The sensitivity of probabilistic record linkage: Estimating the number of "false negatives" in a linkage involving client records of a large domiciliary care organisation

Luke Marinovich
Silver Chain Nursing Association

Stuart Fuller
Data Linkage Unit

Michael Hobbs
University of Western Australia

Gill Lewin
Silver Chain Nursing Association

Abstract

The Continuing Care Linkage Study has been established to examine patterns of health service utilisation, including hospitalisation, domiciliary care, geriatric services and residential care, in persons with continuing disability. As part of this study the Data Linkage Unit of the Department of Health (WA) performed a linkage between clients from Silver Chain Nursing Association (SCNA, Western Australia's largest home care provider) and records from the Western Australian Hospital Morbidity Data System (HMDS). Personal identifiers used in the matching process included family name, given names and date of birth.

Where a client is referred to SCNA from a hospital, this is recorded within the individual SCNA record. It is therefore possible to examine the characteristics of clients who are recorded as having at least one hospital admission according to SCNA records (from referral), but did not link to an actual hospital admission record in the HMDS. This paper will present an analysis of these "false negatives", comparing their characteristics with those for which a matching hospital admission could be found.

A previous linkage of SCNA records with hospital admission records was performed in 1998 without the use of client names. Comparisons will be made between the groups of matched and unmatched records in this linkage and the more recent linkage using full names and dates of birth.

Introduction

The estimation of "false negatives" in data linkage is notoriously difficult. The term "false negative" (or alternatively "Type I error") simply refers to an instance where a true link has not been identified in the linkage process (Arellano and Weber, 1998; Woodward, 1999). Current linkage research conducted in Western Australia (WA) utilising the Hospital Morbidity Data System (HMDS – part of the WA Data Linkage System) and client data from Silver Chain Nursing Association (WA's largest home care provider) has allowed for an estimation of the proportion of false negatives in such a linkage. The HMDS contains records of hospital admissions in WA, while the Silver Chain database contains details of referrals, assessments and service provision, along with demographics and contact details of clients. Linkages between these data sources have been conducted on two separate occasions – the first in 1998 with Silver Chain client names compressed as NYSIIS codes, and the second in 2001 with full names being used in the linkage process. A comparison between these two linkages will be described below to demonstrate the effect of including full name information in the linkage process. A discussion of the issue of false negatives, the process by which false negatives were identified, and an estimate of the proportion of false negatives in the most recent linkage will be presented. A description of the characteristics of Silver Chain clients identified as false negatives in both linkages will also be given.

Comparison between the 1998 and 2001 linkages

The Silver Chain data used in the linkages varied in several important ways, as shown in Table 1. Firstly, the data provided in 1998 had surname information compressed to a 5-character NYSIIS code whereas in 2001, the linkage was conducted with full surnames. Secondly, the 1998 linkage was conducted with just a given name initial, compared with data provided for the 2001 linkage including full given names and preferred names. These two differences represent the most substantial variations between the data sets in terms of the potential impact on linkage. Further, the 2001 linkage also included date of death and transaction date fields not included originally. The only data provided in 1998 and not in 2001, was suburb information (2001 data having only postcode).

	1998 Linkage	2001 Linkage
Full surname		√
NYSIIS code	5 char provided by SCNA	6 char generated by DLU from full surname field
Given names		√
Preferred names		√
Initial	√	√
Gender	√	√
Date of birth	√	√
Umrn	√	√
Suburb	√	
Postcode	√	√
Date of death		√
Transaction date		√

Table 1 Data provided in both data sets

Technical information about how the linkages were conducted is provided in Table 2. In 1998 two runs were conducted, with six passes on the first and five passes on the second. In 2001 three runs were conducted, with eight passes each on the first two runs and two passes on the final run. A 'run' is defined as a program containing a series of passes, where in each pass certain fields contained in both data sets are blocked and matched on, to find links. The difference in the linkage strategy was due to the increase in the number of available fields for blocking and matching of the data.

	1998 Linkage	2001 Linkage
# of Runs	2 runs (6 passes, 5 passes)	3 runs (8 passes, 8 passes, 2 passes)
Blocking Variables	UMRN, nysiis, Day/Month/Year of Birth, Sex	UMRN, Nysiis, Soundex, Day/Month/Year of Birth, Sex, Postcode, Given Name 1, Initial 1
Matching Variables	Sex, DOB, nysiis, initial, postcode	Surname, Given Name 1, Given Name 2, Initial 2, Date Of Birth, Sex, Postcode

Table 2 Comparison between runs, blocking variables and matching variables used

On a more general level, there were other differences between the two linkages. In 1998, the Silver Chain data set contained 82,805 unique Patient ID's, or PID's, while the data set used for the 2001 linkage contained 152,050 unique PID's. The first linkage was completed in September 1998 and the second in November 2001. The 1998 linkage used Morbidity Data from 1980 to 1997 and the 2001 linkage used Morbidity Data from 1980 to 2000.

In comparing the overlap between the two data sets, there were 79,192 PID's common to both the 1998 and 2001 data. The 2001 data contained 72,858 PID's that were not found in the 1998 data, due to the three additional years of data in the newer set. However, there were also 3,613 PID's from the 1998 set not found in the 2001 set. Figure 1 illustrates these features. The relatively small proportion of clients who did not appear in the second set of PID's resulted from data cleaning at Silver Chain, where duplicate PID's had been identified and deleted.

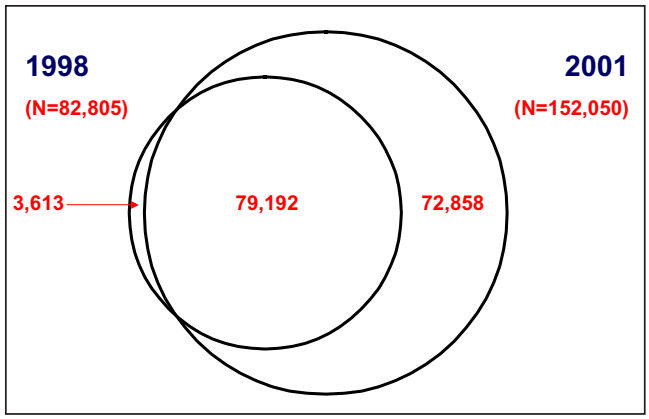

Figure 1 Comparison of the records in the 1998 and 2001 Silver Chain data sets

Number and percentage of PID's linked

Within the 1998 set, 67,517 (81.5%) records linked to HMDS, whereas the 2001 linkage resulted in 126,825 links (83.3%). Using a Venn diagram to show the overlapping data sets, Figure 2 illustrates the number of links found within each data set. Figure 3 illustrates the number and percentage of links within each data set using a horizontal bar chart.

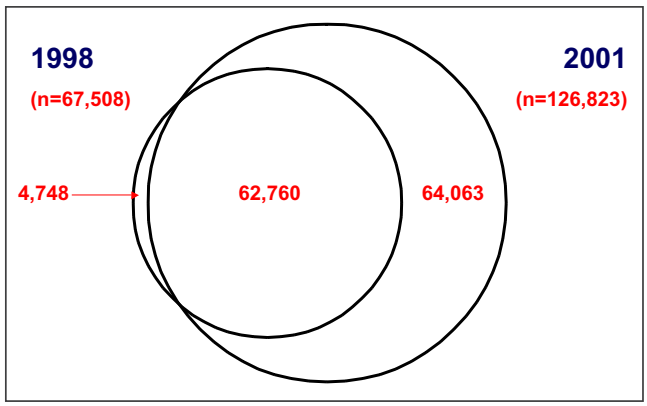

Figure 2 Comparison of the number of links found in all Silver Chain records

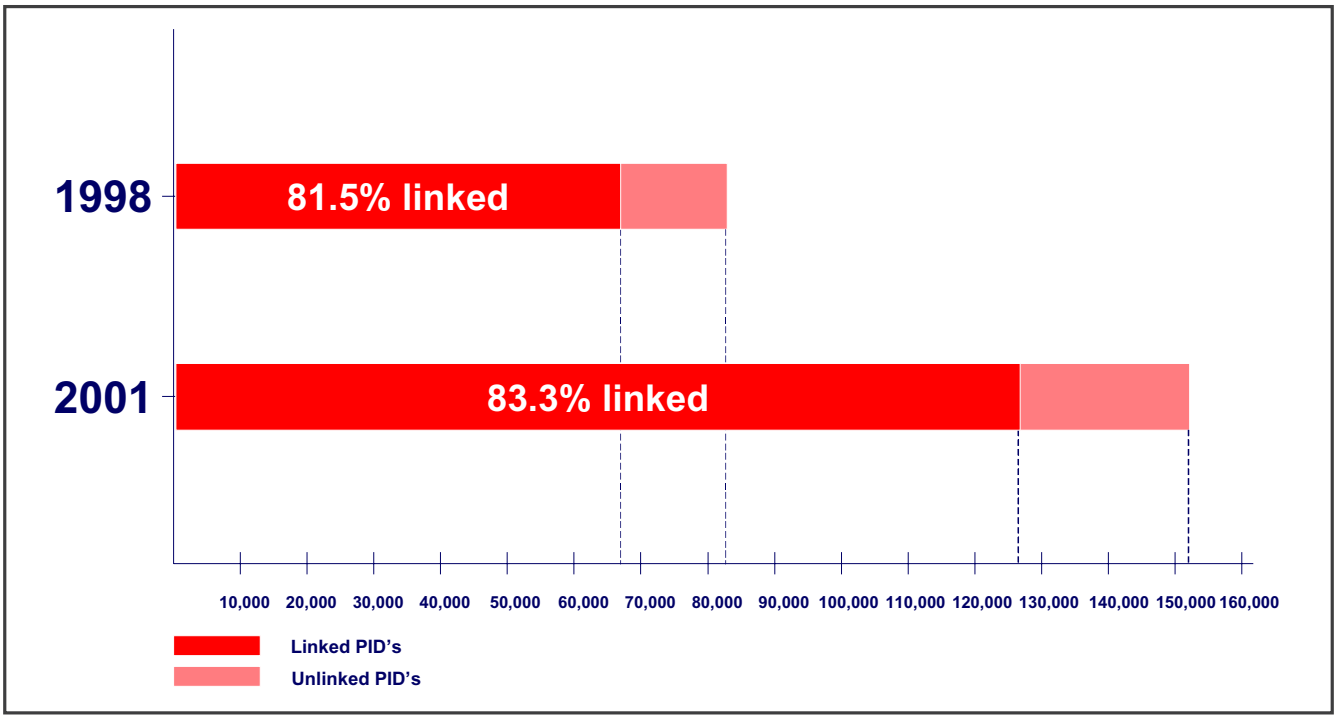

Figure 3 Comparison of all links in 1998 and 2001 Silver Chain records

It would be more appropriate however to compare the group of PID's shared by both data sets. As can be seen in Figure 4, of the 79,192 PID's appearing in both sets, 64,944 (82%) linked in 1998 compared with 70,641 (89.2%) in 2001 – an increase of 5,697 links (7.2%).

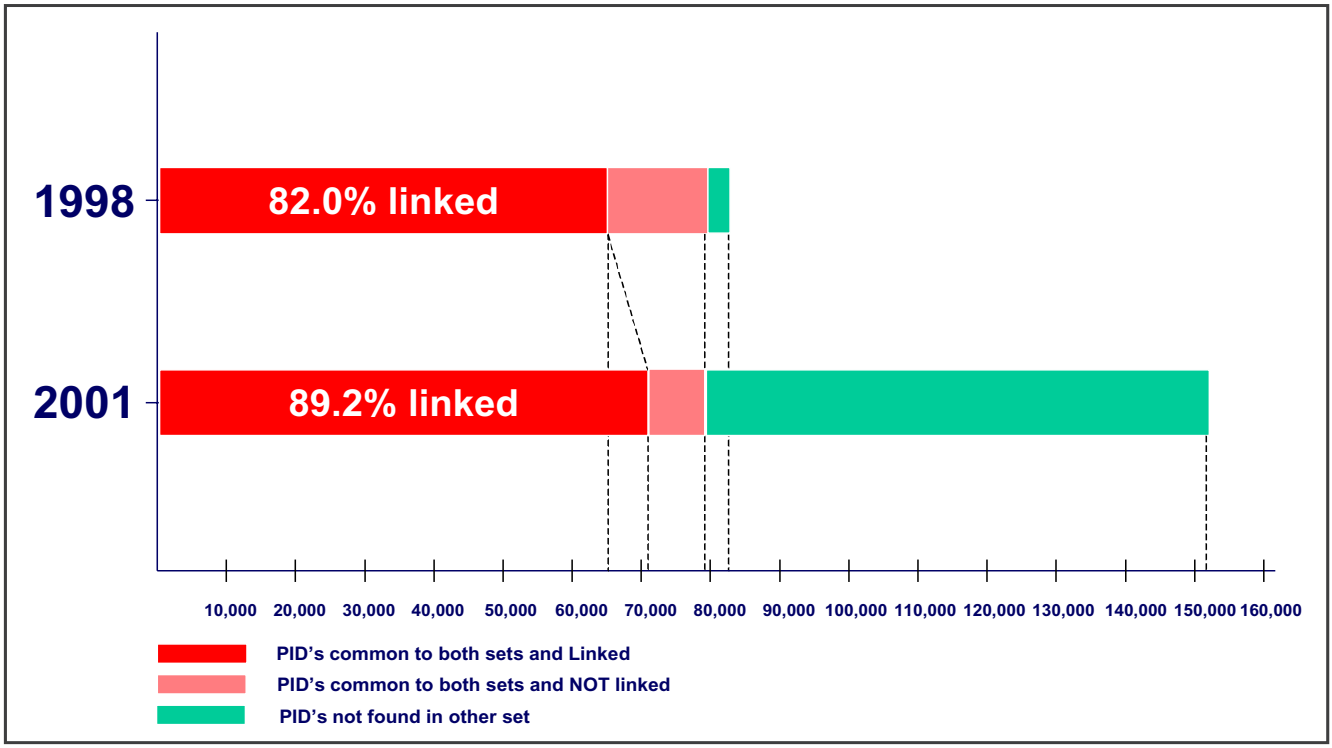

Figure 4 Comparison of the links for the Silver Chain records in both 1998 and 2001

Comparison of links found in 1998 and 2001

The actual number of links found in 1998 was 67,508 compared with 126,823 in 2001, with 62,760 links in common. Among the 79,192 records common to 1998 and 2001 sets of Silver Chain records, 64,950 links to HMDS were found in 1998 compared with 70,715 in 2001 (62,760 in common). In addition, among the links found in 1998, 2,190 links were not found in 2001 (due to data cleaning at Silver Chain), whereas an additional 7,955 links were found in 2001 that were not found in 1998. This represents an overall increase of 5,765 in the number of links found in the second linkage (see Figure 5). From a different perspective there was an overall increase in linkage of 7.2% (see Figure 4).

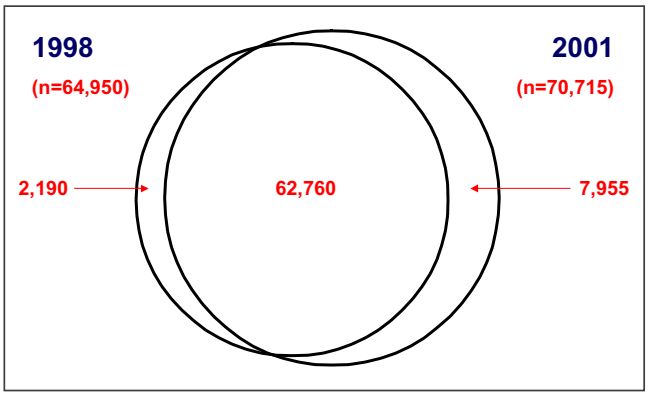

Figure 5 Comparison of 1998 and 2001 sets of linked records

In summary, the use of full surnames, given and preferred names compared to a 5-character NYSIIS code for surname and an initial only for given names resulted in an improved linkage rate and more individual PID's being linked. These additional links that were found in 2001 with the availability of more identifying information were part of the group of missed links (or false negatives) in the 1998 linkage.

Detecting false negatives

A major impetus for conducting the linkage a second time was a concern over the lower than expected linkage rate and therefore the proportion of false negatives suspected in the first linkage. It is usually difficult to determine whether particular unlinked records should in fact have returned links. This linkage between Silver Chain data and hospital admission records (HMDS) however provided an opportunity, as Silver Chain records containing a flag indicating that they had been referred from a hospital would be expected to have at least one Hospital Morbidity record.

The estimation of false negatives in the linkage of Silver Chain and HMDS data thus involves the identification of Silver Chain clients who were recorded as having a referral to Silver Chain from a hospital, but for whom no linkage to a Hospital Morbidity chain was found. In this way, 11,455 such clients (13.8%) were identified as 'false negatives' in the 1998 linkage with 2,083 (1.4%) in the 2001 linkage (see Figure 6). Much of this reduction in the level of missed links may be attributed to the use of more identifying information in the linkage process.

Gender

The false negative group consisted of 45% males and 55% females – a slightly greater ratio of males to females than the data set as a whole (df=3; χ^2=498.022; p<0.001). Although statistically significant, the difference is not substantial in absolute terms, and in the absence of a plausible hypothesis for expecting a difference between the groups it would be unwise to attribute any great meaning to this result.

Age

Figure 7 represents the mean ages of the different groups of clients, based on whether they had a hospital referral in the Silver Chain data set and whether they actually linked to a Hospital Morbidity chain. All mean ages are significantly different from one another (df=3; F=3046.577; p<0.001), but again in practical terms, the difference between the linked groups and the false negatives is not substantial – the mean ages for clients in these groups are all in the sixties, and are within a six year range. However, the group of unlinked PID's without a hospital referral ("true negatives") is interesting in that its mean age is between 12 and 18 years younger than the other groups. From a hospital morbidity point of view, this would seem logical given that this group has between 12 and 18 fewer years in which to accrue hospital morbidity records. Hence, the fact that this is a younger group may be a plausible reason to explain why such clients did not link. From a Silver Chain point of view however, this result is puzzling. It is difficult to envisage a situation whereby younger clients would be referred to Silver Chain for domiciliary care in the absence of a hospital admission. This will be an issue subject to further investigation at Silver Chain.

Deaths

Silver Chain flags deceased clients on the ComCare 3 system, so it is possible to examine the characteristics of false negatives in terms of deaths. A significantly lower proportion of deceased clients were evident in the false negative group – in all, 10.8% of the false negative group had been flagged as being deceased, compared with 22.1% in the set of PID's as a whole (df=3; χ^2=7554.757; p<0.001). It appears from this result that deceased clients actually linked better. One explanation for this may be that clients who are deceased have one extra piece of information available for the linkage process (date of death), and hence may represent more "complete" records. For this reason, it is plausible to suggest that deceased clients may link to HMDS more readily, hence the under-representation of deceased clients in the false negative group.

In summary, the proportion of false negatives in the 1998 linkage was 13.8% (compared with 1.4% in the most recent

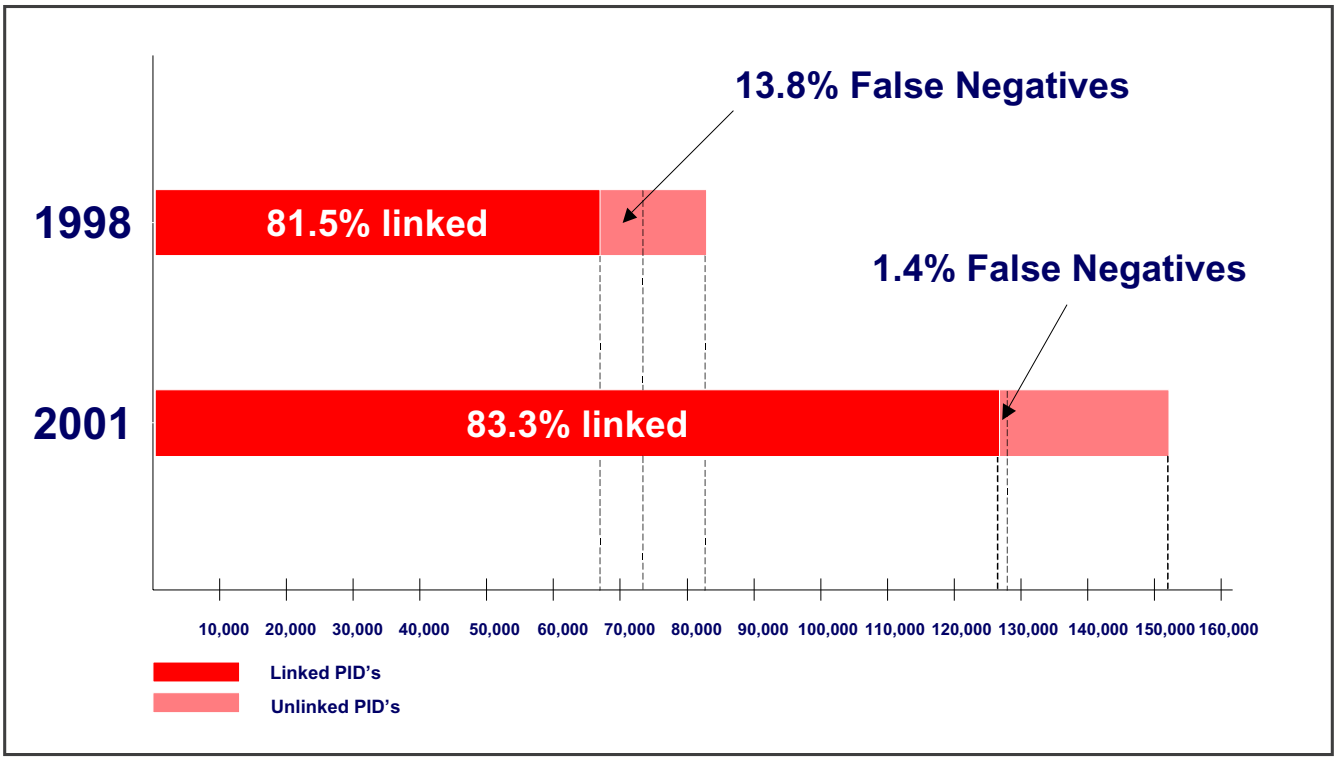

Figure 6 Comparison of the number and proportion of false negatives in 1998 and 2001

An examination of the characteristics of the group of false negative in terms of age, gender and mortality follows.

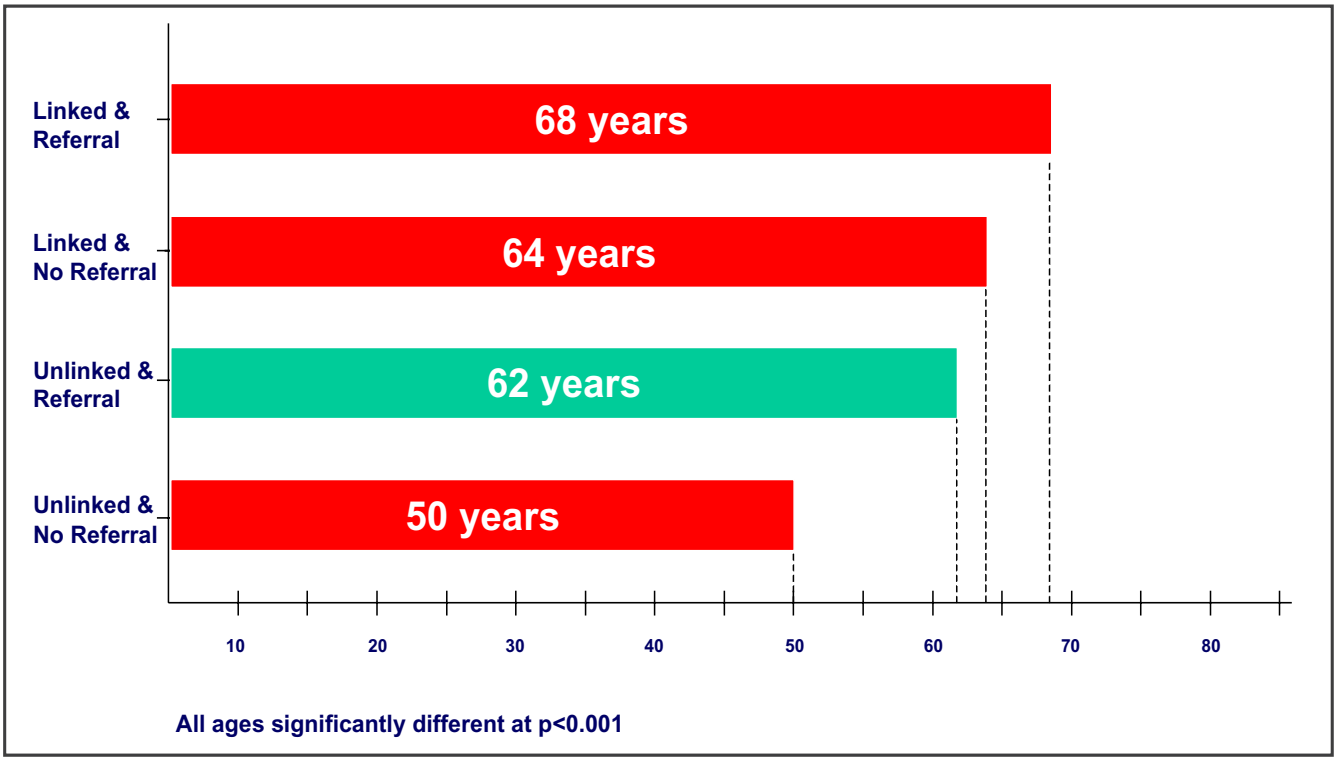

Figure 7 Age characteristics

linkage), indicating that the use of full names instead of NYSIIS codes in the probabilistic linkage process substantially reduces the proportion of false negative links.

Data reliability

There are three main issues that complicate interpretation of the data presented above. Firstly, these analyses are dependent entirely on the source of referral code collected by SCNA to identify clients referred by a hospital. In cross-checking the code for hospital referral against the field denoting the referring organisation, there was a relatively small proportion of clients recorded as being referred from a hospital for whom a hospital was not in fact the referring organisation (for example Aged Care Assessment Teams, nursing homes, and Silver Chain bases). Quite simply, the combination of a hospital referral source code with, for example, a Silver Chain or ACAT referral organisation code could be cause for suspicion about the accuracy of the data, or it may be an accurate representation of the referral process (for example, hospitals may forward referrals to SCNA through an ACAT, which in WA are located within the hospitals themselves). It is impossible to tell.

The second major issue involves the fact that the hospital referral code – logically enough – will only be present if a client is referred to SCNA from a hospital. It is not a flag to indicate whether a particular client has ever had a hospital admission. Therefore, from this aspect it is likely that the estimate of the false negative rate obtained here is an underestimation.

Finally, it is highly likely that referrals from outpatient hospital clinics are recorded on the Silver Chain database as a hospital referral, even though such patients would not necessarily have been admitted to the hospital. In such cases, using the hospital referral flag in Silver Chain data would potentially overestimate the false negative rate.

Conclusions

Despite the difficulties in using the Silver Chain hospital referral flag as a marker for false negatives, the decrease in the proportions of false negatives between the two linkages is substantial. Therefore, some confidence can be placed in the suggestion that this is indicative of an improved linkage in the second instance. Further weight is given to this argument by the increase in the proportion of PID's linked between the two linkages, as well as the increase in the number of links found. Therefore, it seems reasonable to conclude that the inclusion of full given names, preferred names and surnames is preferable to NYSIIS name compression in probabilistic linkage due to the resulting improvement in the quality of the linkage produced.

References

Arellano, M.G. and Weber, G.I. (1998) Issues in identification and linkage of patient records across an integrated delivery system. Journal of Healthcare Information Management, 12(3), 43–52.

Woodward, M. (1999) Epidemiology: Study Design and Data Analysis. Chapman Hall/CRC; London.