Statistical linkage keys: How effective are they?

John Bass

Centre for Health Informatics, Curtin University, Western Australia

Carol Garfield

Data Linkage Unit, Department of Health, Western Australia

Abstract

Existing measures of statistical linkage keys (SLKs) have usually focussed on how well a key represents the source population and on the extent of duplication ie multiple keys for one individual as well as multiple individuals sharing the same key. We really need to know whether the analysis of data linked by deterministic matching of SLKs leads to significantly different conclusions than would be obtained through analysis of "real" linked data. A project using information from the Western Australian Data Linkage Project has been able to provide some answers to this question.

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

The primary aim of the study is to compare the results of typical analyses of linked data from the same set of hospital and death records linked by means of the HACC and SAAP SLKs as well as the WA PID. The effects of increasing the time period and of indigenous status have been examined. Two analyses are presented here – the total number of bed days per patient, and the relative risk of death. The results are significant, indicating that the linkage method should be taken into account when interpreting the results of analyses of linked health data.

Introduction

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

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

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

- Errors in the source information leading to the generation of multiple keys for one individual, e.g. when a surname is misspelt ("Smith"/"Smythe") or when there is a name change (as often occurs at marriage or divorce); and
- Multiple clients sharing similar identifying information leading to the construction of a single linkage key.

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

Existing effectiveness measures of HACC and SAAP SLKs

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

Two SAAP collections made in 1998–1999 and 1999–2000 reported 25% and 21% client refusals with a further 3.5% and 2.5% missing due to insufficient data. Estimates of duplication

rates ranged from 3.3% to 5% (AIHW, 2000a). These estimates were within a level of accuracy acceptable to the SAAP Data and Research Advisory Committee (DRAC).

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

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

Current measures of the effectiveness of SLKs

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

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

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

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

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

Duplication rates for HACC and SAAP SLKs

Duplication rates of the HACC and SAAP keys in the WA study are summarised in Table 1. For each key, the first row

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

Table 1 Duplication rates of HACC and SAAP keys compared to WA PID

shows the percentage frequency of multiple HACC keys for one individual (i.e. one WA PID) while the second row shows the percentage frequency of more than one individual sharing one HACC key. The third row shows the ratio of these two percentages while the fourth row shows their sum.

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

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

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

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

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

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

Number of days in hospital

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

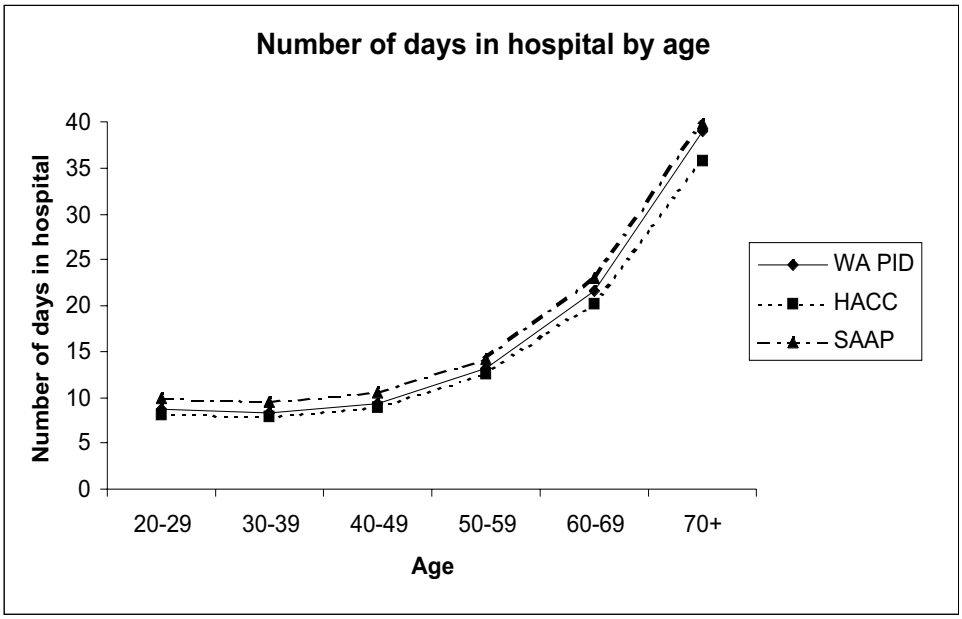

Figure 1 Number of days in hospital by age group according to data linked by HACC and SAAP keys and the WA PID

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

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

These differences may be significant, but are they large enough to make an impact in practical applications? Table 2 shows the average number of days in hospital by age group (together with the 95% confidence limits) for the WA PID, HACC and SAAP linkages.

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

Table 2 Average number of days in hospital by age group (with 95% confidence limits)

Table 2 shows that, in the example of the 70+ years age group, the average number of days in hospital per patient according to the WA PID is 39.0 compared with 35.8 days for data linked by the HACC key and 39.8 days for data linked by the SAAP key. The 95% confidence limits for the average number of days according to the WA PID range from 38.5 to 39.4. This means that we can be 95% certain that the true value of the average (estimated at 39.0) occurs in this range.

Table 3 displays the percentage differences between the average number of days according to the WA PID and HACC keys, the WA PID and SAAP keys, and the HACC and SAAP keys, also indicating which comparisons are significantly different at the 95% confidence level.

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

Table 3 Percentage difference of days in hospital by age group (* = 95% significant)

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

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

What about the expectation that precision would be decreased in data linked by the SLKs? Table 4 shows the standard errors of the average values in Table 2.

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

Table 4 Standard errors of average values in Table 2

The standard errors of the average values do not vary greatly or in a consistent pattern. The HACC averages are generally slightly more precise than the WA PID averages, with the SAAP linkage showing a slightly larger variance. If the data in Table 4 are normalised to remove the effect of differences in the average values, then the WA PID and HACC standard errors are virtually identical with the SAAP data displaying a consistent small (and not significant) increase.

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

Indigenous status

Linkage of data from persons of indigenous Australian descent is often more difficult as compared to linkage of other cultural groups, with frequent name changes and relatively poor recording of dates of birth and other demographic details. Tables 5 through 7 show the results of an analysis of the number of days in hospital per patient by indigenous status rather than by age group.

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

Table 5 Average number of days in hospital by indigenous status (with 95% confidence limits)

The results in Table 5 show that the estimates of number of days in hospital per indigenous patient covered a wide range from just under 28 (WA PID) through about 26 (SAAP) to just over 22 (HACC). The significance and extent of these differences are summarised in Table 6.

Indigenous status	WA PID -> HACC	WA PID -> SAAP	HACC -> SAAP
Not indigenous	–5.7 *	10.6 *	17.2 *
Indigenous	–20.4 *	–5.3	18.9 *
Total	–6.1 *	10.1 *	17.3 *

Table 6 Percentage difference of days in hospital by indigenous status (* = 95% significant)

Indigenous status	WA PID	HACC	SAAP
Not indigenous	0.05	0.05	0.06
Indigenous	0.06	0.06	0.07
Total	0.08	0.07	0.09

Table 7 Standard errors of average values in Table 5

Tables 6 and 7 show a similar pattern in the analysis for indigenous status as that shown by the analysis for age groups, with significant differences between the average values (except for the WA PID/SAAP figures for indigenous patients) and virtually constant precision.

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

Relative risk of death

The quality of the death data linkages was investigated by performing a Cox regression for the WA PID, HACC and SAAP linked data sets to show the relative risk of death by age group, sex and indigenous status. Details of this analysis are provided in Figure 4 and Tables 8, 9 and 10.

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

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

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

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

Table 8 Relative risk of death by age group compared to 20–29 year olds; males compared to females; and indigenous Australians compared to non-indigenous patients (with 95% confidence limits)

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

Table 9 Percentage difference of relative risk of death (* = 95% significant)

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

Table 10 Standard errors of average values in Table 8

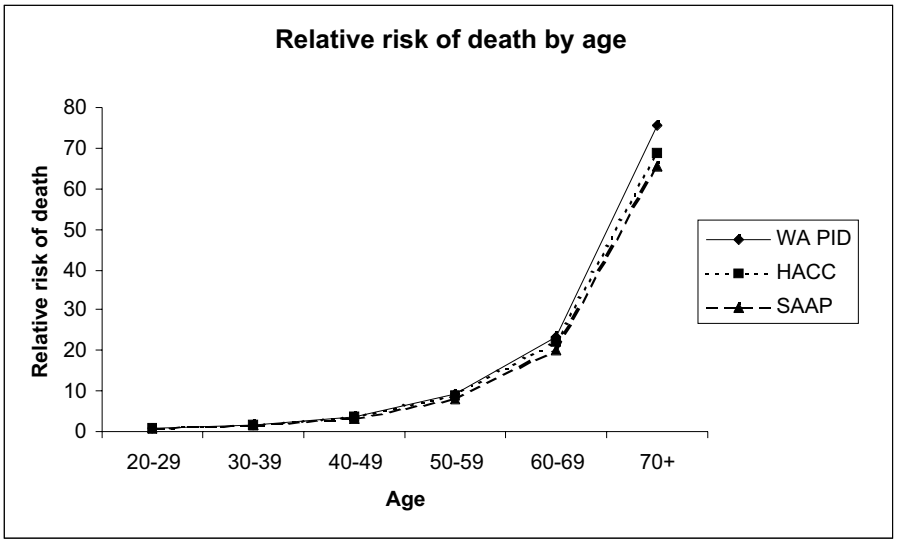

Figure 2 Relative risk of death by age group for data linked by HACC and SAAP keys and the WA PID

Table 10 shows a marked increase in standard errors with increase in age group. This reflects the sharp increase in risk of death among older patients. For the WA PID linkage, the relative risk of death increases by a factor of 45 from the 30–39 age group to the 70+ age group, while the standard error increases by a factor of 35. Taking the increase in risk into account, there is therefore only a small increase in variance among the values for the older patients.

Conclusions

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

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

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

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

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

References

Australian Institute of Health and Welfare 2000a. SAAP National Data Collection Annual Report 1998–99 Australia, SAAP NDCA Report Series 4, AIHW Cat. No. HOU 38, Canberra.

Australian Institute of Health and Welfare 2000b. The use of linkage keys for statistical work in community services: Background paper for the Statistical Linkage Project of the National Community Services Information Management Group, AIHW.

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

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

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