An investigation of the speeding-related crash designation through crash narrative reviews sampled via logistic regression

Cole D. Fitzpatrick (Graduate Research Assistant) [a],[*],
Saritha Rakasi (Graduate Research Assistant) [a],
Michael A. Knodler Jr. (Associate Professor) [b]

[a] *University of Massachusetts Amherst, 139B Marston Hall, 130 Natural Resources Road, Amherst, MA 01003, United States*
[b] *University of Massachusetts Amherst, 214 Marston Hall, 130 Natural Resources Road, Amherst, MA 01003, United States*

ARTICLE INFO

ABSTRACT

Speed is one of the most important factors in traffic safety as higher speeds are linked to increased crash risk and higher injury severities. Nearly a third of fatal crashes in the United States are designated as "speeding-related", which is defined as either "the driver behavior of exceeding the posted speed limit or driving too fast for conditions." While many studies have utilized the speeding-related designation in safety analyses, no studies have examined the underlying accuracy of this designation. Herein, we investigate the speeding-related crash designation through the development of a series of logistic regression models that were derived from the established speeding-related crash typologies and validated using a blind review, by multiple researchers, of 604 crash narratives. The developed logistic regression model accurately identified crashes which were not originally designated as speeding-related but had crash narratives that suggested speeding as a causative factor. Only 53.4% of crashes designated as speeding-related contained narratives which described speeding as a causative factor. Further investigation of these crashes revealed that the driver contributing code (DCC) of "driving too fast for conditions" was being used in three separate situations. Additionally, this DCC was also incorrectly used when "exceeding the posted speed limit" would likely have been a more appropriate designation. Finally, it was determined that the responding officer only utilized one DCC in 82% of crashes not designated as speeding-related but contained a narrative indicating speed as a contributing causal factor. The use of logistic regression models based upon speeding-related crash typologies offers a promising method by which all possible speeding-related crashes could be identified.

Published by Elsevier Ltd.

1. Introduction

Throughout the literature, speed is one of the most important factors in traffic safety. As speed increases, so too does the risk of a crash in both rural and urban areas (Aarts and van Schagen, 2006) as does the severity of crashes involving pedestrians, (Rosén et al., 2011) and not involving pedestrians (Mao et al., 1997). Nearly a third of fatal crashes in the United States are designated as "speeding-related", which is defined by the National Highway Traffic Safety Administration (NHTSA) as "the driver behavior of exceeding the posted speed limit or driving too fast for conditions." (Liu and Chen, 2009). This speeding-related crash designation is critical as the American Association of State Highway Transporta-

tion Officials (AASHTO) Strategic Highway Safety Plan recommends the use of targeted conventional speed enforcement as a strategy to reduce speeding-related crashes (Neuman et al., 2009). This type of strategy requires accurate data related to roadways with a high frequency of speeding-related crashes. However, an inherent challenge with the speeding-related designation is the manner in which it is derived. The law enforcement officer who responds to a crash and completes the subsequent crash report must select one or more *Driver Contributing Codes* (DCCs) which are supposed to explain why the crash occurred. This discretionary decision is often made following an investigation of the scene and interviews with the motor vehicle operator(s) and any witnesses.

Numerous studies have investigated speeding-related crashes, and while none investigated the reliability of the speeding-related designation, each acknowledged the limitations of the designation. For example, the Oregon Department of Transportation conducted a study where high speeding-related crash locations were identified for possible mitigation. In their discussion they note, "the

* Corresponding author.
E-mail addresses: cfitzpat@umass.edu (C.D. Fitzpatrick), srakasi@umass.edu (S. Rakasi), mknodler@ecs.umass.edu (M.A. Knodler Jr.).

analysis relies on crash reports, which are subject to the interpretations of a variety of individuals completing the crash report form. Specifically, the fact that a crash has been identified as speeding-related is not based on a scientific analysis, and may be the result of opinion or best judgment" (Monsere et al., 2006).

The Federal Highway Administration (FHWA) funded a study which developed a speeding-related typology and compared data from two different states which used differing definitions for speeding-related crashes. The study noted several crash characteristics which were more commonly found in crashes designated as speeding-related. Additionally, they concluded that the NHTSA definition was most appropriate for the speeding-related classification. Finally, the report cautioned against the type of analyses which was conducted in Oregon stating, "it is difficult to know whether an identified variable shows a true higher association with speed or whether the association shown is partially due to an officer bias" and "treatment programs oriented to these factors may not be as successful as if oriented to other characteristics where such a bias is not expected" (Council et al., 2010).

In 2014, a Speed Management Plan was developed jointly by NHSTA, FHWA, and the Federal Motor Carrier Safety Administration (FMCSA). The plan sought to reduce speeding-related fatalities and injuries and improve the safety experience for all road users. While the plan recommended a data driven approach using the speeding-related designation, it also cautioned that "the precise role of speeding in crashes can be difficult to ascertain, as speeding is often defined in broad terms. Further, the determination of whether speeding was involved in a fatal crash is often based on the judgment of the investigating law enforcement officer" (Speed Management Program Plan, 2014).

The crash narrative is the responding officer's written account of what occurred before, during, and after the crash. Crash narratives can be used to more thoroughly investigate the cause of a crash as crash narratives often provide information beyond what is captured in the pre-defined fields of the crash report. Examples highlighting the utility of crash narratives are present throughout the traffic safety related literature. In one of the more in depth studies, McKnight and McKnight reviewed 2000 crash narratives to determine if crashes involving younger drivers were due to carelessness or inexperience (McKnight and McKnight, 2003). Crash narratives have also been utilized previously to conduct in-depth investigations of crashes involving military vehicles (Pollack et al., 2013), work zone crashes (Swansen et al., 2013), helmet status in motorcycle crashes (Graves et al., 2015), and distraction-related crashes (Dube et al., 2016).

1.1. Objectives and hypotheses

The primary objective of this study was to improve the identification of speeding-related crashes by investigating commonalities in the types of crashes that are routinely misclassified as either speeding-related or not speeding-related. Logistic regression models based upon established speeding-related crash typologies were developed to predict the probability that a specific crash would be designated as speeding-related. The model outputs were then used to strategically sample crash narratives in order to identify potential crashes where the model prediction disagreed with officer's recorded crash causation (i.e. driver contributing code). The resulting evaluation of crash narratives was based upon two hypotheses that were tested:

Hypothesis 1: model predictions correlate with crash causation determinations resulting from crash narrative reviews.

Hypothesis 2: commonalities exist among the crashes with a misclassified speeding-related designation as determined through the crash narrative reviews.

The resulting output of the hypothesis testing would be an improved methodology to identify speeding-related crashes and any crash commonalities identified from misclassified crashes would be used to improve the classification of speeding-related crashes.

2. Methods

This study consisted of three primary phases. First, a series of logistic regression models were developed to assign a probability that a crash was, or was not, designated via the crash report as being speeding-related. Second, these models were utilized to sample crash reports for subsequent crash narrative reviews by multiple researchers that were unaware of the crash designation (i.e. a double blind narrative review). Finally, based upon the crash narrative review, specific crashes which had crash narratives that did not align with the officer's speeding-related designation were manually reviewed to identify shared characteristics. This section will describe the methods for the three phases of this study.

2.1. Logistic regression model

2.1.1. Data

Three years of crash data from the state of Massachusetts from 2012 to 2014 were obtained. The roadway inventory database, maintained by the Massachusetts Department of Transportation (MassDOT) was utilized in order to link the crash to the roadway on which it occurred. Initially, 373,205 unique crashes were included in the database with an individual entry for each driver involved in the crash. Next, any crashes with an improperly coded driver age (e.g., driver age > 110) or driver sex (driver sex ≠ male or female) were removed from the database. For interstate crashes, entries were removed which had recorded speed limits which differed between the crash report and the roadway inventory. This was not conducted on other functional classifications as the speed limits reported on the crash report were inconsistent with those from the roadway inventory file. Instead, speed limit was not included in these models due to the low confidence in the data accuracy. Finally, only entries involving "Person Number: 1", also known as motor vehicle operator #1 (MV1), were included in the model development. This decision was made to conform to one of the fundamental assumptions of logistic regression models which states that all observations must be independent from one another. MV1 was selected for inclusion in the model as MV1 was more commonly at fault for exceeding the posted speed limit or driving too fast for conditions (DTFFC). Specifically, in 4.2% of all crashes MV1 was at fault due to speeding, compared to only 1.1% of crashes being the fault of MV2-5 for speeding. The crashes were then filtered by the functional classification of the roadway on which they occurred in order to create five logistic models. Multiple models were developed in order to improve the prediction capabilities of the model. The grouping of functional classifications and sample size for each model is presented in Table 1. Altogether, 161,419 crashes, both injury and property-damage crashes, were used to develop five different logistic regression models.

2.1.2. Model variables

The five logistic regression models were developed based upon the speeding-related crash typology from (Council et al., 2010). Two crash characteristics were expressed in different ways in order for the model to better fit the data. First, a crash occurring at night can be identified either by the time at which the crash occurred or the light conditions. Second, the crash type input was either single vehicle crash or first harmful event occurring outside of the roadway. The data field which resulted in a better model fit was selected. It was not possible to use both as the fields described are highly