Measuring data and link quality in a dynamic multi-set linkage system

Diana Rosman, Carol Garfield, Stuart Fuller, Alexia Stoney, Todd Owen and Geoff Gawthorne
Data Linkage Unit, Department of Health (WA)

Abstract

The international quality management standard ISO 9001:2000 consists of a generic set of guidelines to be applied in any industry. In Health Information Management both product (information) and service (information delivery) aspects need to be considered.

The WA Data Linkage System comprises 3.7 million chains of linked records drawn from six population-based data sets and is updated monthly. In addition, links to six additional data sets are updated six monthly and ad-hoc linkages to special cohorts also occur. Although matching algorithms and the levels of clerical review have been tuned to minimise mismatches and missed matches, the system also needs to be efficient to achieve a quality product and a quality service.

In probabilistic record linkage, variations in the values of key variables are tolerated in order to achieve optimum linkage. Consequently, among a group of linked records, variations in the values of key variables may be due to inherent data errors or may signal that incorrect links have been made.

Elsewhere link quality has been assessed by theoretical or empirical methods. Results have been measured against an expected linkage rate or compared with a 'gold' (or even a 'platinum') standard. However, these methods are problematic in a dynamic multi-set linkage system as there is no benchmark for comparison.

In January 2001, a random sample of 5,000 records for admissions to hospital during 1990 was selected. Complete chains of links for each of these were extracted for manual review. It was determined that 28 (0.6%) of the 4854 distinct chains contained one or more records that had been linked incorrectly. One year later the task was repeated and errors were discovered in just 15 (0.3%) of the 4868 selected chains.

This paper proposes a system for on-going monitoring of the quality of the links with rules based on those used by the six expert linkage operators during the two pilot manual linkage audits.

Background

Probabilistic record linkage permits links to be made between records even though there may be variations in the demographic details used in the matching process. Expert judgement is required in developing the linkage strategy, writing the rules and setting the thresholds for acceptance or rejection of potential links. Attention to detail is necessary during the clerical review of doubtful links, particularly where the decision to accept or reject a given link may be based on conflicting information drawn from several data sources over an extended period.

The WA Data Linkage System has been continually refined since its inception in 1995. It now contains 3.7 million linked chains of records drawn from six population based data sources spanning 30 years. In 1999, an estimate of the quality of the links in the system was made using a weighted sample of pairs of potentially matched records and detailed manual checking of pairs with weights between an upper threshold with ~100% valid links and a lower threshold with ~0% valid links. At that time, it was found that the level of false positives (mismatches) was 0.11% and the level of false negatives (missed matches) was also 0.11% (Holman et al, 1999).

Since that time, WA electoral roll records have been incorporated with a resultant improvement in linkage from the confirmation of name and address changes over time. In addition, linkage to midwives notifications has assisted in providing information about name changes through marriage.

A substantial improvement in the timeliness of the system has also occurred with the introduction of monthly updates of linkages to hospital admission and death records. However, the latter development has the potential to temporarily reduce the overall quality of links as each month is initially linked against the most recent 10-year period. Complete 30-year internal linkages are only reviewed at the end of a six-month period.

Elsewhere link quality is often measured against a known standard, eg linkages performed with a reduced set of personal identifiers (ie, name, gender, date of birth and address) may be measured against a linkage using the complete set (Rosman, 1996). However, there is no appropriate benchmark for the WA system, which is considered to be a 'platinum' standard for other linkage systems, and a determination of link quality must rely on detailed (manual) scrutiny of sampled linked chains.

Aim

To examine the factors contributing to linkage errors (mismatches and missed matches) in a dynamic multi-set linkage system.

Motivation

The international quality management standard ISO 9001:2000 recommends that any quality assessment of a system should 'say what it does, do what it says, prove it and improve

it'. It also recommends that all employees should be involved in the assessment of their work.

An assessment of the 'quality' of an information system needs to consider product (information) and service (information delivery) aspects. That is, the accuracy, completeness and timeliness of the information within the system, as well as the delivery of information from the system, need to be considered. For a data linkage system based on demographic information recorded in health records, the accuracy, completeness and timeliness of the links must be assessed separately from the accuracy, completeness and timeliness of the underlying demographic information on which the links are based. Nevertheless, links based on inaccurate or incomplete information are more likely to be problematic.

Link quality assessment

In order to estimate the level of false positive matches (ie, 'mismatches') in the system, a random sample of 5,000 (~1%) hospital admission records for 1990 was extracted in January 2001. All links related to these 5,000 records were also extracted for manual review. Among the 5,000 records, there were 4,854 distinct chains comprising ~180,000 records over the period 1970 to 2000. A detailed clerical assessment was performed by the six members of the WA Data Linkage team. Each team member was responsible for checking approximately 30,000 lines for possible linkage errors. All questionable linkages were set aside and these were discussed by team members at the end of the manual checking process. It was decided that 28 (0.6%) chains contained mismatches.

During the following 12 month period, several new datasets were linked to the system. These included the WA electoral roll, midwives notifications, the MBS/PBS data for Diabetes, the records of the Disability Services Commission and a 1987 survey of aboriginal residents of the Kimberley region. As new information was drawn into the system, many doubtful links involving name and address changes over time were resolved.

The quality assessment process was then repeated in January 2002 with another 5000 randomly chosen records from 1990 hospital admissions (~1%). Among these 5,000 records, there were 4,868 distinct chains consisting of ~180,000 records. After manual review and discussion it was agreed that 15 (0.3%) of the chains contained mismatches.

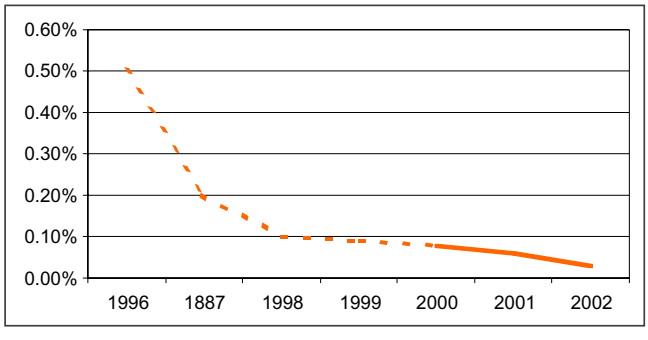

Figure 1 Change in linkage error rate over time

Figure 1 illustrates the expected reduction in error (or improvement in quality) of the data linkage system over time with the dotted line showing the suggested improvement between 1995 and 1999. The actual estimates for 1999, 2001 and 2002 are depicted by the solid line.

Detection of linkage errors

The manual review process took about six person weeks to complete, delaying normal system processing. It was evident that a more targeted approach was needed to detect and correct the small number of linkage errors.

As a first step, an assessment of the extent of variation in demographic information within the linked chains was performed. It was assumed that linkage errors would be more likely to occur where there were conflicting variations in name, date of birth or address over an extended period across several datasets. All linked chains containing a record for admission to hospital in 1990 were selected and the frequency of variations in the values of sex, date of birth and postcode are given in Table 1.

	1	2	>2
gender	236,686	5,242 (2.2%)	
date of birth	203,363	30,190 (12.5%)	8,375 (3.5%)
day	222,870	16,940 (7.0%)	2,118 (1.0%)
month	229,693	11,155 (4.6%)	1,180 (0.5%)
year	222,432	15,028 (6.2%)	2,468 (1.0%)
postcode	100,896	70,878 (29.3%)	70,154 (29.0%)

Table 1 Demographic variants within linked chains

About 15% of the linked chains contained more than one value for the date of birth, with 3.5% having more than 2 values. Most of these variations occurred in the day or the year of birth. As expected given the time span of many of the records, variations in postcode were common, with only 30% of chains having one postcode throughout.

Since variations in demographic variables would be expected in chains containing a large number of records and/or spanning a long period of time, these factors were investigated first. Figures 2 and 3, respectively, show the distributions of the number of records and the time span of the linked chains containing 1990 admissions. The maximum possible time span for morbidity records was 32 years (1969 to 2001).

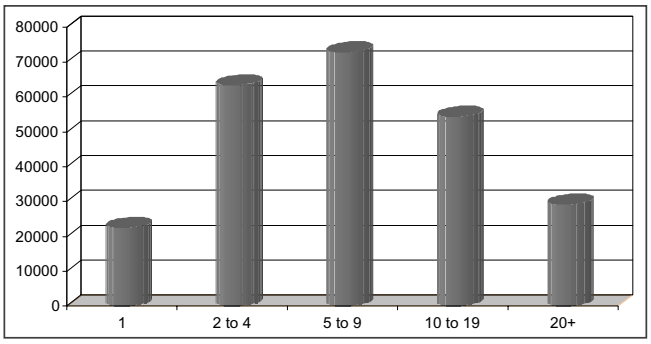

Figure 2 Distribution of number of records in linked chains based on 1990 admissions

As can be seen from Figures 2 and 3, a large proportion of the chains contained more than 10 records and spanned more than 10 years.

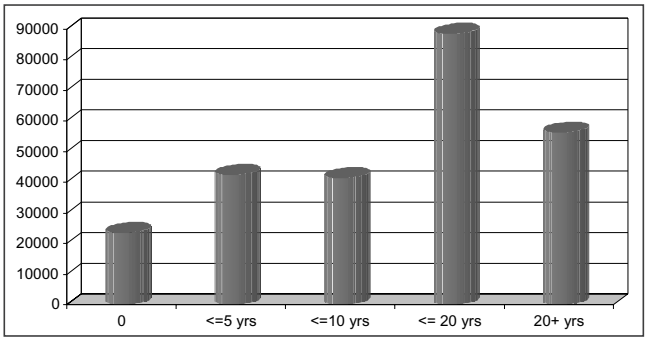

Figure 3 Distribution of the time span within linked chains based on 1990 records

Although variations in date of birth were expected for long chains (ie, many records) covering a long period of time, other factors were of interest. A generalised linear model (GLM) was constructed to model the number of different values for the date of birth (variants) with the number of records, the time span, the number of different postcodes and the year of birth as explanatory factors.

Of the 10% of variation in the number of values of date of birth accounted for by the GLM model, 3% could be attributed to the year of birth and 10% to the number of different postcodes. As expected, most of the variation was attributable to the number of records (27%) and the time span (60%) of each chain.

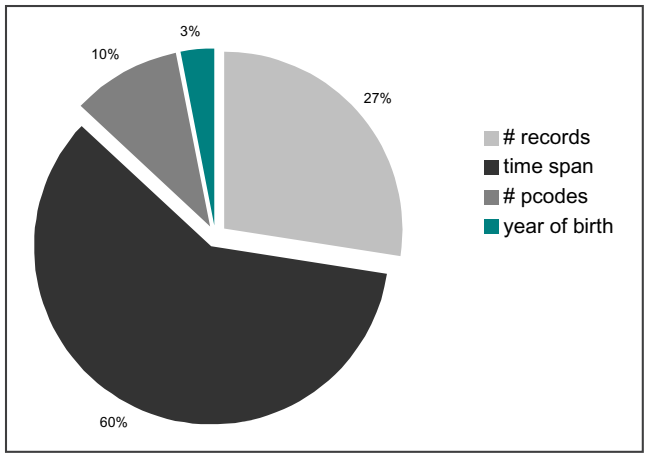

Figure 4 Factors associated with the number of variants in date of birth

A logistic regression model was also constructed to examine the factors contributing to the existence (or not) of different values for date of birth within the linked chains. After adjusting for the number of records, time span, number of postcode values and gender, the negative value for the coefficient of 'year of birth' indicated that chains of records for those born earlier (ie, older persons) were more likely to contain more than one value for date of birth ($p<0.0001$). This model also indicated that there was no difference between males and females ($p=0.2882$) in whether or not there was more than one value for date of birth among their linked records.

	estimate	std error	prob
intercept	18.7511	0.5374	<0.0001
# of records	0.0046	0.0003	<0.0001
time span	0.0063	0.0001	<0.0001
# of pcodes	0.1272	0.0036	<0.0001
gender	-0.0064	0.0060	0.2882
year of birth	**-0.0113**	**0.0003**	**<0.0001**

Table 2 Factors associated with existence of variants in date of birth

The greater likelihood of date of birth variants in older age groups means that date of birth (even in combination with gender and postcode) is a poor discriminator for older people and its use in the absence of full name identifiers for linkage should be avoided.

It is possible that some of the linked chains having a number of different values for date of birth might contain incorrectly linked records. Perhaps this group of linked chains, particularly those for older people, should be examined more closely to

determine whether the variations in date of birth are consistent across other linked datasets.

More generally, any chain with several different values for key demographic variables within one dataset (eg, hospital admissions) but not in others (eg, electoral roll) might signal that an incorrect link (mismatch) exists. Other types of suspicious chains, such as those with dates out of sequence (eg, admission after death) or overlapping admission episodes, could also be routinely targeted for manual inspection.

Within the WA Data Linkage System, the estimated linkage error rate (~3 chains in 1000) in 2002 was much lower than the level of variation in key demographic variables (~1 chain in 10). Targeted sampling to detect this level of linkage error therefore requires a strategy that takes a number of factors into account. For instance, Table 3 illustrates some combinations of variations in key demographic variables in linked chains that might signal linkage errors. Chains with a change in family name as well as a change in given name, or a change in family name and date of birth in one dataset but not in another (parallel) dataset, might be appropriate targets in the first instance.

Other strategies to detect linkage errors could include checks for inconsistencies in dates of admission, treatment and death across several datasets.

	family name	given name	date of birth	address
family name		√	√	
given name			√	√
date of birth				√

Table 3 Targeted detection of mismatches within linked chains

Conclusion

The WA Data Linkage System of 3.7 million chains of linked records drawn from six population-based datasets is updated monthly. Links to six additional datasets are updated six monthly and ad-hoc linkages to special cohorts also occur. Although matching algorithms and the levels of clerical review have been tuned to minimise mis-matches and missed matches, the system also needs to be efficient to achieve a quality product and a quality service.

In January 2002, an internal quality audit was performed by the data linkage team members. As a result it was estimated that the number of chains in the WA Data Linkage System containing one or more false positive matches (mismatches) was 0.3%. This estimate was based on manual scrutiny of a 1% random sample of chains derived from 1990 hospital admission records and is significantly lower than the level of 0.6% estimated in January 2001.

It is clear that the level of possible mismatches in the system is much lower than the underlying variation in reported name, address, date of birth, gender or postcode in the source datasets. Consequently, a system for monitoring the quality of the links will require a targeted approach, supplemented by manual linkage audits.

References

Holman CDJ, Bass AJ, Rouse IL & Hobbs, MST. Population-based linkage of health records in Western Australia: development of a health services research linked Database. ANZJPH 23(5): 453–459, 1999.

Rosman, D.L. The feasibility of linking hospital and police road crash casualty records without names. Accident Analysis and Prevention, 28(2): 271–274, 1996.