important. The preceding analysis of PPV, sensitivity and contested matches indicate that there could be some small biases in the SLK-based linkage when compared with the name-based linkage. Whether these differences will affect analyses in practice is examined below.

The distribution of linked records from the name-based linkage and the SLK-based linkage strategies were examined and compared for a range of variables. While there are standard goodness-of-fit tests for comparing distributions, there are two limitations in using them in this project. First, as sample sizes increase, the closer the distributions need to be in order not to be statistically significantly different. While this is to be expected (and desirable from the point of view of statistical tests), for large sample sizes many tests (such as the Chi-squared test) are, in practice, always statistically significant. Second, when comparing match sets the two distributions being compared are based on highly overlapping datasets, bringing the validity of standard tests into question.

In the current context, we are more interested in whether any differences in linkages make any practical difference in terms of interpretation of analyses and application of results. Therefore results from the two linkage methods were compared to simply gauge whether there were any practical differences between them. It should also be noted that the 'true' distributions and rates are between those reported below using the name-based linkage and the SLK-based linkage as it has been shown that not all name-based links are true and not all SLK only links are false.

4.1 Age and sex

Table 4.1 compares the age and sex profile of links made using the name-based linkage and the SLK-based linkage strategies. Overall, the distributions were very similar for the two linkage strategies. For example, men accounted for the same proportion of links in the two linkage strategies (37.6%). However, there was some very minor variation by age, with the proportions of matched records in the older age groups for women being affected (differences of one tenth of a percentage point).

When linking to deaths, death rates in various population groups are of interest. Of the 415,057 ACCMIS records, the name-based linkage identified a total of 172,766 deaths (Table 4.2). That is, 416 deaths per 1,000 ACCMIS clients. As expected from its 98.5% sensitivity, under the SLK-based linkage strategy the estimated death rate per 1,000 ACCMIS clients was slightly less (412 deaths per 1,000 ACCMIS clients).

Table 4.1: Distribution of matches, by age and sex, by linkage strategy (per cent)

		Name-based linkage	SLK-based linkage
Sex[(a)]	**Age**[(a)]		
Males	0–39	0.0	0.0
	40–49	0.1	0.1
	50–64	1.1	1.1
	65–69	1.2	1.2
	70–79	7.9	7.9
	80–89	18.1	18.1
	90–99	8.9	8.9
	≥100	0.3	0.3
	Total	*37.6*	*37.6*
Females	0–39	0.0	0.0
	40–49	0.1	0.1
	50–64	0.8	0.8
	65–69	0.9	0.9
	70–79	7.3	7.4
	80–89	27.8	27.9
	90–99	23.8	23.7
	≥100	1.6	1.6
	Total	*62.4*	*62.4*
	Total	**100.0**	**100.0**
Total (number)		**172,776**	**170,928**

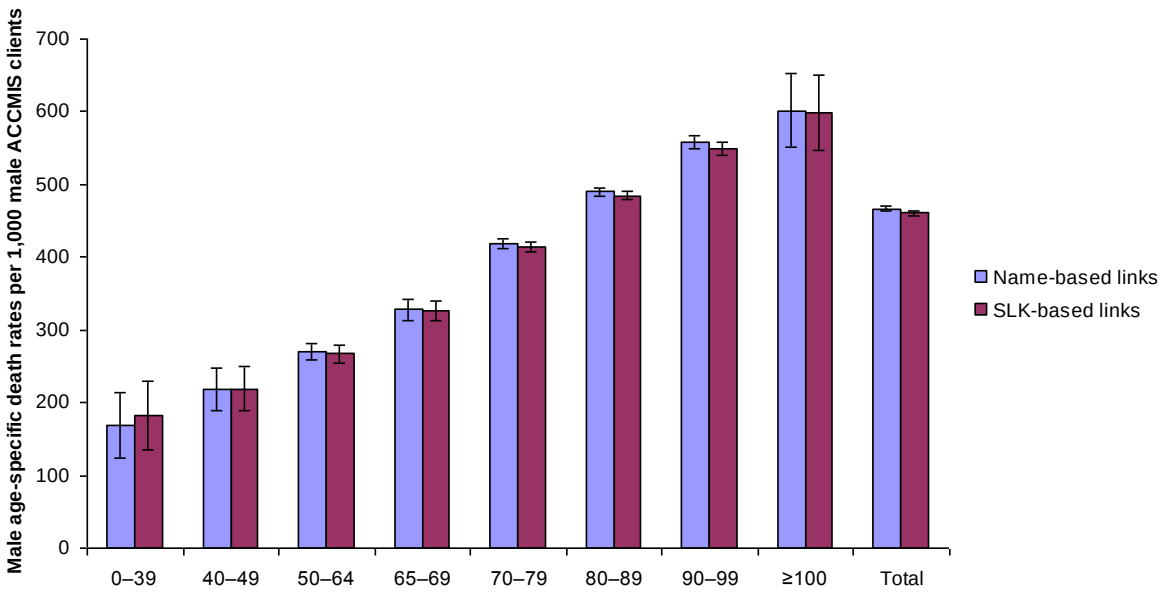
(a) From ACCMIS data.

Furthermore, if we compare death rates across age groups the death rates generally increase with age (Figure 4.1 and Figure 4.2). That is, we see that this conclusion is the same under both linkage strategies. More specifically, if we compare the 95% confidence intervals of the male death rates from the name-based estimates for those aged 70 to 79 with those aged 80 to 89 the confidence intervals do not overlap. This suggests that the death rates between these groups are statistically different. If we compare these groups using the SLK-based estimates the same conclusion is reached. The only group that suggested different results under the two strategies were females aged 0 to 39. Unlike the name-based results, the SLK-based results suggested that this group had a higher death rate than the 40 to 49 age group. However, the females 0 to 39 is a very small group (283) and hence the estimated underlying death rate has high variability under both linkage strategies, as illustrated by the wide 95% confidence intervals in Figure 4.2. The females 40 to 49 group is also quite small (885). Consequently, the highly overlapping confidence intervals for the death rates for the two age groups, for women under both linkage strategies again lead to the identical conclusion: the differences between death rates of the two age groups are not statistically significant — that is, the underlying death rates are the same.

Table 4.2: Age and sex specific death rates per 1,000 ACCMIS clients, by linkage strategy

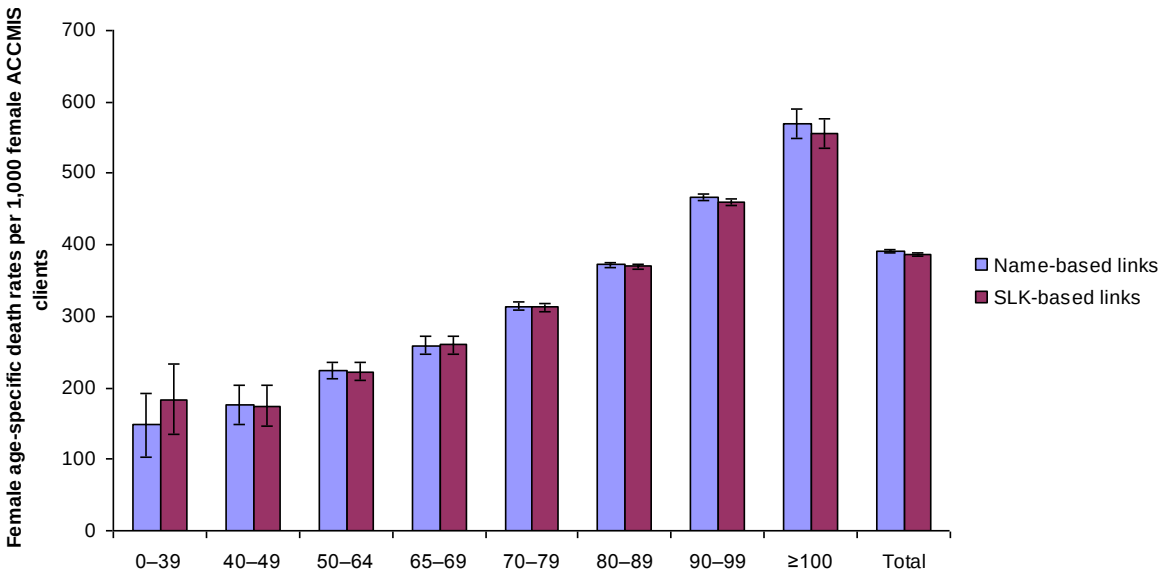
			Name-based linkage		SLK-based linkage	
Sex[(a)]	**Age**[(a)]	**ACCMIS category specific population**	**Number of links**	**Death rate per 1,000 ACCMIS clients**	**Number of links**	**Death rates per 1,000 ACCMIS clients**
Males	0–39	307	52	169	56	182
	40–49	945	206	218	207	219
	50–64	6,941	1,877	270	1,853	267
	65–69	6,368	2,091	328	2,078	326
	70–79	32,466	13,561	418	13,447	414
	80–89	63,766	31,333	491	30,935	485
	90–99	27,495	15,333	558	15,125	550
	≥100	875	527	602	524	599
	Total	*139,163*	*64,980*	*467*	*64,225*	*461*
Females	0–39	283	42	148	52	184
	40–49	885	157	177	155	175
	50–64	6,275	1,405	224	1,398	223
	65–69	6,054	1,569	259	1,573	260
	70–79	40,252	12,635	314	12,590	313
	80–89	129,126	48,129	373	47,753	370
	90–99	88,099	41,056	466	40,452	459
	≥100	4,920	2,803	570	2,730	555
	Total	*275,894*	*107,796*	*391*	*106,703*	*387*
Total (number)		**415,057**	**172,776**	**416**	**170,928**	**412**

(a) From ACCMIS data.



Source: Table 4.2

Figure 4.1: Male age-specific death rates per 1,000 men by linkage strategy, with 95% confidence intervals



Source: Table 4.2

Figure 4.2: Female age-specific death rates per 1,000 women by linkage, strategy with 95% confidence intervals

4.2 Cause of death

In the analysis of aged care data, cause of death may also be of interest and it is important to know if there would be biases in these estimates. Overall, the distributions of causes of death are the same when comparing the two linkage strategies to one decimal place (Table 4.3). Similarly, there is very little variation in the death rates between the two linkage strategies. These results provide confidence that there are no distributional biases between the two linkage strategies in terms of cause of death.

Table 4.3: Main cause of death by linkage strategy

	Name-based linkage			SLK-based linkage		
Cause of death[(a)]	**Deaths**	**Per cent**	**Death rates per 1,000 ACCMIS clients**	**Deaths**	**Per cent**	**Death rates per 1,000 ACCMIS clients**
Certain infectious and parasitic diseases (A00-B99)	2,211	1.3	5	2,166	1.3	5
Neoplasms (C00-D48)	23,990	13.9	58	23,763	13.9	57
Diseases of the blood and blood-forming organs (D50-D89)	605	0.4	1	599	0.4	1
Endocrine, nutritional and metabolic diseases (E00-E90)	7,750	4.5	19	7,637	4.5	18
Mental and behavioural disorders (F04-F05, F052-F99) without dementia	776	0.4	2	764	0.4	2
Diseases of the nervous system (G00-G29, G32-G99) without dementia	5,343	3.1	13	5,291	3.1	13
Dementia (F00-F03, F051, G30, G31)	17,082	9.9	41	16,883	9.9	41
Diseases of the eye and adnexa (H00-H95)	19	0.0	0	19	0.0	0
Disease of the circulatory system (I00-I99)	76,199	44.2	184	75,463	44.2	182
Diseases of the respiratory system (J00-J99)	19,845	11.5	48	19,636	11.5	47
Diseases of the digestive system (K00-K93)	5,332	3.1	13	5,263	3.1	13
Diseases of the skin and subcutaneous tissue (L00-L99)	714	0.4	2	704	0.4	2
Diseases of the musculoskeletal system and connective tissue (M00-M99)	1,836	1.1	4	1,810	1.1	4
Diseases of the genitourinary system (N00-N99)	5,747	3.3	14	5,677	3.3	14
Certain conditions originating in the perinatel period (P00-P96)	2	0.0	0	2	0.0	0
Congenital malformations, deformations and chromosomal abnormalities (Q00-Q99)	226	0.1	1	225	0.1	1
Symptoms, signs and abnormal clinical and laboratory findings, not elsewhere classified (R00-R99)	857	0.5	2	843	0.5	2
External causes of morbidity and mortality (V01-Y98)	3,975	2.3	10	3,914	2.3	9
Missing	267	0.2	1	269	0.2	1
Total	**172,776**	**100.0**	**416**	**170,928**	**100.0**	**412**
ACCMIS population	**415,057**	**..**	**..**	**415,057**	**..**	**..**

(a) From NDI data. The causes of death have been categorised based on the International Classification of Diseases 10th Revision (ICD-10).

4.3 State and territory at death

The importance of postcode data in the SLK-based linkage strategy in conjunction with systematic missing data for this variable in the NDI data suggests that this could be a source of bias in the SLK-based linked data. Table 4.4 compares the percentage of links made using the two linkage strategies by state and territory at death. Overall, the distributions by linkage strategy were very similar. However, there were some differences.

- Western Australia had marginally more links relatively under the name-based linkage strategy when compared with the SLK-based linkage strategy (8.1% compared with 7.9%, respectively, Table 4.4); that is, the SLK-based linkage strategy made relatively fewer links than the name-based strategy. This can be explained primarily by the fact Western Australia does not have postcode recorded on the NDI; however, it has additional address information that could be used in name-based clerical review that was not available during the SLK linkage process.
- Despite Tasmania not having postcode information on the NDI there was no difference in the percentages of links for Tasmania in the two strategies (2.8%, Table 4.4).
- Queensland had slightly more links relatively under the SLK-based linkage strategy as opposed to the name-based linkage strategy (17.8% compared with 17.6%, respectively).

Table 4.4: State and territory at death by linkage strategy

		Name-based linkage			SLK-based linkage		
State and territory[(a)]	ACCMIS category specific population	Per cent of links	Death rates per 1,000 ACCMIS clients	95% confidence interval	Per cent of links	Death rates per 1,000 ACCMIS clients	95% confidence interval
New South Wales	146,660	35.4	418	(412 – 424)	35.5	415	(409 – 421)
Victoria	103,277	24.7	412	(404 – 420)	24.7	408	(400 – 416)
Queensland	72,863	17.6	418	(409 – 427)	17.8	416	(407 – 425)
Western Australia	33,917	8.1	414	(401 – 427)	7.9	400	(387 – 413)
South Australia	40,408	10.0	426	(414 – 438)	10.0	423	(411 – 435)
Tasmania	11,404	2.8	430	(406 – 454)	2.8	413	(390 – 436)
Australian Capital Territory	4,503	1.1	413	(376 – 450)	1.1	403	(367 – 439)
Northern Territory	2,025	0.3	266	(222 – 310)	0.3	247	(205 – 289)
Total	**415,057**	**100.0**	**416**	**(412 – 420)**	**100.0**	**412**	**(408 – 416)**

(a) From ACCMIS data.

Whether these minor differences in match distribution affected state and territory specific death rates was examined. Both Western Australia and Tasmania had relatively large gaps between the name-based and SLK-based death rate per 1,000 ACCMIS clients (414 deaths per 1,000 ACCMIS clients compared with 400 deaths per 1,000 ACCMIS clients, and 430 deaths per 1,000 ACCMIS clients compared with 413 deaths per 1,000 ACCMIS clients, respectively, Table 4.5). This is consistent with lower expected match rates for these two states. Relatively larger differences were also seen for the two territories, where smaller absolute differences in number of links can have a relatively large impact on rates.

The ranking of states and territories by death rate resulting from the two strategies are different (Table 4.5). However, the very similar death rates and their 95% confidence intervals across the states and territories, except for the Northern Territory, suggest that many of the differences are not statistically significant. For example, the name-based death rate estimates for Tasmania and South Australia are ranked highest and second highest (430 deaths per 1,000 ACCMIS clients and 426 deaths per 1,000 ACCMIS clients, respectively) across all states and territories. However, among the SLK-based estimates South Australia is ranked highest (423 deaths per 1,000 ACCMIS clients) while Tasmania is only fourth highest (413 deaths per 1,000 ACCMIS clients). Despite the differences in ranking between the two linkage methods the 95% confidence intervals for the two states overlap under both strategies, so that both sets of estimates suggest the differences are not statistically significant. These results emphasise the importance of examining both statistical significance and likely sources of bias before drawing conclusions from linked data.

Table 4.5: State and territory specific death rates per 1,000 ACCMIS clients, by linkage strategy

		Name-based linkage		SLK-based linkage	
State/territory[(a)]	ACCMIS Category specific population	Number	Death rates per 1,000 ACCMIS clients	Number	Death rates per 1,000 ACCMIS clients
New South Wales	146,660	61,196	417	60,725	414
Victoria	103,277	42,584	412	42,163	408
Queensland	72,863	30,445	418	30,341	416
Western Australia	33,917	14,054	414	13,561	400
South Australia	40,408	17,194	426	17,106	423
Tasmania	11,404	4,904	430	4,715	413
Australian Capital Territory	4,503	1,860	413	1,816	403
Northern Territory	2,025	539	266	501	247
Total	**415,057**	**172,776**	**416**	**170,928**	**412**

(a) From ACCMIS data.

4.4 Summary

The above results together show that, overall, the SLK-based linkage strategy resulted in a matched data set that largely reflected the name-based linkage strategy. Consequently, while biased downward slightly, death rates derived from the SLK-based linkage look very similar to those obtained from the name-based linkage. These results provide further evidence of the utility of the matched data sets obtained using the SLK-based linkage process.

Appendix 1: Data

Data from the ACCMIS were linked to the NDI using both linkage strategies. The PIAC project included ACCMIS program use from 1 July 2002 to 30 June 2006 and deaths from 1 July 2003 to 31 December 2006. The same data were used for the linkage comparison. This section introduces the structure of the two data sets and the scope and quality of the data used in the two linkage methods.

A1.1 Aged and Community Care Management Information System

The Department of Health and Ageing's (DoHA) administrative aged care data is maintained on ACCMIS. This database contains information on client's use of the Residential Aged Care (RAC) program and the community aged care package programs, including Community Aged Care Packages (CACP), Extended Aged Care at Home (EACH) packages, Extended Aged Care at Home Dementia (EACHD) packages and the Transition Care Program (TCP).

Data structure

The ACCMIS system is person-based. Individual clients are identified via name and other demographic data, and given a distinct client identification number (client ID). ACCMIS identifies two types of clients:

1. RAC+: those using RAC or any of the small EACH, EACHD or TCP programs
2. CACP: those using CACP.

ACCMIS clients can fall into one of the following three categories (Figure A1.1):

(A) If a person's program use stays within the four RAC+ programs the person has a single RAC+ client identification number.

(B) If a person has been both a CACP recipient and a client of any of the RAC+ programs then they are identified on the system twice, that is they have two client identification numbers (RAC+ and CACP).

(C) If a person's program use stays within CACP then the person has a single client CACP identification number.

Identification of new and continuing clients in the two program groups is carried out by DoHA staff. Occasionally repeat program use by a client is not identified, resulting in multiple client IDs for an individual within a program (Table A1.2)(AIHW: Karmel 2005).

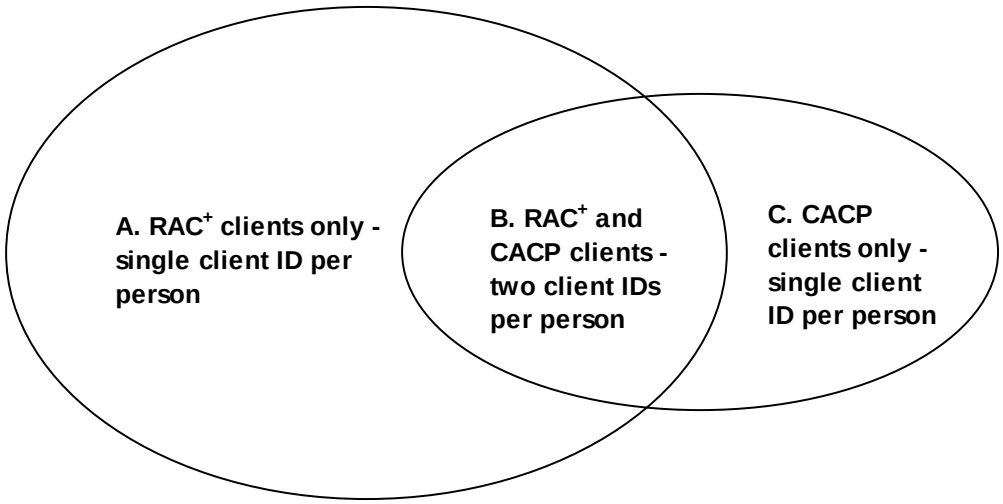


Figure A1.1: Structure of the ACCMIS data in relation to RAC+ and CACP use

Identifying ACCMIS clients

People with two client IDs on ACCMIS, that is clients who had been both a CACP recipient and a client of any of the RAC+ programs, were identified using SLK-based linkage as part of the PIAC project (Karmel et al. 2010). After eliminating duplicate client records for RAC+ and CACP clients and accounting for cross-program use the overall estimated number of people that used ACCMIS between 1 July 2002 and 30 June 2006 was 415,057 (Table A1.1). Of all ACCMIS clients, 81% used RAC+ programs only, 10% of clients used the CACP program only and 9% used both a RAC+ program and CACP program. For data linkage purposes each client appeared only once on the data set being linked.

Table A1.1: Total number of clients on ACCMIS 1 July 2002 – 30 June 2006

	Number	Per cent
A. Clients that had a RAC+ record only	335,029	80.7
C. Clients that had CACP record only	41,874	10.1
B. Clients that had both a RAC+ and CACP record	38,154	9.2
Total number of ACCMIS clients	**415,057**	**100.0**

Data for linkage

People who used an ACCMIS-reported program between 1 July 2002 and 30 June 2006 were included in the data linkage. The data used for linkage were based on:

- First name and surname of client
- DOB.
- Sex.

- State of usual residence.
- Postcode of usual residence.
- Likely date of death (DOD) up to 30 June 2006 (date of last discharge extracted as the likely DOD for those that had last discharge given as death or deceased).

Data quality

Data quality is investigated in terms of the prevalence of missing or poor quality data.

Name, sex and DOB are rarely missing. For example, of the CACP records first name or surname was missing in only 0.01% of cases (Table A1.2). There is some evidence of use of guessed dates of birth (especially 1 January), but only to a small extent with 1 January birth dates being two to three times more common than other 1st of the month DOBs (Table A1.2). The use of possibly guessed DOBs was more prevalent in the CACP than RAC+ data. Looking at the top five DOBs, dates that were 1st of January or July at the beginning of a decade (1920, etc.) do not appear to be overly common when compared with other 1st of January or July DOBs, so that clients with an SLK-581 possibly incorporating a guessed DOB may have reliable year of birth (YOB) information.

A person only had a DOD on the ACCMIS linkage data if they had died during the use of an ACCMIS program and had a DOD before 1 July 2006 otherwise they were recorded on the extract for matching as not having a DOD. DOD was not available or missing at a higher rate in CACP data (87%) compared with RAC+ data (53%). This is most likely because people using the CACP program are in better health generally compared with those using residential aged care. Postcode of usual residence was missing for less than 1% of clients.

Table A1.2: Quality of ACCMIS linkage data, by program, people with ACCMIS-recorded care at some stage 1 July 2002 – 30 June 2006

	RAC+		CACP	
Missing information	**Number**	**Per cent**	**Number**	**Per cent**
First name	—	—	11	0.01
Surname	18	—	1	—
Sex	—	—	—	—
Improbable/missing year of birth for SLK-581	3	—	—	—
1 January YOB	1,777	0.48	644	0.80
1 January decade	239	0.06	124	0.15
1 January dates	*2,018*	*0.54*	*768*	*0.96*
1 July YOB	1,373	0.37	496	0.62
1 July decade	196	0.05	83	0.10
1 July dates	*1,569*	*0.42*	*580*	*0.72*
Other 1st of the month	1,009–1,139	0.27–0.30	220–254	0.27–0.31
Top 5 dates	[(a)]87–126	. .	[(b)]31–49	. .
Next top date	[(a)]86	0.02	[(b)]30	0.04
Postcode	1,909	0.51	619	0.77
DOD[(c)]	197,573	52.88	69,600	86.94
Clients with multiple ACCMIS IDs[(d)]	821	0.22	56	0.07
All client IDs	**373,595**	**. .**	**80,056**	**. .**

(a) Top 5 dates: 1 January 1920, 1919, 1921, 1923, 1915, 1 July 1920 joint 5th, next top date: 1 January 1918.

(b) Top 5 dates: 1 January 1930, 1920, 1924, 1925, 1926, 1 July 1920 joint 5th, next top date: 1 January 1917.

(c) A person only had a DOD on the ACCMIS data if they had died during the use of an ACCMIS program and had a DOD before 1 July 2006 otherwise they were recorded on the extract for matching as not having a DOD.

(d) Duplicates were identified using the method given in AIHW: Karmel 2005. 821 RAC+ client IDs relate to 409 people and 56 CACP client IDs relate to 28 people.

A1.2 National Death Index

The NDI is a database, housed at the AIHW, which contains records of all deaths occurring in Australia since 1980. The data are obtained from the Registrars of Births, Deaths and Marriages in each state and territory. For each deceased person the following variables are stored on the NDI: name (first, middle and surname), DOB (or estimated year of birth), age at death, sex, DOD, address, state and territory of registration and registration number.

Data structure

A person who dies is assigned a record with a unique mortality identifier. A person can only die once and therefore should only have one record on the NDI. However, a person may have multiple records under the same mortality identifier when new or revised data on their death is added to the NDI. Therefore a part of the data cleaning process included identifying and removing these duplicates.

Data for linkage

The NDI data for linkage covered deaths over the period 1 July 2003 – 31 December 2006 (as reported by May 2007). Earlier death records were not considered for matching as people who died in that period were not within the scope of the PIAC project. Neo-natal deaths (based on reported age at death less than one year) were also excluded. This left a total of 495,644 death records for matching.

The data used for linkage was as follows:

- First name and surname of client.
- DOB.
- Sex.
- State of usual residence at death.
- Postcode of usual residence at death (derived from address).
- DOD.

Duplicate records

As stated above, people may have more than one record on the NDI. However, records for the same death are given the same mortality identifier. Linking multiplicity of death records is only an issue if it results in erroneous links. The advantage of retaining repeated records is that the revised information may contain name revisions so that retaining all the records implicitly introduces specific name variation into the linkage. This is particularly important for SLK-based linkage. Overall, 6.2% of NDI records (30,867) over the period of interest contained repeated mortality identifiers (excluding the first occurrence) (Table A1.3).

A duplicate death record was defined as one having the identical first name, surname, DOB, DOD, sex and postcode of residence as another record. Such records do not contain any additional data to aid the SLK-based linkage. Overall, 4.8% (or 23,124) of NDI death records were dropped as duplicates (Table A1.3). In addition, records with a missing mortality identifier were excluded as being of unacceptable quality. This affected 244 records.

After excluding the 23,124 duplicate death records, there were still some repeated death records as evidenced by the higher number of records with non-unique mortality identifiers (30,867). A further 4,302 records had non-unique SLK-581 and DOD combinations, and manual inspection showed that these possible duplicate records were generally for the same person. The additional 2,155 likely death duplicates associated with these people were also excluded from matching as they provided no new name information for the SLK-based matching. After these exclusions, it is estimated that as many as 5,300 people (30,867 – 23,124 – 2,155 – 244 = 5,344) still had more than one death record with either name or DOD differences on the NDI data set used for matching to the PIAC cohort.

Exclusion of duplicate and probably duplicate death records and those with a missing mortality identifier left a total of 470,121 death records for linking to ACCMIS (Table A1.3).

Table A1.3: Repeated death records in NDI data, deaths 1 July 2003 – 31 December 2006, as on NDI database by May 2007

Record type	Prevalence of duplicate death records	
	Number	Per cent
Repeated mortality identifiers	30,867	6.2
Duplicate death records (based on identical first name and surname, DOB, DOD, sex and postcode of residence[(a)]) – dropped repeats	23,124	4.8
Probably repeated death records (based on identical SLK-581 and DOD, without missing data) – dropped repeats	2,155	0.4
Missing mortality identifier	244	0.1
Dropped from data linkage	*25,523*	*5.2*
Remaining records that still had more than one death record with either name or DOD differences on the NDI data set	5,344	1.1
Retained for data linkage	*470,121*	*94.9*
All records	**495,644**	**100.00**

(a) 'Postcode of residence' is the last four characters of the address of residence on the NDI (missing for 12.47% of NDI records, predominantly for Western Australia and Tasmania).

Note: Table excludes births with reported age at death of less than 1 year (to exclude coincident multiple birth neo-natal deaths which can look like duplicates when using DOB, DOD and postcode). This exclusion was highly unlikely to affect linkages to recipients of aged care program services.

Data quality

Overall, 97.7% of records had complete name data and good DOB data. Name was missing in 0.02% of NDI records, DOB data was missing for 0.13% of records, and very few records had both missing name or sex data and missing DOB (Table A1.4). As for ACCMIS, there is some evidence that 1 January and 1 July DOBs were being used as default values, but these types of DOBs occurred in less than 1% of records. A further 1.35% of records had a DOB that was inconsistent with the reported age at death suggesting possibly poor DOB data.

Other data used when linking ACCMIS and NDI records were DOD and postcode of usual residence. The first of these was rarely missing (0.04%). However, just over 12% of records were missing postcode information. This was because postcode information is not generally available on the NDI database for deaths reported in Western Australia and Tasmania.

Table A1.4: Quality of DOB data on the NDI, deaths 1 July 2003 – 31 December 2006

	Number	Per cent
Missing name and sex		
Missing full name	2	—
Missing first name only	75	0.02
Missing surname name only	9	—
Any with missing name	*86*	*0.02*
Any with missing sex (includes a small number with some missing name data)	122	0.03
Missing DOB data		
Missing day	598	0.13
Missing month	590	0.13
Missing year (including <1887, and > 2006)	529	0.11
Missing any part of DOB for SLK-581	*604*	0.13
Inconsistent DOB and reported age at death[(a)]	6,346	1.35
Possibly poor DOB data		
1 January dates (all years)	2,396	0.51
1 July dates (all years)	1,785	0.38
Other 1st of the month	1,242–1,428	0.26–0.30
Mean per day of the year	1,283	0.27
Top 5 dates—all 1 January (1922, 1923, 1920, 1921, 1924)	88–104	0.02–0.02
Next top date (1 January 1925)	83	0.02
Missing DOD	**177**	**0.04**
Missing postcode[(b)]	**58,625**	**12.47**
All	**470,121**	**. .**

(a) More than 1 year different between reported age at death and age derived from DOB and DOD.

(b) Postcode is not generally available on the NDI for deaths reported in Western Australia and Tasmania as address data for these states exclude postcode.

References

AIHW: Karmel R 2005. Data linkage protocols using a statistical linkage key. Cat. no. CSI 1. Canberra: AIHW.

Campbell KM 2005. Rule Your Data with The Link King© (a SAS/AF® application for record linkage and unduplication). Philadelphia.

Jaro M 1989. Advances in Record-Linkage Methodology as Applied to Matching the 1985 Census of Tampa, Florida. Journal of the American Statistical Association 84.

Karmel R, Anderson P, Gibson D, Peut A, Duckett SJ & Wells Y 2010. Empirical aspects of record linkage across multiple data sets using statistical linkage keys: the experience of the PIAC cohort study. BMC Health Services Research 10:41.

Porter E & Winkler W 1997. Approximating String Comparison and its Effect on an Advanced Record Linkage System.

List of tables

List of figures