Road Injury Information Program
Report Series, Number 7

The Linkage of Hospital and Police Information on Road Crash Casualties: An Investigation of Alternative Methods

by

Diana L. Rosman
Road Accident Prevention Research Unit
Department of Public Health
University of Western Australia

Australian Institute of Health & Welfare
National Injury Surveillance Unit

AIHW National Injury Surveillance Unit
Road Injury Information Program

Report documentation page

Report No.	Date	Pages	ISBN
RIIP-7	April 1995	30	0 642 22704 7

Title:
The Linkage of Hospital and Police Information on Road Crash Casualties: An Investigation of Alternative Methods

Author(s):
Rosman D.L.

Performing Organisations:
Road Accident Prevention Research Unit
Department of Public Health
University of Western Australia
Nedlands W.A. 6009
Australia

Sponsor:
AIHW National Injury Surveillance Unit
Mark Oliphant Building
Laffer Drive
Bedford Park S.A. 5042
Australia

Abstract:

By linking together data routinely collected by the various authorities dealing with road crash casualties, comprehensive information can be assembled for the analysis of outcomes of road crashes.

The objective of this study was to assess the feasibility of linking hospital and police records which do not contain casualty names.

A three month sample of records from the Western Australian Road Injury Database (RID), linked previously using named data and the Generalised Interactive Record Linkage System (GIRLS), was used to investigate different linkage methods. The effect of the level of identifying information on the quantity and quality of links was tested by re-linking the records under different conditions. The GIRLS links were used as the 'gold standard' against which all subsequent linkages using different methods were assessed.

It was found that the dates of crash occurrence and hospital admission, even when used in conjunction with age and sex for matching hospital and police records, was insufficient to produce reliable links without the added discriminating power of the family name of the casualty. The false positive rate of linkage of hospital and police records was 56%, under conditions which derived a true positive link rate of 52% using variables other than name.

Based on the results of this study, it was recommended that if hospital and police data are to be linked both data sets should include family name ideally in full form or, less desirably, in coded form.

TABLE OF CONTENTS:

EXECUTIVE SUMMARY:

Introduction:

By linking together data routinely collected by the various authorities dealing with road crash casualties, comprehensive information can be assembled for the analysis of outcomes of road crashes. Where unique identifiers, such as social security number etc, are not available, names, age and other attributes help to discriminate between individuals. For some projects access to names is restricted and linkage is carried out using only other attributes.

Objective:

This study assesses the feasibility of linking hospital and police records which do not contain casualty names.

Methods:

A three month sample of records from the Western Australian Road Injury Database (RID), linked previously using named data and the Generalised Iterative Record Linkage System (GIRLS), which involved considerable manual checking to produce high quality linkage, was used to investigate different linkage methods. The SAS/Links macros (or subprograms), which were installed on the same computer as that used in the creation of the database, were used for the linkage process. These macros were capable of sophisticated rule definition and weight assignment. The effect of the level of identifying information on the quantity and quality of links was tested by re-linking the records under different conditions. These GIRLS links were used as the 'gold standard' against which all subsequent linkages using different methods were measured.

Linkage efficiency is dependent upon rule definition, weight assignment and the choice of identifier. Initial weight assignment is set to reflect the discriminating power of comparison fields. The frequency of occurrence of repeats of full names, phonetically

encoded family names, and age and sex without names was used as a measure of the discriminating power of these identifiers.

Rules incorporating full names, as well as crash/admission date, age, sex and road user type, were written for SAS/Links and the results compared with those from the original GIRLS project. The rules were then modified to replace names with the NYSIIS (New York State Information Interchange System) code of the family name. This coding scheme grouped together similar sounding family names in 'pockets' prior to matching, thus overcoming some problems with spelling variations. An alternative coding method, SOUNDEX, was also tested. Finally all rules with name references were removed and comparisons using the remaining variables performed.

Results:

There were 1090 hospital admission records and 3929 police casualty records for the test period - June, July and August 1988. The GIRLS system had previously found 587 hospital records with a matching police record. There were 24 internal links or 'repeats' among the hospital records and 39 internal links among the police records.

Uniqueness:

The 'uniqueness' of an identifier is defined here to be the number of different values of that identifier expressed as a proportion of the number of records in the dataset. This 'uniqueness' of full name, NYSIIS, SOUNDEX and records without names for the hospital records is displayed in Table 1. A similar set of results for the police dataset is contained in Table 2. This 'uniqueness' of name-based identifiers decreased in both hospital and police data as the number of records increased, but the level of 'uniqueness' was similar for similar sized files in both sources. Thus, for these identifiers, the file size was the main determinant of discrimination. The discriminating power of unnamed records, however, deteriorated with increasing numbers of crashes per day. Thus, for

hospital admission data without names and fewer records for a given period, there were about 91% of 'unique records' compared to 77% for police casualty data independent of the period and file size. (Police data contained records for all casualties of road crashes, not just those admitted to hospital).

Deterministic Linkage:

Using exact matching, 276 (47%) of the 587 hospital to police matches for the three month trial period were produced. Thus, without using some form of probabilistic linkage method, only 25% of the 1090 hospital records of road crash casualties would have been matched to a police record instead of the 54% located using probabilistic methods.

Probabilistic Linkage:

Table 3 contains a summary of the influence on the linkage rate of the type of individual identifiers.

Linkage using full name:

Using SAS/Links macros and all available name identifiers, 580 (99%) of the 587 previous GIRLS links were identified. Of the 7 GIRLS links not found, 4 had spelling variations combined with age or date differences, while the remainder had missing or vastly differing ages. Rules to capture these automatically could not be easily defined without creating a huge set of mismatches. The SAS procedures were able to provide 13 new linked record pairs, 10 of which had family name variations which would affect NYSIIS code pocketing and all weighted name comparisons. Missing values and large date differences accounted for the other three. Application of a new 'one letter' rule (i.e. allowing the spelling of name to vary by a single letter) corrected for the variations in family name.

Linkage using NYSIIS:

SAS with NYSIIS achieved 536 (91%) of the 587 GIRLS links. In addition, SAS found 25 new links increasing the rate to 95% of the 593 links found using names and SAS/Links.

Linkage using SOUNDEX:

The results achieved when SOUNDEX codes were substituted for NYSIIS in the SAS procedures are encouraging. Of the 587 named GIRLS links, 50 (9%) were not among the potential links extracted, but another 36 (6%) not found by GIRLS appeared to be valid. Thus a total of 573 (97%) of the 593 named SAS links were found using only the SOUNDEX code of the family name. The cost of achieving this was that more potentially linked pairs had to be checked manually.

Linkage using unnamed data:

Of the 700 potentially linked pairs extracted without name references, only 308 (44%) were valid links as defined by the GIRLS system using names. Only 2 of the remaining links were among those identified by SAS/Links using full name. Thus more than half of the probable links extracted by these procedures were invalid. This total of 310 links represents 52% of the 593 named SAS links. Reducing the thresholds to extract more probable links only increased the size of the manual checking task without improving the level of valid links.

Discussion:

The SAS/Links procedures (and possibly any other flexible linkage software system) were shown to be capable of finding about 99% of the links produced by GIRLS. In addition, an extra 13 links were detected by SAS. It has also been shown that a substantial number of links could be achieved automatically, provided that the rules and weights were tailored to suit the peculiarities of the data (the level of missing information and the level of miscoding).

Linkage with phonetic coding of the family name has been shown to approximate that produced using the full name. Linkage without names, however, should only be attempted with caution, due to the large number of invalid links interspersed with valid ones. These would be inseparable without a reference standard such as that provided by the GIRLS linkage using full names.

Conclusion:

The dates of crash occurrence and hospital admission, even when used with age and sex for matching hospital and police records are insufficient to produce reliable links without the added discriminating power of the family name of the casualty.

Recommendations:

Based on the results of this study, it is recommended that if hospital and police data are to be matched (or linked), both data sets should include family name ideally in full form or, less desirably, in coded form. Confidentiality can be guaranteed if linkage is carried out on a secure computer system, not connected to an external network. Licence agreements between data custodians and experienced linkage teams would help to formalise obligations in this area and ensure that names and other identifying information is removed before data is analysed.

INTRODUCTION

Background

Record linkage is not a new concept. It is simply the process of bringing together information from two records that are believed to belong to the same individual. As early as 1946, Dunn described the potential for linking records of the 'principal events of life' into a 'book of life' starting with the birth record and ending with the death record. Since then, much work has been done in both Canada and the U.K. on linking medical records to provide more detailed information on the progress of treatment and disease over time. Often such diverse data sources as hospital admission, clinical and laboratory data as well as financial data may be brought together to complete the picture. Family record linkage has been used to identify disease patterns in order to ascertain possible genetic effects of disease. Linkage of medical records to death records is carried out routinely by medical research workers involved in survival analyses. A good example of this is the California Automated Mortality Linkage System (CAMLIS) developed by Arellano et al (1984) which uses a combination of deterministic and probabilistic decision criteria.

However, it is only relatively recently that non-medical data has been used to extend the information in medical records. Researchers examining the causes of injury in general, and road crash trauma in particular, are increasingly looking to other data sources to provide information on the circumstances surrounding the events leading to the injury. In the U.S. Fife (1989) matched Fatal Accident Reporting System (FARS) cases to motor vehicle death records and States et al (1990) used linked police accident reports and hospital records of admitted patients to investigate the New York State safety belt use law. In Australia, a pilot study carried out by Gordon et al (1986) in the Hunter region of New South Wales measured the feasibility of linking various data sources without using names or other unique identifiers. This pilot study concluded that there was a trade-off between the proportion of records linked and the level of confidence that the linked records referred to the same person. They found that a linkage rate of 50% would require acceptance of an error rate of 10%. A more extensive record linkage project was carried

out in Western Australia, where under strict confidentiality provisions, name-identified data was made available for hospital, police, ambulance and death records for casualties of road crashes over a 15 month period. This linkage project used the Generalised Iterative Record Linkage System (GIRLS) developed at Statistics Canada by Hill et al (1981) and is described in detail in Ferrante et al, 1993. Since this project used name-identified data and included considerable manual checking to produce high quality linkage, it was considered to be the 'gold standard' against which to measure the performance of other software and linkage criteria on the same test dataset. The investigation outlined below was carried out using the Road Injury Database and compared different methods of linking road casualty data obtained from hospital and police records.

Objectives

The existence of a linked database and access to the original records provided an opportunity to examine the hospital/police linkage process using the same datasets under different linkage conditions. The objectives of this study were to compare the quantity and quality of the links which could be obtained using another software package and then to compare the results for different levels of identifying information. Since access to this specialised software such as GIRLS is not readily available, this project aimed to test the performance of the more widely available SAS/Links. Data matches produced with the Generalised Iterative Record Linkage System (GIRLS) and full name identifiers were used as the 'gold standard' against which to compare links produced with SAS/Links and full names. Records with only encoded family names and records without name identifiers were then compared with these.

METHODS

Probabilistic Linkage

Record linkage methods which attempt to match exactly on a set of criteria are termed 'deterministic', while those which allow for certain expected variations in some criteria and measure the degree of uncertainty in the match are referred to as 'probabilistic'. The latter

method allows for the matching of records where identifiers are not unique. Since, the same individual may be identified differently in two data sources, due to spelling variations or the inconsistent use of pseudonyms and diminutives of given names, probabilistic linkage techniques rather than deterministic ones are preferable. Small errors in information such as age or date of crash can also be tolerated with this method.

Probabilistic linkage involves the allocation of weights for each record pair based on the outcomes of comparisons between pre-defined linkage criteria or rules. Several passes through the data are performed in order to extract as many links as possible (see Appendix II for further information). For each comparison, a positive weight is assigned if the two quantities agree or partially agree and a negative weight is given if they differ radically. No weight is applied if information is missing. The sum of these comparison weights for each record pair is calculated and if this value is below the pre-defined 'cut-off' value, the record pair is rejected. If a total weight above a much higher 'threshold' is obtained, the record pair is defined as a 'definite' link. Records with values between the 'cut-off' and the 'threshold' are said to be 'possible' links. These 'possible' links are entered into the next pass and the process repeated, until the final remaining pool of 'possibles' is checked manually and obvious mismatches discarded. Rules and cut-off points are adjusted until all links are achieved or the proportion of mismatches to matches becomes unmanageable.

Since the GIRLS system was not operational at the commencement of this study, another linkage system which was also available on the Health Department IBM mainframe was used for this project. The 'Links' macros for the Statistical Analysis System (SAS) were acquired from Roos and Wadja (1986). These had been modified by R. Hockey (Department of Public Health, University of W.A.) to incorporate weight calculations as in GIRLS. The comparison rules had also been extended, so that the 'Links' system in conjunction with the SAS language was capable of quite sophisticated rule definition. The use of a different software tool on the same data using the same criteria enabled a

comparison to be made between the performance of a sophisticated special purpose system such as GIRLS and an enhancement to a more general statistical system such as SAS.

Internal Linkage

Record linkage between data sources such as hospital and police records for road casualties, where an individual may have more than one crash and be admitted to hospital on several occasions for some or all of those crashes, results in what is termed 'many-to-many' linkage. This form of linkage is much more complex than simple 'one-to-one' as in linkage between coroner's reports and death certificates or even 'many-to-one' matching as in linkage between hospital and death records. Care needs to be exercised in crash/admission date comparisons so that each hospital episode is connected with the appropriate crash information. Often manual intervention is the only solution in difficult cases.

Internal linkage of the hospital dataset prior to attempting the between-source linkage, allowed all records belonging to the same individual to be identified. These sets of multiple records for individuals could have resulted from duplicate records being entered incorrectly or from quite valid records of transfers between hospitals or re-admissions at a later date. Similarly, for the police dataset, it was possible for duplicate records to arise where a report was made both by an attending police officer and a crash participant. In addition, the same individual could have been a casualty of more than one crash in the 3 month period. With a period under investigation of more than three months, this becomes a more common occurrence.

Data Sources

The hospital data used here were derived from the Hospital Morbidity System of the Health Department of Western Australia. All hospital discharges for which the external

cause of admission was 'traffic accident' ie E810.0 to 819.9; bicycle crash (E826.1) or late effects of a traffic crash (E929.0) were included.

In W.A. all casualty crashes are required to be reported to the police if they occur on a public roadway. Thus, one would expect to find a matching police report for most of the casualties who were admitted to hospital. For some casualties, such as off-road pedal cyclists who are not involved with another vehicle, reports to the police are not required. Hospital admissions resulting from 'late effects of a traffic crash' would also not be expected to have a matching police report since this condition is not defined until 12 months after the crash event. It has previously been reported by Rosman and Knuiman (1993) that only 64% of those hospital records which could be described as 'reportable' were matched to a police report. This is consistent with the 30% level of under-reporting reported by other investigators (Steadman & Bryan (1988), Harris (1990)).

The police casualty records were extracted from the ROTARS database jointly managed by the Police and Main Roads Departments. These police records were included only if the report stated that an injury had occurred. A manual check of a sample of participants in non-casualty crashes for the month of June 1988 indicated that an additional two percent links would be obtained by extending the police data to all reported crash participants (Ferrante et al, 1993).

The hospital and police records for the period October 1987 to December 1988 had previously been linked during the construction of the Western Australian Road InjuryDatabase (RID). A full description of the original linkage process and the structure of the resultant RID, which also contains ambulance and death records, can be found in Ferrante et al (1993). The hospital and police casualty records for the months of June, July and August 1988 were extracted from the RID for this study.

Uniqueness of identifiers

One factor which is known to influence the ability to link on a given set of criteria is how well those criteria can uniquely define an event or individual. The discriminating power is a measure of the 'likelihood of fortuitous full agreement' (Newcombe, 1989) or the chance of two quite unrelated records being linked because they agree on the variables selected for matching. The number of different values or general frequency of a variable is a factor in the calculation of its global discriminating power.

The relative discriminating power of family name, NYSIIS code and SOUNDEX code (see Appendix I) was examined in the hospital and police datasets. Since it was expected that the size of the dataset would influence these results, subsets of records for periods of 3, 6, 9, 12 and 15 month periods were examined for both the Hospital Morbidity Dataset and the police ROTARS database.

Comparison of SAS/Links and GIRLS

A similar strategy to that described in Ferrante et al (1993) was followed here using SAS/Links on named data. All blanks, hyphens and apostrophes were deleted from family names prior to entering the linkage programs. In order to minimize the number of comparisons required, records were grouped or 'pocketed' and then compared. Several passes through the data were made using different pocketing schemes in order to extract as many links as possible. Details of the various passes are described in the Appendix II.

The RID linkage project had previously linked 587 hospital and police records for the months of June, July and August using the GIRLS system. The SAS/Links routines were used here to re-link the hospital and police records for the same period. Comparison rules and cut-off weights were adjusted until as many as possible of the GIRLS links were found. The aim was to minimize the number of non-links remaining among the set of 'possible' links, while capturing all known GIRLS links. In this way the SAS/Links procedures were calibrated against the GIRLS system for this three month sample period.

When the calibration process had been completed, a comparison of the performance of GIRLS and SAS/Links was made on another two sets of records, which had not been part of the calibration process. Hospital and police records for the month of January, 1988 were extracted and processed using the SAS/Links procedures developed for the June, July and August data. The same four passes were carried out and a comparison made between linked records produced by SAS/Links and GIRLS. In this way, cross validation of the SAS procedures was performed.

Comparison of Identifiers

Hospital and police casualty records for the three month period June, July, August 1988 were re-linked using different identifiers. The groups of linked records, which were produced for each of these linkages, were compared. The first two linkages used phonetic codes of the family name as described in Appendix I.

The first linkage replaced the name with the NYSIIS code of the family name. The same procedure was followed as before, with small changes to the SAS/Links routines so that only the NYSIIS code of family name was used in place of any reference to the family name or given names. For the next linkage, the only change involved replacing the NYSIIS phonetic coding of the family name with the SOUNDEX code. This meant that the new code was used both as the pocketing variable in the first pass and in place of name variables in all comparisons. The final linkage was performed as if there were no name identifiers on either record. All references to names and phonetic codes were deleted and the first pass, which required pocketing on NYSIIS (or SOUNDEX), was bypassed.

Each of these linkages was performed using several passes through the data, as described in Appendix I.

RESULTS

For the period June 1st to August 31st, 1988 the Hospital Morbidity file contained 1090 hospital admission records. For the same period there were 3929 police casualty records in the ROTARS database. The GIRLS system had previously identified 587 (54%) hospital records with a matching police casualty record.

Internal Links

Of the 1090 hospital admissions, 24 were confirmed to be repeated identities. One of these was a re-admission and the remainder related to transfers between hospitals. Internal linkage of police records on the other hand identified 14 duplicate records and 25 records of multiple crashes in the same three month period. With an increased time frame, the possibility of internal links (e.g. duplicate records) increases. This is illustrated in the complete RID which contains 147 repeats and 81 duplicates in the police dataset and 460 re-admissions and 3 duplicates in the hospital dataset for the 15 month period.

Uniqueness of identifiers

The 'general frequencies' for the unpaired records in the hospital dataset are displayed in Table 1. To obtain these values the number of different "unique" identifiers (e.g. the number of family names that were unique) were counted and expressed as a percentage of the total number of records in the file. These values do not give the percentage of unique records in the dataset, since each name has been counted once, including those which occurred more than once. In the discussion that follows, the term 'unique' will be applied to indicate the phenomenon described above.

TABLE 1
Comparison of Discriminators
Percent of each group with 'unique' identifiers
(Hospital Morbidity Dataset)

	PERIOD				
	1988			1987/88	
IDENTIFIER	Apr-Jun	Mar-Aug	Jan-Sep	Dec-Nov	Oct-Dec
Family name alone	77.4%	68.6%	63.5%	59.7%	56.5%
Family name,first name	92.0%	89.0%	86.8%	85.0%	83.7%
Family name,first name, age,sex	94.8%	93.7%	93.0%	92.7%	92.2%
NYSIIS	66.8%	54.5%	48.3%	43.4%	39.9%
NYSIIS,age,sex	94.1%	92.2%	90.4%	89.6%	88.5%
SOUNDEX	56.4%	42.5%	35.4%	30.3%	26.5%
SOUNDEX,age,sex	93.8%	91.8%	89.9%	88.9%	87.3%
Admission date,age,sex	91.3%	90.6%	90.7%	90.5%	90.5%
No. of records	1055	2275	3394	4513	5697

Results in Table 1 show that the proportion of hospital records with different family names decreased from 77% for approximately 1000 records to 56% for 5700 records. In other words, 23% of the 1055 hospital records for April 1988 to June 1988 contain repeats of family names. As expected, there were fewer different NYSIIS codes than complete family names for each file size, with the proportion of different values ranging from 67% down to 40% as the number of records increased. The SOUNDEX code alone was less able to discriminate between individual records than the NYSIIS code. For SOUNDEX, 56% of the 1055 records for the three month period (April to June) and 26% of the 5700 records for the full 15 months were 'unique'.

As expected, when age and sex were used in conjunction with the name identifier or phonetic code to identify an individual record, the relative frequency of 'unique' values

increased markedly. Both NYSIIS and SOUNDEX codes combined with age and sex were 'unique' in 94% of the 1055 records and 88% of the 5697 records.

TABLE 2
Comparison of Discriminators
Percent of each group with 'unique' identifiers
(Police ROTARS Dataset)

	PERIOD				
	1988			1987/88	
IDENTIFIER	Apr-Jun	Mar-Aug	Jan-Sep	Dec-Nov	Oct-Dec
Family name alone	63.9%	56.3%	51.7%	48.2%	45.7%
Family name,first name	92.8%	90.2%	88.5%	86.8%	85.1%
Family name,first name,age,sex	97.5%	97.0%	96.6%	96.4%	96.0%
NYSIIS	47.9%	38.0%	33.1%	29.4%	26.8%
NYSIIS,age,sex	94.9%	92.6%	90.6%	88.9%	87.2%
SOUNDEX	32.7%	22.1%	18.1%	15.1%	13.2%
SOUNDEX,age,sex	93.8%	90.6%	88.1%	85.6%	83.4%
Crash date,age,sex	77.6%	76.7%	77.1%	77.5%	77.6
No. of records	3953	7967	11440	15088	18676

Similar results were found for casualty records in the police ROTARS database. From Table 2 it can be seen that there were proportionately fewer different names and phonetic codes in the police data compared with the hospital data for each period. This is because the police casualty dataset was larger (3953 records for 3 months and 18676 for 15 months) than the hospital admission dataset for each period. However, 64% of family names, 48% of NYSIIS codes and 35% of SOUNDEX codes were 'unique' in each dataset when similar sized files of about 3500 records were compared.

The main point of interest in these results is the lower overall discriminating power of crash/admission date. In Table 1 for the period January to September 1988, 91% of the 3394 hospital records were considered 'unique' when no name identifiers were used in the

linkage process. A similar number of records in the police dataset correspond to the period April to June 1988 (3953). However, only 78% of these records had different crash date/age/sex values under the same conditions.

It can be seen from the last row of both Tables 1 and 2 that there is little variation in 'uniqueness' with file size in either the hospital (~91%) or police (~78%) datasets. It appears that crash date is the dominant variable here. There were 40 casualties per day in the police dataset compared to 12 in the hospital dataset. If the decreasing trend in 'uniqueness' with an increase in the number of casualties were linear, one could expect about half (90% less 39%) of the records to be 'unique' for crash date, age and sex in a dataset with 130 crashes a day. This greatly decreases the usefulness of 'date of occurrence' as a linkage variable for large populations or those with high crash rates.

Deterministic Linkage

A hospital record having identical values for family name, initial, age, sex and date of crash/admission as a police record was considered to be an exact match. This method of matching is often referred to as 'deterministic' linkage. There were 276 such matched record pairs in the 3 month period from June, July and August 1988, representing 47% of the 587 GIRLS links. Thus without using some form of probabilistic linkage only 25%, rather than 54% of the 1090 hospital records would have been matched to a police record.

Probabilistic Linkage

Linkage using Full Name, comparing SAS/Links and GIRLS

The first attempt to link using SAS/Links and full name identifiers without any manipulation of the family name and using default frequency weights for all variables yielded only about 94% of GIRLS links. The SAS procedures were modified, after examining the GIRLS links which had not been captured by this first attempt. Appendix III describes the details of the outcome of each pass.

Using all available name identifiers, 580 (99%) of the 587 previous links were identified. Of the 7 GIRLS links not extracted by SAS/Links, 4 had spelling variations combined with age or date differences, while the remainder had missing or vastly differing ages. Rules to capture these automatically could not easily be defined without creating a huge set of mismatches. The SAS procedures were also able to provide 13 new linked record pairs, 10 of which had family name spelling variations, (eg HAGBOOM - MAGBOOM; BELL - BEZL; GOK - COX; BAYNE - BWYNE; RIDU - RIOU) which would radically alter the NYSIIS code. As NYSIIS was the first pocket variable, this would mean that comparisons between these names would not be made until later. Comparisons between full name, NYSIIS and first 3 letters of family name comparisons in later passes would also discriminate against these links. Application of a 'one letter' rule (wherein the spelling of name was allowed to vary by one letter) corrected for this in 10 cases and the other three were found by hand (gender was missing on two of these three cases and the other had a difference of 33 days between accident date and hospital admission).

In order to test the performance of the SAS procedures using another sample of data not used in the calibration, hospital and police records for January 1988 were extracted and linked using the same procedures. Of the 361 hospital records and 1017 police records, GIRLS had identified 160 (44%) links. The SAS procedures extracted 178 potentially linked pairs, which were checked manually. Of the 165 (46%) valid links, eight were new links not found by GIRLS and there were four false negatives. The remaining 13 were false positives.

Linkage using NYSIIS

Similar procedures to those outlined for named data were performed with the NYSIIS code of family name replacing all references to family name, first name and initial.

For the three months June, July and August 1988, 602 potentially linked pairs of hospital and police records were extracted. Of these 41 were incorrect and 51 of the previous GIRLS links were not identified. Thus SAS with NYSIIS achieved 536 (91%) of the 587 GIRLS links for this period. In addition, SAS found 25 new links totalling 561 links in all. This represents 95% of the 593 links found using SAS/Links and the full name.

Linkage using SOUNDEX

The SOUNDEX phonetic coding scheme is less discriminating than NYSIIS but has the advantage that it is less recognisable as a persons name. Some discriminating power based on name is given without breaching confidentiality requirements.

The results achieved, when the SOUNDEX code was substituted for NYSIIS in the SAS procedures, were encouraging. Of the 631 potentially linked pairs extracted, 58 (9%) were invalid. Of the 587 GIRLS links, 50 (9%) were not among the 631 potential links but 36 (6%) links not found in GIRLS appeared to be valid. Thus a total of 573 links were identified using the SOUNDEX phonetic coding method. This represents 97% of the 593 links found using the full name and SAS/Links. The cost of achieving this was that more potentially linked pairs had to be checked manually.

TABLE 3
Linkage Results for Different Identifiers using SAS/Links

		LINKS IDENTIFIED		
IDENTIFIER	GIRLS links	New links	Missed links	Invalid links
Full name	580 (99%)*	13	7	74
NYSIIS	536 (91%)*	25	51	41
SOUNDEX	537 (91%)*	36	50	58
No name	308 (52%)*	2	279	390

*percentage of GIRLS links

Linkage using unnamed data

Since the police records did not contain such demographic information as occupation, country of birth and marital status (and date of birth was unavailable for casualties other than the driver), the only variables common to both sources on which to match were age, sex, crash/admission date and road user type. The road user type was unknown for about a quarter of the hospital records and this comparison was given a lower weight than the default.

Using procedures as before, with all name rules removed and thresholds adjusted accordingly, 700 potentially linked pairs were extracted. However only 308 (44%) of these were valid links as defined by the GIRLS system. Two of the links were among the extras found by SAS/Links using full names. The remaining 390 were invalid. Thus, more than half of the probable links extracted using these cut-off values and procedures were false positives. This means that without the ability to check linked pairs against a known standard, as has been done here, more would be incorrect than correct. It was also difficult to distinguish between valid and invalid links based on the weight assigned. Increasing the cut-off value would thus reduce the number of valid links as well as the number of invalid ones.

DISCUSSION

In order to maintain the confidentiality of records, all linkage projects in W.A. which use named hospital and police records, are required to use the State Government's IBM mainframe computer at the Health Department. Software for this environment is more expensive than for mini- or micro-computers. The choice for this study was limited to the Generalised Iterative Record Linkage System (GIRLS), purchased at a cost of $15,000 from Statistics Canada in 1989, and an enhanced set of Links macros for SAS. It has been shown here that these macros, written for a general purpose statistical package,

perform as well as the system designed specifically for record linkage. The level of disagreement of about one percent (Table 3) is within operator error.

It must be acknowledged that use of these macros requires some programming expertise and a detailed knowledge of the idiosyncrasies of the data sources. The GIRLS system, on the other hand, is more user-friendly. Although record linkage strategies need to be well understood and rules must be written in a strict format, the same level of expertise in procedural programming is unnecessary.

Although the results presented here indicate that about 95% of the previously linked records could be found using phonetic coding of the family name (NYSIIS or SOUNDEX) when used in conjunction with age, sex and crash/admission date, the question of data volume should be considered. By examining a larger dataset than that offered by the 3 month trial period, it was shown that the discriminating power provided by such variables as crash date, age and sex depended on the number of crashes per day as well as the size of the files. The potential for conflict between phonetic codes also increased with file size. This occurs not only because a larger sample is more likely to include more casualties with similar names, but because the potential for casualties to be involved in several crashes increases. Duplicate records also contribute to this problem.

There were some minor differences between the pool of links produced by GIRLS and that produced by SAS/Links. The specific reasons for this one percent difference is detailed above. However, the 'one letter' rule, devised to capture some spelling variants which had been linked by GIRLS, produced an extra 13 links not previously found, for the 3 month study period. Some of these minor differences were likely to have been operator decisions resolved manually.

It has also been shown here that a substantial amount of linkage can be achieved automatically, if the operator has intimate knowledge of the characteristics of the data

sources. It is evident that the software used for record linkage needs to be capable of sophisticated rule definition and that these rules need to be sensitive to the data being linked. Trial linkage runs may be needed to 'tune' the procedures, while at the same time providing initial estimates of the linked agreement and disagreement weights for each rule.

More complex name comparisons may be required where phonetic coding and the 'one letter' rule (i.e. where the spelling of name is allowed to vary by one letter) are insufficient. For instance, family names of WERENREICH on the hospital file and LENREICH on the police file were found. The rarity of Germanic names in Australia and the similarity of the last part of the name makes a link intuitively likely. In another case, knowledge of first name fashions and matching dates enabled resolution where the first names of a father and son (ALLAN and CALLAN) had been reversed in the police data.

Without access to named data or at least some encoded version of the family name, it is difficult to produce quality linkage of data such as police crash reports, which may contain more missing data and spelling variations than is usual in health data. It appears from this study that about 50% of the links produced without names could be false positives or mismatches. In addition, about half of the 'true' links would not be discovered due to insufficient information. For internal linkage, which is essential to identify repeat crashes and re-admissions, the lack of any name information would become more serious as the crash/admission date would no longer be a strong indicator of a link. A further disadvantage to linking without names is that it is very difficult to manually check possible links. Manual checking with names may be the only way to distinguish between possible links with equivalent weights or where there are errors in ages and crash/admission dates.

It is likely that there are undiscovered record pairs in the dataset among those rejected through lack of information or low weights. Also, if the date of crash was incorrectly stated and fell after the admission date, a match would never be considered and potential

links would be lost. In all decisions about links, a conservative approach has been taken, so that there is confidence in any results derived from this linked dataset.

Care was also taken when carrying out the four linkage operations not to introduce any extraneous variation by altering the linkage procedures other than to vary the comparison rules dealing with names etc. It is possible that a different weight assignment regime could be developed to minimize the level of invalid links while maximizing the level of valid ones. Although not fully explored, this is unlikely to be successful, since it was not possible to achieve any improvement in the quantity or proportion of valid links by adjusting the thresholds and cut-off values for rejected, possible and definite links.

The sets of linked records produced with and without names did not differ in their age/sex or road-user-type distributions. The added discriminating power of *date of birth* over *age* of casualties, used in conjunction with *crash/admission date*, would add discriminating power where names were not available. Unfortunately *date of birth* was only available on the police records for drivers and thus the use of *date of birth* instead of *age* for this study would have produced a biased set of links with drivers overrepresented.

CONCLUSIONS

This investigation has shown that linkage success is probably not dependent on the software used, provided that the rules for comparison can be tuned to suit the data sources. Very similar linked datasets were produced by SAS/Links and GIRLS on named data. It seems that calibration against a test dataset is necessary to achieve optimum results. This could be done manually if an existing linked datset were not available.

Phonetic codes have also been shown to be almost as efficient as the full name in producing quality links. By reducing the level of identifying information to just the phonetic code of the family name, about 95% of all GIRLS named links were found. If data providers are reluctant to release named or even phonetically encoded data, more

sophisticated coding schemes could be used to ensure confidentiality without sacrificing discriminating power.

In contrast to these positive results is the poor performance of the linkage on unnamed data. The crash date, even if used in conjunction with age and sex, is not a good discriminator, especially for situations where there is a large number of crashes per day. For large datasets the ability to manually check doubtful links is important and this is very difficult without personal identifiers such as names or phonetic codes.

A further concern in using unnamed data is that an 'acceptable' linkage rate between hospital and police records may be achieved without the operator being aware that about half of these links could be false positives and a similar number of false negatives may never be detected. This can only be checked where there is access to a 'gold standard' such as that provided by the original GIRLS linkage on named data.

Another concern of most record linkage practitioners is that fully automated linkage systems may mask some of these problems. Even if the age/sex/road-user-type distributions under different criteria are similar, incorrectly linked records could lead to incorrect crash details being associated with injury outcomes.

RECOMMENDATIONS

Based on the results of this study, it is recommended that if hospital and police data are to be matched (or linked), both data sets should include family name ideally in full form or, less desirably, in coded form. Confidentiality can be guaranteed, if linkage is carried out on a secure computer system, not connected to an external network. Licence agreements between data custodians and experienced linkage teams would help to formalise obligations in this area and ensure that names and other identifying information is removed before data is analysed.

REFERENCES

Arellano MG, Petersen GR, Petitti DB and Smith RE. The California Automated Mortality Linkage System. Am J. Public Health 1984; 74(1):1324-1330.

Dunn HL. Record linkage. Am J. Public Health 1946; 36:1412-1416.

Ferrante AM, Rosman DL and Knuiman MW. The Construction of a Road Injury Database. Accid. Anal. Prev. 1993; 25(6): 659-665.

Fair ME and Lalonde P. Application of Exact ODDS for partial agreements of names in record linkage. Computers and Biomedical Research 1991; 24: 58-71.

Federal Office of Road Safety Road crashes resulting in hospitalisation - Australia 1990.

Fife D. Matching fatal accident reporting system cases with National Center for Health Statistics motor vehicle deaths. Accid. Anal. Prev. 1989; 21(1): 79-83.

Gill LE & Baldwin JA. Methods and technology of record linkage. some practical considerations, in: Textbook of Medical Record Linkage (Baldwin JA, Acheson ED, Graham WJ,eds) Oxford: Oxford University Press, 1987: 39-54.

Harris S. The real number of road traffic accident casualties in the Netherlands: a year-long survey. Accid. Anal. & Prev. 22(4): 371-8, 1990.

Henderson J, Goldacre MJ, Graveney MJ & Simmons HM. Use of medical record linkage to study re-admission rates. British Medical Journal 1989; 299:709-713.

Hill T & Pring-Mill F. Generalised Iterative Record Linkage System: GIRLS (Glossary, Concepts, Strategy guide, User guide). Statistics Canada, Ottawa: 1981.

Howe GR & Lindsay J. A generalized iterative record linkage computer system for use in medical follow-up studies. Computer and Biomedical Res. 1981; 14: 327-340.

Newcombe HB. Handbook of Record Linkage. Methods for health and statistical studies, administration, and business. Oxford University Press. Oxford: 1988.

Newcombe HB, Fair ME & Lalonde P. The use of names for linking personal records. J. Am Stat Assn 1992; 87:(420) 1193-1207.

Roos LL & Wadja A. Record Linkage Strategies. Meth. Inform. Med. 1991 30: 117-123.

Roos LL, Wadja A. & Nicol FP. The art and science of record linkage: methods that work with few identifiers. Comput Biol Med 1986; 16: 45-47.

Rosman DL & Knuiman MW. A Comparison of Hospital and Police Road Injury Data. Accid. Anal. Prev. 1994 in press.

States JD, Annechiarico RP, Good RG, Lieou J., Andrews M, Cushman L. & Ingersoll G. A time comparison study of the New York State Safety Belt Use Law utilizing hospital admission and police accident report information. Accid. Anal. Prev. 1990; 22(6): 509-521.

Steadman, LA & Bryan, RJ Cost of Road Accidents in Australia. Bureau of Transport & Communication Economics, Occasional Paper 91, 1988.

APPENDIX I

Phonetic Codes

Since it is often difficult to gain access to official records containing names, an investigation of the feasibility of linking hospital and police records using a phonetic coding of the family name in place of the family name and given name(s) was performed. It is possible that the use of phonetic codes may overcome some of the reservations of data custodians while still retaining enough identifying information to facilitate the linkage process.

The reason for using phonetic coding systems is usually to bring together variant spellings of what are essentially the same names. These coding systems suppress the vowel information due to instability and replace similar sounding consonants by a standard character or group of characters. There are several alternative methods, two of which have been considered here. The NYSIIS (New York State Intelligence Information System) coding scheme uses up to 6 characters to define a name. The vowel positions are retained by replacing all vowels or groups of vowels with the single character 'A' and similar consonants eg K - C; M - N; or KN - N are standardized. The SOUNDEX system is less specific in that all vowels are dropped and more consonants grouped together (eg B,P,F,V). The final code contains up to 4 characters the first of which is the first letter of the family name and the remainder are numbers. (Newcombe, 1988).

For this experiment, it is assumed that only phonetic coding of family name was available, with no indication of first name or initial. In fact, it may have been possible to obtain greater discriminating power, especially between family groups if some part of the first name were available.

APPENDIX II

Methods

Details for each pass using named data

The first pass used the NYSIIS (New York State Intelligence Information System) phonetic code of the family name as the pocketing variable and only records falling in the same pocket were compared. (This coding scheme is described in Newcombe (1988)). Variables used for comparison were the family name, first name or initial, age, sex, road user type and accident date. Sophisticated variations of first and second given names as described by Newcombe et al (1992) were not necessary as most of the police records only contained initials rather than full given names.

Default frequency weights were used for name, NYSIIS and sex, but the frequency weights for age and accident date were modified to take into account the relative distance between values in the two sources. For instance, a large negative weight was given if the date of admission to hospital preceded the accident date on the police record. An exact match or an admission one day after the crash date were given almost equivalent weights to allow for late night crashes. Differences in dates of more than a month were given zero weight. Weights for age were allocated to reflect the fact that the age of children could be known more precisely than older adults. So that age differences of 3-5 years could be tolerated in a casualty aged 50 but not in one aged 15.

Records not linked in the first pass were processed in the second pass where the crash/admission date was used as the pocket variable. Some permutations of the family name were allowed here, to take into account some known spelling variations - eg DURNIN and DURWIN. Thus if the family names on two records being compared had more than 6 letters and only one of these was different, the names were treated as equivalent and weighted accordingly. In addition, since it was known from previous linkage of the same data, that transcription errors in the police records sometimes led to

'10' being entered as '70' or '23' as '53' for age, this could be taken into account in the weighting scheme.

The third pass again used dates as the pocketing variable, but only compared hospital and police records if the crash occurred on the day before the hospital admission date. The weighting scheme used was the same as the second pass.

Only records which had not previously been linked but which matched exactly on age and sex were entered into the fourth pass. Crash dates, family name, first name, and road user type were used for weighting within these groups

APPENDIX III

Results

Details for each pass using named data

The first comparison sequence or 'pass' extracted 566 records whose weight exceeded the threshold of 180. Of these, 541 were links previously defined by the GIRLS linkage, 7 were new links not found previously and 18 were mismatches. The records whose total weight was below the threshold on the first pass were submitted to the next pass.

The 'threshold' chosen for each pass depended on the number of comparisons to be performed. A higher value was needed in the first pass, which pocketed by NYSIIS code and compared family name, the first three letters of family name, the first given name, its first 3 letters and initial, than the second pass which had fewer comparisons.

For the second pass, a threshold of 70 was required to achieve the 28 GIRLS links with an extra 3 links being identified. There were 54 mismatches collected in this pass due to the low cut-off weight. From examination of the GIRLS links it was evident that some links whose dates agreed exactly were being given low weights because of name spelling variations and large age differences. A threshold of 70 captured all of these. The remaining record pairs whose total weight in the second pass was less than 70 were then entered into the next pass.

For the third pass, where crash dates were pocketed with hospital admissions a day later, the cut-off weight of 100 produced 7 GIRLS matches, 2 extras and 4 mismatches. The fourth pass, which pocketed on age and sex alone, had a cut-off weight of 110 and only gained a further 2 GIRLS links, with one additional link not found previously and 3 mismatches.

To check whether any more links could be found, the threshold for the first pass was lowered from 180 to 160, producing 33 extra 'probable' links. If this was followed

through, 33 fewer records would have entered into later passes. Since eleven of the twelve old GIRLS links had been captured later in the second or third pass and the only old GIRLS link not found later had an age difference of 19 years, this adjustment was not made. Since the threshold weight for these later passes had already been set to capture all GIRLS links possible, the number of links was considered to be optimum.