Accident Analysis and Prevention 130 (2019) 84–90

Contents lists available at ScienceDirect

Accident Analysis and Prevention

journal homepage: www.elsevier.com/locate/aap

Crash data quality for road safety research: Current state and future directions

Marianna Imprialou*, Mohammed Quddus

Transport Studies Group, School of Civil and Building Engineering, Loughborough University, Loughborough LE11 3TU, United Kingdom

ARTICLE INFO

Article history:
Received 30 July 2016
Received in revised form 13 February 2017
Accepted 22 February 2017
Available online 3 March 2017

Keywords:
Accident analysis
Accident data
Crash data quality
Crash police reports
Crash severity
Road safety

ABSTRACT

Crash databases are one of the primary data sources for road safety research. Therefore, their quality is fundamental for the accuracy of crash analyses and, consequently the design of effective countermeasures. Although crash data often suffer from correctness and completeness issues, these are rarely discussed or addressed in crash analyses. Crash reports aim to answer the five "W" questions (i.e. When?, Where?, What?, Who? and Why?) of each crash by including a range of attributes. This paper reviews current literature on the state of crash data quality for each of these questions separately. The most serious data quality issues appear to be: inaccuracies in crash location and time, difficulties in data linkage (e.g. with traffic data) due to inconsistencies in databases, severity misclassification, inaccuracies and incompleteness of involved users' demographics and inaccurate identification of crash contributory factors. It is shown that the extent and the severity of data quality issues are not equal between attributes and the level of impact in road safety analyses is not yet entirely known. This paper highlights areas that require further research and provides some suggestions for the development of intelligent crash reporting systems.

1. Introduction

The "garbage in garbage out" (GIGO) principle suggests that the quality of input data is directly related with the outputs of an analysis (Oliveira et al., 2005). Data quality is however a subjective measure that refers to the level of appropriateness of data for a specific use (e.g. Juran and Godfrey 1999; Herzog et al., 2007). The intended use largely determines data collection methods as well as the contents and the details of the included attributes. As a consequence, when datasets are used for purposes different from the primary, data quality might not be ideal. One example of such datasets that may have multiple uses is road crash reports collected by public authorities (in most cases the local police). Crash reports typically provide sufficient information for their primary use to provide information when civil claims arise and to develop regional and national safety performance statistics. However, due to the lack of alternatives and despite the limitations, police crash reports are also the main source of data for road safety research (OECD/ITF, 2015).

As crashes are one of the main externalities of transport systems (Johansson et al., 2014), a significant amount of resources is invested every year for their mitigation. The number of crashes has indeed been reduced over time, especially in the developed countries, although the decrease may not be attributed entirely to improvements in road infrastructure, traffic conditions or driving behaviour but it might be also related with improved vehicle quality and medical services. The development of measures for crash prevention largely relies on the outcomes of road safety analyses. Whether focused on road users or the environment, road safety analyses aim to unveil the conditions that are more likely to lead to crashes; that is why they are usually data demanding. It is unclear to which extent crash records are accurate enough to be considered "fit for purpose" for road safety analyses, as little research has been done towards this direction. There are indications though that crash data have considerable shortcomings, which imply a potential distortion of research outcomes and a subsequent negative impact on developing countermeasures.

The two main problems of crash data relate to completeness and accuracy. Crash under-reporting is a well-recognised and studied problem of road safety research globally (Alsop and Langley, 2001; Salifu and Ackaah, 2012; Watson et al., 2013). The reporting authorities are not always responsible for this issue; there are various reasons why a percentage of crashes are not included in

* Corresponding author.
E-mail address: M.Imprialou@lboro.ac.uk (M. Imprialou).

https://doi.org/10.1016/j.aap.2017.02.022
0001-4575/© 2017 Elsevier Ltd. All rights reserved.

the official records. For instance, police might not have been notified of the crashes either because the users agree to sign private settlements for insurance purposes or there was no third party participation (i.e. single vehicle crashes) or there were not obvious injuries just after the crash (Amoros et al., 2006; Barancik and Fife, 1985). Although there are variations in the estimated levels of under-reporting across countries, there is an agreement that selection biases are mainly related with injury severity and road user type. Obviously fatal or serious crashes are more likely to be reported to the police, while a large proportion (approximately two-thirds and beyond) of slight and property damage crashes remain unreported (Abay, 2015; Alsop and Langley, 2001; Amoros et al., 2006; Elvik and Mysen, 1999). Similarly, motorcyclist and cyclist injuries have the highest under-reporting rates among road users (Salifu and Ackaah, 2012; Watson et al., 2015). The impact of under-reporting on the outcomes of road safety analyses is not just a speculation. Missing crash records have been found to affect both injury severity analyses (Yamamoto et al., 2008; Yasmin and Eluru, 2013; Ye and Lord, 2011) and crash frequency models (Ma, 2009) due to under-representation of some injury categories.

On the other hand, reported crash data are not perfect either. Misreporting and incompleteness are two serious issues of crash databases. The problem is more evident, and possibly more serious, when the key variables that are used for integrating crash datasets with other explanatory datasets (e.g. traffic data) or the variables that describe the outcomes and the circumstances of crashes (e.g. severity) are inaccurate or missing. Road safety research has evolved rapidly in terms of traffic data quality and methodological approaches (Mannering and Bhat, 2014). However, crash data seem not to have followed a similar trend in terms of quality and might be inadequate for current sophisticated analyses that are not necessarily capable of addressing this issue. Errors and inaccuracies in crash datasets are challenging, if not impossible, to be identified and corrected by data users and that partially explains why research on the particular area is limited.

This paper summarises findings of existing research on crash data limitations in order to highlight potential issues that could emerge in multiple different types of road safety analyses. This will be enhanced using evidence from four crash datasets from different countries. The purpose of this paper is to increase awareness of crash data limitations, to pinpoint areas that require further research and to provide ideas about future crash reporting systems.

2. The five W's of crash reports

Crashes are rare and often multi-causal events that follow a unique sequence of events. It is therefore not always straightforward to identify their exact causes. To capture as much useful information about crash occurrences as possible, local authorities develop specialised report forms (paper-based or electronic) that need to be completed after a crash. The information can be categorised using the five "W" questions:

- "*Where?*": crash location;
- "*When?*": crash time;
- "*What?*": crash severity;
- "Who?": involved users (and vehicles) and;
- "Why?": crash contributing factors.

Depending on their scope and perspective, road safety analyses may employ a plethora of different combinations of information included in crash reports. These five categories are not equally useful for all analyses; for example, in a network-level crash frequency model individual driver-related data might not be considered, while in a study on the crash characteristics of different driver profiles, crash time and location might not be important. Additionally, some of the variables that consist these categories are significantly easier to be reported (e.g. driver's gender) compared to others (e.g. crash location), therefore misreporting and missing data problems are not equally distributed between variables. The following sections will summarise some of the most significant flaws of crash data that have been discussed in existing literature.

2.1. Crash location and time

Location and time are two fundamental elements of a crash report. For crash analyses that aim to explain fully or partially crash occurrence through environmental conditions (e.g. traffic, weather, road geometry) these variables are of particular importance because they enable linkage of crash datasets with datasets that include environmental conditions information. One major issue that arises when data linkage is required is the lack of common and comparable spatial and temporal attributes between datasets. This happens because most of the datasets are held by different organisations that typically are not road safety-focused, so direct matching is rarely possible. For instance, crash locations might be reported with a set of coordinates, traffic data might be reported with reference to the links of a base map, geometry data might be represented by another more detailed base map and finally, weather data might be obtained from the closest weather stations that are represented by a set of coordinates (examples of detailed description of data linkage for crash modelling can be found in Abdel-Aty and Pemmanaboina (2006) and Imprialou et al. (2016)). The combination of multiple datasets with different spatial and temporal attributes becomes even more challenging when these datasets contain inaccuracies.

Errors in reported crash location and time add complexity in data pre-processing, especially for studies that heavily rely on these attributes.[1] The required level of details for crash location and time varies between different types of studies. For example, in order to develop annual crash frequency models each crash that occurred during the study year should be allocated to one road link (or section) which is relatively simple to be identified. On the other hand, in real-time crash prediction models where the exact pre-crash conditions should be examined, apart from the location, that should be precise enough to identify upstream and downstream traffic sensors, the exact time of the crash is also required.

Crash location is reported with multiple different systems around the world such as linear referencing, offset from junction, coordinates and address. Considerable inaccuracies in crash locations have been reported for all systems (e.g. Dutta and Noyce 2005; Burns et al., 2013; Brown et al., 2015). For instance, Miler et al. (2016) found that a 33.5% of the crashes of a relatively large database (8550 observations) had inaccurate crash location attributes. The inaccuracies may be due to human error, equipment failure (when GPS is used), limited training of personnel or can be inherent to the reporting method (e.g. Brown et al., 2015; Imprialou et al., 2015). Existing crash mapping algorithms employ a broad range of methods and attributes of crash records in order to correct erroneous crash locations. Early crash mapping algorithms primarily employ reported road name, road type and simple geometric approaches such as buffer zones (Austin, 1995) or shortest distance from the reported crash location to the road network (Loo, 2006). More recent algorithms consider additional attributes from crash reports such as vehicle direction just before the crash and

[1] Studies that employ disaggregated information on the traffic and environmental conditions that are related with individual crashes such as real-time crash prediction models (Hossain and Muromachi, 2013) or condition-based count modelling (Imprialou et al., 2016).

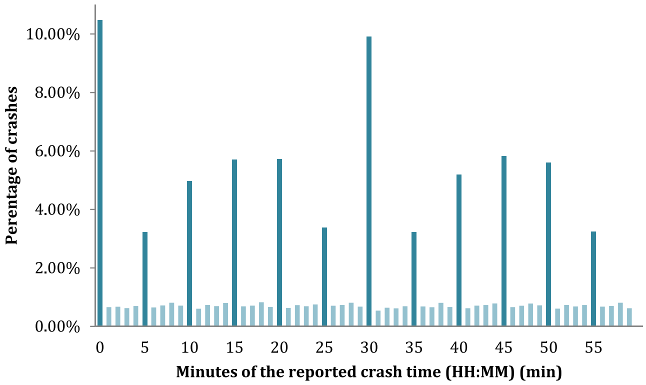

Fig. 1. Distribution of minutes of the reported crash times in STATS 19 database.

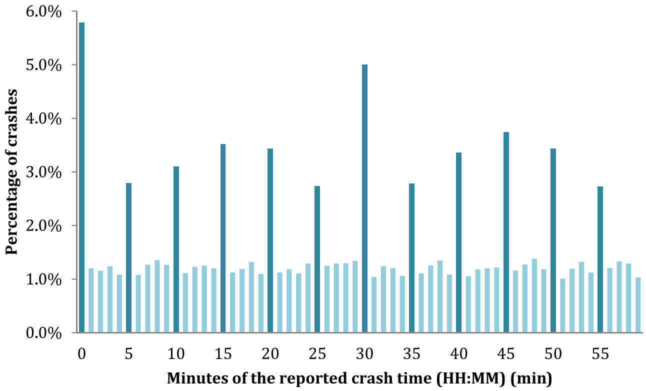

Fig. 2. Distribution of minutes of the reported crash times in FARS database.

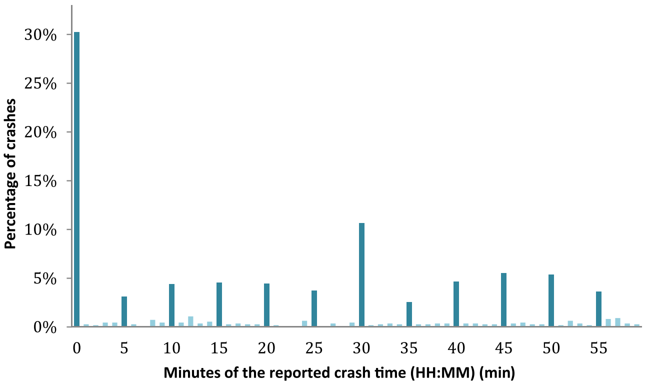

Fig. 3. Distribution of minutes of the reported crash times in the Australian Road Deaths Database.

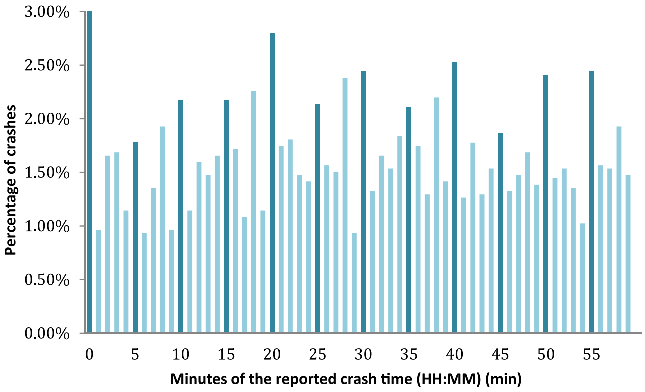

Fig. 4. Distribution of minutes of the reported crash times in Attica Tollway crash database.

employ advanced statistical or Artificial Intelligence (AI) concepts. For example, Artificial Neural Networks (Deka and Quddus, 2014) and Fuzzy Logic (Imprialou et al., 2014) have been applied to correct freeway crash locations achieving matching accuracy of 98.4% and 98.9% respectively (for a detailed overview of crash mapping algorithms the reader is referred to Imprialou et al. (2015)).

Despite the fact that substantial improvements in crash location have been achieved through crash mapping algorithms, the vast majority of crash analyses do not perform this kind of data pre-processing. This may be due to the fact that crash mapping algorithms are relatively complex to develop and evaluate (e.g. the validation of some of the aforementioned crash mapping algorithms requires some hundreds of manual identifications of crash locations). Another reason could be that the impact of using less accurate crash location is not entirely known.

There is a gap in knowledge on whether crash data with inaccurate crash locations are appropriate for road safety analyses. There are, however, indications that crash locations affect the evaluation of countermeasures (Brown et al., 2015) as well as the estimation of coefficients of crash frequency models (Imprialou, 2015). The impact of location inaccuracies is likely to be more significant for studies that require more detailed location information on pre-crash conditions.

Many crash reports do not contain minute-level time accuracy as reported crash time might rely on witnesses' statements which means that time is by default rounded (Golob and Recker, 2003). Figs. 1–4 show the distribution of the reported crash times by minute (reported in hours-minutes format, HH:MM) of four official crash datasets:

a) STATS19 the official Police crash database in the UK that contains all reported crashes (fatal, serious or slight) that occurred in the country during 2014 (Department for Transport, 2011a).
b) FARS the US fatality analysis reporting system that contains information for all fatal crashes during 2014 extracted from multiple official documents including Police reports (U. S. Department of Transportation, 2010)
c) Australian Road Deaths Database that includes Police reports for all fatal crashes during 2015 (BITRE, 2016)
d) Attica Tollway crash database that includes crashes of all severities occurred during 2009 on a 65 km urban motorway (Attica Tollway) in Athens, Greece (Odos, 2017).

The periodic spikes of the distributions show that a disproportionate number of crashes have been reported at times when the minute (MM) indication ended with zero or five; 66.4%, 42.4%, 83% and 27.8% for datasets a, b, c and d respectively. It is noteworthy that in Attica Tollway crash database that has the lowest percentage, crash times are determined with the aid of automatic incident detection cameras. The assumption that crash time has a uniformly distributed error of 2.5 min is probably incorrect though, as a large proportion of crashes are reported at the first and/or the thirtieth minute (e.g. 41% of all crashes in the Australian Road Deaths Database, Fig. 3).

The impact of misreported time might be significant for studies that take into account instant traffic changes such as real-time crash prediction. So far, in studies that require the identification of pre-crash conditions many researchers use the traffic conditions several minutes (typically 5–15 min) prior to the reported crash time (e.g. Xu et al., 2013; Abdel-Aty et al., 2012). This approach reduces the problem if the error of the reported time is known and positive (i.e.

the crash time is assumed to be recorded at a later time in relation to the actual time) that for the aforementioned reasons might not be true. If a crash report includes crash time that is earlier than the one that the crash actually occurred, the selected pre-crash conditions might not represent the conditions of interest, but the conditions that occurred as a result of the crash.

Abdel-Aty et al. (2005) suggest to employ the speed of the shockwave that propagates backwards upstream of a crash as the upstream traffic starts decelerating after the occurrence. Shockwave speed can be estimated by dividing the distance between two contiguous traffic measurement stations upstream of the crash location (e.g. loop detectors) with the time difference of shockwave arrival at these locations. If the shockwave speed is known, the time when a shockwave arrived at a specific network location can be estimated. Therefore, the time that the shockwave arrived at the crash location is assumed to be the actual crash time. Although theoretically valid, this method requires the availability of spatially and temporally disaggregated traffic data that are may not be available, especially for urban road networks. Moreover, this correction method assumes that crash location is accurate, which may not always be the case. To the best of the authors' knowledge, the impact of inaccurate crash time in crash analyses has not been formally evaluated yet, so research is required in this unexplored area.

2.2. Crash severity

Not all crashes have equal societal impact and therefore public authorities typically prioritise crashes by severity in their crash mitigation strategies (e.g. Vision Zero introduced by Sweden) (Whitelegg and Haq, 2006). Crash outcomes in terms of their severity are therefore particularly interesting for road safety research because they are related with the effective allocation of resources for crash prevention. Crash severity corresponds to the maximum injury severity of all involved casualties. Severity characterisation by the police is typically based on fixed criteria such as the level of personal injuries and the length of stay in hospital (e.g. typically a crash is reported as fatal if it results in the death of at least one of the casualties within a 30-day period after the crash) (U. S. Department of Transportation, 2010) but the definitions are not uniform all over the world which suggests a lack of consistency and comparability between studies. Also, in many places around the world property damage only crashes are not recoded by the police which also leads to remarkable data losses. For example in the US Highway Safety Information System (HSIS) collisions are divided into five main severity categories (i.e. fatal, incapacitating, non-incapacitating, minor injury and property damage only) (Council and Mohamedshah, 2009) whereas in UK STATS 19 crashes are only split into three categories (fatal, serious and slight) (Department for Transport, 2011b) despite the fact that property damage only crashes in the UK are estimated to be approximately 94% of all crashes (Department for Transport, 2013).

Moreover, the fixed criteria that police uses for injury severity classification are not always sufficiently accurate, leading to potential misclassifications as it has been proven by a number of studies that conducted comparisons of crash reports with hospital records from multiple countries (Morris et al., 2003; Watson et al., 2015; Yannis et al., 2014). Generally, more serious crashes tend to be more accurately reported in terms of severity but also of other reported attributes (such as speed just before the crash) (Chung and Chang, 2015). While the identification of fatally injured and non-injured users involved in a crash is more straightforward, the classification between different (non-fatal) severity levels has been found to be challenging (Farmer, 2003). Dove et al. (1986) suggested that one third of the reported as "severe" injuries were over-classified and another third were under-classified. This is consistent with the findings of other studies that found that approximately one third of injuries characterised by the police as "incapacitating" were medically classified as "minor" (Popkin et al., 1991; Sherman et al., 1976). However, McDonald et al. (2009) report that a 15% of reported as "slight" injury crashes were found to be in fact "life threatening" according to hospital data.

Injury severity misclassification is not random as it has been found to be related with specific crash or user characteristics (for instance, sensitive user injuries tend to be over-classified) (Amoros et al., 2006). In addition to selection bias arising from the crash under-reporting, classification bias might further affect analyses that use crash severity in order to explain crash occurrences such as severity modelling or multivariate count regression models. To overcome this problem some researchers have suggested linkage of crash data with hospital data prior to any analyses (e.g. Watson et al., 2015) that might be quite effective but possibly time and data demanding.

2.3. Crash involved users

Road users involved in a crash can be either drivers or vehicle passengers or pedestrians (termed here as: *other road users*). As driver errors are considered to be associated with approximately three quarters of all crashes (Department for Transport, 2015; Stanton and Salmon, 2009) many studies analyse driver characteristics in order to identify and explain cohort-related differences in the frequency and type of crashes. Although typically not related with the cause of crashes, the characteristics of other road users are also particularly interesting for road safety research, especially when they belong to special user groups (e.g. children).

Police crash reports normally include demographic information of the drivers and other users as well as seat belt usage and their positions in the vehicle. Involved user data are affected by selection bias because specific user groups have been found to be consistently under-reported. For instance, crashes involving younger users tend to be less frequently reported compared to crashes involving older ones (Amoros et al., 2006; Janstrup et al., 2016). Literature does not highlight any serious misreporting issues for user demographics which might be explained by the fact that demographic characteristics are straightforward to be reported and can be also corrected after leaving the crash scene if necessary. Although the literature on this topic is not extensive, seat belt usage rate was found to be 13–18% over-reported in data collected over a 15-year period (1993–2007) from the US Crashworthiness Data System of the National Automotive Sampling System (NASS-CDS) (Viano and Parenteau, 2009). Similarly to other attributes, seat belt use is likely to be more accurately reported in crashes that resulted in at least one fatality (Cummings, 2002).

Examining user demographics of FARS and STATS 19, it was revealed that the latter dataset includes flaws even in the aforementioned "straightforward-to-report" attributes. Missing values for driver age and gender were less than 2% in FARS but 11% of all driver demographics had one or more missing values in STATS 19. Missing values can affect the outcomes of analyses and the list-wise approach of deleting incomplete observations is not a panacea when the "missing completely at random" (MCAR) assumption is not true (Heitjan and Basu, 1996). Moreover, as it can be seen in Fig. 5, reported ages in STATS19 included some periodic spikes in ages between 25 and 55 years old. Although not as substantial as with crash times, the shape of this distribution implies that there is some rounding in the ages of the involved drivers that could poten-

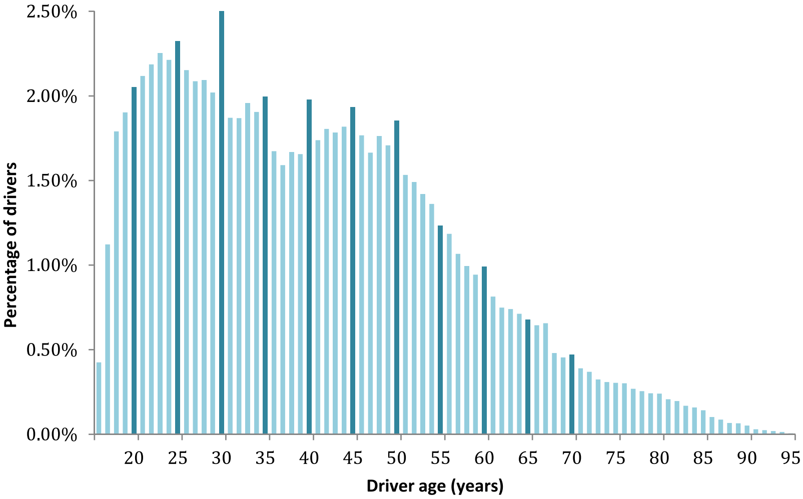

Fig. 5. Distribution of reported driver ages in STATS 19.

tially impact analyses relying on this data. However, FARS did not appear to have a similar pattern.[2]

2.4. Crash contributory factors

Understanding crash causation is probably the most critical element for developing crash preventive measures. Most police crash reports include information on the potential reasons that lead to crashes as these were evaluated by police officers that attended the scene after a crash occurrence. The reports typically include a fixed list of potential contributory factors related with the road environment, drivers, vehicles and weather conditions and the officers need to select (and sometimes prioritise) those that seem more relevant to the circumstances of a crash. This is not an easy task to complete; due to the inherent complexity of a crash mechanism understanding and reporting their underlying causes in minimal time using solely a generic, pre-designed form is probably an unrealistic goal. Officers due to time restrictions, in addition to lack of experience, are likely to report the minimum permitted number of contributory factors that may be inaccurate or incomplete. The extent of these errors is practically unknown because, unlike to some other crash attributes, the evaluation and correction of these potential errors are not easy and require substantial amounts of resources and evidence. In-depth crash investigations, where crash investigation teams visit the crash scene and collect data independently of the police and evaluate the possible reasons for accidents, can significantly enhance the quality of contributory factors (e.g. Beanland et al., 2013; Flannagan et al., 2015). This data collection method is available in many countries around the world and can contribute to the development of methods for improving the quality of crash reports (e.g. Couto et al., 2016). It should however be noted that an in-depth investigation is typically implemented on a small scale due to the high operational cost and therefore it is unlikely to replace the most commonly used crash reporting system. In one of the few studies that compare contributory factors identified by the police and by independent specialist teams after visiting the crash site, it was found that there were significant differences between them (Montella 2011). Specifically, contributory factors in police crash reports were mainly focused on drivers' errors while transport experts were concentrated on the interactions between vehicles and the road environment.

An important issue with many contributory factors is the lack of objective and generally accepted descriptors and this is more evident in driver-related factors. For example, fatigue and sleepiness could play a substantial role in crashes, but their definition is not always clear and the perceptions of drivers and police officers might be different (Dobbie, 2002; Stutts et al., 2003). Because of this and also due to the fact that some of the drivers involved in fatigue-related crashes have fatal injuries, fatigue is considered to be generally under-reported (Corfitsen, 1999; Michalaki et al., 2015; Sagberg, 1999). One solution to this problem could be self-reported contributory factors (when possible) as they are more likely to be more accurate than those included in crash reports (Stutts et al., 2003).

It is debatable whether self-reporting would be suitable for contributory factors that are related to drivers' unlawful behaviours such as driving under the influence (DUI). That is because drivers might not want or be able to recall whether they were impaired by a substance (e.g. alcohol, cannabis etc.) (OECD, 2010). Although DUI is a contributory factor that is relatively easy to be assessed objectively, not all involved drivers are tested (ETSC, 2006; Orsay et al., 1994). In the UK, for example, only 54% of drivers involved in crashes had a breath test afterwards between 2003 and 2015 (Department for Transport, 2015). This could be due to difficulty of officers in identifying potentially impaired users, unavailability or malfunction of necessary equipment, time pressure if one or more of the involved users is seriously or fatally injured and even refusal of the involved drivers. Additionally, even if a test takes place, the results might not be included in the actual report (ETSC, 2006).

3. Future directions to improve crash data

Crash reports were, have been and will remain valuable for all road safety analyses as long as researchers are aware of their limitations and use them with caution. Summarising the previous sections, crash reports are likely to contain missing or inaccurate information mainly in crash location and time, severity, participants' characteristics and contributory factors. The issues are not equally serious for all attributes and across datasets. However, poor quality crash data can hinder or damage safety analyses and consequently affect the evaluation and development of successful road safety interventions.

[2] The difference of error between the two datasets may be related with the fact that FARS includes only fatal crashes which are expected to be more carefully reported.

3.1. Topics for further research

The impact of some of the aforementioned crash data flaws on road safety analyses is not yet entirely understood and so more research is needed. Further attention should be also given to the development of transferable methods for addressing crash data inaccuracies that can be broadly applied by safety researchers and practitioners. These efforts can significantly improve the quality and reliability of the datasets however, post-processing approaches are more likely to address some but not all the limitations. To significantly enhance the quality of crash data, radical improvements in crash reporting systems are essential.

Evaluating existing literature on crash data quality few issues that the road safety research community needs to be aware of have been identified:

- Impact of inaccurate reported crash locations on the outcomes of crash analyses (e.g. risk mapping, hot-spot identification, link-based crash frequency models, real-time crash prediction)
- Impact of incorrect reported crash time on the outcomes of crash analyses that consider pre-crash traffic/geometric/weather conditions (e.g. real-time crash prediction, condition-based modelling)
- Impact of imprecise crash severity on the outcomes of crash analyses (e.g. severity modelling, risk mapping)
- Impact of unreliable user information on the outcomes of crash analyses (e.g. behavioural studies, cohort analyses, evaluations of specific users)
- Evaluation of the accuracy of crash database attributes (including reported crash contributory factors) through comparisons with detailed data sourcing from in-depth crash investigations or naturalistic studies
- Methods for crash data post-processing (e.g. crash mapping) for improving data quality and analytical methods for handling crash data reporting errors
- Development of advanced and more accurate crash reporting methods.

3.2. Intelligent crash reporting systems

Current paper-based crash reports are eventually going to be incomplete or erroneous to an extent despite improving officers' training and applying crash data post-processing. For eliminating the chances of mistakes, the ultimate goal for the road safety research community is the development of a seamless crash database where all components can be integrated automatically. A system like that might be entirely applicable only in the era of automated transport systems (e.g. connected and autonomous vehicles) where in-vehicle sensor units (e.g. radars integrated with high accuracy GPS) will have the capability to capture the trajectories of all surrounding vehicles and crashes (if any) will be automatically reported without any human intervention. Although such transport systems represent a major innovation for the automotive industry, their potential impact with respect to timing, uptake and penetration remains uncertain, especially in the near future.

However, the development of more intelligent and integrated crash reporting systems is possible with the technology that is currently available; in fact, a number of crash reporting systems with intelligent features are emerging around the world. The first requirement of a seamless crash database is an official base-map that would store all the road-related attributes (such as traffic and geometry) with common indexes, so linkage would be direct and accurate. Public authorities by reporting crashes with the aid of GPS-based applications would enable automatic identification and linkage of the crash locations with the road environment conditions. Advanced GPS-applications are already being used in several new crash reporting systems such as Traffic and Criminal Software (TRACS) in the US, Collision Recording and Sharing (CRASH) in the UK, PRC and ReGIS in Italy (Department for Transport, 2011c; Montella et al., 2017; Ogle, 2007). Improved accuracy of crash locations combined with detailed real-time traffic, weather and geometric information would enable the identification of the actual conditions responsible for the crash occurrence. The system would be enhanced substantially if the time of crash is reported as the time that the police received the call automatically so the error in this information could be captured more easily (as it is currently applied in Saudi Arabia (Altwaijri, 2013)). An alternative to that would be the use of the reported time from in-vehicle systems that call automatically the nearest emergency centre such as the eCall initiative introduced by the European Union (European Commission, 2016).

Crash severity and casualty information should be also enhanced and corrected using applications that will enable the linkage of hospital records and crash databases. Demographic information and crash history of all drivers of a country or region could be stored in a database and could be identified by the crash reporting application automatically using the driving licence number, magnetic strips or barcodes that are applied in New York, US (Ogle, 2007). Crash contributory factors remain one of the most challenging information to be accurately reported however the use of recordings from on-board and dash-board in-vehicle cameras or event data recorders can significantly enhance the initial statement about the causes of a crash made by the officers. To that end, when such data will be available they should be stored and processed in order to identify contributory factors that were not known or noticed when the report was first completed at scene.

Crash reporting systems that incorporate more technological applications do have limitations. Reliance on technology may lead to systematic errors or data losses in the case of equipment failure or misuse and might even be vulnerable to hacking. Moreover, purchasing and developing new crash reporting systems as well as training of data collecting teams is costly. The time required to fill in crash report forms is also likely to be increased, at least when a new system is newly introduced (Montella et al., 2017). The vision for prefect crash databases may be utopic, however now it is more timely and necessary than ever to work on improving crash data in anticipation of new and advanced road safety analyses that will lead to the development of more successful preventive measures and technologies.

Acknowledgement

We would like to thank Attica Tollway Operations Authority (http://en.aodos.gr/) for providing us with crash data.

References

Abay, K.A., 2015. Investigating the nature and impact of reporting bias in road crash data. Transp. Res. Part A Policy Pract. 71, 31–45.

Abdel-Aty, M.A., Pemmanaboina, R., 2006. Calibrating a real-time traffic crash-prediction model using archived weather and ITS traffic data. IEEE Trans. Intell. Transp. Syst. 7, 167–174.

Abdel-Aty, M., Pande, A., Hsia, L.Y., Abdalla, F., 2005. The potential for real-time traffic crash prediction. ITE J. 75, 69.

Abdel-Aty, M.A., Hassan, H.M., Ahmed, M., Al-Ghamdi, A.S., 2012. Real-time prediction of visibility related crashes. Transp. Res. Part C Emerg. Technol. 24, 288–298.

Alsop, J., Langley, J., 2001. Under-reporting of motor vehicle traffic crash victims in New Zealand. Accid. Anal. Prev. 33, 353–359.

Altwaijri, S., 2013. Analysing traffic crashes in Riyadh City using statistical models and geographic information systems. {\copyright} Saleh Altwaijri.

Amoros, E., Martin, J.-L., Laumon, B., 2006. Under-reporting of road crash casualties in France. Accid. Anal. Prev. 38, 627–635.

Austin, K., 1995. The identification of mistakes in road accident records: part 1, locational variables. Accid. Anal. Prev. 27, 261–276.

BITRE, 2016. Australian Road Deaths Database [WWW Document]. Dep. Infrastruct. Reg. Dev, URL https://bitre.gov.au/statistics/safety/fatal_road_crash_database.aspx (Accessed 6.15.16).

Barancik, J.I., Fife, D., 1985. Discrepancies in vehicular crash injury reporting: northeastern Ohio trauma study IV. Accid. Anal. Prev. 17, 147–154.

Beanland, V., Fitzharris, M., Young, K.L., Lenné, M.G., 2013. Driver inattention and driver distraction in serious casualty crashes: data from the Australian national crash in-depth study. Accid. Anal. Prev. 54, 99–107, http://dx.doi.org/10.1016/j.aap.2012.12.043.

Brown, K., Sarasua, W.A., Ogle, J.H., Mammadrahimli, A., Chowdhury, M., Davis, W.J., Huynh, N., 2015. Assessment of crash location improvements in map-based geocoding systems and subsequent benefits to geospatial crash analysis. In: Transportation Research Board 94th Annual Meeting, Washington, DC.

Burns, S., Miranda-moreno, L., Stipancic, J., Saunier, N., Ismail, K., 2013. An accessible and practical geocoding method for traffic collision record mapping: a quebec case study. Transp. Res. Rec. (2460), 39–46.

Chung, Y., Chang, I., 2015. How accurate is accident data in road safety research? An application of vehicle black box data regarding pedestrian-to-taxi accidents in Korea. Accid. Anal. Prev. 84, 1–8, http://dx.doi.org/10.1016/j.aap.2015.08.001.

Corfitsen, M.T., 1999. Fatigueamong young male night-time car drivers: is there a risk-taking group? Saf. Sci. 33, 47–57.

Council, F., Mohamedshah, Y.M., 2009. Highway Safety Information System Guidebook for the Illinois State Data Files [WWW Document]. Fed. Highw. Adm, URL http://www.hsisinfo.org/guidebooks/illinois.cfm (Accessed 7.10.16).

Couto, A., Amorim, M., Ferreira, S., 2016. Reporting road victims: assessing and correcting data issues through distinct injury scales. J. Saf. Res. 57, 39–45.

Cummings, P., 2002. Association of seat belt use with death: a comparison of estimates based on data from police and estimates based on data from trained crash investigators. Inj. Prev. 8, 338–341, http://dx.doi.org/10.1136/ip.8.4.338.

Deka, L., Quddus, M., 2014. Network-level accident-mapping: distance based pattern matching using artificial neural network. Accid. Anal. Prev. 65, 105–113.

Department for Transport, 2011. STATS 20 – Instructions for the Completion of Road Accident Reports from non-CRASH Sources.

Department for Transport, 2011. STATS19 road accident injury statistics –report form [WWW Document]. URL https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/230590/stats19.pdf.

Department for Transport, 2011. Collision Recording And SHaring (CRASH) [WWW Document]. Natl. Arch. URL http://webarchive.nationalarchives.gov.uk/20110503151558/http:/dft.gov.uk/pgr/statistics/committeesusergroups/crash.

Department for Transport, 2013. Reported Road Casualties in Great Britain 2012 Annual Report: A valuation of road accidents and casualties in Great Britain in 2012.

Department for Transport, 2015. Reported Road Casualties Great Britain:2014 Annual Report. London, UK.

Dobbie, K., 2002. Fatigue-related crashes: An analysis of fatigue-related crashes on Australian roads using an operational definition of fatigue.

Dove, A.F., Pearson, J.C., Weston, P.A., 1986. Data collection from road traffic accidents. Emerg. Med. J. 3, 193–198, http://dx.doi.org/10.1136/emj.3.3.193.

Dutta, A., Noyce, D., 2005. Impact of raising speed limits on traffic safety.

ETSC, 2006. Road Accident Data in the Enlarged European Union: Learning from Each Other. European Transport Safety Council, Brussels.

Elvik, R., Mysen, A.B., 1999. Incomplete accident reporting: meta-analysis of studies made in 13 countries. Transp. Res. Rec., 133–140.

European Commission, 2016. eCall: Time saved = lives savedNo Title [WWW Document]. URL https://ec.europa.eu/digital-single-market/en/ecall-time-saved-lives-saved (Accessed 7.12.16).

Farmer, C.M., 2003. Reliability of police-reported information for determining crash and injury severity.

Flannagan, C.A.C., Balint, A., Klinich, K.D., Sander, U., Manary, M.A., Cuny, S., McCarthy, M., Phan, V., Wallbank, C., Green, P.E., et al., 2015. Comparing motor-vehicle crash risk of EU and US vehicles.

Golob, T.F., Recker, W.W., 2003. Relationships among urban freeway accidents, traffic flow, weather, and lighting conditions. J. Transp. Eng. 129, 342–353.

Heitjan, D., Basu, S., 1996. Distinguishing missing at random and missing completely at random. Am. Stat. 50, 207–213.

Herzog, T.N., Scheuren, F.J., Winkler, W.E., 2007. Data quality and record linkage techniques. Springer Science & Business Media.

Hossain, M., Muromachi, Y., 2013. Understanding crash mechanism on urban expressways using high-resolution traffic data. Accid. Anal. Prev. 57, 17–29.

Imprialou, M.-I.M., Quddus, M., Pitfield, D.E., 2014. High accuracy crash mapping using fuzzy logic. Transp. Res. Part C Emerg. Technol. 42, 107–120.

Imprialou, M.-I., Quddus, M., Pitfield, D., 2015. Multilevel logistic regression modeling for crash mapping in metropolitan areas. Transp. Res. Rec. J. Transp. Res. Board, 39–47.

Imprialou, M.-I.M., Quddus, M., Pitfield, D.E., Lord, D., 2016. Re-visiting crash–speed relationships: a new perspective in crash modelling. Accid. Anal. Prev. 86, 173–185.

Imprialou, M.-I., 2015. Developing accident-speed relationships using a new modelling approach. {\copyright} Maria-Ioanna Imprialou.

Janstrup, K.H., Kaplan, S., Hels, T., Lauritsen, J., Prato, C.G., 2016. Understanding traffic crash under-reporting: linking police and medical records to individual and crash characteristics. Traffic Inj. Prev. 17, 580–584, http://dx.doi.org/10.1080/15389588.2015.1128533.

Johansson, O., Pearce, D., Maddison, D., 2014. Blueprint 5: True Costs of Road Transport. Routledge.

Juran, J.M., Godfrey, A.B., 1999. Quality Handbook. Republished McGraw-Hill.

Loo, B.P.Y., 2006. Validating crash locations for quantitative spatial analysis: a GIS-based approach. Accid. Anal. Prev. 38, 879–886, http://dx.doi.org/10.1016/j.aap.2006.02.012.

Ma, P.E., 2009. Bayesian analysis of underreporting Poisson regression model with an application to traffic crashes on two-lane highways. Transportation Research Board 88th Annual Meeting.

Mannering, F.L., Bhat, C.R., 2014. Analytic methods in accident research: methodological frontier and future directions. Anal. Methods Accid. Res. 1, 1–22, http://dx.doi.org/10.1016/j.amar.2013.09.001.

McDonald, G., Davie, G., Langley, J., 2009. Validity of police-reported information on injury severity for those hospitalized from motor vehicle traffic crashes. Traffic Inj. Prev. 10, 184–190, http://dx.doi.org/10.1080/15389580802593699.

Michalaki, P., Quddus, M.A., Pitfield, D., Huetson, A., 2015. Exploring the factors affecting motorway accident severity in England using the generalised ordered logistic regression model. J. Saf. Res. 55, 89–97.

Miler, M., Todić, F., Ševrović, M., 2016. Extracting accurate location information from a highly inaccurate traffic accident dataset: a methodology based on a string matching technique. Transp. Res. Part C Emerg. Technol. 68, 185–193.

Montella, A., Chiaradonna, S., Criscuolo, G., De Martino, S., 2017. Development and evaluation of a web-based software for crash data collection, processing and analysis. Accid. Anal. Prev, http://dx.doi.org/10.1016/j.aap.2017.01.013.

Montella, A., 2011. Identifying crash contributory factors at urban roundabouts and using association rules to explore their relationships to different crash types. Accid. Anal. Prev. 43, 1451–1463.

Morris, A., Mackay, M., Wodzin, E., Barnes, J., 2003. Some injury scaling issues in UK crash research.

OECD, 2010. Drugs and Driving: Detection and Deterrence. OECD Publishing, Paris.

OECD/ITF, 2015. Road Safety Annual Report 2015, International Transport Forum. OECD Publishing, Paris.

Attiki Odos, 2017. Attiki Odos-Brief Description [WWW Document]. URL http://en.aodos.gr/description/ (Accessed 2.10.17).

Ogle, J.H., 2007. Technologies for Improving Safety Data. Transportation Research Board.

Oliveira, P., Rodrigues, F., Henriques, P.R., 2005. A formal definition of data quality problems. In: Tenth International Conference on Information Quality (ICIQ'05), MIT, Cambridge, MA USA.

Orsay, E.M., Doan-Wiggins, L., Lewis, R., Lucke, R., RamaKrishnan, V., 1994. The impaired driver: hospital and police detection of alcohol and other drugs of abuse in motor vehicle crashes. Ann. Emerg. Med. 24, 51–55.

Popkin, C.L., Campbell, B.J., Hansen, A.R., Stewart, R.R., 1991. Analysis of the accuracy of the existing KABCO injury scale.

Sagberg, F., 1999. Road accidents caused by drivers falling asleep. Accid. Anal. Prev. 31, 639–649.

Salifu, M., Ackaah, W., 2012. Under-reporting of road traffic crash data in Ghana. Int. J. Inj. Control. Saf. Promot. 19, 331–339.

Sherman, H.W., Murphy, M.J., Huelke, D.F., 1976. A reappraisal of the use of police injury codes in accident data analysis. Proceedings: American Association for Automotive Medicine Annual Conference, 128–138.

Stanton, N.A., Salmon, P.M., 2009. Human error taxonomies applied to driving: a generic driver error taxonomy and its implications for intelligent transport systems. Saf. Sci. 47, 227–237.

Stutts, J.C., Wilkins, J.W., Scott Osberg, J., Vaughn, B.V., 2003. Driver risk factors for sleep-related crashes. Accid. Anal. Prev. 35, 321–331, http://dx.doi.org/10.1016/S0001-4575(02)00007-6.

U. S. Department of Transportation, 2010. FARS Analytical User's Manual 1975–2012.

Viano, D.C., Parenteau, C.S., 2009. Belt use: comparison of NASS-CDS and police crash reports. Traffic Inj. Prev. 10, 427–435.

Watson, A., Watson, B.C., Vallmuur, K., 2013. How accurate is the identification of serious traffic injuries by Police? In: The Concordance Between Police and Hospital Reported Traffic Injuries, Proceedings of the 2013 Australasian Road Safety Research, Policing & Education Conference.

Watson, A., Watson, B., Vallmuur, K., 2015. Estimating under-reporting of road crash injuries to police using multiple linked data collections. Accid. Anal. Prev. 83, 18–25.

Whitelegg, J., Haq, G., 2006. Vision zero: adopting a target of zero for road traffic fatalities and serious injuries. The Stockholm Environment Institute, Stohcklom, Sweden.

Xu, C., Wang, W., Liu, P., 2013. A genetic programming model for real-time crash prediction on freeways. IEEE Trans. Intell. Transp. Syst. 14, 574–586.

Yamamoto, T., Hashiji, J., Shankar, V.N., 2008. Underreporting in traffic accident data, bias in parameters and the structure of injury severity models 40, 1320–1329. 10.1016/j.aap.2007.10.016.

Yannis, G., Papadimitriou, E., Chaziris, A., Broughton, J., 2014. Modeling road accident injury under-reporting in Europe. Eur. Transp. Res. Rev. 6, 425–438.

Yasmin, S., Eluru, N., 2013. Evaluating alternate discrete outcome frameworks for modeling crash injury severity 59, 506–521.

Ye, F., Lord, D., 2011. Investigation of effects of underreporting crash data on three commonly used traffic crash severity models 51–58.