Journal of Safety Research xxx (2017) xxx–xxx

Contents lists available at ScienceDirect

Journal of Safety Research

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

Understanding factors associated with misclassification of fatigue-related accidents in police record

Yanyan Li, [a,*] Toshiyuki Yamamoto, [b] Guangnan Zhang [c,d]

[a] Department of Civil Engineering, Nagoya University, Furo-cho, Chikusa-ku, Nagoya 464-8603, Japan
[b] Institute of Materials and Systems for Sustainability, Nagoya University, Furo-cho, Chikusa-ku, Nagoya 464-8603, Japan
[c] Institute of Guangdong, Hong Kong and Macao Development Studies, Sun Yat-Sen University, Xingang Xi Road, Guangzhou, China
[d] Center for Studies of Hong Kong, Macao and Pearl River Delta, Sun Yat-Sen University, Xingang Xi Road, Guangzhou, China

ARTICLE INFO

Article history:
Received 7 July 2017
Received in revised form 7 September 2017
Accepted 5 December 2017
Available online xxxx

Keywords:
Fatigue
Police report
Misclassification
Association rule
Logistic regression model

ABSTRACT

Introduction: Fatigue is one of the riskiest causes of traffic accidents threatening road safety. Due to lack of proper criteria, the identification of fatigue-related accident by police officers largely depends on inferential evidence and their own experience. As a result, many fatigue-related accidents are misclassified and the harmfulness of fatigue on road safety is misestimated. *Method:* In this paper, a joint model framework is introduced to analyze factors contributing to misclassification of a fatigue-related accident in police reports. Association rule data mining technique is employed to identify the potential interactions of factors and logistic regression models are applied to analyze factors that hinder police officers' identification of fatigue-related accidents. Using the fatigue-related crash records from Guangdong Province during 2005–2014, factors contributing to the false positive and false negative detection of the fatigue-related accident have been identified and compared. *Results:* Some variables and interactions were identified to have significant impacts on fatigue-related accident detection. *Conclusions:* Based on the results, it can be inferred that the stereotype of certain groups of drivers, crash types, and roadway conditions affects police officers' judgment on fatigue-related accidents. *Practical applications:* This finding can provide useful information for training police officers and build better criteria for fatigue identification.

1. Introduction

Fatigued driving is a serious problem threatening road safety around the world. Police records from different countries indicate a range of 1%–4% incidence of fatigue/sleep-related crashes of all registered crashes (Radun & Radun, 2009; Traffic Management Bureau, Ministry of Public Security, PRC, 2008). However, several questionnaire-based surveys suggest that the role of fatigue in a traffic accident is misestimated. The National Sleep Foundation (2008) reported that approximately 32% of respondents in the *Sleep in America* poll had driven while fatigued at least once a month. In China, a survey conducted in Guangdong province in 2007 also showed that 9.3% of drivers had driven while fatigued in the past 30 days (Yan, Ma, & Xu, 2010). The difference between police reports and surveys implies that police reports could have significantly misestimated the harmfulness of fatigue in road safety. One of the possible reasons is that police officers are not as aware to the presence of fatigue and have difficulties in identifying fatigue-related accidents (Robertson, Holmes, & Vanlaar, 2009).

Among all causes of traffic accidents, fatigue-related accidents are easily neglected or misclassified due to the difficulty in observing and identifying driver fatigue (Filtness, Armstrong, Watson, & Smith, 2015; Radun, Ohisalo, Radun, Wahde, & Kecklund, 2013). Unlike drunk driving crashes, no blood or breath test can be applied to quantify driver's fatigue level at the crash scene (DaCoTA, 2012; Pack et al., 1995). As a result, there is currently no standard methodology for identifying fatigue as the cause of the crash (Crummy, Cameron, Swann, Kossmann, & Naughton, 2008; Filtness et al., 2015) and defining fatigue-related accident largely relies on inferential evidence or experience. For example, police officers may consider a crash to be fatigue-related when the following conditions appear (Horne & Reyner, 1995, 1999; NCSDR/NHTSA, 1998): occur late at night or mid-afternoon; single vehicle run off the roadway; occur on a high-speed road; absence of skid marks or braking. Some fatigue-related accidents were determined even by eliminating other causes of accidents (e.g., speeding, drunk driving, etc.).

To assist in identifying fatigue in an accident, proxy measurements are developed aiming to improve reporting accuracy of fatigue-related accidents (Filtness et al., 2015). In Australia, the Australian Transport Safety Bureau (ATSB, 2006) has developed the proxy definition for fatigue/sleep-related accident, and five jurisdictions in Australia have already incorporated proxy definition into their reporting process. In

* Corresponding author.
E-mail addresses: lyy901207@gmail.com (Y. Li), yamamoto@civil.nagoya-u.ac.jp (T. Yamamoto), sysuzgn@gmail.com (G. Zhang).

https://doi.org/10.1016/j.jsr.2017.12.002

Queensland, for example, fatigue can be considered as a contributor to a crash when it fitted the proxy definition: single-vehicle crashes in more than 100 km/h speed zones which occur during midnight and in the afternoon, or where a vehicle runs out of roadway and the driver does not try to avoid the accident (Armstrong, Filtness, Watling, Barraclough, & Haworth, 2013; Filtness et al., 2015). Although these proxy definitions are based on experience or scientific research, they are criticized for being too specific (Armstrong et al., 2013; Crummy et al., 2008) and may provide misleading instructions for police officers. A questionnaire-based study conducted in Australia by Crummy et al. (2008) found that only a small proportion of participants that actually had a fatigue/sleep-related crash was correctly identified by ATSB proxy definitions (ATSB, 2006).

Reliable and accurate records are essential for assessing the scope of fatigue-related accident problems, as well as monitoring and evaluating the effectiveness of intervention measures. A survey in Ontario showed that 56.6% of traffic police felt that they did not receive enough training to identify drivers who were fatigued or drowsy, or determined the role of fatigue in a crash (Robertson et al., 2009). Although several risk factors identified by prior research and public belief are believed to contribute to fatigue-related accidents, few works have been done to prove whether these factors are useful for police officers to identify fatigue-related accidents. That is, some of the factors believed to be associated with fatigue-related accidents are not helpful in judging whether an accident is fatigue-related, and may even lead to incorrect classification of the cause of accidents. Therefore, in this study, we proposed an analysis framework based on existing crash data to identify factors that easily make fatigue-related accidents misclassified by police officers, examine the interactive effects of those factors, and provide better inference for determining fatigue-related accident by removing some misleading terms, which help to improve enforcement strategies.

The paper is organized as follows: In Section 2, the detail of analysis strategies will be discussed; Section 3 describes the dataset and variables used in this study; the results will be presented in Section 4; Section 5 will further discuss the results; in the last section of the paper, a conclusion will be given.

2. Methodology

2.1. Objectives and research strategy

This study aims at investigating potential factors that hinder police officers' identification of fatigue-related accidents. However, some factors have individual effects as well as combined effects on the determining of fatigue-related accidents. Classic logistic regression model lacks appropriate criteria to incorporate interactions between independent variables when there are a large number of variables to be considered. Instead, ignoring interactions may cause biased estimation. Therefore, our strategies for this analysis are: (a) association rule data mining technique is applied to identify important interactions between factors, which helps overcome the disadvantage of classic logistic regression model in selecting appropriate interactions; (b) incorporating the interactions identified by association rules, binary logistic regression models are applied to find out factors that hinder police officers from correctly identifying fatigue-related accidents.

2.2. Association rule analysis

Regression models in road safety research focus on establishing and analyzing relationships between "dependent" and "independent." It is also important to take the correlation between "independent" variables into consideration since it may hamper the statistical analysis (Pande & Abdel-Aty, 2009). With the increasing number of independent variables, however, the number of interactions will grow at an accelerated rate. Thus, the methodology for identifying potential interaction among a large number of crash-related factors is needed. Association rule data mining technique can potentially identify relationships that are not well known from current research works and have been used in traffic safety research (Das & Sun, 2014; Montella, Aria, D'Ambrosio, & Mauriello, 2011; Pande & Abdel-Aty, 2009; Weng, Zhu, Yan, & Liu, 2016). Some studies have combined association rule data mining technique with logistic regression model for other purposes (Kamei, Monden, Morisaki, & Matsumoto, 2008; Shaharanee, Hadzic, & Dillon, 2009), but few of them use association rule analysis as a tool for selecting potential interactions among variables. Changpetch and Lin Q9
(2013a, 2013b) proposed a model selection method procedure for logistic (Changpetch & Lin, 2013a) and multinomial logit model (Changpetch & Lin, 2013b), which help to improve the classic model by considering potential interactions.

In this study, association rule analysis is performed using a priori algorithm according to the methodology introduced by Agrawal, Imieliński, and Swami (1993). A rule is defined as an implication of the form "A → B", where A is the antecedent (left-hand-side, LHS) and B is the consequent (right-hand-side, RHS). It is important to note that the rule should not be interpreted as a direct causation, but as associations between variables (Montella, Aria, D'ambrosio, & Mauriello, 2012; Pande & Abdel-Aty, 2009). Three measures are commonly used in filtering rules: support, confidence, and lift. Support measures the frequency of LHS and RHS appearing in the dataset and is calculated as follows:

$$\text{Support}(A \rightarrow B) = P(AB) \tag{1}$$

where P(AB) represents the probability of case containing A and B at the same time. Confidence determines how frequently RHS appears given that LHS occurs:

$$\begin{aligned}\text{Confidence}(A \rightarrow B) &= P(B|A) \\ &= P(AB)/P(A)\end{aligned} \tag{2}$$

where P(A) is the probability of case containing A. Lift is a measure of the statistical dependence of the rule. A lift value that is smaller than 1 indicates negative independence between LHS and RHS, a value equal to 1 indicates independence, and a value that is greater than 1 indicates positive interdependence (Montella et al., 2012). Higher lift value indicates stronger associations. Lift is defined as follows:

$$\begin{aligned}\text{Lift}(A \rightarrow B) &= P(B|A)/P(B) \\ &= P(AB)/P(A)P(B)\end{aligned} \tag{3}$$

To make sure that the identified rules are reasonable and accurate, the minimum threshold values for these three indexes need to be specified. Since there are no clear criteria for choosing threshold values, different studies employed different threshold support and confidence values (De Oña, López, & Abellán, 2013; Montella et al., 2011, 2012; Pande & Abdel-Aty, 2009) based on the nature of the data (balanced Q10
or not) and sample size (small or large databases). For example, Pande and Abdel-Aty (2009) set 0.009 and 0.1 for them, respectively. Thus, in this study the minimum threshold values for support, confidence and lift are set as follows: Support ≥0.01, Confidence ≥0.1, and Lift ≥1.2. It also needs to be emphasized that only rules with two items in the LHS are selected for ease of interpretation. We first generate rules with non-fatigue to fatigue crash or fatigue to non-fatigue crash in RHS from all the generated rules. Then, all the selected rules are descending ordered by confidence, and the top 10 are changed into interactions and incorporated into logistic regression models as inputs.

2.3. Binary logistic regression model

Two assessment results will be recorded for normal procedure crash records in the database. The one recorded by the police officer at the crash scene is denoted as on-site assessment. Normally, *on-site assessment* was determined by a quick check at the crash scene, surrounding

Please cite this article as: Li, Y., et al., Understanding factors associated with misclassification of fatigue-related accidents in police record, *Journal of Safety Research* (2017), https://doi.org/10.1016/j.jsr.2017.12.002