ROAD INJURY DATA IN SOUTH AFRICA - AN ASSESSMENT OF THE NATIONAL ROAD COLLISSION DATABASE

M SINCLAIR

Department of Civil Engineering, Stellenbosch University, Private Bag X1, Matieland, 7602
Tel. 021 8083838; msinclair@sun.ac.za

ABSTRACT

Many road authorities, traffic engineers and road safety experts in South Africa believe that the country's collision data is unreliable and incomplete. In the face of challenges around data collection and dissemination, the Road Traffic Management Corporation (RTMC) valiantly manages to compile and release annual assessments of the status of collisions, casualties and other indicators of road traffic across all nine provinces. This annual analysis, which becomes the key source of intelligence upon which many policies and initiatives are constructed, is inescapably limited by the quality of the source data. To date, the imperfections embedded in the data behind these reports have not been fully understood. By analysing the RTMC database (1st December 2001 to 7th April 2010), this report presents an overview of the types of data that are most frequently missing, incomplete or questionable**.** The analysis allows us to understand exactly how incomplete the underlying datasets are, and facilitates a more informed assessment of the generalisability and reliability of findings that emerge from them.

1 INTRODUCTION

1.1 Traffic injuries globally

In 2002, road-traffic injuries ranked as the tenth leading cause of death in the world (World Health organisation, 2002). In 2004 that ranking had been upgraded to seventh, and it is expected that by 2030 road injuries will rank as the fifth highest leading cause of death (WHO, 2009). While other causes of premature mortality worldwide are showing signs of improvement, road injuries are instead escalating in importance.

The aggregate rates of road traffic fatality per 100 000 population have been found to be lowest in high income countries in the European region (in the 2002 WHO study this was calculated to be 11.0 per 100,000), whereas the highest rates were reported in the low-income and middle-income countries in eastern Mediterranean (26.4) and African regions (28.3) (Ameratunga, Hijar and Norton, 2006). However, estimates for 35 of the 110 countries used to calculate these regional figures were based on incomplete data. The most complete datasets were provided by high-income countries which presented with the lowest fatality rates, whereas little or no data were available from the poorest areas and countries in the world, believed to have the highest fatality rates. Attempts to better quantify collision and death rates per individual country on the continent of Africa in particular have been frustrated by the problems of under-reporting, poor quality of reporting, disorganized record storage and the reluctance of some countries to engage in global studies.

South Africa boasts one of the most comprehensive collision recording systems on the continent, yet even this system is criticised routinely for containing 'poor quality' data. The purpose of this paper is to examine the overall structure and broad content of the database, to look at what the database does well, and highlight areas in which errors may exist.

1.2 The value of a collision database

Reliable information on the scale of traffic injuries, and on the nature of traffic collisions that generate those injuries, are fundamental requirements for the development of appropriate and workable road safety strategies in any country. Road injuries are a largely preventable public heath issue in the developing world, but are only preventable if the determinants of road trauma are available, and are effectively utilised.

A prerequisite for making improvements in road safety is to understand the scale and the nature of the problem - to evaluate the factors that are found to have been associated with collisions in the past. The development of a database on collisions thus has the potential to allow an evidence-based approach to preventing road traffic collisions in the future. A robust database can be used at a number of levels: at national level it can be used to inform policy and set national budgets required to enhance safety; at regional or provincial level it can help regional authorities determine their own road safety priorities, and at local level it can assist local authorities determine the location and causes of the main road safety problems on their road network.

Generally a database is the consequence of two separate processes; the recording of the collision locally and the subsequent integration of local data into a national database. Once this data is sufficiently robust, reliable and accurate, it can be used for analysis and dissemination.

The WHO recognised that the quality or completeness of collision reporting is influenced by a number of factors, including the following:

- The sector responsible for recording the information
- The proportion of collisions involving vulnerable road users such as pedestrians (pedestrians being routinely under-reported)
- Poor or absent links between police, transport and health data systems (WHO Global Status Report on Road Safety, p30).

A further factor in data quality is that of database design, and crucially, the inclusion of data-quality testing within that design.

1.3 Legal responsibilities for collision recording in South Africa

The legal duty to record and retain collision information in the South African context is not clearly established in legislation, but is rather implied in the National Road Traffic Act, 1996 (Act 93 of1996), Section 61(f), which describes the duty of the driver to report an injury collision to a police or traffic officer within 24 hours. The implication in terms of subsection (1)(f) is that the Police or other authority referred to will then keep the records as needed in terms of prescript.

Regarding the collation of local data into a single database, the Road Traffic Management Corporation (RTMC) - which was established in terms of Section 3 of the Road Traffic Management Corporation (RTMC) Act, No. 20 of 1999 - is intended to achieve (among others) the following objectives:

- enhance the overall quality of road traffic service provision and, in particular, to ensure safety, security, order, discipline and mobility on the roads
- improve the exchange and dissemination of information on road traffic matters
- stimulate research in road traffic matters and effectively utilise the resources of existing institutes and research bodies.

To this end, the organisation's structure includes the functional unit of 'Research and Development', whose purpose is to monitor rates and trends in road traffic activities and achievement of goals. The RTMC is thus recognised as the central authority that bears responsibility for maintaining a national collision database and for analysing and disseminating data to all relevant levels.

2 DATABASE STRUCTURE

Any collision recording system is very much the product of the data entered on the Collision reporting form. In other words, the quality of the whole is dependent on the quality of the source data itself. The starting point for a good reporting form is that it needs to contain sufficient data so that the following questions can be answered with little effort:

- **Where** (precisely) did the collision occur?
- **When** did the collision occur?
- **Who** was involved?
- **Why** and **how** did the collision take place? Did any environmental factors play a role, and if so, what were they?
- Finally, what was the **outcome** of the collision in terms of injury and damage?

While these key questions look simple, the amount of data required to answer each satisfactorily is fairly extensive. In terms of who is involved, for example, the database needs more than a name. Descriptive information such as age, gender, their role (as driver/passenger/pedestrian), their origin or home address etc., are also necessary information if the data is going to be any use in subsequent analysis of cause and prevention. For legal purposes, information such as date of birth, driver vehicle details, vehicle ownership etc are also key areas of information that need recording.

While there are very many different models of data recording available worldwide, most of the databases that work most effectively are based on key common data categories. As a bare minimum, the following are suggested by the Transport Research Laboratory:

Table 1. Recommended factors for collision database (after TRL)

General details **Details/circumstances** Police reference Date and time Region/province Police station reference number Severity Collision type Number of vehicles involved Number of people injured Contributory factors	**Road type** Class of Road Road name or number Number of carriageways Speed limit Junction type Road width Presence/absence road shoulder	**Environmental features** Light Weather Road lighting Road surface condition Road surface quality Junction control Geometry (Curvature, incline) Presence of road works	**Precise location** Map reference XY coordinates (preferably) or alternative such as Kilometre post Plain language location Plain language description (free text) Sketches
Vehicle/Driver details Vehicle type Vehicle manoeuvre Vehicle damage Length of skid marks		Driver age Driver sex Licence number Seatbelt/helmet Alcohol/drugs suspected**?**	
Casualty details Type of road user Age Sex Severity of injury Brief description of injury Passenger location		Passenger location pre-and post-collision Pedestrian location pre and post collision Pedestrian manoeuvre	

The complexity of the relationships between the categories in the table above is such that the database needs to be constructed as a relational database.

In a relational database all data are represented in tables of rows (an individual record) and columns (attribute fields that contain data items). A column, or values in multiple columns, is used to define a unique primary key. Two or more tables are related by this common primary key, and are joined by this to form a new table.

The relational approach is based on mathematical theories of relational algebra - each table represents a set and therefore cannot include any row whose entire contents are duplicated. This allows one-to-many relations to be effectively managed; for example, one collision may involve three people. The information regarding the accident is given a unique reference; this reference is the primary key that is also assigned to the three individuals. Information regarding the accident can be joined to the three individuals, rather than recording the accident information three times.

A relational database also has the advantage of layering huge quantities of data. Combined with the reduction of variables to numerics (e.g. 0, 1 or 2 instead of male, female and unknown, for example) this minimises the size of the overall database which makes analysis and handling of data faster and more efficient.

Key to the success of a relational database, however, is the appropriate selection of a data coding technique. There are four classes of ways in which data can be entered into most relational databases:

- Dates – these can include the time as part of the date field
- Strings - text which can include numbers combined with letters
- Numbers – numeric values only either as integers or decimal (floating point) numbers
- Boolean logical fields - e.g. true/false

Used correctly, these classes of data entry can provide an accurate and easily interrogated set of data. Mistakes in the selection of class can, however, create opportunities for error. The quality of data entry in strings, for example, is particularly important. Typing mistakes - even the inclusion of a ‘space’ before the entry begins, can cause the entry to be missed in data searches (the entry ‘Unknown’ is not the same as the entry ‘UNKNOWN’). Where codes are used inappropriately, for example where the options given are either positive or negative, without making provision for an ’unknown’ option, can create restrictions to collecting comprehensive data.

Some of the categories of information, particularly free text and sketch-based data, cannot be easily coded or given straightforward values for comparison.

Most of the data required is factual, and thus verifiable. However even the best databases incorporate entries that are based on opinion – in particular the causation factors which are of necessity completed once the collision has occurred and needs to be sourced either from witnesses or from evidence on the scene. The inclusion of subjective information presents some difficulty, not least because the possibility that the views could be challenged in court can have the effect of making police officers reluctant to commit themselves unless their opinion is substantiated by other hard evidence, which is not often available. In the UK, to get around this difficulty, the collision recording form (the T1a form) instructs officers to select up to 6 contributory factors (from a total of a possible eighty-five) but also to indicate whether that chosen factor was ‘very likely’ or ‘possible’. This has proven to be a more comfortable method for police officers to convey their opinion.

3 RESEARCH METHODOLOGY

The Road Traffic Management Corporation provided a copy of their Access database - ‘Arrive Alive Version 8’ - to Stellenbosch University. The database contains data from 1st December 2001 to 7th April 2010. The database includes 151,383 complete data records relating to 69,143 traffic collisions. All of these collisions resulted in the death of at least one person, hence their inclusion in the database. Within this database, following from the legislation related to collisions in SA, collisions are largely referred to as ‘Accidents’ though references also exist to ‘Crashes’.

The main interface of the database is a pre-designed form, through which a series of predefined queries can be run upon a user selected date range. Predefined queries have the advantage of ensuring that analysis that is carried out by different analysts is consistent. However, there are two particular disadvantages associated with them. First, is that any gaps or errors hidden within the principal tables remain hidden, and logical checks are impossible to carry out. The second disadvantage is that new queries cannot be set up within the interface tables that make up the database. Only by looking at the original tables that form the underlying data is it possible to do analysis outside of the predefined queries.

The methodology thus revolved around examining the database tables outside of the pre-defined queries and examining the contents of each in terms of completeness, internal consistency and cross correlation with other related data tables.

4 RESULTS

4.1 Overall structure

The AA database is presented as a Microsoft Access database. Data can be considered as a series of *Principal Tables* within which key accident information is stored, and *Secondary Tables* which are supporting Look Up Tables which provide explanations of fields within the Principal tables. The structure is shown in Figure 1 below:

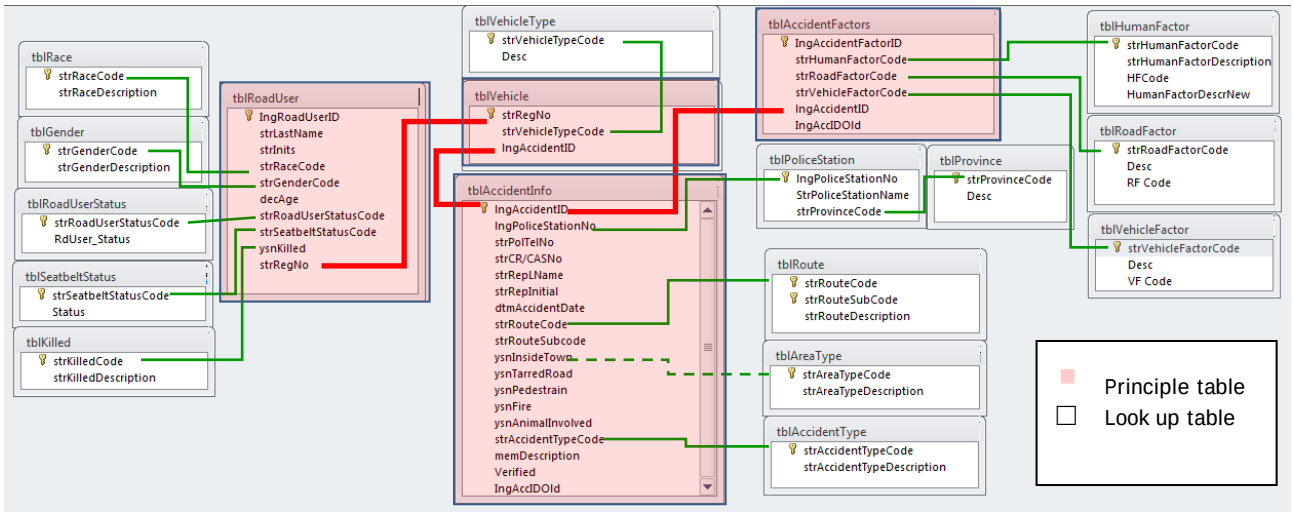

Figure 1: Arrive Alive (version 8) Database Structure

The four principal tables contain information specific to four aspects of each collision.
tblAccidentInfo – captures the recording of the accident
tblVehicle – details of the vehicles involved in each accident
tblRoadUser – demographics of the individuals involved with accidents
tblAccidentFactors – listing of causal factors of each accident.

The principal table, to which each of the other three are linked and without which the information contained in the other three becomes unusable, is the general Accident information table tblAcidentInfo. The most important field of this table is the *IngAccidentID*, or the unique reference number allocated to each entry, through which connections to the vehicle, injured road users, and causation factors can be established.

All four principal tables contain reference to old data, and refer to an old and new coding system. While there is no reason to suspect that the conversion from old to new is done incorrectly in the predefined queries, the parallel existence of two numbering systems does facilitate opportunity for error.

Before the contents of each of the principal tables are examined in detail, it is useful to consider application of code types especially with their role in the *Secondary Tables*.

4.2 Coding used

In order to limit size of the principal databases, values that are repeated many times are given a Code rather than the full explanation or description, and additional 'look up tables' contain text values of these repeated values which are then joined together by having a common *key field*. For example in the Principal table **Road User** the attribute 'Gender' is recorded as values 1, 2 or 3 and the supporting table 'tblGender' provides the information that gives meaning to the Gender Code, e.g. value 1 = "Unknown", 2 = "Male" and 3 =

"Female". This system works very well when values are appropriately attributed to the possible answers. When an insufficient range of values is provided (e.g., not providing a value option for 'Unknown'), major errors can result.

Table 2 shows the use of common fields used within one of the principal tables ('Accident Information') to maximise the amount of information available. In the fields where only 'Yes' or 'No' options are given the absence of values for 'unknown' information can be seen as a potential shortcoming. Such limited choice selection, which does not provide an option of 'not known', can reduce accuracy.

Table 2: Example Field Types Allocated To Attribute Fields

Attribute Fields	Field Type	Field Values / Comments
IngAccident	Number	A sequential reference number for each record
Police Station	Number	Number - Through look up table 'Police Station' enables a join to determine the Province
Police Phone No.	Number	Dialling code and telephone number of station.
Reporter Name	String	Name of person responsible for taking call
Accident Date	Date	Given as a single item, but can be separated to YYYY-MM-DD and the Hour (some contain Minute)
Route Codes	String	Route Code – road name e.g. N2
Inside Town	Boolean	True/False value set as Yes / No
Tarred Road	Boolean	Yes / No
Pedestrian	Boolean	Yes / No
Fire	Boolean	Yes / No
Animal Involved	Boolean	Yes / No
Accident Type Code	Set up as string but numerical values entered	Contains 25 different definitions types of accident causes –e.g. Turn in face of oncoming traffic' these also include whether pedestrian or animal involved.

4.3 Details of the four principal tables

4.3.1 Accident Information – tblAccidentInfo :

The number of records is 69,143, but the highest sequential reference number is 71,032 suggesting that 1,889 entries have been deleted at some point. For the purposes of further analysis, the total entries in this section will be considered to be 69,143. The table hosts most of the administrative related information, such as reporting police station (and contact details), name of person who reported collision, the unique accident reference field IngAccidentID which is allocated automatically once an original report is created, and general details about the collision including location, date, time. The report also includes general attributes, including Accident Type, whether a pedestrian was involved and whether a fire ensued.

Entries are recorded for 100% of some of these categories, - IngAccident (reference number, Case number (from the local police station), and Police station number. Other factors have a very high completion rate, including Accident date (99.95%) and time (98.7%).For the fields with a Boolean range, completion rates are 100%, indicating that the fields are most likely prepopulated with one of the two values as default.

The field related to Accident location reflects a 99% completion rate, which is very high. However the quality of the entries makes the data almost impossible to use. The responses take the form of Route Code – road name e.g. N2, with a Route Subcode – road section e.g.W2. The location data is based on descriptions only, and start and end points of each route are not defined. The 'many to many' relationship means that multiple joins are needed to get correct description applied to each route code/subcode pair. As it stands the location data is of little value in pinpointing collision clusters.

4.3.2 Road User - tblRoadUser :

This is the database where demographics of accident victims, their role within the accident and the outcome in terms of accident fatality are recorded. The table contains details of 151,383 individuals who were killed or injured in the 69,143 fatal collisions identified in the Accident information table. This table also contains ghost entries, in that there are more listings than Accident reference numbers listed, suggesting that some 1,386 records have been deleted, though no explanation for this is given. Some details of the Road User table are described below.

Field = '*Name*'

This has a high level of completion at 99.7% for surname and 79.5% for initials. Under the surname, however, the word UNKNOWN has been used with numerous variations in spelling, reducing the total known name count to 81.6%. However, the completion rate is satisfactory and it is clear that information that is not known has been entered as such.

Race is also completed to an extremely high degree - 99.91% complete, so blank values account for less than 1 percent. 'Unknown' is provided as an option, and 7.14 of entries were marked as such.

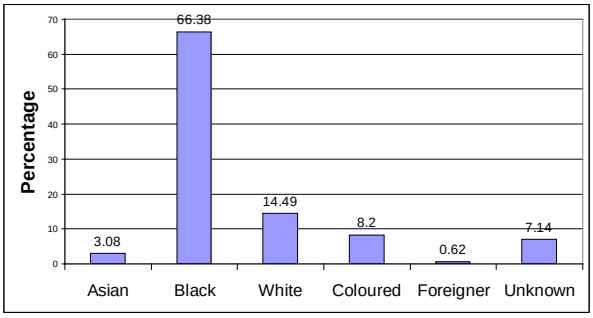

Figure 2: Race

Field = '*Road User Status*'

This section has a 100% completion rate, and data indicates a dominance of drivers in the road users killed or injured. However, it is impossible to establish the accuracy that field input. The fact that there are more drivers identified than vehicles in the *tblVehicle* table suggests some level of inaccuracy as one vehicle can only have one driver. Further, the fact that the 'driver' category appears to contain names of people who did not die as a result of their injuries - whereas the 'pedestrian' and 'passenger' group are recorded as all having been killed - suggests a fundamental imbalance in the contents of the data.

Field = '*Gender*'

Again, the dataset appears to be almost complete, with a 99.96 % completion rate (0. 04% of records are left blank). There is an unexpectedly high number of unknowns – nationally the average of 'Unknowns' total 7.9% but on a provincial basis this ranges from 5.1 % (Eastern Cape) to 11.9% (Western Cape). The dataset confirms a preponderance of males (at a national average of 76.3%) with a far lower incidence of injury to females (15.7%).

Field = '*Seat Belt Status*'

At first glance, the amount of data relating to seatbelt wearing is fairly high, with records filled in for 99.84% of entries. However, in this field the status was entered as 'Unknown' in 56% of records. This field also identifies Pedestrians which comprise 22.2% of records, so the eligible road users (i.e. drivers and passengers) where seat belt status was recorded represent only the remaining 21.6% of the total – of which 17.26% were recorded as not wearing seatbelts, and 4.36% recorded as having worn seatbelts.

When looking at the relationship between drivers killed and seatbelt usage, the database indicates that seatbelt data is effective for drivers. The percentage of drivers killed whilst not wearing a seatbelt is 39.4%, but this is reduced to 22.0% when a seat belt is worn. However this is based on an analysis of 20,444 drivers from a recorded 89,882 -

statistically questionable results. Efficacy of seatbelts on passengers cannot be made as all passengers are shown to have been killed, regardless of seatbelt status.

Field = '*Killed/Not Killed*'

All Road Users, except 'Driver' are given the status of Killed. This was confirmed by running a report from the pre-defined query to show pedestrian fatalities between 1/1/2000 and 31/12/2009 which this returned 32,864 records (urban 14,765 and rural 18,099)

When the same query is compiled on separate tables this shows that these records represent all of the road users, i.e. there are no pedestrians recorded as having the status Not Killed.

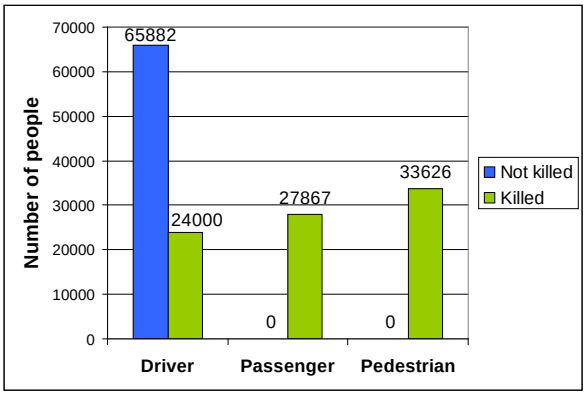

Figure 3: Numbers Killed & Not Killed

It is thus impossible to draw any conclusions about the effect of road accidents on Pedestrians or Passengers, as either (i) the table is correct and these road users only make it into the database if they are killed, or (ii) the default setting for this field is 'Killed' and these have not been changed for Passengers or Pedestrians

4.3.3 Vehicle data – tblVehicle :

This table is used to find out particulars of a vehicle but it also enables reporting information that is held within the **AccidentInfo** to be linked to the victim (**RoadUser**) as it contains the primary fields *IngAccidentID* (to link to AccidentInfo), and s*trRegNo* (string-Registration Number) to link to the **RoadUser** table. The strRegNo field contains a number or errors. In particular, in some records the registration number has been overwritten by the field IngAccidentID, thus losing the registration information, and making further searches impossible. There are over 1,100 road users who cannot be linked to the AccidentInfo database as there is not a corresponding strRegNo value.

4.3.4 Accident factors – tblAccidentFactor :

The tables relating to accident factors comprise three areas of investigation; the role of the driver, the vehicle and the road environment. The data inputter is only able to select one option from each of these categories. The setup of this section is immediately problematic. In the Human Factors, for example, a selection must be made from a list of 15 possible human errors which in reality seldom occur in isolation. This immediately undermines the value of this section.

In each of the three sections there is no option to select 'No human factors relevant', or 'no vehicle factors relevant. The inputter is forced to leave the section blank if there was no contributory effect from the factor in question. This makes it impossible to determine whether data has been omitted intentionally or erroneously. In the human factors section, 9.1% of entries had been left blank; in the vehicle section 92% were blank, and in the road factor section 89.7 % of records were blank. It is impossible, without the provision of a clear option of 'irrelevance' to establish a confidence level for this data.

The bigger problem here is, however, the fact that the forced selection of one factor among many possible relevant ones for each collision creates an inherent and significant level of unreliability within this entire section. The data becomes, to all intents and purposes, incredible, and any analysis from it needs to be treated with caution.

5 DISCUSSION

Data quality is ultimately a consequence of data input accuracy, completeness, and consistency – as well as how relevant these are in addressing the objectives of the database. The achievement of the input factors is addressed briefly below.

5.1 Accuracy

There are a number of factors affecting accuracy in this database. The first is that of *lack of data*, the second that of *misuse of input options*. Both are well illustrated in the '*Age'* field in the **RoadUser** table. Here, age data was is not recorded in 20.7% of cases. Further, looking in detail at the data that has been entered, it seems likely that unknown ages often appear to be entered as zero (0.0) rather than being left blank. The value of 0.0 accounts for 21.5% of all entries, this distorts the age profile of entries, see Figure 4 below.

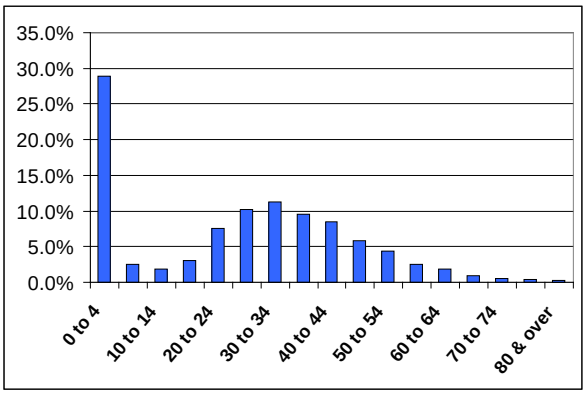

Figure 4: Ages as recorded (excluding unknowns)

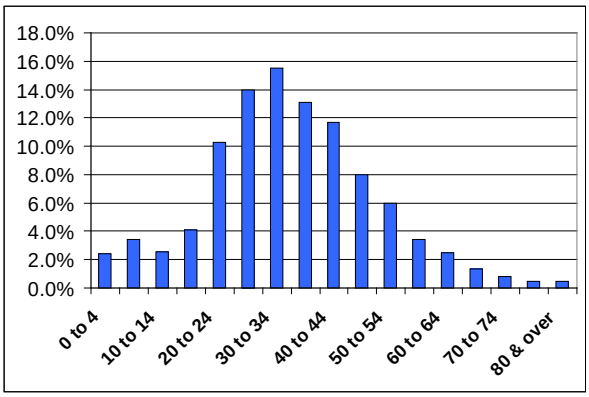

Figure 5: Ages with '0.0' removed

If the Age 0.0 and blank values are ignored then children up to age 19 constitute 12.5% of all casualties where age is known. While this cannot be entirely accurate, as fatal injuries will occur to children under 1 year, the resultant graph is a far closer approximation of the expected distribution than the data as it currently stands.

This highlights a second problem affecting accuracy, which is that of **coding**. Coding allows inputters to limit the values of valid responses to particular question, by placing restrictions on the data type, length, range of acceptable values etc. This, done well, can maximise consistency which in turn can improve the data mining and analysis potential. Coding can also create problems if answers have a default which needs to be changed it is impossible to verify, in those cases, whether the final tally is accurate or whether other problems are hidden. For example if the default has been used when the current information was not known, or whether the inputter omitted to make an active selection.

Finally, some of the coding imposes unhelpful restrictions on answers– for example, the use of the Boolean range for 'Inside/Outside' town is of questionable value. It does not take into account per-urban contexts or rural hamlets, and assumes a clear distinction which is often unrealistic.

5.2 Consistency

Inconsistency in individual data entries does exist – particularly in string entries, and these could be avoided by ensuring tighter input rules, conventions and quality control by checking for blank fields that have not been addressed. Obvious inconsistencies of this type are currently left uncorrected, which creates problems for data mining. Further, these is clear evidence that some of the answer options were misused (the use of the zero 0.0 in the age field, for example) and no tests have been built in to establish the relative appropriateness of similar responses in different fields.

There is some inconsistency in where accident information should be captured within the current reporting system. For example most causal information is included within the AccidentFactors table, but the table AccidentInfo also contains some features of the collision such as whether a pedestrian was involved or whether a fire ensued.

There is also a worrying level of inconsistency in some of the correlation tests done between fields which cannot be explained by the available data. A good example of this is the role of animals in collisions. There are three opportunities for the involvement of animals to be recorded, all found in different parts of the database. Role of animals is recorded in *AccidentFactors* database as *RoadFactor* Code 11, but zero records have been allocated this cause. The role of animals is recorded in *AccidentInfo* in True/False field *'AnimalInvolved'* - which gives 100 accident events. The role of animals is also recorded in *AccidentInfo* through the *AccidentType* field (Code 22) which gives 279 accidents. This example highlights problems of consistency where causal factors are captured and the logical errors that this creates.

On the other hand, correlation tests for pedestrian numbers indicate a higher degree of accuracy. In *AccidentInfo*, under the section *Pedestrians*, 47% of entries confirmed that a pedestrian had been involved. The total number of pedestrians counted here was 36,629. In the *RoadUser* table, 22.27% are recorded as being pedestrian, a number of 36,626.

5.3 Completeness

The database examined shows overall high levels of data completeness, though no tests for data accuracy and relevance have been built into the system to help identify incorrect data. At a superficial level, much of the database appears to be well populated, indicating good administrative procedures. In most cases, missing information cannot be easily interpreted. In some cases data is missing because it is not available, in others because the posted answers are irrelevant, and in others because of human error. These cannot be distinguished from one another.

The setup of the database itself has unintended limitations on data quality, in particular the unnecessary restriction on the selection of factors involved means that data input into the form is at best partial. Also, the original police Accident Report Forms do not themselves offer the same choices for attending officers, and so the inputters are selecting single factors from an accident description. This is a huge source of error and misinterpretation.

6 CONCLUSION

Referring to the five key questions that were referred to at the start of this paper (the ‘where’, ‘when’, ‘who’, ‘why’ and ‘outcome’), problems with the database currently undermine the confidence in the answers that can be mined for all of these questions. The ‘where’ will continue to be problematic until a more accurate and systematic application of coordinate points is rolled out. The ‘who’ is tarnished by problems with implausible age and race data. The ‘why’ and ‘how’ is particularly poorly completed because of fundamental problems with the selection options in the database structure. The ‘outcomes’ are limited to ‘killed/not killed’ and so miss out on a huge amount of key casualty information that could be used for analysis and casualty prevention. This is a limitation which should be addressed in the future . No information – apart from a single option regarding drunk driving – is collected with respect to alcohol use. Only the question of ‘when’ appears to be answered sufficiently robustly, with a high percent of valid entries, However, inaccuracies are not tested for, so confidence levels remain uncertain.

The database structure should be reassessed to ensure that the field types are able to capture what is needed, to allow for internal quality control checks and that the links between tables minimise the opportunity for errors. One example is the need to use the table *Vehicle* as an intermediary to link the tables *AccidentInfo* and the table *RoadUser,* as it is only the table *Vehicle* that includes both linking fields – this is termed a transitive dependence. If the Vehicle table is edited then cross references can be broken, but this step could be removed by including the field AccidentInfoID within the RoadUser table.

Awareness needs to be raised with those responsible for data capture about the need for data accuracy – the effects that using 0.0 in a number field when an age is not known, or of the importance of consistent use of string fields. Fundamental problems with data collection - at the roadside or police station by police and traffic officers - needs to be addressed and improving the quality of data collection through enhanced and ongoing training needs to be prioritised.

Ongoing criticisms of South African collision data undermine the confidence of users of the RTMC data. This analysis has shown that questions about the quality of the data are valid, and that areas of missing or anomalous data are reason for concern.

The challenge for the RTMC is to prioritise data quality assessment as an integral part of their data management process. The initial data collection form and processes of data completion by primary and secondary inputters needs to be re-examined. Coding systems in place are often loose and create opportunities for ambiguity. The options open to selection are sometime too restrictive, and there is great opportunity for misinterpretation by the imputers having to make decisions about what is and isn’t relevant in the selection of causation factors. The RTMC currently invests huge amounts of efforts and resources into the management of the database – some fundamental changes would ensure far higher levels of accuracy, and consequently better quality casualty analysis in the future.

REFERENCES

Ameratunga, S, Hijar, M and Norton, R, 2006: Road-traffic injuries: confronting disparities to address a global-health problem; www.thelancet.com Vol. 367 May 6, 2006.

Healey, R.G. 1991. "Database management systems". P251-67 in 'Geographic Information Systems & Science'. Wiley.

World Health Organisation, 2009, Global status report on road safety.
World Health Organization, 2002. A 5-year WHO strategy for road traffic injury prevention. http://www.who.int/violence_injury_prevention/publications/road_traffic/5yearstrat/en/ index .html.