Achieving quality of Canadian crash data

Aline Chouinard
Canada
Transport Canada
aline.chouinard@tc.gc.ca

Jean-Francois Lecuyer
Canada
Transport Canada
jean-francois.lecuyer@tc.gc.ca

1. Introduction

Transport Canada maintains a large database of all reportable traffic collisions that occur on Canada's roads. This database, called the National Collision Database (NCDB) [1], contains data on all motor vehicle collisions reported annually to Transport Canada by the 13 provinces and territories (i.e. jurisdictions)

One of the strategic goals of Canada's Road Safety Vision 2010 is to improve the quality [2] of the data in the National Collision Database (NCDB).

2. Purpose of the study

The purpose of this paper is to present Transport Canada's strategy to improve the quality of the NCDB data.

3. What is data quality?

Data quality is defined by seven properties [3]:

- Relevance: Why do we collect this data? Does it meet our needs?
- Uniformity: Are all jurisdictions using the same definitions?
- Completeness: Are there missing fields or incomplete records? Complete data means that all data elements are provided and that there are no records with missing values.
- Timeliness: Are deadlines to submit the data being met?
- Accuracy: Does the data represent the reality?
- Comparability: How does the data compare to external sources of similar data?
- Coherence: Are the values of the different data elements consistent? Are the data records consistent with the values of the data elements? For example: Is the road surface wet when the weather is reported as "raining"? Does every vehicle have a driver?

4. Data and methods

We initially tackled completeness, timeliness and accuracy.

4.1 Completeness

The completeness of NCDB [4] data was measured for 2005 by calculating the percentage of "truly available" data elements in the NCDB data dictionary [5]. We say that a data element is "truly available" for a jurisdiction if the data element is provided and has fewer than 20% "unknowns". As an example, in NCDB there are 68 data fields. If 5 fields are not provided and another 13 fields have more than 20% unknowns, the completeness rate (i.e. 50/68) is 74%.

4.2 Timeliness

The timeliness of NCDB data for the years 2005 and 2006 was measured as the number of days the data was delivered past the deadline for each province/territory. A negative value for timeliness means that the data was delivered before the deadline.

4.3 Accuracy

NCDB accuracy was measured by linking NCDB with the data from the in-depth collision investigation files of Transport Canada's Collision Investigations and Research Division (CIRD) and comparing the data elements that are common to both databases.

The most recent five years of NCDB data, 2001 to 2005, were linked to the collision investigations (CI) data. The following CI studies were chosen because they matched the time frame:

- ACR5 and ACR6: Studies on air cushion restraint systems (air bags) (629 cases)
- ASF3 and ASF4: Special studies (212)
- SID4 and SID5: Side impact studies (140)

The collisions, vehicles and occupants in NCDB were linked to the CI data using the methodology described in Lecuyer et al. (2009) [6].

5. Results

5.1 Completeness

We obtained the "Truly available" scores by jurisdiction for fatal and serious injury collisions and all collisions. The completeness of NCDB by jurisdiction varies between 49% and 93% for all collisions and between 56% and 93% for fatal and serious injury collisions. The completeness of NCDB for fatal and serious injury collisions is higher than or equal to the completeness for all collisions for each jurisdiction.

5.2 Timeliness

We measured the timeliness of the NCDB data for the years 2005 and 2006. The deadline for receiving the data is August 31st of the year following the year in which the collisions occurred. "Days from Deadline" have negative values when the data was early and positive values when the data was late.

In 7 of the 13 jurisdictions, the 2006 data was timelier than the 2005 data. Moreover, the 2006 data was early for eight jurisdictions, in comparison to 7 in 2005.

5.3 Accuracy

5.3.1 Injury outcome

Table 1 shows that the NCDB injury outcome data concurred with the CI data for 74% of the occupants. The agreement was 62% for occupants with no injuries. The agreement was higher, at 78%, for occupants with injuries. For fatalities, the agreement was 91%. Among the eight fatalities in the CI data that are not fatalities in NCDB, six victims should have been counted as fatalities in NCDB because they died on the scene (or less than 30 days after the collision). Thus NCDB seems to underestimate the number of fatalities by approximately 7 to 8%.

Table 1: Comparison of injury outcome

Injury outcome from CIRD	Injury outcome from NCDB					
	No injury	Minimal -Minor	Serious	Fatality	Unknown	Total
No injury	220	96	2	0	36	354
Injury	114	369	124	0	29	636
Fatality	2	1	4	82	4	90
Total	336	466	130	82	66	1080

5.3.2 Restraint use

Table 2 shows that for injured and fatally injured occupants only, the two databases concurred with regard to restraint use for 67% of the occupants. The agreement is much higher for restrained occupants (77%) than for unrestrained occupants (23%). Thus the number of restrained occupants seems to be overestimated by approximately 12% in NCDB.

Table 2: Comparison of restraint use for injured/fatally injured occupants

Restraint use from CIRD	Restraint use from NCDB					
	Belted	Unbelted	Child restraint	Other/ Unknown	Not applicable	Total
Restrained	448	2	14	122	14	600
Unrestrained	59	28	0	32	4	123
Unknown	2	0	0	1	0	3
Total	509	30	14	155	18	726

5.3.3 Side impact collisions

The CI side impact studies concerned side impact collisions between two vehicles or between a vehicle and a fixed object. Based on the combination of the variables "Collision configuration" and "First impact location" from NCDB (excluding one jurisdiction, which does not provide the variable "First impact location" in NCDB), the two databases concurred for 54% of the collisions. This indicates that the variables "Collision configuration" and "First impact location" are often not correctly coded in NCDB and, as a result, only half of the side impact cases can be identified in NCDB.

5.3.4 Air bag deployment

The CI air bag studies concerned only those collisions during which the air bag deployed. Based on the variable "Air bag deployment" from NCDB, the air bag deployed in only 17% of the collisions according to NCDB (two jurisdictions do not provide this variable in NCDB, so they were excluded). Thus there were cases of air bag deployment that were not coded as such in NCDB.

5.3.5 Accuracy by jurisdiction

We measured the accuracy of the specific NCDB variables studied by jurisdiction. The accuracy for "Restraint use" and "Side impact collisions" is particularly low for some jurisdictions. The accuracy for "Restraint use" varies between 25.5% and 83.3% across the jurisdictions, whereas the accuracy for "Side impact collisions" is between 38.9% and 70%. However the accuracy for "Air bag deployment" is low for all jurisdictions, varying between 3.0% and 47.1%. On the other hand, the accuracy for "Number of occupants in the vehicle" is high for all jurisdictions, varying between 89.5% and 95.3%.

5.3.6 Accuracy by collision severity

Table 3 shows that the accuracy of NCDB increases with collision severity for each of the variables considered except "Number of occupants in the vehicle". The accuracy of NCDB for "property damage only" collisions is particularly poor for "Injury outcome", "Restraint use" and "Air bag deployment".

Table 3: Accuracy of NCDB by collision severity

Collision severity in NCDB	Variable					
	Injury outcome	Restraint use	Age	Number of occupants in the vehicle	Side impact collisions	Air bag deployment
Property damage only	55.2%	32.8%	83.2%	93.1%	50.0%	14.6%
Injury collision	75.0%	67.9%	84.0%	91.9%	50.0%	15.7%
Fatal collision	89.6%	75.5%	87.3%	90.4%	66.7%	37.0%
Total	73.6%	61.9%	84.4%	92.1%	53.8%	17.0%

6. Conclusion

Although the CI files do not constitute a representative sample of crashes, it allowed an analysis of the quality of NCDB. The completeness, timeliness and accuracy of NCDB were measured for each province/ territory. The results show that:

- The completeness of NCDB varies from 56% to 93% among the province/ territories for fatal and serious injury collisions.
- Small jurisdictions tend to send more timely and complete data and there appears to be a trade-off between the timeliness and completeness of the data in many jurisdictions.
- The number of fatalities seems to be underestimated by 7 to 8% in NCDB.
- The number of restrained injured or fatally injured occupants seems to be overestimated by approximately 12% in NCDB. As well, restraint use is missing or unknown in 30% of the cases where restraint use was known in the CI data.
- The combination of "Collision configuration" and "First impact location" is recorded correctly in NCDB for only 54% of side impact cases.
- "Air bag deployment" is reported only 17% of the time in NCDB.
- The accuracy of NCDB is low for "property damage only" collisions compared to injury or fatal collisions, particularly for the variables "Injury outcome" and "Restraint use".

7. Next steps

A study is underway to document the relevance of the NCDB data elements. Also, discussions on the uniformity of NCDB are being pursued. Studies to assess the coherence and comparability of the NCDB data will also be undertaken shortly. In addition, we will repeat the analysis presented here in 5 years. Work is underway to compare the NCDB data with the coroner files on alcohol and drug use and with data from event data recorders.

References

[1] Transport Canada, National Collision Database (NCDB), 2001-2005.

[2] Transport Canada, Road Safety Vision 2010, TP13347, June 2002

[3] Herzog, T.N., Scheuren, F.J., Winkler, W.E., Data Quality and Record Linkage Techniques, Springer, New York, 2007

[4] Johanson, V., Data Elements "Truly" Available in 2005 NCDB, PowerPoint Presentation, March 1, 2008

[5] Transport Canada, NCDB Data Dictionary, 2005

[6] Lecuyer, J., Chouinard, A., Hurley, R., Measuring the accuracy of the National Collision Database against in-depth collision investigations files, Proceedings of the 19th Canadian Multidisciplinary Road Safety Conference, Saskatoon, June 2009