Comparing name-based and event-based strategies for data linkage

A study linking hospital and residential aged care data for Western Australia

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*.

Please note that as with all statistical reports there is the potential for minor revisions of data in *Comparing name-based and event-based strategies for data linkage: a study linking hospital and residential aged care data for Western Australia* over its life. Please refer to the online version at <www.aihw.gov.au>.

DATA LINKAGE SERIES
Number 3

Comparing name-based and event-based strategies for data linkage

A study linking hospital and residential aged care data for Western Australia

Rosemary Karmel
Australian Institute of Health and Welfare

and

Diana Rosman
Department of Health, Western Australia

2007

Australian Institute of Health and Welfare
Canberra

AIHW cat. no. CSI 3

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 74024 693 4

Suggested citation

AIHW: Karmel R & Rosman DL 2007. Comparing name-based and event-based strategies for data linkage: a study linking hospital and residential aged care data for Western Australia. Data linkage series no. 3. Cat. no. CSI 3. Canberra: Australian Institute of Health and Welfare.

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

Community Services Integration and Linkage Unit
Australian Institute of Health and Welfare
GPO Box 570
Canberra ACT 2601

Phone: (02) 6244 1000

Published by the Australian Institute of Health and Welfare
Printed by

Contents

Acknowledgments

This study could not have been carried out without the vital input of a number of people. The contributions of several people within the Australian Institute of Health and Welfare are especially acknowledged. In particular, the event-based linkage used in the study builds on an idea originally developed by Diane Gibson for linking event data when name data are not available. In addition, Peter Braun put considerable time and effort into preparing the residential aged care datasets used in the data linkage and analysis. Valuable comments on drafts of the report were provided by Phil Anderson and Jonas Lloyd.

We also gratefully acknowledge the staff of the Health Information Linkage Branch (formerly the Data Linkage Unit) in the Western Australian Department of Health on whom the accuracy and completeness of the Western Australian Data Linkage System relies. Within the Health Information Linkage Branch we thank Emma Brook who undertook the linkage to the Residential Aged Care records and Carol Garfield who provided technical support and managed the linkage and extraction processes. John Bass made significant contributions to the project as facilitator, coordinator and custodian of the linkage keys for the cross-jurisdictional arrangements.

Discussions with Dr Terry Neeman, consultant statistician with Covance, and Greg Griffiths concerning the analysis and presentation of comparison results were also very fruitful.

This project was funded primarily by the Australian Institute of Health and Welfare and the Western Australian Department of Health. The Australian Health Ministers' Advisory Council through its Statistical Information Management Committee also contributed funds towards this project.

Symbols in tables

–	Nil or rounded to zero
. .	Not applicable
n.p.	Not publishable

Abbreviations

ABS	Australian Bureau of Statistics
ACAT	Aged Care Assessment Team
ACCMIS	Aged and Community Care Management Information System
ACT	Australian Capital Territory
BSESLA	basic SLA group E linkage strategy
CPC	constrained postcode E linkage strategy
CPC_f	constrained postcode E linkage strategy, includes 2-digit postcode matching
CPC_s	constrained postcode E linkage strategy, not including 2-digit postcode matching
CSLA	constrained SLA group E linkage strategy
$CSLA_f$	constrained SLA group E linkage strategy, includes 2-digit postcode matching
$CSLA_s$	constrained SLA group E linkage strategy, not including 2-digit postcode matching
DOB	date of birth
DoHA	Australian Government Department of Health and Ageing
E linkage	event-based data linkage (name data not used)
HiL	Health Information Linkage Branch, Western Australian Department of Health
ICD-10-AM	International Classification of Diseases, 10th revision, Australian modification
MDC	major diagnostic category
NHMD	National Hospital Morbidity Database
N linkage	name-based data linkage
NSW	New South Wales
NT	Northern Territory
PPV	positive predictive value
Qld	Queensland
RAC	residential aged care
RCS	resident classification scale (RCS1–4 = high care, RCS5–8 = low care)
SA	South Australia
SLA	statistical local area
stat. adm.	statistical admission into hospital episode
stat. sep.	statistical separation from hospital episode
Tas	Tasmania
US	United States of America
Vic	Victoria
WA	Western Australia
WADLS	Western Australian Data Linkage System
YOB	year of birth

Key points

Background to the study:

1. Statistical analysis of movement from hospital to residential aged care would provide a strong evidence base to inform policy development for the hospital–aged care interface. Linking currently available data for such analysis has the advantages of reduced cost, timeliness and no additional data collection imposition.
2. Data linkage of records for individuals is commonly carried out using detailed demographic data, including name and/or a person identification number. While neither name data nor a unique person identifier are available on the national datasets for both the hospital and residential aged care systems for such data linkage, a range of demographic and event data is available.
3. Two earlier projects by the Australian Institute of Health and Welfare investigated the theoretical and practical feasibility of a linkage strategy (termed event-based linkage) which uses demographic data (excluding name) and geographic and event data to link episode records from the hospital and residential aged care national databases.

Current study:

4. The current study directly compares results from linking hospital and residential aged care data using a name-based linkage strategy without specific event information, with those from a group of event-based linkage strategies which did not use any name information. For this comparison, unidirectional hospital-to-residential-care event links were selected, from the full set of annual events in Western Australia, independently using the name-based and event-based linkage strategies.
5. Prior to detailed assessment, the Institute's event-based linkage was refined through initial comparisons with the name-based linkage established in the Western Australian Department of Health, with the results being used to distinguish between effective and poorly performing match tactics, thereby preventing the collection of large quantities of incorrect, or false, links.
6. Within transition destinations, links identified by these refined event-based strategies were highly reliable: overall 98% of links identified by the event-based strategies were true matches when compared with links from the name-based strategy.
7. As expected from earlier studies, the relatively high incidence of missed links (8–14% among preferred strategies) resulted in the event-based linkage strategies underestimating the volume of flow from hospital to residential aged care. There is also the potential for bias if these links relate to a particular subset of the population.
8. However, this study confirmed that the event-based linkage strategies resulted in linked data that were representative of the name-based linked data in terms of the distributions across key variables and within types of transitions. In particular, illustrative examples looking at patterns of use and characteristics of people making particular types of transitions between the two sectors show that analyses using linkage results derived from the name-based and event-based strategies lead to very similar conclusions.
9. Overall, the reliability of the links identified by event-based strategies—especially when movement is known to have occurred, dates are expected to coincide and region information is reliable—suggests that such an approach could be used in other areas. In

particular, although this has not yet been tested, a similar strategy could be used to derive whole-of-stay hospital records for cases when patients move within the hospital sector. In such cases, date of transfer and region information (postcode) should be highly consistent on the before- and after-transfer records.

10. The detailed technical processes and close collaboration required for this project have confirmed that detailed knowledge of both the service systems and the data collection practices within those systems are essential when identifying transition events using any data linkage method.
11. Key findings for event-based linkage are that:
    - Using detailed comparisons of name-based and event-based links, and expanding the event data used to identify links, it has been possible to improve the general performance of the event-based linkage strategy as put forward in the initial feasibility study; in particular, event-based linkage strategies that adjust the linkage procedures according to available information (constrained event-based linkage) perform better than those that apply the same approach to all records (basic 1event-based linkage).
    - Conservative event-based linkage more often tends to miss matches than make false matches, with inconsistent event date and/or region data on the two databases being the main causes of missed links. This sensitivity of event-based linkage varies with both the strategy and transition destination.
    - Within event-based linkage, broader geographic matching can be used for areas with small populations, but not among those with denser populations. Within these limits, methods using broader regions when matching perform better than those using smaller regions.

A broad summary of the methodology and results of the study follows.

Summary

Background

The interface between acute hospital care and residential aged care has long been recognised as an important issue in aged care services research. Despite this, existing national data provide very poor information on the movement of clients between the residential and acute care sectors. Current national datasets on the two sectors are derived from administrative collections, and have been designed primarily to provide information on the specific program which they describe, rather than to provide information on program interfaces or system-level information.

Data linkage is a statistical tool that can be used to link data from different sources, thereby expanding the types of statistical investigations that can be carried out (including analysis of movement over time) without increasing the reporting load of service providers or requiring special surveys. Given the existence of national datasets for the two sectors, this suggests that data linkage could be used to develop datasets suitable for investigating movement between the hospital and aged care sectors.

While neither name nor a common person identification number are available for linking data from the two sectors, some demographic data are available. In addition, information on transition dates, that is entry and exit dates, is available for all periods of hospitalisation and use of residential aged care. To see if such event data are sufficient to allow data linkage, investigations into the feasibility of linking hospital morbidity and residential aged care datasets using a combination of demographic and event data were conducted by the Australian Institute of Health and Welfare (AIHW). Findings from the feasibility study suggested that the set of linked client records resulting from an event-based anonymous linkage strategy could provide a valuable source of information on the client characteristics and service use patterns associated with movements between the two sectors (AIHW 2003).

In the feasibility study, matching was based on date of birth, sex, region of usual residence, and hospital separation and residential aged care (RAC) entry dates. In the absence of validation against a gold standard linkage, doubts concerning the efficacy of the strategy were raised because of the lack of either name or a common person identifier on the two datasets. This issue has been addressed using two distinct methods: first, by employing statistical theory to investigate the effectiveness of the AIHW event-based strategy in a range of linkage situations (AIHW: Karmel 2004); second, by measuring any differences between the two approaches through direct comparisons between a name-based linkage and the event-based linkage. It is these comparative analyses that are the subject of this report. Opportunities to compare such linkages against a 'gold standard' are rare but have been undertaken previously by comparing deliberatively reduced identifiers within fully identified data extracts (for example, Rosman 1996).

In the current analysis, we compare two distinct strategies for linking hospital and RAC events. The first is a name-based process carried out by the Health Information Linkage Branch (HiL) of the Western Australian Department of Health, here termed N linkage. The second is the AIHW event-based linkage strategy—E linkage—which matches event records

using date of birth, sex, geographic region of usual residence and event dates. The scope of the comparisons was limited to movements from hospital to RAC in Western Australia during the financial year 2000–01. Western Australia was chosen for this study because of its unique position in having a well-established data linkage system containing links within and between a number of health data collections which was recently expanded to include links to RAC data.

In summary, the report:

- examines ways to refine the original E linkage strategy
- examines missed and false links made by E linkage (when compared with N linkage)
- compares the E and N links as a whole, looking for similarities and differences in distribution across variables of interest
- compares results from analyses using E and N linked datasets to identify the types of analyses for which the results are comparable and those for which there are differences that need to be taken into account.

Ethical approval for the project was obtained from the AIHW Ethics Committee, and permission to use the relevant data was obtained from both the Steering Committee for Cross-Jurisdictional Linked Administrative Health Data (a joint Australian and Western Australian government committee) and the Confidentiality of Health Information Committee of the Western Australian Department of Health.

Context

The purpose of the data linkage described here was to identify movements of people from hospital into RAC. Before discussing the linkage it is important to understand both the circumstances under which people move between the two sectors, and the data available for linkage and analysis.

Movement between services

When people finish an episode of care in hospital there are a number of possible outcomes: they may transfer within the hospital system, transfer to a residential health care service, go back to their home in the community, go home to RAC, enter RAC (either temporarily for respite care or permanently), or their hospital episode may have ended with their death. Similarly, people entering RAC may be coming from a range of settings: they may be moving from their home into RAC to live either permanently or for respite care, they may be returning after an episode in hospital (commonly while on RAC hospital leave, but possibly also while on social leave from RAC), or they may be entering residential care—either as a permanent admission or for short-term care (a respite care admission)—at the end of a period of hospitalisation (Section 2).

When looking at movement from hospital into RAC, the aim of the data linkage is to link people's exits from hospital with their relevant entries into residential aged care, thereby allowing analysis of patterns of movements between the two sectors.

Data

Datasets were limited to events within Western Australia occurring in 2000–01 (and a few days either side). In addition, as 95% of admissions into RAC are for people aged 65 years and over (AIHW 2002), the data were limited to events for people who were aged at least 65 years by 30 June 2001. Two extracts were obtained for both the hospital and RAC data: one to be used to establish links (a linkage dataset), and a second to be used for analysing link bias (an analysis dataset) (Section 3).

Hospital data

Hospital event data were extracted from the Hospital Morbidity Data System, which is managed and maintained within the Information Collection and Management Directorate of the Western Australian Department of Health. This system supplies similar data for the National Minimum Dataset held by the AIHW. The extract for this project was organised and extracted through the processes established for the Western Australian Cross Jurisdictional Data Linkage Protocol. The extract contained one record per hospital discharge, but with statistical discharges and same-day events removed. These inpatient hospital episodes included both public and private hospital separations for all of the Western Australian population for the period of interest. The hospital data items for linkage and those for analysis were supplied separately.

For the E linkage strategy (carried out by the AIHW) a special extract of hospital separations was derived. This contained only data that are available nationally on the National Hospital Morbidity Database (NHMD) so that the E linkage would reflect results that would be obtained if the strategy were to be applied to link NHMD and RAC data nationally or in other jurisdictions. The hospital linkage dataset contained only the information required for establishing and checking E links, including date of birth, sex, postcode of usual residence, admission and separation dates, and modes of hospital admission and separation.

Two sets of records were excluded from the hospital dataset used for E linkage: hospital episodes that ended with a transfer within the hospital system, and same-day hospital episodes (that is, without staying overnight). The former were excluded as these people do not leave the hospital system at the end of the episode of care, while the latter were excluded to avoid unnecessary erroneous matches as people are unlikely to go from hospital into residential aged care on a single day. For N linkage, it was not necessary to exclude within sector transfers as the person-based linkage allows such transfers to be combined into a single period of hospitalisation.

In 2000–01, in Western Australia there were just over 86,200 hospital separations with a non-transfer discharge that lasted one or more nights. Note that to allow for small gaps between hospital separation and RAC admission (and for small differences in recording dates), separations for a few days either end of the financial year were also included in the dataset for linkage.

To allow detailed comparison of the linked datasets resulting from the N and E strategies, a dataset specifically for analysis was also derived. This analysis dataset contained the information to be used when investigating possible analytical bias, and included age, sex, region of usual residence, country of birth, marital status, modes of hospital admission and separation, length of stay, and hospital diagnosis and procedure variables.

Residential aged care data

The RAC data included all permanent and respite admissions and hospital and social leave events for 2000–01, totalling slightly more than 19,600 events for Western Australia. The data were derived from the Australian Department of Health and Ageing's Aged and Community Care Management Information System (ACCMIS).

As for the hospital data, two extracts were obtained: one for linkage and one for analysis. While the N linkage process is based on established person links for people using hospital and RAC services, the appropriate event links still had to be derived. Therefore, both the N and E linkage used the same RAC event linkage file to establish final event links.

The RAC linkage file, containing data for establishing and checking event links, included date of birth, sex, postcode of usual residence (for leave events this is the RAC facility's postcode), country of birth, marital status, event type, admission and discharge dates, leave event start and end dates, place and date of Aged Care Assessment Team (ACAT) assessment, and mode of discharge. The RAC analysis file, with information to be used when investigating possible analytical bias, contained age, sex, region of residence before current event, region of RAC facility, country of birth, marital status, event dates, care needs, place and date of ACAT assessment, data on the previous RAC admission event (if applicable), and mode of discharge.

The linkage strategies

Name-based (N) linkage

The name-based linkage was undertaken by the Health Information Linkage Branch (HiL) within the Western Australian Department of Health using the two-phase cross-jurisdictional linkage protocol described by Kelman et al. (2002). The linkage process uses as much personal information as is available to create and load links as they are discovered. The links are stored in the Western Australian Data Linkage System (WADLS) which is a dynamic on-line system able to be accessed and updated continuously by linkage staff within HiL (Section 4). Links between a large variety of health-related datasets are established, stored and maintained in this system. Linkages to RAC records for 1990–2003 were performed during 2004. Each link created between a RAC client record and any of the other sources contributing to the WADLS records was loaded into quarantined tables within the WADLS so that access could be managed separately from the main system, but links could be updated at a later stage. For the RAC linkage, demographic information included surname, given names, sex, date of birth and address. Information on the dates of events within the RAC was not available for linkage. Separate linkages between RAC and Western Australian electoral, ambulance, hospital, emergency and death records were performed.

Using HiL's hospital–RAC person links, all relevant hospital and RAC event records were retrieved from their respective data custodians and then matched to form combined strings of hospital events and RAC events for the same person. The most appropriate hospital–RAC event link was then chosen by measuring the closeness of hospital and RAC event dates. Where there was a choice between matches to different RAC event types, matches to RAC hospital leave had priority over matches to admissions.

Overall, the N strategy resulted in 8,106 links between hospital separations and entry (or re-entry) into RAC.

Event-based (E) linkage

The event-based E linkage strategies link records by using limited demographic information in conjunction with event dates (Section 5). Additional data on event characteristics can also be used. In this study, the information used to establish links included: date of birth, sex, postcode of usual residence, postcode of hospital, episode start and end dates, hospital episode admission and separation modes, and time and place of assessment for entry into RAC. The resulting linked dataset varies depending on whether all this information is used when identifying links, and how particular data items are used to specify the linkage process.

Overall, for this project three types of E linkage were investigated:

- constrained matching within statistical local area (SLA) group (abbreviated to CSLA)
- constrained matching within postcode (CPC)
- basic matching within SLA group (BSESLA).

To ensure that the privacy of individuals was maintained, the E linkage was carried out within the AIHW using the Institute's protocol for protecting privacy when carrying out data linkage, and data were protected according to standard AIHW procedures (AIHW 2006).

Constrained E matching

The purpose of constrained E matching is to find the best match using all available event date information and event descriptors. To achieve this, matching procedures are tailored specifically for comparisons between different subsets of RAC and hospital events defined in terms of their type and/or admission and separation characteristics. Matches are then established using date of birth, sex, region, and event dates. Knowledge of both the service systems and data collections is used to determine the specific comparisons, with the specific matching processes varying depending on: whether a hospital episode began with a within-sector transfer, the patient's reported destination after discharge from hospital, the type of RAC event (admission or leave event), and the place and time of ACAT assessment for a RAC admission. Because two dates are available for RAC hospital leave, and the related hospital episode may end in a number of ways, match procedures for these events are the most complicated (see Appendix 2).

Constrained E matching is carried out in two stages. Initial matches are selected using 1-to-1 probability matching. Relatively broad match criteria are used to identify possible matches between particular subsets of RAC and hospital data, with the selected RAC record match for a particular hospital record being that with the highest probability of matching the hospital record given the data used for matching; for this project, 12 hospital–RAC dataset pairs were used and each pair was compared using up to seven distinct match passes. Some variation in date of birth, region and event dates is allowed when establishing matches. Finer match rules are then applied to select the final matches, using, first, deterministic rules to exclude poor matches and then specified match priorities to choose between duplicate links.

Two variants of constrained E linkage were considered, differing only in the size of the region used when establishing links. CSLA linkage is based on matching within SLA groups, where an SLA group is that set of SLAs that overlaps a postcode and two postcodes are said to match if they have a common SLA in their SLA groups. CPC linkage restricts matching to regions based on the four digits of the reported postcode, and so generally uses smaller match regions than CSLA.

Before refining the constrained strategies using information from initial comparisons with N linkage, the CSLA and CPC strategies resulted in 7,802 and 7,781 linked events, respectively.

Basic E matching

In addition to constrained linkage, basic E linkage was used. Under this strategy, the same linkage process is used for all hospital and RAC events, irrespective of their admission and separation characteristics. As a consequence, match data are limited to information available for all events, namely date of birth, sex, postcode of usual residence and a single movement date (that is, hospital separation matching a RAC entry date). Matches are selected using 1-to-1 probability matching, although only minor variation from exact matching is allowed (in event dates).

Basic matching using SLA group as the match region (BSESLA) was used in the initial feasibility study of hospital–RAC event-based linkage (AIHW 2003). It is therefore important to compare the results from this approach with those obtained using the person-based linkage. It is also useful to compare the effectiveness of the BSESLA strategy with that of the constrained approaches to measure the gains in moving from a simple strategy to a more complex one.

Using BSESLA linkage resulted in 6,693 event matches.

Comparing linkage strategies

When linking records, four outcomes are possible: a true link, no link, a false link (false positive) and a missed link (false negative). The correspondence between two strategies can be gauged by seeing how many of the links are the same and how many are different. While some constraints were imposed on the HiL by the RAC data provider when linking Western Australia RAC and health data, the use of name and address in N linkage, and the availability of name and address reporting history across a range of health service events, results in this linkage being highly reliable. Consequently, in this study it was used as the reference standard against which the E linkage results were compared, that is, to determine whether an E link was 'true' or 'false'. Linkage checks indicate a very low error rate in the established WADLS.

For this project, two key measures were used when comparing matches:

- Positive predictive value (PPV): the percentage of E links that are true links
  = E true links/E links
- Sensitivity: the percentage of all links that are identified by the E linkage strategy
  = E true links/N links.

Using N linkage as the reference standard, N-only links represent those links missed by E linkage, while E-only links represent false matches made under the strategy.

Refining constrained E linkage

As stated above, constrained E linkage was carried out by matching between 12 hospital–RAC dataset pairs, with up to seven different (and pair-specific) match criteria being applied to each pair. The reliability of a particular match criterion can be gauged by comparing the resulting E links with N links. In particular, the PPV of E matches made via specific match

criteria can be used to detect particular match passes that result in unacceptably high levels of false matches. Such passes can then be dropped from the constrained E linkage strategy without compromising the reproducibility of the linkage process for other jurisdictions (Section 6).

To this end, poorly performing matching algorithms used in the constrained E linkage were identified using the condition that the underlying PPV should be above 60% with 95% probability. As a result, a number of match criteria were tightened or dropped altogether. In addition, analysis of the distribution of the population across geographic areas indicated that allowing substantial variation in the region match by relaxing the match regions to those defined by the first two digits of a postcode could lead to relatively low PPVs for some of the resulting regions in the more populous states.

After refining the constrained strategies as above, the CSLA strategy resulted in identifying 7,595 transitions from hospital into RAC when broad region matching was allowed (that is, matching within regions identified by the first two digits of the person's postcode) and 7,253 when it was excluded. For CPC linkage, the corresponding numbers were 7,587 and 7,078 respectively. For both the SLA- and postcode-based constrained linkage approaches, removing match passes that lead to unacceptably high false match rates improved their overall PPV by two percentage points (up from slightly under 96% to almost 98% for both strategies).

Missed links

After refining the constrained strategies, compared with N linkage the E strategies still missed some links and falsely identified others, with missed links considerably outnumbering false links (688 versus 177 for CSLA linkage with broad region matching—see Table S.1). Poor region matching was the main reason for missing links to RAC admissions under the CSLA strategy. This was a less important, but still significant, reason for missing matches among RAC leave events. Because of the very limited geographic data available on the hospital and RAC datasets for use in E linkage, matches missed due to poor region matching on the two datasets cannot be retrieved by adjusting the E linkage strategy.

Missed matches to RAC leave events were primarily the result of poorly matching event dates for related hospital and RAC events. The large gaps between the recorded dates for the linked events in these missed matches (commonly 3 or more days) indicate that it would not be possible to adjust the CSLA matching strategy to allow capture of these matches without risking the introduction of large numbers of false matches.

False matches

The majority of false E matches is for admissions: for CSLA almost two-thirds of the small number of false matches were for admissions, with another third being for RAC hospital leave events and a small number relating to RAC social leave.

To ensure that CSLA false links were not the result of person links being missed by N linkage, the reliability of the person links on which N linkage was based was re-examined by reviewing HiL's person links related to the CSLA-only links. Investigations by HiL of the client links implied by the CSLA-only links led to HiL identifying just two additional person links between hospital and RAC clients that had previously been missed.

Many CSLA-only links have exact matches on both date of birth and event dates. This suggests these false matches are caused by similar hospital and/or RAC activity by similar people (in terms of date of birth and sex) living in a particular region. Comparisons indicate

that one of the most effective ways of reducing the number of false matches made under the CSLA strategy would be to reduce the size of the geographic region used in matching. Analysis of CSLA-only links indicates that this would be more effective for links to RAC admissions than for links to RAC leave. However, overly narrowing the geographic matching criteria could result in dropping many more true links than false links (see below).

Quality of E linked datasets

There are five E linkage strategies that could be used in different situations—depending on the regional distribution of the population under study and the human resources available to undertake the data linkage:

- $CSLA_s$—constrained SLA group linkage strategy, not including matching within expanded regions defined by the first two digits of the postcode (termed 2-digit postcode matching)
- $CSLA_f$—constrained SLA group linkage strategy, including 2-digit postcode matching
- CPC_s—constrained postcode linkage strategy, not including 2-digit postcode matching
- CPC_f—constrained postcode linkage strategy, including 2-digit postcode matching
- BSESLA—basic SLA group linkage strategy.

These five strategies resulted in very similar total positive predictive values but a range of sensitivities (Table S.1) (Section 7): overall, both the refined constrained and basic E linkage strategies had PPVs of around 98%. Excluding matching on 2-digit postcode had very little effect on the PPV of the constrained strategies. However, disallowing such matches noticeably reduced the total number of matches (by 342 and 509 matches for CSLA and CPC, respectively) and consequently reduced the sensitivity of the strategies—from 91.5% down to 87.6% for CSLA, with a similar shift for CPC. Nonetheless, the sensitivity of the constrained strategies, both with and without 2-digit postcode matching, was considerably higher than that for BSESLA: the constrained E linkages resulted in over 5% more links than the basic strategy.

If all the E linkage strategies result in linked datasets that are equally representative of movements of people from hospital to residential aged care, the above results indicate that if 2-digit postcode matching can be included there is little to choose between the $CSLA_f$ and CPC_f strategies. In this case, since straight postcode matching is much easier both to understand and carry out than SLA group matching, the CPC_f strategy would be preferred over the $CSLA_f$ strategy. However, if 2-digit postcode matching cannot be included, the $CSLA_s$ strategy would be preferred because of its greater sensitivity while having almost the same PPV as CPC_s matching. Because of its relatively poor sensitivity, BSESLA matching would only be used if there were insufficient resources to undertake the more complex constrained matching. Analyses suggest that in this case basic matching using 3-digit postcode could provide similar results without introducing the complication of SLA group matching (see Section 5).

Table S.1: Positive predictive value and sensitivity of event-based E linkage strategies, using name-based N linkage as the reference standard

Match strategy		True links (A)	False links (B)	Missed links (C)	Total links (D = A+B)	PPV (A/D)	Sensitivity (A/F)	Relative size (D/F)
			Number				Per cent	
Name-based N linkage	**(F)**	**8,106**	. .	. .	. .	. .	. .	**100.0**
Event-based E linkage								
$CSLA_s$		7,100	153	1,006	7,253	97.9	87.6	89.5
$CSLA_f$		7,418	177	688	7,595	97.7	91.5	93.7
CPC_s		6,936	142	1,170	7,078	98.0	85.6	87.3
CPC_f		7,418	169	688	7,587	97.8	91.5	93.6
BSESLA		6,539	154	1,567	6,693	97.7	80.7	82.6

Source: Table 6.17.

The similar total PPVs but varying sensitivities of the E linkage strategies raises the question of whether differences in their linked datasets are in size only, as the consistent PPVs suggest, or whether there are some underlying distributional differences in the matches made under the various strategies.

Detailed analyses of the various sets of links show that the PPV and, in particular, the sensitivity of E links vary with RAC event type (Table S.2). Differences were also apparent for categories within a range of other variables. However, much of this variation can be explained by the mix of RAC event types within particular groups. When modelling the propensity of E linkage to miss N links within RAC event type, only a small number of variables were found to have statistically significant effects, with more variables being identified for the less sensitive basic linkage than for constrained linkage.

In practice, the dominant effect of RAC event type on the efficiency of E linkage is largely mitigated by the logical requirement of separate examination of different transition types, that is, movement into permanent admissions, respite admissions, RAC leave returns and the community. Furthermore, given that the E linkage strategies are much more likely to miss N links than make false matches, analysis indicates that E links provide a good basis for examining the demographic profile of people undertaking various types of transitions. There was some evidence, however, that there could be some small regional and/or hospital episode differences in the profile of N and E links for RAC admissions and RAC leave events. Such differences are likely to be greater for basic than for constrained E links.

Noting the above, the question then arises as to whether differences in the profiles of N and E links affect the utility of the E linked datasets for looking at movements between sectors.

Table S.2: Positive predictive value, sensitivity and relative size, by RAC event type and E linkage strategy, using name-based N linkage as the reference standard

	RAC event type				
Match strategy	**Permanent admission**	**Respite admission**	**Hospital leave**	**Social leave**	**Total**
Event-based E linkage strategy			**PPV (per cent)**		
$CSLA_s$	95.1	98.2	98.7	95.5	97.9
$CSLA_f$	94.8	97.4	98.7	95.5	97.7
CPC_s	95.0	98.0	98.7	99.3	98.0
CPC_f	95.0	97.2	98.7	99.3	97.8
BSESLA	95.5	97.8	98.5	89.2	97.7
			Sensitivity (per cent)		
$CSLA_s$	75.0	81.1	92.5	91.3	87.6
$CSLA_f$	88.5	86.4	93.3	91.3	91.5
CPC_s	69.5	78.5	91.7	90.7	85.6
CPC_f	88.9	87.0	93.1	90.7	91.5
BSESLA	65.2	73.8	86.4	91.9	80.7
			Relative size (per cent)		
$CSLA_s$	78.9	82.6	93.8	95.7	89.5
$CSLA_f$	93.3	88.7	94.6	95.7	93.7
CPC_s	73.2	80.2	92.9	91.3	87.3
CPC_f	93.6	89.4	94.3	91.3	93.6
BSESLA	68.3	75.5	87.7	103.1	82.6
Name-based N linkage (number)[(a)]	**1,723**	**852**	**5,370**	**161**	**8,106**

(a) Analysis indicated that for a small number of links the event match chosen by the N linkage strategy was not the preferred link. In particular, for 18 matches (0.2% of N links) the preferred link was to a RAC hospital leave event rather than the chosen (earlier) admission event for the same person.

Source: Table 7.3.

Utility of E linked data for analysis

When undertaking analysis of transitions, it is the combined effect of missed N links and false E links that determines the overall utility of an E-linked dataset. Examination of all records linked under the particular strategies shows that, overall, the E linkage strategies resulted in linked data that largely reflected the N linkage match set in terms of the distributions across key variables (Section 8). That is, while not exactly the same, the E linkage match sets' distributions looked highly similar to those for the N linkage match set. In this respect, constrained methods using SLA group when matching performed better than those using straight postcode, and constrained matching (even without using 2-digit postcode matching) performed at least as well as or better than basic E matching.

Many analyses of movement between hospital and RAC will want to compare people who have moved between the two sectors with those who have not. In this case, both linked and unlinked records are examined. In this study, examples of such analyses considered three

broad groups of analysis: movement from hospital, movement into RAC, and an example looking at a specific issue—dementia. In all cases, analysis was carried out taking into account the type of transition into RAC, thereby removing one of the greatest sources of possible bias identified in the earlier analyses. Comparisons were limited to the three E linkage strategies $CSLA_s$, CPC_s and BSESLA. The 2-digit postcode E linkage strategies were not considered as they could not be used for national analysis due to the high population density in some regions in the larger states.

In terms of practical utility, analysis by post-hospital destination or source of RAC admission indicated that, as expected from the sensitivity estimates, the E linkage strategies underestimate the volume of movement between hospital and RAC, with permanent RAC admissions being particularly affected (see relative size in Table S.2). Nevertheless, illustrative examples looking at patterns of use and characteristics of people moving between the two sectors—such as those in Tables S.3 and S.4—show that analyses using links derived from the N and E strategies lead to very similar conclusions. Examination of results also indicate that, irrespective of the linkage strategy, care needs to be taken when drawing conclusions as some differences may not be statistically different due to small numbers in some cross-classifications.

Table S.3: Analysis example 1: Median length of hospital episode, by transition type and sex, for name-based N linkage and event-based $CSLA_s$ linkage, separations for people aged 65 years and over, 2000–01

	Name-based N linkage			Event-based $CSLA_s$ E linkage		
Movement type	**Males**	**Females**	**All**	**Males**	**Females**	**All**
	Median length of hospital episode (days)					
Returning to permanent RAC[(a)]	6	6	6	6	6	6
Hospital to permanent RAC[(a)]	33	31	32	32	29	30
Hospital to respite RAC[(a)]	16	14.5	15	16	15	15
Hospital to community/other[(b)]	4	4	4	4	4	4
Died in hospital[(b)]	8	8	8	8	8	8

(a) Based on linked hospital and RAC records. See also notes (a) and (b) to Table 8.4 for additional information.

(b) Unlinked hospital separations. Deaths are based on reported hospital mode of separation.

Note: Table excludes same-day hospital episodes, statistical discharges and transfers to other hospitals. Length of stay excludes days on leave from hospital.

Source: Table 8.6.

Table S.4: Analysis example 2: Care level and dementia status for RAC entries, by transition type, for name-based N linkage and event-based CSLA$_s$ linkage, 2000–01

	Name-based N linkage					Event-based CSLA$_s$ E linkage				
	Care level					Care level				
Transition type	High	Low	All	Total		High	Low	All	Total	
	Row %			Col. %	N	Row %			Col. %	N
Return from hospital[(a)]										
With dementia	64.5	35.5	100.0	31.0	1,711	64.9	35.1	100.0	30.7	1,583
Without dementia	33.6	66.4	100.0	69.0	3,802	33.3	66.7	100.0	69.3	3,575
All	*43.2*	*56.8*	*100.0*	*100.0*	*5,513*	*43.0*	*57.0*	*100.0*	*100.0*	*5,158*
Into permanent RAC from hospital[(a)]										
With dementia	83.1	16.9	100.0	45.9	769	83.9	16.1	100.0	45.9	598
Without dementia	72.6	27.4	100.0	54.1	905	74.5	25.5	100.0	54.1	705
All	*77.4*	*22.6*	*100.0*	*100.0*	*1,674*	*78.8*	*21.2*	*100.0*	*100.0*	*1,303*
Into respite RAC from hospital[(a)]										
With dementia	32.5	67.5	100.0	24.6	209	32.4	67.6	100.0	24.7	173
Without dementia	19.1	80.9	100.0	75.4	640	19.6	80.4	100.0	75.3	526
All	*22.4*	*77.6*	*100.0*	*100.0*	*849*	*22.7*	*77.3*	*100.0*	*100.0*	*699*

(a) Based on linked hospital and RAC records. See also notes (a) and (b) to Table 8.4 for additional information.

Note: Diagnosis of dementia includes diagnoses of dementia and Alzheimer's disease (ICD-10-AM Ed. 1 categories F00–F03, and G30—see NCCH 1998). Table excludes 115 cases with missing RCS, and all unlinked RAC hospital leave events.

Source: Table 8.17.