LINKING ONTARIO COLLISION, VEHICLE REGISTRATION AND TRAUMA DATA

Glenn Robbins
Transport Canada
Canada
Paper Number 98-S6-W-39

ABSTRACT

A feasibility study was conducted to link the police reported traffic collision reports with both trauma data and vehicle registration data. For the purposes of this study, the data was limited to that from the Province of Ontario. The study examines the benefits of being able to do such linking and the limitations of the data sources that affected the number of successful matches. The study also investigates some of the barriers of obtaining and linking the data on a regular basis such as ownership, security and confidentiality.

OVERVIEW

Certain research questions that arise sometimes cannot be answered using a single data source. In order to answer complex questions it is advantageous to link databases together. For example, there was interest in examining the fitment of seat belts in light duty vehicles by analyzing injuries that were sustained by the seat belt wearing occupants of vehicles involved in a collision. The requirement was to determine if there was significantly more injuries caused by seat belt wearing in similar type crashes across different makes and models of vehicles.

In order to draw useful conclusions from this type of analysis there was a requirement to obtain several pieces of information: the type of collision that occurred (e.g. side impact, rear-end etc.), the type of injury that was sustained to occupants of the involved vehicles, whether the occupant was restrained, and the type of vehicle (including make, model and model year) in which the person was an occupant. All of this information was not contained in a single database. However, the existence of separate databases which together could contain all the necessary elements, presented the opportunity to attempt to electronically link them.

This paper examines existing databases that could be used in such analysis, describes how they might be linked and reports on the attempt to link them. This study does not analyze specific safety related issues such as the question posed above. The scope was simply to look at the process of linking the selected databases and discuss the issues related to such a process.

SOURCES OF DATA

For the purposes of this study, data was limited to the province of Ontario for the years 1991 to 1994, inclusive. This limitation was required due to the availability of required data. The identified sources were: collision data (Ontario Motor Vehicle Accident Report System (OMVARS)), hospital discharge data (Ontario Trauma Registry Comprehensive Data Set (OTRCDS)) and vehicle registration data (Ontario's Vehicle Registration Database (OVRD)).

Collision Data

Transport Canada creates a national collision database each year based on electronic data supplied by each of the ten provinces and two territories called the Traffic Accident Information Database (TRAID). Although each of the jurisdictions may code and record their collision report information differently, a protocol had been designed to compile all the collision data into a similar format. The initial idea was to use the Ontario data as compiled in the TRAID database as the basis for the collision data required for this study. However it soon became clear that the use of the data contained within the Ontario Motor Vehicle Accident Reporting System (OMVARS) would be useful in its original format for the following reasons:

1) The system records the date of birth for all drivers, TRAID has age only. Date of birth is more specific than age of person and therefore, at least for drivers, would help enable more accurate matching as date of birth is contained with the selected hospital data (OTRCDS).

2) TRAID does not store the vehicle plate number or the Vehicle Identification Number (VIN). However, the OMVARS records the license plate number and jurisdiction (province, territory, state, country) of each involved vehicle. The plate number could potentially be used to link into a registration database to get specific vehicle information.

3) The OTRCDS is designed to collect and store collision related information. Much of that information is collected using codes similar to the OMVARS.

There is some ability to segment the persons involved in collisions by the severity of the injuries sustained (this is true in both TRAID and OMVARS). Although the variable 'injury severity' is separated into three broad injury related categories (minimal, minor, major), it does provide a reasonable method to segregate the injury data. It is expected that those coded with a 'major' injury should be a reasonable estimate of those persons who were admitted to hospital for at least one night as the OTRCDS is limited to such cases. The number of major injuries recorded by the OMVARS is listed in Table 1 below.

Table 1
Number of Major Injuries
As Reported in Ontario
(Motor Vehicle Accident Reporting System)

Year	Major Injuries
1991	7,005
1992	6,690
1993	6,644
1994	6,023
Total	26,362

Hospital Patient Discharge Data

For the purposes of this study, the Ontario Trauma Registry Comprehensive Data Set (OTRCDS) was used as the source for hospital based discharge data. This database is a compilation of patients who were admitted to one of twelve 'lead' trauma hospitals in the province of Ontario. Although this database contains only a portion of persons admitted to all Ontario hospitals, there are certain advantages that make it better than more typical hospital discharge data for linking. Some of these advantages are :

1) The database is designed to capture collision related information such as the vehicle type, location of vehicle impact, seating position and collision configuration. These type of data elements make it extremely useful when trying to link with standard collision data reporting systems.

2) Injury information is collected using the Abbreviated Injury Scale (AIS). This coding scheme is very familiar to the analysts at Transport Canada as it is used primarily in the road safety community for examining the extent of injuries that happen in motor vehicle collisions.

3) The date and time of injury is recorded rather than just time and date of admission to hospital. It is possible that the time of admission may lag several hours after the actual time of the injury whereas the time of injury should be close , if not identical to, the time of the collision reported in the OMVARS.

4) The data is recorded in such a way to make it possible to differentiate persons who were involved in a traffic related incidents to those who were not. The database uses the coding system for external causes of injury (E-Code) as defined by the World Health Organization's Internal Classification of Diseases. In this study, cases that had an ecode of 810 - 829 were selected.

Collection of data for the OTRCDS began in 1991. At the time of this study, the last complete year of data was 1994. It is evident that the quality of the data improved over those years, especially for collision specific variables such as collision configuration which has improved in reporting from 26% in 1991 to 60% in 1994. Similar improvements are evident across all collision specific variables.

Due to concerns over confidentiality there were certain variables not available to the author for this study that may have otherwise been useful in making successful linkages. Three in particular were the person's name, the collision report number and the geocode. All of these could be extremely useful in the linking process. It is encouraging that the latter two have been included in the design of the OTRCDS.

Table 2
Persons Admitted to Lead Trauma Centers
Injuries due to Motor Vehicle Collision

Year	Trauma Records
1991	816
1992	1,192
1993	1,357
1994	1,369
Total	4,734

Registration Data

Although the generic vehicle type (e.g. car, tractor trailer, school bus etc.) and model year are captured in the collision data, there is extremely limited specific vehicle information collected in motor vehicle accident reports. In the OMVARS, the vehicle make and model information is limited to a total of eight characters and there is no standard for how these eight characters are coded. Therefore a variety of codes and sometimes extremely cryptic combinations exist in the collision data. Combination of characters such as 'FORDESCRT', 'FORDESCOR', 'FRDESCRT', ESCRTFRD, 'ESCORT" and even 'FORD" are all coded for 'Ford - Escort'. In such a format there is no allowance for other information such as size, series etc. This limitation makes doing any sort of analysis on specific make and model extremely difficult, if not impossible.

The OMVARS does capture the vehicle license plate number and jurisdiction. Therefore, theoretically, it should be possible to extract this information and try to verify which vehicle is registered to a specific plate. The registration file is suppose to contain a Vehicle Identification Number (VIN). The VIN, when decoded, has the potential to give specific information such as vehicle make, model, series and other vehicle specific information.

There were two concerns about this process. First, in order for such a algorithm to work, the registration database must keep a history of the vehicles that have been registered to a certain plate. This is imperative as the linking to the registration file is done after the collision has occurred (in some cases several years). The vehicle registered to a license plate may have changed over time. This is possible as persons change vehicles either by desire (sell old vehicle, replace it with new) or by necessity (vehicle is involved in collision and is unrecoverable). Persons are permitted to register subsequent vehicles to the same plate. If no history is maintained then it is impossible to determine the VIN that was registered to the plate at the time of the collision. This could have a very profound influence on any analysis done as vehicle may have originally been a Chevrolet Caprice but now the plate is registered to a Ford F-10 pickup. The Ontario registration database is designed to maintain such a history.

The second concern is that the VIN that is registered to the plate is valid. Since the VIN is a 17 character code there was concern over the likelihood of invalid VIN's entered into the registration system. If the VIN is invalid it will be impossible to decode. Of the VIN's of light duty vehicles (automobiles, light trucks and vans) involved in fatal or non-fatal injury collisions during the years 1991-1994 inclusive that were entered into the decoding software 'VINDICATOR', about 85% were able to be decoded. Whether this is similar to the registration file as a whole was unable to be determined in time for this paper.

It is important to realize that there are some vehicles from other provinces and countries involved in collisions that occur in the province in Ontario. Fortunately, on average, between 1991 and 1994 about 95% of all vehicles involved in injury and fatality producing collisions in Ontario were registered in Ontario. The other 5% were ignored for the purposes of this study.

LINKING COLLISION DATA TO HOSPITAL DATA

Potential Linking Variables

All variables that are similar in both databases are perspective linking variables. As mentioned earlier, an advantage of using the OTRCDS is that it is designed to capture collision related variables such as collision configuration, person position, type of vehicle etc. that would not normally be captured as part of hospital discharge data. In order for such data to be useful there but be a reasonable level of recording of this data. In Table 3 is a list of the potential linking variables and the proportion of observations where the information was recorded.

An additional concern is that although similar data elements are collected in both the collision and hospital databases, different codes are sometimes used to capture the same information. Before the linking process can begin, this differing codes require modification if the process is going to be successful at all. Although for the most part the OTRCDS tried to be similar to the OMVARS it sometimes varied.

Table 3
Limitations of Potential Linking Variables

	Data Sources	
Data Element	**Trauma Registry**	**Collision Data (Ontario)**
Date of Collision	Coded as date of injury	100% Complete
Time of Collision	Hour & Minutes (9% Unknown)	Hour Only 0.5% Unknown
Person Gender	0.5% Unknown	0.5% Unknown
Person Age	0.1% Unknown	0.5% Unknown
Person Date of Birth	0.2% Unknown	Available for drivers only
Type of Collision	68% Unknown	0.1% Unknown
Type of Vehicle	60% unknown	In order to 'match' some 'translation' required.
Vehicle Damage Location	76% Unknown	2.5% Unknown
Person Position	43% Unknown	0.1% Unknown
Person Injury Severity	At least one night in a lead Trauma Hospital	Coded as 'Major Injury'
Person Ejected	44% Unknown	0% Unknown

Although not an explicit variable within the OTRCDS, the limitation is that it contains only persons who were admitted for at least one night. Therefore, to maximize the potential number of successful and accurate linkages, the records from the collision data were limited to major injuries. In addition, the number of cases contained in the OTRCDS are less than the total of all hospital admissions as there are only twelve lead trauma hospitals recording the data in this format. Therefore, total number of hospital records is a significantly small proportion of all persons sustaining major injuries as reported in the OMVARS (see Table 4 below).

Table 4
Compare Number of Hospital Discharge Records To 'Major Injury' Records

Year	Hospital Discharge Records	Major Injuries	Proportion of Major Injuries
1991	816	7,005	11.6%
1992	1,192	6,690	17.8%
1993	1,357	6,644	20.4%
1994	1,369	6,023	22.7%
Total	4,734	26,362	18.0%

Linking Process

A deterministic method was used for linking the databases. That is, the selected variables from each of the two databases (OTRCDS and OMVARS) had to have identical values in order to be linked. In addition, only a single record within each of the data sources could contain the same values as otherwise there was no constructive way to choose one of the records over the other. An extremely conservative approach would have been to use all the potential linking variables and simply select only cases that are matched in one iteration. However, in light of the number of observations contained within OTRCDS that had incomplete information recorded in the collision related variables, this was not practical.

The challenge was to minimize the possibility of incorrectly linking records, while at the same time maximizing results. The larger the number of variables used to match the records, the higher the probability of a correct link. The fewer variables used to match, the higher the probability of 'duplicates' and also the higher the probability of incorrectly linking cases.

For each iteration the process is the same (summarized in Table 5). First, the variables to be used for matching are selected. Each of the databases (hospital and collision) are separately sorted by the variables selected in step 1. After the sorting is complete, it is determined which of the records is unique within the database and which are duplicates. The result of this step is four distinct data sets, two for each of the hospital and collision databases (Hospital - duplicates of selected variables, Hospital - no duplicates, Collision - duplicates of selected variables, Collision - no duplicates). The cases that were determined as containing duplicate values are excluded from any possibility of matching during the current iteration as it is impossible to differentiate which of the records should be matched to any records with similar characteristics in the other database.

The fourth step was to merge together the two data sets containing no duplicates. Instances where the values of the selected variables match identically to a record in the other data set are linked. Records that are linked are output to the file containing all successful matches. The observations contained within the duplicate subset and non-matched subset are recombined for subsequent iterations.

With each iteration the variables selected for matching are changed. The challenge for such a process is to decide at which point the observations that have been linked are likely to be incorrect matched. That is, are there two few variables being used so that observations that are 'matched' do not really represent the same case.

Table 5
Basic Procedures in the Linking Process

Procedure	Explanation	Data Set(s) Created
Select Variables	Select variables from those listed in Table 3 to be used for matching.	
Sort Data	Sort hospital and collision that are not yet matched by the selected variables	1) Collision Data - sorted 2) Hospital Data - sorted
Separate Unique Records From Duplicate Records	If duplicates of variables to be used for matching exist then all similar records are excluded from current iteration for a potential match.	3) Collision Data - Duplicates 4) Collision Data - Non-duplicates 5) Hospital Data - Duplicates 6) Hospital Data - non-duplicates
Merge data sets by selected variables	Using only the proportion of the data with no duplicates, merge data sets 4) and 5) by the selected variables. Successful matches are added to linked records. Unsuccessful matches are retained for subsequent iterations	7) Linked Records 8) Collision Data - not linked 9) Hospital Data - not linked
Prepare records for subsequent iterations	Prepare remainder of records for subsequent iterations. Bring together subsets 3 and 8 and 5 and 9. Start next iteration.	

Results

The results of linking the collision and hospital databases are listed below in Table 6. On average there is about 65% of the records in the trauma data matched to the collision data. The fewest number of variables used during the process were the date, time, hour, gender and age of the person. Depending on the level of expertise with the data and comfort with linking, this may actually be too many or too few variables.

Table 6
Number of Trauma Records Linked to Collision Data

Year	Trauma Records	Collision Records	Number Matched	Proportion of Trauma Records Matched
1991	816	7,005	550	67.4%
1992	1,192	6,690	797	66.9%
1993	1,357	6,644	877	64.6%
1994	1,369	6,023	892	65.2%
Total	4,734	26,362	3,116	65.8%

LINKING COLLISION DATA TO REGISTRATION DATA

The end result desired by linking the collision data to the registration database is to be able to determine vehicle specific information about the vehicle involved in the collision. In order to do this there was a two step process. The first step was to determine the date and the license plate number of the vehicles involved in the selected collisions. Both had to be available for a possible match to take place. Therefore, if a plate for a vehicle was not present or invalid in the collision database then it would be impossible to link it to the registration file. The second step is to decode the VIN's that are registered to each of the plates. VIN's that cannot be decoded are considered invalid and also affect the final results. Only when there is a valid VIN that has been 'decoded' can the link be considered successful.

Results

On average, the proportion of the number of vehicles involved in fatal and injury producing collisions that had a license plate recorded in the OMVARS that has a license plate recorded was around 92% over the four years. The VIN was decoded for only light duty vehicles due to a limitation the software being use for this process. On average about 85% of the VIN's were being successfully decoded. Although the VIN was decoded, it was necessary to try to determine whether the VIN actually represented the vehicle that was in the collision. The only way to do this was to use the vehicle model year and the cryptic vehicle make information on the collision data. As it would have taken considerable time to verify all cases, about 50 per year were spot checked. There was close to 100% agreement.

FINAL RESULTS

Table 7
Number of Persons Linked To Make and Model
(VIN's decoded only for Light Duty Vehicles)

Year	Vehicle Type	Number matched	Vehicle Make and Model Matched	Proportion of Vehicle Occupants Matched
1991	Light Duty Vehicle	393	323	82.2%
	Other Vehicle	74		
	Pedestrian	83		
	Total	550		
1992	Light Duty Vehicle	588	497	84.5%
	Other Vehicle	112		
	Pedestrian	97		
	Total			
1993	Light Duty Vehicle	649	536	82.5%
	Other Vehicle	115		
	Pedestrian	113		
	Total	877		
1994	Light Duty Vehicle	688	564	82.0%
	Other Vehicle	99		
	Pedestrian	105		
	Total	892		

In order to determine the overall success of this process it is important to examine the number of hospital records that were able to be linked to persons coded in the collision database and, where these persons were occupants of vehicles, what type of vehicle they were in at the time of the collision. As can be seen in Table 7 about 82% of all light duty vehicle occupant records that were linked to hospital information were also linked to detailed vehicle information.

DISCUSSION

The overall success of linking the particular databases used in this study depended greatly on the linking together of the hospital data and the collision data. The availability of the VIN's was fairly high and was really an insignificant factor in the final number of records matched. Only matching about 65% of hospital records to collision records may seem disappointing. However, the potential for a higher ratio of matches is excellent due to the existence of other very useful variables such as person name, collision report number and geographical area that were unavailable for this study.

Certainly the actual number of records that were linked with both the vehicle and hospital information were fairly low, especially considering the type of research question posed at the beginning of the paper. By the time an analyst starts to segment a few hundred records across dozens if not hundreds of vehicle models, the number of records for each specific vehicle make and model is going to be extremely small and any resulting analysis likely to be inconclusive.

The potential of such linking on a larger scale cannot be overlooked. In addition to the question that was posed at the beginning of this paper, many other types of road safety related questions could be answered by linking such databases. One of the main ones could be the overall cost of collisions. The availability of hospital data could help analysts better calculate the costs (at least in terms of hospital stay and procedures required to help rehabilitate the injured persons). In addition, the OTRCDS is designed to capture the emergency services that are used and the amount of time required for each of these services. In a much broader sense, such linking could also help decide policy direction by helping determine the type of collisions that are causing the most severe injuries and helping organizations prioritize resource allocation for specific research.

All that said, there are obstacles and challenges to doing database linking. The more important ones are probably ownership and confidentiality. In cases where there are different owners of each of the databases there may be some contention as to who should be allowed to do the linking, who should be responsible for storing and maintaining the data, and who should be able to use the data on an ongoing basis. These issues are many times not separate from confidentiality, as certain laws may give certain rights to some who collect data but not to others. Also, in this electronic age some jurisdictions are bringing in legislation to limit the amount of linking of separate databases to protect the rights of individual privacy.

This was the author's first attempt in linking databases. I think this study has proven that such linking is, at least, technically possible. It is obvious that the larger the quantity of similar variables and the higher the quality of the information contained on the candidate databases, the more successful the linking will be. However, for linking to be successful and useful on a much larger scale will require the cooperation and goodwill of database owners, database operators and also legislators.

REFERENCES

Association for the Advancement of Automotive Medicine. 1990. *The Abbreviated Injury Scale, 1990 Revision.* Des Plaines, IL : Association for the Advancement of Automotive Medicine.

Highway Loss Data Institute. 1995. *Vindicator User's Manual, Vindicator 96 - Release No. 1.* Arlinton, VA : Highway Loss Data Institute.

Ministry of Transportation for Ontario. 1987. *Motor Vehicle Accident Report System.* Downsview, ON : Ministry of Transportation for Ontario.

Ontario Trauma Registry. 1995. *Ontario Trauma Registry Data Dictionary*, Don Mills, ON :Ontario Trauma Registry.

Transport Canada - Road Safety : Evaluation and Data Systems. 1995. *TRAID User's Guide.* Ottawa, ON : Transport Canada.

World Health Organization. 1977. *Manual of the International Statistical Classification of Diseases, Injuries, and Causes of Death.* Geneva : World Health Organization.