Linking SAAP, child protection and juvenile justice data

Technical report

DATA LINKAGE SERIES
Number 14

Linking SAAP, child protection and juvenile justice data

Technical report

Australian Institute of Health and Welfare
Canberra
Cat. no. CSI 14

> **The Australian Institute of Health and Welfare is a major national agency**
> **which provides reliable, regular and relevant information and statistics**
> **on Australia's health and welfare. The Institute's mission is**
> ***authoritative information and statistics to promote better health and wellbeing*.**

> **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

Abbreviations

AIHW	Australian Institute of Health and Welfare
CDSMAC	Community and Disability Services Ministerial Advisory Council
CP	child protection
JJ NMDS	Juvenile Justice National Minimum Data Set
NDC	national data collection
NIS	notifications, investigations and substantiations
SAAP	Supported Accommodation and Assistance Program
SAAP NDC	SAAP National Data Collection
SLK	statistical linkage key
SLK-581	statistical linkage key comprising five letters of name, eight digits for date of birth and one digit for sex.

Symbols

—	nil or rounded to zero
. .	not applicable
n.a.	not available

1 Introduction

The Community and Disability Services Ministers' Advisory Council (CDSMAC) funded the Australian Institute of Health and Welfare (AIHW) to investigate the feasibility of linking data collections for three community sector services: child protection, juvenile justice and the Supported Accommodation Assistance Program (SAAP).

The resulting report, *Linking SAAP, child protection and juvenile justice data collections: a feasibility study* (AIHW 2008b), found that both the juvenile justice and SAAP data collections were suitable for linkage, as would be the unit-record child protection data collection, when implemented, and that analysis of such a linked data set would provide valuable information for improving services for young people and for implementing, monitoring and evaluating targeted interventions to prevent young people from entering these sectors.

Following the results of the feasibility study, the CDSMAC agreed to fund the implementation of the linkage with the available data. This report describes the technical details of the process used to link the three community sector data collections. Derivation of data to be used in analysis is also discussed. A companion report, *Children and young people at risk of social exclusion: links between homelessness, child protection and juvenile justice* (AIHW 2012), presents the analysis of the linked data.

2   Linkage using statistical linkage keys

The linkage method used in this project involved statistical linkage keys. A statistical linkage key (SLK) is a string of characters and numbers that contains sufficient information to link records for statistical analysis, but does not necessarily contain sufficient information to identify individuals with certainty. They are typically used to link data sets where information such as full name is not available or where additional privacy protection is desired.

One of the more commonly used SLKs is SLK-581, which is used in a number of community services and other collections to identify records that belong to the same person for the purposes of statistical analysis. The SLK-581 comprises:

- Five letters of a person's name—second, third and fifth letters of family name and second and third letters of given name.
- Eight digits of date of birth—ddmmyyyy.
- One digit for sex (1=male, 2=female).

For example, Mary Brown born on 1 July 1958 would have an SLK-581 of RONAR010719582.

SLKs can be used wherever records relating to the same person need to be linked and a unique identifier is not available, or where full name and demographic data are not available for data linkage—either because these data are not collected or because of privacy considerations. Unique identifiers are usually not available because they have not been collected or because they have not been provided to the linking agency. In the current context, SLKs are used because neither full name nor unique identifiers are available.

The simplest way to use SLKs is to link using a deterministic method, which means that where a particular value for an SLK appears in multiple data sets, the corresponding records are assumed to relate to the same person. Thus, there are four possible results from deterministic linkage:

1. True positive: the records relate to the same person (match) and they share the same SLK (link).

2. True negative: the records do not relate to the same person (no match) and the records do not have the same SLK (no link).

3. False negative: the records relate to the same person (match) but the records do not have the same SLK (no link). This can occur where a person has changed their name or their date of birth has been incorrectly recorded in one of the data sets.

4. False positive: the records do not relate to the same person (no match) but the records have the same SLK (link).

These possibilities are illustrated in Table 2.1.

Table 2.1: Examples of possible results using deterministic linkage with the SLK-581

Data set 1	Result	Data set 2
True positive (or true match)		
Mary Brown	Match=yes	Mary Brown
RONAR010719582	Link=yes	RONAR010719582
True negative (no match)		
Mary Brown	Match=no	Elizabeth Brown
RONAR010719582	Link=no	RONLI010719582
False negative (missed match)		
Mary Brown	Match=yes	Mary Smith (nee Brown)
RONAR010719582	Link=no	MIHAR010719582
False positive (false match)		
Mary Brown	Match=no	Maria Brownlow
RONAR010719582	Link=yes	RONAR010719582

A deterministic method of data linkage using SLKs will have a high rate of success where the SLK used is adequately distinctive (low rate of false positives and high rate of true negatives) and where the population being linked tends to consistently report components of the linkage key (high rate of true positives and low rate of false negatives).

Previous work has shown that SLK-581 can be adequately distinctive for use in simple deterministic linkage (that is, linking records that match exactly on SLK-581), and in populations with few name changes and where components such as date of birth are consistently reported, linkage with the SLK-581 will be sufficiently accurate for statistical purposes (AIHW: Ryan et al. 1999). However, it is not reasonable to assume that the populations being linked in this project will have few name changes, as these populations include children in the child protection system who might be adopted and families who might be escaping from family violence. Further, about 25% of the SAAP National Data Collection (NDC) records have incomplete SLK-581s, which means that these records cannot be linked using simple deterministic matching.

An alternative to using a simple deterministic linkage method is the step-wise deterministic linkage method developed by the AIHW (AIHW 2011c; Karmel et al. 2010). This method aims to reduce the number of false matches (where the records of multiple distinct people are falsely assumed to belong to a single distinct person) and reduce the number of missed matches (where the records of a single distinct person are falsely assumed to belong to multiple distinct people). In this method, SLKs containing the components of the SLK-581 can be used, as well as SLKs containing additional variables such as postcode and Indigenous status. These components are systematically varied to account for people whose keys vary across the data collections. This enhanced linkage method allows for missing data and increases the number of true matches and reduces the number of false matches (caused by separate individuals sharing the same SLK) but requires considerably more resources than simple deterministic linkage.

A recent study (AIHW 2011a) found that this step-wise method was effective in identifying true links among aged care data sets. In that study, nearly all (99.7%) of the links identified by this method were also identified by the name-based linkage strategy with clerical review (considered to be 'true' links), while the method identified 98.5% of all links made by the name-based linkage strategy.

3 Data

Data used in this project come from four sources:

Supported Accommodation and Assistance Program National Data Collection (SAAP NDC): this data collection contains information from specialist agencies that provide services to people who are homeless or at risk of homelessness under the Supported Accommodation and Assistance Program.

Juvenile Justice National Minimum Data Set (JJ NMDS): this data set contains information on young people who are supervised either in the community or in detention by juvenile justice agencies.

Department of Human Services, Victoria: the Victorian Government supplied data on children and young people who had been the subject of a notification, investigation, or substantiation relating to a child protection issue.

Department of Health and Human Services, Tasmania: the Tasmanian Government supplied data on children and young people who had been the subject of a notification, investigation, or substantiation relating to a child protection issue.

3.1 Data availability

Data for this project comprised SAAP NDC data for 2006–07 to 2008–09, JJ NMDS for 2000–01 to 2009–10 (for most states and territories) and child protection (CP) data for 1990–91 to 2008–09 (Victoria) and 2004–05 to 2006–07 (Tasmania). The data included from the three data collections are given by financial year in Table 3.1.

Table 3.1: SAAP NDC, JJ NMDS and CP data availability

State or territory	SAAP NDC	JJ NMDS	CP
NSW	2006–07 to 2008–09	2000–01 to 2009-10	n.a.
Vic	2006–07 to 2008–09	2000–01 to 2009-10	1990–91 to 2008–09
Qld	2006–07 to 2008–09	2000–01 to 2009-10	n.a.
WA	2006–07 to 2008–09	2000–01 to 2007–08	n.a.
SA	2006–07 to 2008–09	2000–01 to 2009-10	n.a.
Tas	2006–07 to 2008–09	2000–01 to 2009-10	2004–05 to 2006–07
ACT	2006–07 to 2008–09	2004–05 to 2009-10	n.a.
NT	2006–07 to 2008–09	2000–01 to 2007–08	n.a.

SAAP NDC

The SAAP NDC is held by the SAAP National Data Collection Agency at the AIHW. Data for 2006–07 to 2008–09 were included in this project as they contain the standard SLK-581 and were of good quality. Only records where the client had provided consent for the collection of SLK-581, along with other personal data items on circumstances, were included in this project. For the years of interest, between 85% and 90% of SAAP client records had consent collection-wide (AIHW 2008a, 2009, 2010). Children who accompanied a SAAP client to a SAAP agency at any time during that client's support period were also included in the data

linkage. More than 90% of records for accompanying children had consent. Data for 2005–06 also contain the standard SLK-581, but had data quality issues so were excluded from this project. Data before 2005–06 contain an SLK with fewer components and therefore are not suitable for this project. In the remainder of this report, the term 'SAAP client' includes both SAAP clients and accompanying children.

Child protection data

The Victorian Department of Human Services and the Tasmanian Department of Health and Human Services provided data on child protection notifications, investigations and substantiations directly to the AIHW for inclusion in this project. The Victorian data are for 1990–91 to 2008–09 and the Tasmanian data for 2004–05 to 2006–07. The AIHW collects data on child protection from all states and territories; however, these data are in aggregate form and therefore not suitable for this project. The AIHW is working with states and territories to develop a national unit-record collection for child protection, and these data would be suitable for future linkage.

JJ NMDS

The JJ NMDS is administered by the AIHW under the auspices of the Australasian Juvenile Justice Administrators. Data for most states and territories are available from 2000–01 to 2009–10. However, complete detention data are only available from 2006–07 for Tasmania, data for 2000–01 to 2002–03 are unavailable for the Australian Capital Territory, and both Western Australia and the Northern Territory have not supplied data for 2008–09 or 2009–10 (see section 3.4 in AIHW 2011b for more information).

3.2   Linkage issues

Linkage key components

This project used a linkage method that generates a number of linkage keys for each person, with each key based on particular components (see Section 4 for further details of the method). The data available to be included as linkage key components for the data collections used in this project were letters of family and given names, date of birth, sex, Indigenous status, last known suburb or town and last known postcode (Table 3.2).

Table 3.2: Source of linkage key components

Component	Symbol	Details	Source in data collection			
			JJ NMDS	**SAAP NDC**	**CP Vic**	**CP Tas**
Letters of family name	F3	Second, third and fifth letters of family name	Client file (letters of name)	Alpha code	Alpha code	Alpha code
Letters of given name	G2	Second and third letters of given name	Client file (letters of name)	Alpha code	Alpha code	Alpha code
Day of birth	D2	From date of birth	Client file (date of birth)	Date of birth	Date of birth	Date of birth
Month of birth	M2	From date of birth	Client file (date of birth)	Date of birth	Date of birth	Date of birth
Year of birth	Y4	From date of birth	Client file (date of birth)	Date of birth	Date of birth	Date of birth
Sex	S1		Client file	Q1	Sex	Sex
Indigenous status	I1	Indigenous, non-Indigenous or unknown	Client file	Q5	Client file	Client file
Suburb/town/locality name	TS	Reported character string	Order and detention files	Q14	NIS file	NIS file
Postcode	P4	Digits 1–4	Order and detention files	Q14	NIS file	NIS file
Postcode (alternative 1)	P3	Digits 1–3	Order and detention files	Q14	NIS file	NIS file
Postcode (alternative 2)	P2	Digits 1–2	Order and detention files	Q14	NIS file	NIS file
Postcode (alternative 3)	P1	Digit 1	Order and detention files	Q14	NIS file	NIS file

Person identifiers

To link records for individuals across data sets, an individual needs to be defined within each data set. In the JJ NMDS and child protection data collections, each unique person within a state or territory is given a person identifier. This identifier is used to define a person. However, no unique person identifier is available for the SAAP NDC. For this project, a person identifier for the SAAP NDC records was constructed using the state of the SAAP agency plus the SLK-581 components. Therefore, it is possible that multiple distinct people will be defined as a single person if they reported the same SLK-581 values and attended a SAAP agency in the same state or territory. Also, a SAAP client may be given more than one person identifier if they reported different data, including missing values, for SLK-581 on different occasions.

The number of person identifiers (that is, 'people') in the state-level data sets ranged from 406 in the Australian Capital Territory JJ NMDS data set to 99,008 in the Victorian SAAP NDC data set (Table 3.3). None of the individual state and territory data sets in these data collections are routinely merged or linked, so individuals in one state and territory may also be in another state and territory for the same data collection. Therefore, this linkage project not only identifies the number of links between data collections (for example, between the SAAP NDC and the JJ NMDS) but also the number of links within data collections (for example, between the New South Wales JJ NMDS data set and the Victorian JJ NMDS data set).

Table 3.3: Number of person identifiers

	NSW	Vic	Qld	WA	SA	Tas	ACT	NT
SAAP NDC	90,675	99,008	58,356	30,720	38,306	11,705	5,785	9,716
CP	n.a.	65,176	n.a.	n.a.	n.a.	14,884	n.a.	n.a.
JJ NMDS	17,672	10,922	11,749	7,350	5,947	1,845	406	1,111

Changes in linkage key component values

It is possible for values for linkage key components such as name and last known home suburb and postcode to change over time. Also, it is possible for any or all components to be incorrectly recorded. Where these changes are recorded in a data set, it is possible for an individual (as defined by a person identifier) to have multiple values for the same linkage key component.

In the SAAP NDC, information is collected for each attendance or support period per SAAP agency. Therefore, a person as defined by this project could have multiple values for each linkage key component.

In the SAAP NDC data used in this project, each person identifier had, on average, between 1.2 and 1.4 different versions of the linkage key components reported (Table 3.4). Also, some of the records with the same person identifier could have missing data for one or more of the items used as components. For most of the components, the proportion of records with missing values was less than 10% (Table 3.5). However, the proportion of records with missing town/suburb or postcode was between 10% and 20% for most states and territories. Overall, between 73% and 83% of person identifiers had at least one record with no missing values for any of the linkage key components.

Table 3.4: Person identifiers by average number of records, SAAP NDC

	NSW	Vic	Qld	WA	SA	Tas	ACT	NT
Total number of person identifiers	90,675	99,008	58,356	30,720	38,306	11,705	5,785	9,716
Average number of reported versions of linkage key component data per person identifier	1.3	1.4	1.3	1.3	1.4	1.3	1.3	1.2

Table 3.5: Proportion of records with missing values, SAAP NDC

Key component	NSW	Vic	Qld	WA	SA	Tas	ACT	NT
Family name	3.4	4.2	7.9	2.0	2.4	5.4	1.7	1.6
Given name	2.9	2.4	7.0	1.1	1.6	2.6	1.4	1.1
Day of birth	8.1	7.6	6.9	3.8	6.3	5.8	5.2	6.4
Month of birth	7.0	6.3	5.5	2.9	4.9	5.0	3.4	4.7
Year of birth	4.1	3.8	3.7	1.9	2.8	3.4	1.9	2.1
Sex	1.2	0.8	1.0	0.9	1.2	0.9	1.0	1.1
Indigenous status	6.8	8.8	4.2	4.4	5.3	8.8	4.8	3.5
Town or suburb	11.8	10.0	14.9	11.1	14.4	9.6	20.0	5.9
Postcode	10.6	8.7	12.4	10.2	14.1	8.7	19.5	5.4
Person identifiers with at least one complete record (i.e. no missing components) (%)	**76.2**	**76.1**	**72.5**	**81.8**	**78.1**	**77.4**	**75.0**	**82.9**

Note: Children accompanying SAAP clients were more likely to have missing data for SLK-581 components than SAAP clients.

For both the Victorian and Tasmanian CP data sets, a person can have only one value for letters of family and given names, date of birth, sex and Indigenous status. For the available Victorian CP data, there was an average of 1.6 records with differences in linkage key component data per person identifier; the Tasmanian data had little variation (an average of 1.0 versions) (Table 3.6). For the Victorian data, relatively few records had missing data—less than 1% for letters of name, date of birth and sex (Table 3.7). For the Tasmanian data, 30% of records had missing date of birth, and 90% had missing Indigenous status. Due to the high proportion of records with missing Indigenous status, just 10% of Tasmanian person identifiers had at least one complete record, compared with 89% of Victorian person identifiers.

Table 3.6: Person identifiers by average number of records, CP data

	Vic	Tas
Total number of person identifiers	65,176	14,884
Average number of reported versions of linkage key component data per person identifier	1.6	1.0

Table 3.7: Proportion of records with missing values, CP data

	Vic	Tas
Family name	0.0	0.0
Given name	0.0	0.0
Day of birth	0.7	30.1
Month of birth	0.7	30.1
Year of birth	0.7	30.1
Sex	0.8	3.5
Indigenous status	3.1	89.8
Town or suburb	13.5	9.1
Postcode	13.4	8.9
Person identifiers with at least 1 complete record (i.e. no missing components) (%)	**89.1**	**9.7**

In the JJ NMDS, information on client characteristics (letters of name, date of birth, sex and Indigenous status) are updated each year, while information on the town or suburb and postcode of last known home address are provided for each order and detention period, so a person identifier can have multiple values for each linkage key component.

The average number of linkage key component versions recorded per person identifier ranged from 1.2 to 2.0 (Table 3.8), while the proportion of person identifiers with at least one record with complete data for all components ranged from 54% to 93% (Table 3.9). The amount of missing values was highest for the last known home address components, with between 19% and 50% of records missing the town/suburb or the postcode.

Table 3.8: Person identifiers by average number of records, JJ NMDS

	NSW	Vic	Qld	WA	SA	Tas	ACT	NT
Total number of person identifiers	17,672	10,922	11,749	7,350	5,947	1,845	406	1,111
Average number of reported versions of linkage key component data per person identifier	1.7	1.9	1.9	2.0	1.8	1.2	1.5	1.3

Table 3.9: Proportion of records with missing values, JJ NMDS

	NSW	Vic	Qld	WA	SA	Tas	ACT	NT
Family name	0.2	0.0	0.0	0.0	0.1	0.1	0.0	0.0
Given name	0.2	0.0	0.0	0.0	0.1	0.0	0.0	0.9
Day of birth	0.0	0.0	0.0	0.0	0.3	0.0	0.0	0.0
Month of birth	0.0	0.0	0.0	0.0	0.3	0.0	0.0	0.0
Year of birth	0.0	0.0	0.0	0.0	0.3	0.0	0.0	0.0
Sex	0.0	0.0	0.0	0.5	0.5	0.0	0.0	0.0
Indigenous status	13.9	3.4	1.0	6.2	8.6	23.7	0.5	0.5
Town or suburb	23.9	50.0	30.1	19.7	22.7	40.7	40.6	20.5
Postcode	23.8	49.9	29.2	18.7	22.6	40.7	40.6	39.0
Person identifiers with at least one complete record (i.e. no missing components) (%)	**82.9**	**77.2**	**93.4**	**91.1**	**87.0**	**53.7**	**86.9**	**64.7**

4 Linkage

4.1 Method

As discussed earlier, this project used step-wise determinist linkage. This method achieves its three aims (using all available data, having few false links and having few missed links) through the use of a step-wise approach that involves linking using the most precise linkage key before using the next most precise linkage key, and so on. To do this, all possible linkage keys (that is, combinations of linkage key components) need to be tested for suitability for use and an order of use had to be identified. This involved:

1. identifying all possible linkage keys

2. determining the 'precision' of the key, so that poor keys could be excluded and to ensure that stronger (that is, more accurate) keys were used before weaker keys

3. testing the order of the keys and determining the point at which additional keys added little value to the linkage.

Identifying possible linkage keys

As noted previously, the data sets in this project all shared a number of variables or components that could be used to create linkage keys:

- letters of family and given name
- date of birth (day, month and year of birth)
- sex
- Indigenous status
- town/suburb of last known home address
- postcode of last known home address (all 4 digits, digits 1-3, digits 1-2 and first digit).

Different linkage keys were formed by systematically including and excluding the various key components. Doing this for the available components results in 1,280 possible linkage keys—including those which consisted of just a single component.

Determining the precision of the linkage keys

Using data from the JJ NMDS, these 1,280 keys were assessed for their ability to distinguish between individuals. Those keys that were expected to have at least 10% as many different values in the population as SLK-581 were then selected for further assessment; that is, keys were retained if they were expected to have a 'relative distinctiveness' greater than 0.1 when compared with SLK-581. Out of the 1,280 possible keys, 334 met this criterion.

These 334 keys were then tested using data from the JJ NMDS, which was chosen because, unlike the SAAP NDC, it contains an identifier that enables people to be uniquely identified within each state and territory. Additionally, these data allow for changes in client information (such as changes in letters of name between years) to be assessed.

The data set used to test the keys contained the most recent complete record for each young person within each state and territory (so for each linkage key, each person identifier had only one combination). The resulting data set had 57,002 records. The value of a particular

linkage key was derived for each person identifier from the component values and measures of the key's utility for linking records were generated by state and territory. Three measures were used to examine the utility of the keys and to determine their order of use:

1. Average number of person identifiers attached to each key value. A key with a lower average number of person identifiers per key value was considered to be a more precise key than one with a higher average number.

2. The maximum number of person identifiers attached to a key value. A key with a lower maximum number of person identifiers per key value was considered to be a more precise key than one with a higher maximum number.

3. The distinctiveness of the key relative to the SLK-581. A key with higher 'relative distinctiveness' was considered to be a more precise key than one with a lower 'relative distinctiveness'.

Keys were ordered by the average number of person identifiers per key value (lowest to highest). If this measure was the same for two (or more) keys, the maximum number of person identifiers per key value (lowest to highest) was then used to order the keys. If the quality of two keys could still not be distinguished, relative distinctiveness was then used to order the keys (highest to lowest).

Testing the order of the linkage keys

To test the order of the keys, data sets were created by extracting all the records for a particular state or territory from the JJ NMDS, duplicating the records and then randomly allocating each record to one of three data sets. The 334 keys were then used in the order determined above to link these three data sets for each state and territory. After each pass with a particular key, links were removed and the remaining unlinked records were returned for possible linking with the next key. In the final linked data set, each link was considered correct if the person identifiers matched, and considered incorrect if they did not.

To determine the point at which additional keys did not add value to the linkage, the harmonic mean was generated using the cumulative per cent of correct links (number of correct links divided by number of links) and the cumulative per cent of possible links (number of correct links divided by the maximum number of possible links).

This test was run for all states and territories using JJ NMDS data, and the results confirmed the order of the keys as effective and showed that keys after key 162 added little value to the linked data set (see Table 4.1).

Table 4.1: Results of the test of the order of the linkage keys, selected keys

Key no.	Key components	NSW	VIC	Qld	WA	SA	Tas	ACT	NT
		Possible links (cumulative per cent)							
1	F3|G2|D2|M2|Y4|S1|I1|TS|P4	74.96	64.55	78.41	83.92	77.60	57.22	81.33	60.43
100	_|G2|D2|M2|Y4|S1|I1|TS|P2	85.48	66.61	80.91	93.21	85.28	72.49	81.93	81.77
161	F3|G2|_|M2|Y4|_|I1|_|P3	96.27	98.66	99.65	99.60	96.99	90.99	100.00	98.80
162	F3|G2|D2|M2|Y4|S1|_|_|_	99.91	99.61	99.87	99.62	99.04	100.00	100.00	99.04
163	F3|G2|D2|M2|Y4|_|I1|_|P1	99.91	99.61	99.87	99.63	99.04	100.00	100.00	99.04
164	F3|G2|D2|M2|Y4|_|I1|_|_	99.91	99.61	99.87	99.65	99.04	100.00	100.00	99.04
165	F3|G2|D2|M2|_|S1|_|_|P2	99.91	99.61	99.87	99.65	99.04	100.00	100.00	99.04
200	F3|G2|_|_|Y4|_|I1|TS|P3	99.91	99.62	99.87	99.65	99.13	100.00	100.00	99.04
334	F3|_|_|M2|_|_|I1|TS|P4	99.94	99.62	99.92	99.68	99.23	100.00	100.00	99.28
		Correct links (cumulative per cent)							
1	F3|G2|D2|M2|Y4|S1|I1|TS|P4	99.97	99.84	99.97	99.92	99.46	100.00	100.00	100.00
100	_|G2|D2|M2|Y4|S1|I1|TS|P2	99.94	99.77	99.92	99.63	99.12	100.00	100.00	99.71
161	F3|G2|_|M2|Y4|_|I1|_|P3	96.27	98.66	99.65	99.60	96.99	90.99	100.00	98.80
162	F3|G2|D2|M2|Y4|S1|_|_|_	99.92	99.61	99.91	99.55	99.06	100.00	100.00	99.76
163	F3|G2|D2|M2|Y4|_|I1|_|P1	99.91	99.61	99.87	99.63	99.04	100.00	100.00	99.04
164	F3|G2|D2|M2|Y4|_|I1|_|_	99.91	99.61	99.87	99.65	99.04	100.00	100.00	99.04
165	F3|G2|D2|M2|_|S1|_|_|P2	99.92	99.60	99.91	99.52	99.06	100.00	100.00	99.76
200	F3|G2|_|_|Y4|_|I1|TS|P3	99.92	99.60	99.88	99.49	99.02	100.00	100.00	99.76
334	F3|_|_|M2|_|_|I1|TS|P4	99.17	99.51	99.44	98.91	98.48	99.94	100.00	99.40

(continued)

Table 4.1 (continued): Results of the test of the order of the linkage keys, selected keys

Key no.	Key components	NSW	VIC	Qld	WA	SA	Tas	ACT	NT
					Harmonic mean				
1	F3|G2|D2|M2|Y4|S1|I1|TS|P4	85.68	78.41	87.89	91.23	87.18	72.79	89.70	75.34
100	_|G2|D2|M2|Y4|S1|I1|TS|P2	92.15	79.89	89.41	96.31	91.68	84.05	90.07	89.86
161	F3|G2|_|M2|Y4|_|I1|_|P3	98.06	99.16	99.78	99.58	98.01	95.28	100.00	99.28
162	F3|G2|D2|M2|Y4|S1|_|_|_	99.92	99.61	99.89	99.59	99.05	100.00	100.00	99.40
163	F3|G2|D2|M2|Y4|_|I1|_|P1	99.92	99.61	99.89	99.59	99.05	100.00	100.00	99.40
164	F3|G2|D2|M2|Y4|_|I1|_|_	99.92	99.61	99.89	99.59	99.05	100.00	100.00	99.40
165	F3|G2|D2|M2|_|S1|_|_|P2	99.92	99.61	99.89	99.59	99.05	100.00	100.00	99.40
200	F3|G2|_|_|Y4|_|I1|TS|P3	99.91	99.61	99.88	99.57	99.08	100.00	100.00	99.40
334	F3|_|_|M2|_|_|I1|TS|P4	99.55	99.57	99.68	99.30	98.85	99.97	100.00	99.34

Linking the SAAP NDC, CP and JJ NMDS data sets

The eight SAAP NDC data sets (one per state and territory), two CP data sets (Victoria and Tasmania) and eight JJ NMDS data sets (one per state and territory) were linked using 162 keys (see Appendix). All 18 data sets were first linked using key 1. Any conflicts (where a particular combination related to multiple person identifiers from the same state or territory, or where a particular person identifier linked to two or more different person identifiers in the same state or territory) were resolved through random selection. All 18 data sets were then linked using key 2, with updates to previous links made before new links were added to the linked data set. The process continued for the remaining 160 keys.

4.2   Creation of analysis data sets

To create the data set used for analysis, data for each of the people in the final linked data set were extracted from the SAAP, child protection and juvenile data sets. Some of the resulting information about services may conflict, which may indicate issues with either data collection or with the data linkage. For example, it is not possible for a young person to be detained in more than one state or territory at the same time. Similarly, it is not possible for a young person to be detained and be receiving services from a SAAP agency. However, people who access SAAP services as accompanying children are recorded on the same form as their parent or guardian, even though an accompanying child may not have received those services for the same length of time as the parent or guardian.

Overall, there were relatively few issues with the juvenile justice data (see below), which suggests that the linkage is of high quality. Some SAAP support and accommodation records were removed, but this was mainly due to issues with the data collection methods used in the SAAP NDC.

SAAP data

Data for SAAP support periods (which contain information on accommodation dates) were extracted from the SAAP client and accompanying children data sets. Forms for SAAP support periods are submitted at the end of each financial year, even though the support may not yet have ended, and therefore support periods that extend beyond the end of the financial year will have more than one record. In an effort to remove these duplicate records, records were removed where the support period was marked as ongoing at 30 June of the financial year and the support period end date was before 30 June 2009 (that is, the latest SAAP data) and the SAAP identifier (the agency identifier plus the SLK-581) had at least two records in the data set.

Due to data issues with the SAAP data, a conservative approach was taken when preparing the analysis data set to ensure that any observed effects were not overestimated. First, SAAP support period records were deleted where a particular period overlapped with another support period for the same person. Four per cent of people receiving SAAP services and 17% of all SAAP support periods were removed due to date conflicts.

Data on SAAP accommodation were then extracted from the SAAP support records after the deletion of possible duplicates. Records were deleted where a particular period overlapped another accommodation period for the same person. Five per cent of people accommodated through SAAP and 60% of accommodation periods were removed due to date conflicts. A further 1% of people accommodated through SAAP and 1% of accommodation records were

removed because the end date was before the start date. Most of these records were removed due to date conflicts with other accommodation periods for the same person, and manual inspection suggests that many of these conflicts were due to start or end dates that were changed between years.

Records were also deleted where a particular accommodation period overlapped with a juvenile detention period for the same person. Only a small percentage of records were removed due to such conflicting dates. Conflicts with detention dates are possible because, for accompanying children, accommodation dates are recorded for the entire family, not just the accompanying child, who may have entered detention while the rest of the family remained in SAAP accommodation.

Child protection data

Data for child protection notifications, substantiations and investigations were extracted from the data supplied by the Victorian and Tasmania governments for this project. As it is possible for notifications, substantiations and investigations to occur while a child or young person is receiving SAAP services or under juvenile justice supervision, it was not necessary to identify records that overlapped with other SAAP or juvenile justice records.

Child protection records were classified as: substantiated (where the outcome of the investigation into the notification resulted in a substantiation), not substantiated (where the outcome of the investigation into the notification did not result in a substantiation), or not investigated. Notifications that were not investigated include notifications where it was determined that no investigation was required and notifications for which an investigation was not able to be commenced or completed or was finalised at the end of the reporting period.

Juvenile justice data

Selected juvenile justice data were extracted from the 2007–08 version of the JJ NMDS for Western Australia, the Australian Capital Territory and the Northern Territory for years where data are not available in the 2009–10 version of the JJ NMDS (Table 4.2). Data for the remaining states and territories and for the Australian Capital Territory for 2008–09 and 2009–10 were extracted from the 2009–10 version of the JJ NMDS.

Table 4.2: Juvenile justice data extracted from the JJ NMDS

State or territory	JJ NMDS 2009–10	JJ NMDS 2007–08
NSW	2000–01 to 2009–10	n.a.
Vic	2000–01 to 2009–10	n.a.
Qld	2000–01 to 2009–10	n.a.
WA	n.a.	2000–01 to 2007–08
SA	2000–01 to 2009–10	n.a.
Tas	2000–01 to 2009–10	n.a.
ACT	2008–09 to 2009–10	2003–04 to 2007–08
NT	n.a.	2000–01 to 2007–08

Detention episodes were created for each person by removing all records for a particular person where a particular detention period in one state or territory overlapped with a

detention period in another state or territory. No individuals were removed as a result of this data cleaning, and only six individual detention periods were deleted.

Community-based supervision episodes were created for each person using information for community-based supervised orders and detention periods as outlined in 'Section 3.5 Methods' in *Juvenile justice in Australia 2009–10* (AIHW 2011b).

Demographics

After the linkage of the individual data sets, it is possible for each person to have multiple sets of demographic information. For this project, data from the JJ NMDs is considered the most 'trustworthy', as the data are cleaned on submission and any changes in demographic information from a previous year are verified by the submitting jurisdiction. A single set of demographic information (date of birth, sex and Indigenous status) for each person was selected using the following rules:

- Where the person had a juvenile justice record, demographic information was selected from the most recent year of juvenile justice data. Where there were multiple sets of demographic information in the most recent year (for example, from Victoria and Queensland), one set of information was randomly selected.
- Where the person did not have a juvenile justice record, one set of demographic information was randomly selected from the available SAAP and child protection data.

Appendix

Table A1: Keys used in linking the SAAP NDC, CP and JJ NMDS data sets by order of use

Key order	Key components	Key order	Key components
1	F3|G2|D2|M2|Y4|S1|I1|TS|P4	36	F3|G2|D2|_|Y4|_|I1|TS|P4
2	F3|G2|D2|M2|Y4|S1|I1|TS|P3	37	F3|G2|D2|_|Y4|S1|_|TS|P3
3	F3|G2|D2|M2|Y4|S1|I1|TS|P2	38	F3|G2|D2|M2|_|S1|_|TS|P4
4	F3|G2|D2|M2|Y4|S1|I1|TS|P1	39	F3|G2|D2|M2|_|S1|I1|TS|P2
5	F3|G2|D2|M2|Y4|S1|I1|TS|_	40	F3|G2|D2|M2|_|S1|_|TS|P3
6	F3|G2|D2|M2|Y4|S1|_|TS|P4	41	F3|G2|D2|M2|_|S1|I1|TS|P1
7	F3|G2|D2|M2|Y4|S1|_|TS|P3	42	F3|G2|D2|_|Y4|S1|_|TS|P2
8	F3|G2|D2|M2|Y4|S1|_|TS|P2	43	F3|G2|D2|M2|_|S1|I1|TS|_
9	F3|G2|D2|M2|Y4|S1|_|TS|P1	44	F3|G2|D2|_|Y4|S1|_|_|P4
10	F3|G2|D2|M2|Y4|S1|_|TS|_	45	F3|G2|D2|M2|_|_|I1|TS|P4
11	F3|G2|D2|M2|Y4|_|I1|TS|P4	46	F3|G2|D2|_|Y4|S1|I1|TS|P1
12	F3|G2|D2|M2|Y4|_|I1|TS|P3	47	F3|G2|D2|M2|_|S1|_|TS|P2
13	F3|G2|D2|M2|Y4|S1|I1|_|P4	48	F3|G2|D2|_|Y4|S1|I1|TS|_
14	F3|G2|D2|M2|Y4|_|I1|TS|P2	49	F3|G2|D2|_|Y4|_|I1|TS|P3
15	F3|G2|D2|M2|Y4|_|I1|TS|P1	50	F3|G2|D2|_|Y4|_|_|TS|P4
16	F3|G2|D2|M2|Y4|_|I1|TS|_	51	F3|G2|D2|M2|_|_|I1|TS|P3
17	F3|G2|D2|M2|Y4|_|_|TS|P4	52	F3|G2|D2|M2|_|S1|_|TS|P1
18	F3|G2|D2|M2|Y4|_|_|TS|P3	53	F3|G2|D2|M2|_|S1|_|TS|_
19	F3|G2|D2|M2|Y4|S1|_|_|P4	54	F3|G2|D2|_|Y4|_|I1|TS|P2
20	F3|G2|D2|M2|Y4|_|_|TS|P2	55	F3|G2|D2|M2|_|_|_|TS|P4
21	F3|G2|D2|M2|Y4|_|_|TS|P1	56	F3|G2|D2|M2|_|_|I1|TS|P2
22	F3|G2|D2|M2|Y4|_|_|TS|_	57	F3|G2|D2|M2|Y4|S1|I1|_|P2
23	F3|G2|D2|M2|Y4|_|I1|_|P4	58	F3|G2|D2|M2|_|S1|I1|_|P4
24	F3|G2|D2|M2|Y4|_|_|_|P4	59	F3|G2|D2|_|Y4|_|I1|_|P4
25	F3|G2|D2|M2|Y4|S1|I1|_|P3	60	F3|G2|D2|M2|_|_|_|TS|P3
26	F3|G2|D2|M2|Y4|S1|_|_|P3	61	F3|G2|D2|M2|_|_|I1|TS|P1
27	F3|G2|D2|_|Y4|S1|I1|TS|P4	62	F3|G2|D2|_|Y4|S1|_|TS|P1
28	F3|G2|D2|M2|Y4|_|I1|_|P3	63	F3|G2|D2|M2|_|_|I1|TS|_
29	F3|G2|D2|_|Y4|S1|_|TS|P4	64	F3|G2|D2|_|Y4|S1|_|TS|_
30	F3|G2|D2|_|Y4|S1|I1|TS|P3	65	F3|G2|D2|_|Y4|_|_|TS|P3

Table A1 (continued): Keys used in linking the SAAP NDC, CP and JJ NMDS data sets by order of use

Key order	Key components	Key order	Key components
31	F3|G2|D2|M2|Y4|_|_|_|P3	66	F3|G2|_|M2|Y4|S1|I1|TS|P4
32	F3|G2|D2|M2|_|S1|I1|TS|P4	67	F3|G2|D2|M2|_|_|_|TS|P2
33	F3|G2|D2|_|Y4|S1|I1|TS|P2	68	F3|G2|D2|M2|_|S1|_|_|P4
34	F3|G2|D2|_|Y4|S1|I1|_|P4	69	F3|G2|D2|_|Y4|_|_|TS|P2
35	F3|G2|D2|M2|_|S1|I1|TS|P3	70	F3|G2|_|M2|Y4|S1|I1|TS|P3
71	F3|G2|D2|M2|_|_|_|TS|P1	112	_|G2|D2|M2|Y4|S1|_|TS|P3
72	F3|G2|D2|M2|Y4|S1|_|_|P2	113	F3|G2|_|M2|Y4|S1|_|_|P4
73	F3|G2|D2|M2|_|_|_|TS|_	114	_|G2|D2|M2|Y4|_|I1|TS|P2
74	F3|G2|D2|_|Y4|_|I1|TS|P1	115	_|G2|D2|M2|Y4|S1|_|TS|P2
75	F3|G2|D2|_|Y4|_|I1|TS|_	116	F3|G2|_|M2|Y4|_|I1|_|P4
76	F3|G2|D2|_|Y4|_|_|_|P4	117	F3|G2|D2|_|Y4|S1|_|_|P3
77	F3|G2|D2|M2|_|_|I1|_|P4	118	F3|G2|D2|_|_|S1|I1|TS|P4
78	F3|G2|_|M2|Y4|S1|_|TS|P4	119	F3|G2|D2|M2|_|S1|_|_|P3
79	F3|G2|_|M2|Y4|S1|I1|TS|P2	120	_|G2|D2|M2|Y4|_|I1|TS|P1
80	F3|G2|_|M2|Y4|S1|I1|TS|P1	121	_|G2|D2|M2|Y4|S1|_|TS|P1
81	F3|G2|_|M2|Y4|S1|I1|TS|_	122	_|G2|D2|M2|Y4|_|I1|TS|_
82	F3|G2|_|M2|Y4|_|I1|TS|P4	123	_|G2|D2|M2|Y4|S1|_|TS|_
83	F3|G2|_|M2|Y4|S1|_|TS|P3	124	F3|G2|D2|_|Y4|_|I1|_|P3
84	F3|G2|D2|M2|Y4|_|I1|_|P2	125	_|G2|D2|M2|Y4|_|_|TS|P4
85	F3|G2|D2|M2|_|_|_|_|P4	126	_|G2|D2|M2|Y4|_|_|TS|P3
86	F3|G2|D2|_|Y4|_|_|TS|P1	127	F3|G2|_|M2|Y4|_|_|_|P4
87	F3|G2|D2|_|Y4|_|_|TS|_	128	_|G2|D2|M2|Y4|S1|I1|_|P4
88	F3|G2|_|M2|Y4|_|I1|TS|P3	129	F3|G2|D2|M2|_|_|I1|_|P3
89	F3|G2|_|M2|Y4|S1|_|TS|P2	130	F3|G2|D2|_|_|S1|_|TS|P4
90	F3|G2|_|M2|Y4|S1|_|TS|P1	131	F3|G2|D2|_|_|S1|I1|TS|P3
91	F3|G2|_|M2|Y4|S1|_|TS|_	132	_|G2|D2|M2|Y4|_|_|TS|P2
92	F3|G2|_|M2|Y4|_|_|TS|P4	133	F3|G2|D2|_|_|S1|I1|TS|P2
93	_|G2|D2|M2|Y4|S1|I1|TS|P4	134	F3|G2|D2|M2|_|_|_|_|P3
94	_|G2|D2|M2|Y4|S1|I1|TS|P3	135	_|G2|D2|M2|Y4|_|_|TS|P1
95	F3|G2|_|M2|Y4|_|I1|TS|P2	136	F3|G2|D2|_|Y4|_|_|_|P3
96	F3|G2|_|M2|Y4|_|I1|TS|P1	137	F3|G2|D2|_|_|_|I1|TS|P4
97	F3|G2|_|M2|Y4|_|I1|TS|_	138	F3|G2|D2|_|_|S1|_|TS|P3
98	F3|G2|D2|M2|Y4|_|_|_|P2	139	F3|G2|D2|_|_|S1|I1|_|P4
99	F3|G2|_|M2|Y4|_|_|TS|P3	140	F3|G2|D2|_|_|S1|_|TS|P2
100	_|G2|D2|M2|Y4|S1|I1|TS|P2	141	F3|G2|D2|_|_|S1|I1|TS|P1
101	F3|G2|D2|_|Y4|S1|I1|_|P3	142	F3|G2|D2|_|_|S1|I1|TS|_
102	F3|G2|_|M2|Y4|S1|I1|_|P4	143	F3|G2|D2|M2|Y4|S1|I1|_|P1

continued

Table A1 (continued): Keys used in linking the SAAP NDC, CP and JJ NMDS data sets by order of use

Key order	Key components	Key order	Key components
103	F3|G2|_|M2|Y4|_|_|TS|P2	144	F3|G2|D2|M2|Y4|S1|I1|_|_
104	F3|G2|_|M2|Y4|_|_|TS|P1	145	F3|G2|D2|_|_|_|I1|TS|P3
105	F3|G2|_|M2|Y4|_|_|TS|_	146	F3|G2|D2|_|_|S1|_|_|P4
106	_|G2|D2|M2|Y4|S1|I1|TS|P1	147	F3|G2|D2|_|_|_|I1|TS|P2
107	_|G2|D2|M2|Y4|S1|I1|TS|_	148	F3|G2|_|M2|Y4|S1|I1|_|P3
108	F3|G2|D2|M2|_|S1|I1|_|P3	149	F3|G2|D2|_|_|_|_|TS|P4
109	_|G2|D2|M2|Y4|_|I1|TS|P4	150	F3|G2|D2|_|_|S1|_|TS|P1
110	_|G2|D2|M2|Y4|S1|_|TS|P4	151	F3|G2|D2|_|_|S1|_|TS|_
111	_|G2|D2|M2|Y4|_|I1|TS|P3	152	F3|G2|D2|M2|_|S1|I1|_|P2
153	F3|G2|D2|_|_|_|I1|_|P4	158	F3|G2|_|M2|Y4|S1|_|_|P3
154	F3|G2|D2|_|_|_|_|TS|P3	159	F3|G2|D2|M2|Y4|S1|_|_|P1
155	F3|G2|D2|_|_|_|I1|TS|P1	160	F3|G2|D2|_|Y4|S1|I1|_|P2
156	F3|G2|D2|_|_|_|_|TS|P2	161	F3|G2|_|M2|Y4|_|I1|_|P3
157	F3|G2|D2|_|_|_|I1|TS|_	162	F3|G2|D2|M2|Y4|S1|_|_|_

Glossary

Combination: the result of combining the components required for particular key. For example, the combination for the SLK-581 [F3G2D2M2Y4S1] for Mary Brown born 1 July 1958 is RONAR010719582.

Component: the variables that form a particular linkage key. For example, the SLK-581 has the following components: letters of family name, letters of given name, date of birth and sex.

Data collection: the set of data relating to a specific sector. In this project, there are 3 data collections: the SAAP NDC, the child protection data and the JJ NMDS.

Data set: the set of data relating to a particular state or territory within a data collection. In this project, there are 18 data sets: 8 data sets for the SAAP NDC, 2 data sets for the child protection data and 8 data set sets for the JJ NMDS.

Letters of family name: the combination of the second, third and fifth letters of a person's family name. For example, 'Thompson' results in HOP; 'Brown' results in RON. Non-alphabetic characters (for example, blank spaces or hyphens) are ignored when counting the position of each character. Where a family name contains fewer than five letters, a '2' is substituted for the missing letters.

Letters of given name: the combination of the second and third letters of a person's given name. For example, 'Elizabeth' results in LI; 'Robert' results in OB. Non-alphabetic characters (for example, blank spaces or hyphens) are ignored when counting the position of each character. Where a given name contains fewer than three letters, a '2' is substituted for the missing letters.

Statistical linkage key (or linkage key): a variable that consists of the concatenation of specified components. For example, a linkage key formed from the concatenation of the date of birth component and the sex component would be the linkage key [D2M2Y4S1].

The following components were used in this project:

Component	Abbreviation
Letters of family name	F3
Letters of given name	G2
Day of birth	D2
Month of birth	M2
Year of birth	Y4
Sex	S1
Indigenous status	I1
Suburb/town/locality name	TS
Postcode	P4
Postcode (first 3 digits)	P3
Postcode (first 2 digits)	P2
Postcode (first digit)	P1

References

AIHW (Australian Institute of Health and Welfare) 2008a. Homeless people in SAAP: SAAP National Data Collection annual report 2006-07 Australia. SAAP National Data Collection Agency report series no. 12. Cat. no. HOU 185. Canberra: AIHW.

AIHW 2008b. Linking SAAP, child protection and juvenile justice data collections: a feasibility study. Data linkage series no. 5. Cat. no. CSI 5. Canberra: AIHW.

AIHW 2009. Homeless people in SAAP: SAAP National Data Collection annual report 2007–08, Australia. SAAP National Data Collection Agency report series no. 13. Cat. no. HOU 191. Canberra: AIHW.

AIHW 2010. Government-funded specialist homelessness services: SAAP National Data Collection annual report 2008-09 Australia. SAAP National Data Collection Agency report series. Cat. no. HOU 219. Canberra: AIHW.

AIHW 2011a. 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.

AIHW 2011b. Juvenile justice in Australia 2009–10. Cat. no. JUV 8. Canberra: AIHW.

AIHW 2011c. Pathways in Aged Care: program use after assessment. Data Linkage series no. 10. Cat. no. CSI 10. Canberra: AIHW.

AIHW 2012. Children and young people at risk of social exclusion: linking homelessness, child protection and juvenile justice data. Data linkage series no. 13. Cat. no. CSI 13. Canberra: AIHW.

AIHW: Ryan T, Holmes B & Gibson D 1999. A National Minimum Data Set for Home and Community Care. Aged care series. Cat. no. AGE 13. Canberra: AIHW.

Karmel R, Anderson P, Gibson D, Peut A, Duckett S & 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.
<http://www.biomedcentral.com/1472-6963/10/41>

List of tables

Related publications

This report, *Linking SAAP, child protection and juvenile justice data: technical report,* describes the methods used to link Supported Accommodation Assistance Program, child protection and juvenile justice data. A companion report presents the analysis of the linked data:

- AIHW 2012. Children and young people at risk of social exclusion: links between homelessness, child protection and juvenile justice. Data linkage series no. 13. Cat. no. CSI 13. Canberra: AIHW.

A report exploring the feasibility of linking these three data collections was published in 2008:

- AIHW 2008. Linking SAAP, child protection and juvenile justice data collections: a feasibility study. Data linkage series no. 5. Cat. no. CSI 5. Canberra: AIHW.

The following publications on data linkage may also be of interest:

- AIHW 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.
- 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.

AIHW publications can be downloaded for free from the AIHW website <http://www.aihw.gov.au/publications/>. The website also includes information on ordering printed copies.

Following the release of a study exploring the feasibility of linking three community-sector data collections, the Australian Institute of Health and Welfare was funded to link child protection, juvenile justice and Supported Accommodation Assistance Program data. This report describes the process used to link these collections.

The analysis of the linked data is in a companion report, *Children and young people at risk of social exclusion: links between homelessness, child protection and juvenile justice.*