Probabilistic Linkage of Victorian Injury Data Records

D’Elia, A. and Newstead, S.
Monash University Accident Research Centre

Abstract

Research utilising the available mass databases of real world road crashes compiled by Police, government and insurers in Victoria provides much useful information for injury prevention research purposes. The record linkage of these datasets and hospital injury records has the potential to maximise the use of available data by researchers to extend the understanding of the causes, outcomes and costs of road traffic and other injuries. This project explored the feasibility of linking Police-reported motor vehicle crash, Transport Accident Commission (TAC) insurance, in-depth crash data and hospital injury datasets using probabilistic linking routines. It was found that the probabilistic linkage of hospital data to other injury datasets was not feasible without identifying information, however the probabilistic linkage of Police and TAC data did identify potential alternative measures of serious injury that could be used for road safety research.

Keywords

Record Linkage, Probabilistic, Victoria

Introduction

Research utilising the available mass databases of real world road crashes compiled by Police, government and insurers in Victoria, all held and accessed by the Monash University Accident Research Centre (MUARC), provides much useful information for injury prevention research purposes. Separately the data sources have been critical for a wide range of important research carried out by MUARC that has led to the development of new countermeasures and evidence-based preventive policies and programs. However, the linkage of these datasets with in-depth crash data and hospital administrative health (injury) records has the potential to maximise the use of available data by researchers to extend the understanding of the causes, outcomes and costs of road traffic and other injuries.

As has occurred in Western Australia, research based on linked data has the potential to enormously enhance the evidence base for road and non-road injury prevention policy making, preventive measures and evaluation of intervention programs. Further benefits include assessment of data quality issues through cross comparison and assessment of the sampling frames and coverage within each dataset. Effective linking of the datasets would potentially lead to more robust and better qualified results from research that would previously have only analysed the datasets individually.

The feasibility of linking the following datasets was assessed in this project:

State Traffic Accident Record (STAR): This dataset is an extract from VicRoads’ Road Crash Information System (RCIS) and contains records of every road crash involving injury reported to Police in Victoria. Victoria Police collect the data; it is further enhanced by VicRoads who supply de-identified data to MUARC. The data covers crash circumstances, road user and vehicle information in great detail but only has limited information on the injuries sustained by persons involved in the crash. MUARC holds complete records from 1987 onwards.

Transport Accident Commission (TAC) Claims Data: This dataset covers all claims made to the TAC for third party injury compensation for injuries arising from transport accidents. The database has limited detail on crash circumstances and the road users involved, but has ICD (International system for the Classification of Diseases) codes for up to the worst 5 injuries assessed from the claim along with less
detailed codes on body region of injury by severity. MUARC holds complete de-identified data on TAC claims from 1987 onwards.

Australian National Crash In-depth Study (ANCIS): In-depth data on vehicle crashes have been collected at MUARC since 1988 and have proven valuable for a variety of analyses conducted for the Australian Federal Government and the auto industry. The ANCIS study collects great detail on occupant injury, crash circumstances and vehicle damage. Injury is coded using the Abbreviated Injury Scale (AIS). The sampling frame for ANCIS generally covers vehicles aged 10 years old or less that are involved in a crash that resulted in injury to a vehicle occupant that required hospital admission. The ANCIS dataset also includes cases recruited as part of the Holden Crash Investigation project.

Victorian Admitted Episodes Database (VAED): The VAED (injury subset) held by MUARC (and provided by the Department of Human Services (DHS)) contains de-identified records of injury hospital admissions to Victorian public hospitals from 1987 and private hospitals from 1994. VAED data are coded using ICD-10 Australian Modification coding system and includes demographic data (age, gender, postcode of residence, country of birth) and cause, activity, location, human intent and diagnoses codes.

The first major objective of the project was to link records from ANCIS with corresponding records in the TAC and Police mass crash data sources. This would allow data quality issues in the mass crash data sources, such as accident details and the critical issue of injury coding to be assessed. The second major objective of the project was to link injury hospital admission records with the mass crash data records to further validate and enhance the information derived from the analysis of the mass crash data records. This would allow the potential of the linked dataset for a broader variety of injury research to be examined, contingent on the success of linking the hospital data.

Method

Linkage was performed on records relating to individuals. Each dataset was provided without identifying information such as name, address and date of birth, necessitating the use of probabilistic linkage techniques. A probabilistic approach utilises less stringent matching criteria compared with a deterministic one, where a pair of records would only be linked if an exact match occurs on the selected identifiers. The method used for this project required an exact match for the key (non-identifying) variables, but included additional matching stages where selected variables were dropped from the matching “key” in order to improve the overall match rate. This was balanced against achieving as accurate a match as possible, with additional variables used to reject linked records that did not meet specified criteria.

Ethics approval for linkage of the STAR, TAC and ANCIS datasets was obtained from the Monash University Standing Committee on Ethics in Research involving Humans (SCERH). Further ethics approval was then obtained from both the DHS Human Research Ethics Committee (HREC) and the Monash University SCERH for linkage of the VAED.

The general approach for record linkage across datasets included identifying the variables in each dataset that are common, assessing the completeness and accuracy of each common variable, and reviewing their suitability for use in a matching key. The STAR and TAC datasets were linked first, followed by the ANCIS dataset and then the VAED. Linkage was performed for the period 2000 to 2004 inclusive, except for the VAED where it was performed for 2004 only.

Linkage of the STAR and TAC Datasets

For linkage of the STAR and TAC datasets, fifteen variables were identified as common. These variables were classified as relating to accidents (e.g. accident number, number of vehicles involved), vehicles (e.g. registration number, vehicle make) or individuals (e.g. road user type, sex, age). As the TAC dataset covers all claims made to the TAC for injuries arising from transport accidents, it was assumed that for every case listed in the TAC dataset, there existed a corresponding case within STAR apart from claims made for interstate crashes.
Of the common variables, Police accident number (ACCNO) provided the clearest link between the TAC and STAR datasets. In particular, it provided a unique way to match a specific crash in each dataset. Once a specific crash was identified, registration number (REGNO) provided the most unique way of distinguishing between vehicles so that the next step in identifying individual claimants could occur. This was achieved through the use of the common variable USERCODE. For the TAC dataset, USERCODE was based on a variable that identified seating position for vehicle-related claimants or road user type for non-vehicle claimants (e.g. pedestrian). For the STAR dataset USERCODE was based on several variables identifying similar information. It should be noted that the common variable REGNO was only used for matching registration numbers for vehicle-related cases. Matching of non-vehicle cases occurred via blank string values so as not to reduce the overall match rate. Finally, the common variable SEX was added to ensure accuracy, particularly for non-vehicle cases.

Matching of the STAR and TAC datasets was performed using SPSS 15.0 in three stages with the following common variables as matching keys:

Stage 1
ACCNO, REGNO, USERCODE, SEX

Stage 2
ACCNO, USERCODE, SEX

Stage 3
ACCNO, REGNO, SEX

Stage 2 attempted to address the issue of missing values under REGNO, whilst Stage 3 attempted to deal with missing values under USERCODE. Only cases not matching in a prior stage were matched in each subsequent stage.

Each stage of the matching process used the common variables AGE and ACCDATE to reject linked records that did not meet specific criteria. For ACCDATE this occurred where the difference in the date of the accident was equal to or greater than 31 days, and for AGE, rejection of matched cases occurred where the age difference was equal to or greater than 4 years. For example, at Stage 1 around 2% of initially linked cases were rejected. Figure 1 shows a flow diagram for linkage of the STAR and TAC datasets. The file extensions *.sav and *.sps represent SPSS data files and syntax files respectively.

Linkage of the ANCIS Dataset

The ANCIS dataset provides detailed data for a relatively small number of crashes. It was assumed that for every Victorian case listed in ANCIS, there existed a corresponding case within STAR. Inclusion of ANCIS cases in a general linked dataset involved their initial linkage to STAR. The subsequent matching of TAC cases is based on the protocol for the linkage of STAR and TAC datasets as outlined above.

Linkage of the STAR and ANCIS datasets was performed in two stages using the following common variables as matching keys:

Stage 1
ACCNO, USERCODE, SEX, AGE

Stage 2
VIN, USERCODE, SEX, AGE

where VIN is the Vehicle Identification Number.
Figure 1: Matching protocol for linking the STAR and TAC datasets
Overall, the linkage process was designed to generate accurate linkage rather than a high match rate thus allowing for meaningful cross-comparisons to be made between the ANCIS dataset and the STAR and TAC datasets. The inclusion of AGE in the matching key (requiring an exact match), together with the relatively small number of ANCIS cases helped to ensure that this occurred. In principle, the linkage of data would first utilise the variable ACCNO to identify a particular crash and then VIN to identify a particular vehicle (REGNO is unavailable in the ANCIS dataset). However a linkage stage that included both variables was deemed to be unnecessary leaving two stages in the matching process as described below.

Stage 1 linked cases with matching accident numbers whilst allowing for linkage where the vehicle identification number was missing. This stage also took into account cases with incorrect or non-applicable values for the vehicle identification number. Stage 2 matched cases where an exact match was found for VIN and allowed for linkage where the accident number had not been recorded. This constituted a large proportion of cases in the ANCIS dataset. Following the use of ACCNO in Stage 1 and VIN in Stage 2, the common variable USERCODE was used to identify a particular occupant via seat position, followed by the variables SEX and AGE. Consistent with the aim of achieving accurate linkage rather than a high match rate, each stage of the matching process used the common variable ACCDATE to reject matched cases that did not meet the strictest possible criterion, namely where the difference in the date of the accident was greater than 1 day.

Linkage of the VAED

For the purpose of linking the VAED to the STAR, ANCIS and TAC datasets, the VAED (injury subset) was obtained for the period July 2003 to June 2005 inclusive, including re-admission data. A matching protocol was then developed to link 2004 data from the VAED. Prior to linking the VAED, the relevant VAED cases were selected from the data provided. As a probabilistic matching process was required due to the use of de-identified data, it was critical that the number of cases was reduced in order to obtain the highest level of accuracy possible. This was achieved using the following process:

- The period January 2004 to January 2005 inclusive was selected;
- Land transport accidents (and unspecified transport accidents) were flagged; and
- Cases with a TAC account classification were also flagged of which 15% were related to non-land transport accidents.

In general, non-land transport accidents with a TAC account classification appeared to be re-admissions. As the VAED is a separations database, it was not possible to link re-admission cases to earlier, original hospital admissions in the absence of identifying information, e.g. a person’s name and address. Therefore, these cases could not be included in the VAED data subset for matching purposes. In addition, the account classification variable contained many missing values. Overall, this resulted in the selection of land transport accidents (and unspecified transport accidents) for matching purposes.

Before linking the selected VAED hospital admission records, it was necessary to supplement the matched STAR-ANCIS-TAC dataset with any unmatched (leftover) TAC cases. However, it was only necessary to add the relevant unmatched TAC claims, namely those with hospital admission. In the linkage of STAR and TAC data, TAC claims involving both hospital and non-hospital admission were matched, however as described above, for the linkage of hospital data it was critical to keep the number of cases to a minimum in order to obtain the highest level of accuracy possible, with unmatched TAC claims not involving hospital admission no longer required. The development of a matching protocol for the linkage of VAED data to the mass crash data records is described below.

Possible variables in the VAED that could be utilised in a matching key included:

- admdate – admission date;
- age – age in years;
- campname – name of hospital;
- pstcode – patient postcode; and
- sex – sex of patient.

All of these variables were essentially complete, i.e. contained no missing values.

As described above, leftover TAC hospital cases were added to the original linked STAR-ANCIS-TAC data file which effectively created the same number of missing values for each of the STAR variables. Clearly in identifying common variables for matching between the supplemented STAR-ANCIS-TAC and the VAED datasets, the use of TAC rather than STAR variables was preferable considering that the relevant TAC variables contained few if any missing values. However the use of several matching stages meant that a matching key could still include a common variable derived from STAR in initial steps where the aim was to produce a higher degree of accuracy. In particular, the TAC claims data did not contain the name of the admitted hospital, therefore the STAR variable listing the hospital name was utilised to create the common variable HSP.

Linkage of the supplemented STAR-ANCIS-TAC and VAED datasets was performed in six stages using the following common variables as matching keys:

Stage 1
ADMDATE HSP PSTCODE AGE SEX

Stage 2
ADMDATE PSTCODE AGE SEX

Stage 3
ADMDATE1 HSP PSTCODE AGE SEX

Stage 4
ADMDATE1 PSTCODE AGE SEX

Stage 5
ADMDATE PSTCODE SEX

Stage 6
ADMDATE1 PSTCODE SEX

Each stage of the matching process was designed to become slightly less restrictive. Stages 1 and 2 required an exact match for hospital admission date. It should be noted that hospital admission date from the VAED was matched to date of accident from the supplemented STAR-ANCIS-TAC data file. In cases where admission was listed as the day after an accident, Stages 3 and 4 used identical matching keys, however shifted the date for matching appropriately (ADMDATE1). Stages 1 and 3 both used the variable HSP for accuracy, with this variable being dropped in Stages 2 and 4 respectively due to missing information. Stages 5 and 6 did not include HSP and loosened the restrictions on age however did reject matched cases where an age difference of greater than 1 year occurred.

Results

STAR and TAC Datasets – Match Rates

Table 1 shows the match rates achieved for the years 2000 to 2004 inclusive for the STAR and TAC datasets. The match rate is defined as the number of matched cases divided by the number of cases in the TAC dataset, expressed as a percentage. It is reasonable to expect that for every TAC claimant, a corresponding case is contained in the STAR data file. It follows that for the probabilistic matching process utilised here, match rates lower than 100% are due to missing information and inaccuracies in the data.
Table 1: Match Rates for 2000 to 2004

<table>
<thead>
<tr>
<th>Year</th>
<th>Match Rate (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>2000</td>
<td>75</td>
</tr>
<tr>
<td>2001</td>
<td>75</td>
</tr>
<tr>
<td>2002</td>
<td>76</td>
</tr>
<tr>
<td>2003</td>
<td>72</td>
</tr>
<tr>
<td>2004</td>
<td>68</td>
</tr>
</tbody>
</table>

Cross-Comparison of the STAR, TAC and ANCIS Datasets

The linkage of records from ANCIS with corresponding records in the TAC and STAR datasets allowed data quality issues in the mass crash data sources, such as accident details and the critical issue of injury coding to be assessed. This was achieved through the cross-comparison of STAR and TAC variables with corresponding ANCIS variables in the matched STAR-ANCIS-TAC dataset. In a preliminary step, the quality of linkage between the STAR and TAC datasets was assessed through a comparison of Police-collected accident details, namely the number of vehicles involved in an accident (NOVEH) and the Definitions for Classifying Accidents (DCA) code. As the corresponding variables in each dataset come from the same source and have not been utilised in the matching process, it may be expected that they have the same values where a correct match has occurred. Equal values occurred for 96% of matched cases for NOVEH and 77% of matched cases for DCA.

Comparison of NOVEH and DCA was also performed for the linked STAR-ANCIS-TAC dataset. For the matched STAR-ANCIS-TAC cases, equal values occurred between STAR and ANCIS for 80% of cases for NOVEH and 42% of cases for DCA. NOVEH and DCA were originally chosen as they were present in each of the STAR, TAC and ANCIS datasets, however for the ANCIS dataset the values of these variables were determined by the in-depth crash investigators post-accident rather than coming from Police reports as per the STAR and TAC datasets.

With regards to injury coding, it was useful to compare variables in the TAC dataset that could be used as measures of serious injury with corresponding variables that were available in the ANCIS dataset. A report by Hoareau et al. [1] analysing injury severity trends in TAC claims between 2000 and 2004, referred to hospitalisation as derived from Police crash reports as a relatively coarse definition of serious injury. A number of other potentially finer serious injury measures derived from the TAC data included: average length of hospital stay, Maximum Abbreviated Injury Score (MAIS), Injury Severity Score (ISS) and ICD Based Injury Severity Score (ICISS).

The measures MAIS and ISS were both available in the ANCIS dataset, hence a cross-comparison of these alternative measures of serious injury was performed. For the year 2000, 45 ANCIS cases were successfully linked to corresponding STAR and TAC cases in the combined STAR-ANCIS-TAC dataset. Figures 2 and 3 show cross-comparisons for MAIS and ISS respectively. A good correlation ($R^2 = 0.72$) exists between the TAC and ANCIS datasets for the injury measure ISS with a weaker correlation ($R^2 = 0.50$) present for MAIS. We note that a broader assessment of TAC injury severity measures involving many more cases, i.e. between matched TAC and VAED data was contingent on the successful linkage of the VAED to the STAR, ANCIS and TAC datasets.
Linkage of the VAED

A poor match rate of 36% was achieved for the linkage of the VAED to the supplemented STAR-ANCIS-TAC dataset. This indicated that the probabilistic linkage of the VAED to the STAR-ANCIS-TAC dataset was not feasible without identifying information. Rather, the effective matching of VAED data to other Victorian injury data records seemed to require a matching process based on person details.
Discussion and Conclusions

Cross-comparison of the injury severity measure ISS between the ANCIS and TAC datasets ($R^2 = 0.72$) suggested that the use of TAC claims data merged with STAR could provide an alternative and finer measure of serious injury for road safety research to hospitalisation as derived from Police crash reports. However, the use of the common variable ACCNO is limited as a matching variable due to the number of missing values present in the TAC dataset, leading to a reduction in the maximum possible match rate. To improve the overall match rate, the TAC would need to ensure that all claims contain a valid and verified Police accident number.

An alternative to the linkage of the STAR and TAC datasets would be for the TAC to match claims data to Police data in-house. In-house linkage has been performed by matching Police-reported crash data from the Victoria Police Traffic Incident System (TIS) with TAC claims data and STAR. The TAC have reportedly achieved a match rate of 85%, with the non-matched claims explained by interstate and public transport cases.

These results suggest that linkage of Police-reported crash data with the TAC claims dataset is feasible and results in a combined database more capable of measuring injury outcome in detail. This is particularly significant considering that the probabilistic linkage of hospital admissions data was not found to be feasible without identifying information. Although the TAC claims dataset was put together primarily to manage claims, a new project has been started that seeks to collaborate with the TAC in order to develop a linked dataset for use in road safety research. This will include a review of the injury coding practices of the TAC and a process of consultation with the key agencies to automate and streamline the linkage process into an ongoing system.

Future work will identify the most appropriate measure or measures of serious injury from the linked dataset to apply to a suitably long time period for use in road safety performance monitoring and other projects, e.g. time series modelling. In particular, operational biases inherent in serious injury measures derived from Police crash reports (currently taken to hospital, previously being admitted to hospital) should be overcome. Indeed, the use of an alternative and finer measure or measures of serious injury should more closely reflect real serious injury trends and provide a more satisfactory key outcome measure for Victoria’s arrive alive 2008 – 2017 road safety strategy.

The second major objective of the project was to link injury hospital admission records with the mass crash data records to further validate and enhance the information derived from the analysis of the mass crash data records. However this objective was not achieved. It was found that the probabilistic linkage of the VAED to other Victorian injury data records was not feasible, with other issues such as transfer and re-admission cases making linkage of de-identified data challenging. For example, the VAED is a separations database and these cases appeared as new additional cases. Clearly identifying information would be required to link relevant cases within the VAED prior to the linkage to other Victorian injury data records. This is particularly important if comparisons between datasets are to be made, e.g. ICD codes between the VAED and TAC claims data.

In order to link hospital data with mass crash data sources, it is recommended that Victoria set up a Data Linkage Unit (DLU) similar to that in Western Australia. Each year in WA, the DLU links road crash reports to hospital morbidity and mortality data using personal identifiers. Main Roads WA provides the identified road crash data to the DLU and the identifiers are removed after linkage. This dataset is then used to produce the annual report on road crashes in Western Australia [2]. The linkage of hospital and Police crash datasets using identifying information has also occurred in NSW. The Injury Risk Management Research Centre matched records using information that was common to the datasets, such as names, residential addresses and dates of birth. When hospital inpatient separations data was restricted to records relating to motor vehicle traffic crashes on public roads a match rate of 69% was achieved [3]. In addition to linking road crash and hospital data, a Western Australian style data linkage unit using identifying information would also enable the linkage of Victoria’s core population health datasets. This currently occurs in Western Australia with access possible for bona fide research and evaluation projects that have been approved by an institutional ethics committee.
Acknowledgements

This project was funded through the Monash University Accident Research Centre’s Baseline Research Programme for which grants have been received from the following Victorian government agencies: the Department of Justice, the Roads Corporation (VicRoads) and the Transport Accident Commission (TAC).

References


2. Western Australian Police Service, Main Roads Western Australia, Insurance Commission of Western Australia, Department of Health, Department for Planning and Infrastructure and Office of Road Safety (2004). Road crash data sharing between government agencies, Report for the Road Safety Council.