Accident Analysis and Prevention 150 (2021) 105864

Contents lists available at ScienceDirect

Accident Analysis and Prevention

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

Injury risk assessment based on pre-crash variables: The role of closing velocity and impact eccentricity

Michelangelo-Santo Gulino [a,*], Leonardo Di Gangi [b], Alessio Sortino [b], Dario Vangi [a]

[a] Department of Industrial Engineering, Università degli Studi di Firenze, Via di Santa Marta, 3, 50139 Firenze, Italy
[b] Department of Information Engineering, Università degli Studi di Firenze, Via di Santa Marta, 3, 50139 Firenze, Italy

ARTICLE INFO

Keywords:
Velocity change (ΔV)
Crash Momentum Index (CMI)
A priori analysis
Feature ranking
Predictive models
Machine learning

ABSTRACT

Thorough evaluations on injury risk (IR) are fundamental for guiding interventions toward the enhancement of both the road infrastructure and the active/passive safety of vehicles. Well-established estimates are currently based on IR functions modeled on post-crash variables, such as velocity change sustained by the vehicle (ΔV); thence, these analyses do not directly suggest how pre-crash conditions can be modified to allow for IR reduction. Nevertheless, ΔV can be disaggregated into two contributions which enable its *a priori* calculation, based only on the information available at the impact instant: the Crash Momentum Index (CMI), representing impact eccentricity at collision, and the closing velocity at collision (V_r).

By extensively employing the CMI indicator, this work assesses the overall influence of impact eccentricity and closing velocity on the risk for occupants to sustain a serious injury. As CMI synthesizes indications regarding ΔV, its use can be disjointed from the ΔV itself for the derivation of high-quality IR models. This feature distinguishes CMI from the other eccentricity indicators available at the state-of-the-art, allowing for the contribution of eccentricity on IR to be completely isolated. Because of this element of originality, special attention is given to the CMI variable throughout the present work.

Based on data extracted from the NASS/CDS database, the influence of the CMI and V_r variables on IR is specifically highlighted and analyzed from several perspectives. The feature ranking algorithm ReliefF, whose use is unprecedented in the accident analysis field, is first employed to assess importance of such impact-related variables in determining the injury outcome: if compared to vehicle-related and occupant-related variables (as category and age, respectively), the higher influence of CMI and V_r is initially highlighted. Secondly, the relevance of CMI and V_r is confirmed by fitting different predictive models: the fitted models which include the CMI predictor perform better than models which neglect the CMI, in terms of classical evaluation metrics. As a whole, considering the high predictive power of the proposed CMI-based models, this work provides valuable tools for the *a priori* assessment of IR.

1. Introduction

Synthetic data, collected in in-depth accident databases, represent a key element to tune models capable of describing the complexity of real impacts. According to a well-established scheme (Kullgren, 2008), road accidents can be approached from several standpoints by as many dose–response models: three information – responses – represented by the frequency of accidents (Bagdadi, 2013; Wang et al., 2019; Papadimitriou et al., 2019), injury frequency (Ye et al., 2013; Rahman Shaon et al., 2019; Chang et al., 2019; Chiou and Fu, 2013), and injury risk (IR) for the vehicles' occupants (Andricevic et al., 2018; Ding et al., 2019) are dependent on variables linked to the event, whose combination defines the so-called impact severity – dose. Identifying the variables with the most significant influence on accidents is essential to guide interventions aimed at increasing road safety as a whole. Based on the injuries which most frequently occur at a specific site, its dangerousness can be evidenced and changes to the infrastructure proposed to lower the route's speed limit (Anastasopoulos and Mannering, 2011). Furthermore, through specifically developed IR models, the injury outcome of a crash can be inferred at its occurrence based on information from event data recorders (Nishimoto et al., 2019, 2017; Kusano and Gabler, 2014): this application of IR models allows to automatically and promptly select the

* Corresponding author.
E-mail address: michelangelo.gulino@unifi.it (M.-S. Gulino).

https://doi.org/10.1016/j.aap.2020.105864
Received 3 May 2020; Received in revised form 28 September 2020; Accepted 22 October 2020
Available online 29 December 2020

most suitable first intervention vehicle to handle the emergency, providing for a potential reduction in serious injuries and fatalities up to 20% (Nishimoto et al., 2017). This solution currently stands as a priority for the european scientific community: IR models increase the effectiveness of automated emergency call systems (eCalls), which are part of the mandatory instrumentation of newly type-approved vehicles since 2018.[1]

For what specifically concerns the safety offered by vehicles, feedback regarding the typical crash conditions is fundamental to tune and enhance procedures (as the 40% offset EuroNCAP protocol) for crashworthiness assessment by lab tests (Bareiss et al., 2018; Kullgren et al., 2010; Metzger et al., 2015). The recent focusing of the industrial and scientific interest towards Advanced Driver Assistance Systems (ADAS), and more in general on active safety, revealed further and numerous practical applications of IR models; for instance, it is possible to: (a) study the level of safety guaranteed in different critical scenarios by an ADAS in development, as its technical characteristics (e.g., activation time, depth of field) or the opponent type are varied (Rosén et al., 2010); (b) assess the benefits which can be obtained in terms of injury outcome for the ego vehicle's and opponent vehicles' occupants, as a function of the market penetration rate of already developed ADASs (Sander and Lubbe, 2018); (c) propose real-time steering and braking interventions which minimize IR for the involved subjects, in correspondence of inevitable collision states (Vangi et al., 2019d). Therefore, the identification of the variables which make up the impact severity and their correlation with IR represents a key requirement for efficient progress towards active safety enhancement (Jeong et al., 2018).

Statistical techniques as the logistic regression or the probit model are typically used for the development of IR models (Savolainen et al., 2011): information regarding IR is particularly interesting if it relates to the probability of sustaining an injury associated with a Maximum Abbreviated Injury Scale (MAIS) equal to or higher than three (MAIS 3+), indicating a serious injury in at least one body region. Recently, increasing attention has been given to machine learning techniques like random forests or neural networks, particularly suitable for modeling complex phenomena such as impact mechanics (Iranitalab and Khattak, 2017; Abdelwahab and Abdel-Aty, 2001; Li et al., 2018). Analyses focused on variables' significance (only in statistical models) and regressions' predictive power demonstrated that impact severity is represented by the combination of numerous factors. Evidence exists indicating that occupant-related features such as seating position (Mitchell et al., 2015; Viano and Parenteau, 2008), gender and age (Newgard, 2008; Carter et al., 2014) play an important role in the associated injury outcome; for what regards the vehicle properties, the scientific literature is extensive and addresses category (Abdelwahab and Abdel-Aty, 2001; Khattak and Rocha, 2003; Wenzel and Ross, 2005), mass, and size (Wood and Simms, 2002). However, it is not uncommon to retrieve works highlighting a more significant influence of such factors if compared to other studies. For instance Atkinson et al. (2016) note that, unlike some previous works, the seating row represents a feature which significantly influences IR; likewise, while Vadeby (2004) observes an IR model whose fitting quality is not particularly affected by the introduction of the occupant's gender and age, Newgard (2008) highlights that these two factors must be considered to assess the risk of sustaining serious injuries. The explanation is that these features are undoubtedly related to the injury outcome, but to a lesser extent in comparison to impact characteristics. The leading factor which determines the injury degree for the occupants is, in fact, the force (or, equivalently, the acceleration) they experience during the impact; it is thence not surprising that several studies mainly identify impact severity only through the velocity change sustained by the vehicle in the crash (ΔV, proportional to acceleration) and the vehicle area affected by the intrusion. Besides, these two variables alone are sufficient to obtain accurate results in terms of fitting quality on the injury outcome (Kusano and Gabler, 2014; Jurewicz et al., 2016).

Terms like ΔV are identified as post-impact variables, i.e., they can be estimated only if the impact phase between the subjects is reconstructed. Such a detailed reconstruction is computationally expensive: the time required by modern calculators to accomplish the task ranges between minutes for reduced vehicle models and hours/days for complete models (like finite element models or multibody systems) (Vangi et al., 2019c). Nevertheless, the reconstruction of a large amount of present and future conflict scenarios is necessary to fully assess the benefits of introducing new ADASs into the market (Sander and Lubbe, 2018); similarly, in the case of ADAS systems on-board the vehicle with a logic based on the minimization of IR (Vangi et al., 2019d), the best intervention must be determined in real-time starting from the results associated with all possible activations. The greater the number of critical conditions to simulate, the more definite the limitation related to the high calculation time. Nevertheless, the main drawback of post-impact variables as ΔV is that they represent a consequence of the impact; thence, they do not provide direct suggestions on how to modify the vehicles and the infrastructure to reduce the potential IR before collision occurrence. With a view to the overall improvement in road safety, it is thus crucial to rely on *a priori* variables which allow retrieving IR based on pre-impact phase information. To fulfill this requirement, a recent article (Vangi et al., 2019b) proposes the disaggregation of ΔV in two contributions: closing velocity at collision (V_r) and Crash Momentum Index (CMI), such that $\Delta V = \mathrm{CMI} \cdot V_r$. While V_r is clearly an *a priori* variable, CMI has both an *a posteriori* and an *a priori* formulation: ΔV and consequently IR can be derived *a priori* by employing the latter formulation, without necessarily reconstructing the impact phase.

The disaggregation of ΔV in the two contributions is associated with additional advantages: while V_r is an indicator of the effective speed (maximum potential severity), CMI represents impact eccentricity; the higher the CMI, the less eccentric (or oblique) the collision. Therefore, the aim of the disaggregation is not to minimize the role of ΔV in the description of IR: ΔV is still a solid reference, and IR models based on CMI and V_r implicitly contain information regarding ΔV; nevertheless, ΔV disaggregation is essential to separately study the effects of impact speed and eccentricity on IR. Eccentricity has, in fact, a substantial influence on the injury outcome in a crash: closing velocity being the same, an increase in eccentricity decreases the rate of kinetic energy converted into translation (lower ΔV). In extremely eccentric impacts, however, occupants interact in a complex way with the restraint systems and the compartment interiors: occupants' motion is determined by the rotation of the vehicle around its center of mass (Fay et al., 1996). Literature studies are available which aim at assessing the influence of eccentricity on injury, by referring to specific fields of the Collision Deformation Classification (CDC) like the Principal Direction Of Force (PDOF) (Viano and Parenteau, 2008; Pal et al., 2018; Forman and McMurry, 2018; Lai et al., 2012). However, unlike CMI, the CDC alone cannot be used to assess the injury outcome of a hypothetical collision because it does not provide any indication on the associated ΔV. The further information available from the CDC is limited to the extent and location of the intrusion area: CDC is an entirely *a posteriori* variable, resulting in the impossibility of unambiguously determining the point of impact (POI), the arms of the forces, and the actual impact eccentricity as a consequence.

The objective of the present work is to assess the influence of V_r and CMI on impact severity: this allows justifying their joint use to determine ΔV, which is fundamental for *a priori* studies on IR; specifically, because of its elements of originality, special attention is given to the CMI indicator. To achieve a result as reliable and complete as possible, both well-established regression models (like logistic regression) and ma-

[1] Regulation (EU) 2015/758 of the European Parliament and of the Council, concerning type-approval requirements for the deployment of the eCall in-vehicle system based on the 112 service and amending Directive 2007/46/EC – 29 April 2015.

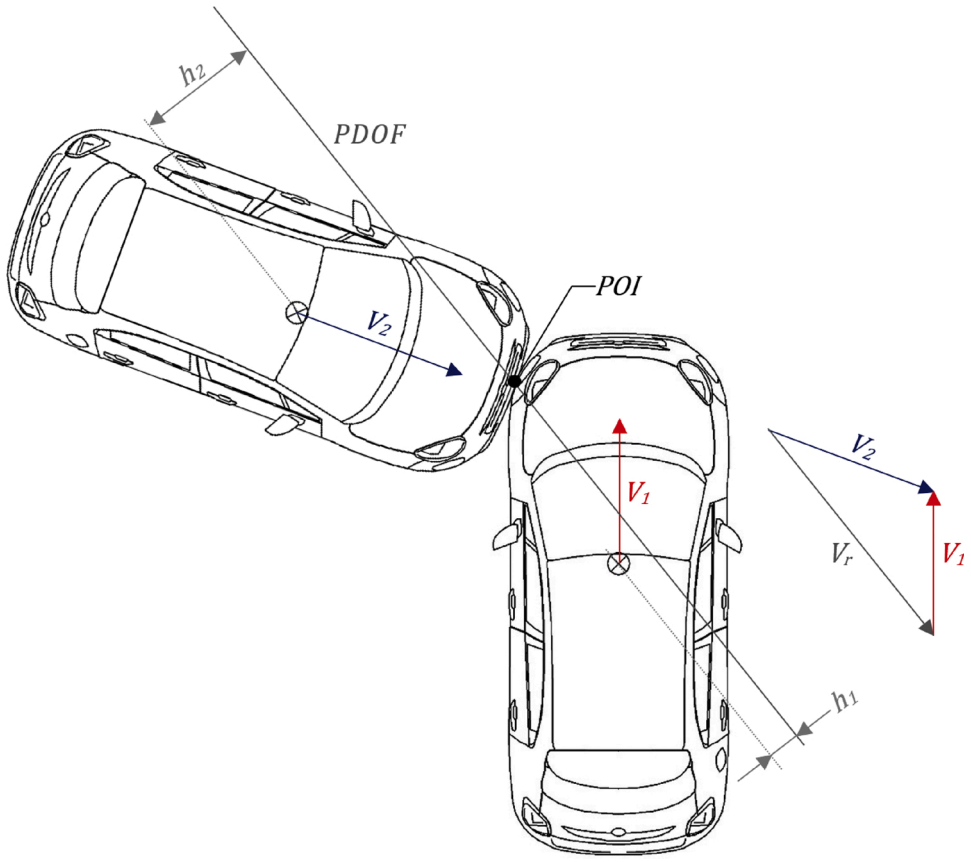

Fig. 1. Planar visualization of an impact between vehicles, whose eccentricity is expressed by the arms of forces h_1 and h_2.

chine learning methodologies are employed to assess such an influence. Support Vector Machines (SVM) in particular, although employed in diverse fields of road safety (Wang et al., 2019; Iranitalab and Khattak, 2017; Chen et al., 2016; Li et al., 2012; Jianfeng et al., 2019), has a unique prior application for the development of IR models (Kusano and Gabler, 2014). An element of absolute novelty in the specific area is the use of a feature ranking technique known as ReliefF: unlike the most widely used Classification And Regression Tree (CART) (Kwon et al., 2015), ReliefF provides results independent of the considered regression model; ReliefF allows to quickly, automatically and objectively determine CMI and V_r contribution to impact severity, also in comparison to the further parameters involved in the modeling. As a consequence, the relative importance of impact characteristics can be established considering the additional features involved, related both to occupants and vehicles. Conversely, features such as the environment (e.g., lighting, weather, and road surface conditions (Alogaili and Mannering, 2020; Fountas et al., 2020; Shannon et al., 2020)) or the sobriety level of the driver (Fountas and Anastasopoulos, 2017) are fundamentally uninformative in the present context: attention is specifically focused on the features available at the collision instant, regardless of the root causes which led to the impact; to exemplarily clarify, if a specific value of speed at the collision instant characterizes the crash, observation of appropriate or adverse environmental conditions is not expected to result in a different injury outcome.

2. Material and methods

The present section highlights the theoretical concepts required to analyze the influence of V_r and CMI on IR. Specifically, the *a* posteriori and the *a* priori formulations of CMI are highlighted first: these allow its value to be derived from, respectively, data contained in in-depth accident databases (for *a* posteriori evaluations on IR) and impact configuration between vehicles (for *a* priori analysis on IR); the set of real-world accidents is subsequently presented on which the analysis has been performed (after extraction of V_r and CMI values associated with the vehicles). Finally, feature ranking techniques and regression models are discussed, which have been employed to thoroughly determine the influence of V_r and CMI on IR.

2.1. Crash Momentum Index (CMI)

The *a* posteriori (or prospective) formulation of CMI has been initially proposed by Huang for front-to-front impacts between two vehicles (Huang, 2002) and has been later extended to any type of collision by Vangi (2014): CMI for the *j*th vehicle ($j = 1,2$) is represented by $\mathrm{CMI} = \Delta V/V_{\mathrm{rPDOF}}$, where V_{rPDOF} is the component along the PDOF of the closing velocity between the two vehicles at the collision instant. To compute the value of V_{rPDOF}, it must be mentioned that the following vector equality must be fulfilled whichever the impact configuration between the vehicles (Vangi, 2020) (vector quantities are highlighted in bold):

$$\overline{\boldsymbol{V}}_{\boldsymbol{r}} - \boldsymbol{V}_{\boldsymbol{r}} = \boldsymbol{\Delta V_1} - \boldsymbol{\Delta V_2} \tag{1}$$

$\overline{\boldsymbol{V}}_{\boldsymbol{r}}$ in Eq. (1) represents the post-impact closing velocity between the two vehicles; in case no sliding between the contact surfaces occurs, the vehicles move at the same velocity at the instant of disengagement ($\overline{\boldsymbol{V}}_{\boldsymbol{r}} = 0$): since the two vectors $\boldsymbol{\Delta V}_1$ and $\boldsymbol{\Delta V}_2$ have the same direction which corresponds to the PDOF (because of momentum conservation), $\boldsymbol{V}_r$ lies along the PDOF too; it can be derived that V_{rPDOF} is equivalent to V_r and CMI for the *j*-th vehicle can be expressed as in Eq. (2):

$$\mathrm{CMI}_j = \frac{\Delta V_j}{V_r} \tag{2}$$

Table 1
General overview of the dataset.

MAIS 3+	Occupant age	Vehicle category	Registration year	Impact side	ΔV (km/h)	V_r (km/h)	CMI
Minimum: 0.000	Minimum: 2.00	4-door sedan, hardtop: 2064	Minimum: 1974	Back Side Impact: 234	Minimum: 6.00	Minimum: 20.04	Minimum: 0.1930
1st Quartile: 0.000	1st Quartile: 21.00	Hatchback: 196	1st Quartile: 1999	Far Side Impact: 581	1st Quartile: 17.00	1st Quartile: 41.86	1st Quartile: 0.3855
Median: 0.000	Median: 33.00	Pickup: 417	Median: 2003	Front Impact: 2600	Median: 23.00	Median: 54.23	Median: 0.4402
Average: 0.107	Average: 37.11	Station Wagon: 118	Average: 2002	Near Side Impact: 671	Average: 26.02	Average: 58.43	Average: 0.4407
3rd Quartile: 0.000	3rd Quartile: 51.00	Suv: 972	3rd Quartile: 2006		3rd Quartile: 32.00	3rd Quartile: 69.69	3rd Quartile: 0.4931
Maximum: 1.000	Maximum: 96.00	Van: 319	Maximum: 2016		Maximum: 97.00	Maximum: 215.44	Maximum: 0.6915

Since ΔV_j is typically available from in-depth accident databases, the value of V_r must be first estimated for the prospective calculation of CMI by Eq. (2). Starting from the effective mass m_1 and m_2 of vehicle 1 and vehicle 2 respectively (comprising the mass of the occupants) and from the energy E_{a1} and E_{a2} they lost in the impact, V_r can be obtained by Eq. (3):

$$V_r = \sqrt{\frac{2\cdot(E_{a1}+E_{a2})\cdot(m_1+m_2)}{m_1\cdot m_2\cdot(1-\varepsilon^2)}} \tag{3}$$

ϵ in Eq. (3) represents the restitution coefficient referred to the POI, which can be in turn estimated starting from the V_r value employing the following law proposed by Antonetti (1998):

$$\varepsilon = 0.5992\cdot exp\left(-0.2508\cdot V_r + 0.01934\cdot V_r^2 - 0.001279\cdot V_r^3\right) \tag{4}$$

Because E_{a1}, E_{a2}, m_1 and m_2 are typically available for collisions collected in in-depth accident databases, V_r value can be retrieved by iteratively solving the system constituted of Eqs. (3) and (4). CMI for the two vehicles can be thus obtained from an *a* posteriori analysis of the accident; it must be noted that, starting from the formulation in Eq. (4), ϵ becomes negligible (0.1) for values of V_r higher than 35 km/h.

For what regards the *a* priori formulation of CMI, let us refer to the planar visualization in Fig. 1 of an impact whose eccentricity is indicated by the arms of forces h_1 and h_2 for the two vehicles. CMI for the vehicles can be expressed in the *a* priori formulation as in Eq. (5) (Vangi, 2014):

$$\mathrm{C\,M}I_1 = \frac{\gamma_1\cdot\gamma_2\cdot(1+\varepsilon)}{\gamma_2+\gamma_1\cdot m_1/m_2}, \quad \mathrm{C\,M}I_2 = \frac{\gamma_1\cdot\gamma_2\cdot(1+\varepsilon)}{\gamma_1+\gamma_2\cdot m_2/m_1} \tag{5}$$

In Eq. (5), $\gamma_j = k_j^2/(k_j^2+h_j^2)$ where k_j and h_j respectively represent the radius of gyration and the arm of forces for the *j*-th vehicle, while ϵ can be derived referring to Eq. (4) (based on the *a* priori parameter V_r). Starting from Eqs. (2)–(5), ΔV is no more an *a* posteriori parameter but can be foreseen for both vehicles once CMI and V_r are known (i.e., starting from information related to impact configuration, inertial properties of the vehicles and their speed).

As reported in Eq. (5) and shown in Fig. 1, CMI is directly linked to the arms of the forces and to impact eccentricity. As the arms of the forces are dependent on PDOF and POI, it can be argued that the same information on eccentricity may be obtained from the CDC typically reported in in-depth accident databases: the CDC, additionally to synthesizing information regarding the intrusion area, provides the categorization of PDOF into 12 integer values according to a visualization that reflects the hours on a clock. Unlike CMI, nevertheless, the CDC is subject to some limitations that prevent its application to *a* priori analyses on the injury outcome associated with a collision:

- PDOF alone does not provide any indication on the ΔV that the vehicle experiences, unlike CMI when jointly employed with V_r (as a consequence of Eq. (2));
- the information relating to the POI, on which determination of the arms depends, is not easily available *a* posteriori by the CDC except by approximate methods: since the CDC provides categorized indications on the location and extent of the intrusion area, the POI may correspond to any point inside such region; the arms of the forces are thus ambiguously defined and known only *a* posteriori once the intrusion area has been identified. This ambiguity is also amplified by the categorization of the PDOF: while the CDC field dedicated to PDOF involves an inherent uncertainty of $\pm 15^\circ$ on the real direction of the impact forces, CMI is a continuous variable that can define all intermediate degrees of eccentricity.

It is interesting to note that, for what regards derivation of IR models, the CDC is subject to a further limitation if compared to CMI: by not containing ΔV-related information, differently from CMI, the CDC must necessarily be used jointly with ΔV itself for the definition of high-quality IR models. As a consequence, contribution to eccentricity which can be attributed to the CDC mixes with the one intrinsically contained in ΔV (and represented by CMI): this superposition of effects resulting from the joint use of CDC and ΔV does not allow to completely isolate the individual contribution of eccentricity and analyze its effect on IR.

2.2. Dataset

To determine the effect of V_r and CMI (among several other variables) on the risk to sustain serious injuries, the study focuses on accidents occurred in the 2004–2015 period which are collected in the National Automotive Sampling System – Crashworthiness Data System (NASS/CDS[2]). The NASS/CDS in-depth accident database is enriched annually with information about almost 5000 traffic collisions, which took place in different areas (known as primary sampling units) of the US territory; the necessary condition for cataloging the accident within the NASS/CDS database is that at least one vehicle is towed away from the accident site. In addition, among all accidents in the US, only a few are selected according to a scheme which aims at making data representative of the national statistics. By providing freely available information on occupant and vehicle characteristics, impact-related variables and the resulting injury outcome, the NASS/CDS database plays a leading role in accident investigations.

To obtain a dataset which is suitable for subsequent processing, only vehicle-to-vehicle impacts are extracted from the NASS/CDS database; furthermore, given the need to estimate V_r and consequently CMI, all collisions for which the terms in Eqs. (2)–(5) are not available have been excluded. Rollover events are also neglected since different dynamics rule over the injury mechanism for this type of impact (Atkinson et al., 2016). Conversely, in line with the state-of-the-art (Forman and

[2] https://www.nhtsa.gov/research-data/national-automotive-sampling-system-nass.

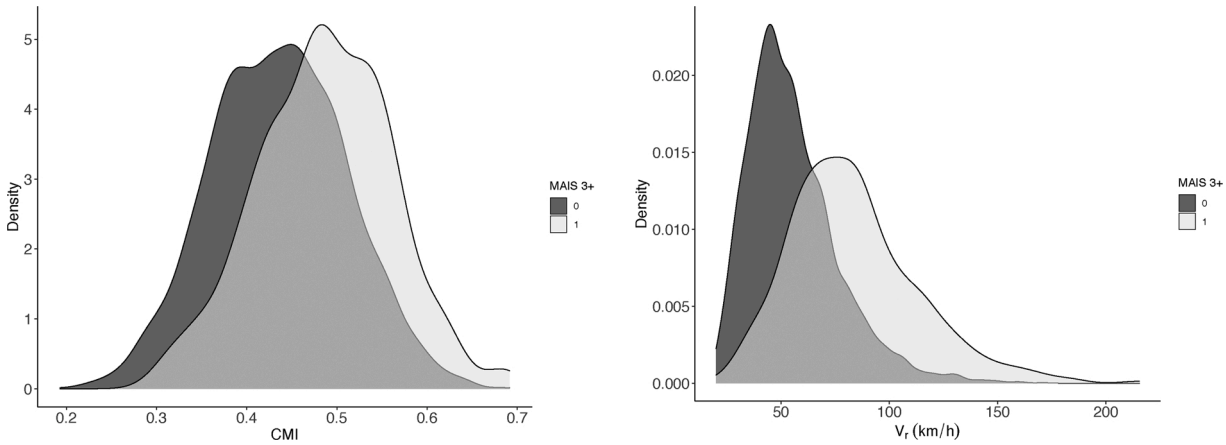

Fig. 2. Density plots of CMI (left) and V_r (right) variables, grouped by MAIS 3+ outcome.

McMurry, 2018), only the events in which belt use is reported for the vehicles' occupants are considered: the superior performance offered by seat belts in reducing serious injuries is nowadays an established concept (Høye, 2016; Boakye et al., 2019; Imler et al., 2010). The remaining data, which are susceptible to cataloging errors (Schlögl and Stütz, 2017) or internal incoherence (Vangi et al., 2019a) were further filtered to increase the reliability of results; referring to the guidelines prescribed by Vangi et al. (2019a), only the events fulfilling the following criterion are preserved:

$$\frac{m_{1m}}{m_{2M}} \leq \frac{\Delta V_2}{\Delta V_1} \leq \frac{m_{1M}}{m_{2m}}. \tag{6}$$

The quantities m_{1m} and m_{2m} in Eq. (6) are the minimum possible masses for the vehicles based on their values collected in the database (curb weight to which the tank capacity is subtracted); m_{1M} and m_{2M} are the maximum possible masses for the vehicles (curb weight to which are added the known mass of occupants and a mass of 100 kg for each occupant whose mass is unknown). This check allows excluding from the analysis all those cases for which the momentum conservation law is not fulfilled with sufficient accuracy; for more details, refer to previous work (Vangi et al., 2019a).

As an overview, in Table 1 are reported the empirical moments corresponding to the variables involved in the analysis; moments are evaluated at the original scale before standardization, required to avoid scale effects. The total number of considered accidents is 1452 for a total of 4086 occupants, each of them reporting – response "1" – or not – response "0" – a MAIS 3+ injury. Based on the data provided by the National Highway Traffic Safety Administration (NHTSA) regarding national statistics and weighting criteria for NASS/CDS cases,[3] it is derived that this set of 1452 accidents represents an actual sample of 5,122,110 vehicle-to-vehicle crashes. As shown in Table 1, the "occupant age" variable is treated like a continuous variable rather than a categorical variable, since a categorization would cause a decrease in the prediction quality (Newgard, 2008). The "vehicle category" has been also considered, because different body frame types can be expected to provide diverse levels of passive protection; from this standpoint, observations associated with coupé vehicles have been excluded as they are limited in number (six occurrences). "Registration year" of the vehicle is also accounted for, as a preliminary indicator of its crashworthiness. The impact-related variable "impact side" is determined based on the intrusion area only for frontal and rear-end impacts; in case the lateral area is involved, the impact is classified as "near side" if intrusion occurs on the same side of the occupant and "far side" otherwise (Sobhani et al., 2011). For convenience, the ΔV variable is also

[3] DOT HS 811 807 – NASS-CDS: Sample Design and Weights.

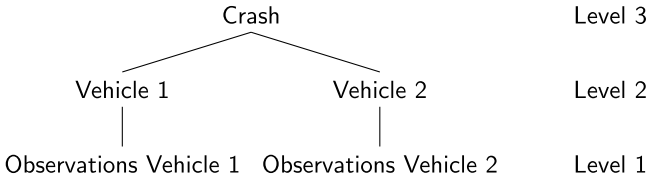

Fig. 3. 3-level hierarchy data structure.

highlighted in Table 1; nevertheless, only the V_r and CMI components of ΔV are considered in the described analysis. Additional insights regarding the dataset are provided in Fig. 2: the distributions of CMI (left) and V_r (right), fundamental elements of the present analysis, are plotted based on the associated value of MAIS 3+ outcome. To increase readability, probability densities are reported rather than frequencies: referring to Table 1, the dataset is unbalanced (about 11% of positive MAIS 3+ examples), with a relevant discrepancy between the number of "0" and "1" outcomes. Despite its simplicity, the visualization of Fig. 2 provides valuable indications, since it preliminarily denotes that serious injuries tend to occur at high values of both V_r and CMI. For more details on data distribution taking each considered variable as a discriminant, refer to Appendix A.

Although the level of unbalancing is not high enough to be in a situation of rare events, such a phenomenon must be taken into account to build a correct predictive model. Additionally, the dataset features a three-level hierarchical structure, depicted in Fig. 3: the instances of the dataset are nested within vehicles and, in turn, vehicles are nested within accidents.

2.3. Theory

In the present section, the main concepts and ideas behind the methodologies involved in the analysis are outlined. First, the ReliefF algorithm (Kononenko, 1994) for feature ranking is highlighted. Second, the statistical (logistic and mixed effects logistic) and machine learning (SVM and random forest) models involved in the analysis are presented; in particular, emphasis is given to the way these models deal with the class unbalance problem. For each described model, the corresponding hyperparameters are reported (i.e., parameters that must be set before the learning process begins). It is essentially possible to distinguish between two different types of hyperparameters: those which deal with the problem of unbalancing (weights) and the ones focused on the improvement of prediction ability.

2.3.1. ReliefF

The use of ReliefF is unprecedented in the field of accident investigation. ReliefF algorithm belongs to the class of Relief algorithms

(Robnik-Šikonja and Kononenko, 2003; Urbanowicz et al., 2018) and is probably one of its most famous variants. ReliefF algorithm is model-independent and employed before the fitting of whichever classical supervised learning predictive models (e.g., CART, logistic regression, SVM, neural network) to improve their predictive performance: ReliefF filters the relevant pieces of information in a dataset, ranking each variable by importance in the determination of the outcome; such ranking is supportive in formulating appropriate, parsimonious, and effective predictive models. Although it is typically employed in classification problems, a version of the algorithm exists in the context of regression problems. In the case of binary classification (as for a MAIS 3+ outcome), given an instance R_i of the dataset, two different sets of instances (nearest hits and nearest misses) are created: a first one which contains k neighbors of R_i belonging to the same class of R_i, and a second one which collects k neighbors of R_i belonging to the other class. The main concept of the algorithm is to penalize the variables that tend to have different values in the nearest hit set and to reward the variables that tend to differ in the nearest miss set. In Algorithm 1, the original formulation proposed by Kononenko (1994) is reported.

Algorithm 1. ReliefF

Input: for each instance a vector of attribute values and the class value
Output: the vector $W \in \mathbb{R}^{a+1}$ of estimations of the qualities of attributes
Set all weights $W[A] := 0.0$;
for $i = 0$ **to** m **do**
  randomly select an instance R_i;
  find k nearest hits H_j;
  for $C \neq class(R_i)$ **do**
    find k nearest misses $M_j(C)$;
  for $A = 0$ **to** a **do**
    $$W[A] := W[A] - \sum_{j=1}^{k} \frac{\mathrm{diff}(A,R_i,H_j)}{mk} + \sum_{C \neq \mathrm{class}(R_i)} \frac{\left[\frac{P(C)}{1-P(\mathrm{class}(R_i))} \sum_{j=1}^{k} \mathrm{diff}(A,R_i,M_j(C))\right]}{mk}$$

The outer iteration in Algorithm 1 is repeated m times, where m represents the total number of processed instances employed to update the weight vector W. Algorithm 1 also depends on the hyperparameter k, i.e., the number of neighbors from the same and opposite class that contribute to update the vector of weights W. Given a variable A, its quality estimation is updated depending on the current instance R_i and the values of variable A of k nearest hits and k nearest misses $M_j(C)$. Each term of the summation over classes is weighted by the factor $\frac{P(C)}{1-P(\mathrm{class}(R_i))}$, where $P(C)$ is the prior probability of class C and $1 - P(\mathrm{class}(R_i))$ is the sum of probabilities of the misses classes. Specifically, the computation of $W[A]$ depends on the function $\mathrm{diff}(A, I_1, I_2)$, where I_1 and I_2 are two general instances. For categorical and continuous variables, $\mathrm{diff}(A, I_1, I_2)$ is calculated as in Eq. (7) and in Eq. (8), respectively:

$$\mathrm{diff}(A, I_1, I_2) = \begin{cases} 0, & \text{if } \mathrm{value}(A, I_1) = \mathrm{value}(A, I_2) \\ 1, & \text{otherwise.} \end{cases} \tag{7}$$

$$\mathrm{diff}(A, I_1, I_2) = \frac{|\mathrm{value}(A, I_1) - \mathrm{value}(A, I_2)|}{max(A) - min(A)} \tag{8}$$

2.3.2. Logistic regression

Logistic regression is a widely used statistical model belonging to the family of Generalized Linear Model (GLM) (Agresti and Kateri, 2011; Hosmer et al., 2013; McCullagh, 2019); it is undoubtedly the most recognized tool in the literature to deal with IR estimation problems, also thanks to its computational simplicity and ease of use (an extremely concise list of examples is represented by the works by Andricevic et al. (2018) regarding frontal collisions, Nishimoto et al. (2019) on pedestrians, and Bareiss et al. (2018) on near side crashes). In logistic regression, a single outcome random variable Y_i $(i = 1, \ldots, n)$, whose realized values are indicated by y_i, follows a Bernoulli probability function that takes value 1 with probability π_i and 0 with probability $1 - \pi_i$. These probabilities π_i vary over observations as a sigmoid function of a linear predictor $\eta_i = x_i^t \beta$, formally a scalar product between the vector of features $x_i = (1, x_{i1}, \ldots, x_{ik})^\top$ referring to the ith observation and a vector of coefficients $\beta = (\beta_0, \beta_1, \ldots, \beta_k)^\top$ to be estimated from the set of data $\{x_i, y_i\}^n$:

$$Y_i | x_i \sim \mathrm{Bernoulli}(\pi_i), \tag{9}$$

$$\pi_i = \frac{1}{1 + exp(-\eta_i)}, \tag{10}$$

$$\eta_i = ln\left(\frac{\pi_i}{1 - \pi_i}\right) = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik}. \tag{11}$$

Coefficients are usually estimated by maximizing the likelihood function $L(\beta)$. Assuming independence over the observations, the likelihood can be written as follows:

$$L(\beta) = \prod_{i=1}^{n} \pi_i^{y_i} (1 - \pi_i)^{(1-y_i)}. \tag{12}$$

The maximum likelihood criterion is based on the maximization of the likelihood function. The log-likelihood (rather than the likelihood itself) is however typically maximized, being less inclined to numerical underflow. The maximization of the logistic log-likelihood is straightforward, since it can be solved as a convex optimization problem (Boyd and Vandenberghe, 2004). In prediction-focused analyses, a regularized version of the logistic regression, known as the ℓ_2 regularized logistic regression, is often preferred. The regularization term, in this case the

sum of the squared of the coefficients, can be used to train models that generalize better on unseen data, allowing to prevent overfitting issues. Coefficients are estimated by optimizing the objective function in Eq. (13):

$$\mathcal{L}(\beta) = \sum_{i=1}^{n} w_{+} y_i log(\pi_i) + w_{-}(1-y_i) log(1-\pi_i) + c \parallel \beta \parallel_2^2 \tag{13}$$

In Eq. (13), w_{+}, w_{-} are pre-fixed weights assigned respectively to positive and negative examples, inversely proportional to class frequencies to deal with the unbalancing problem; conversely, c is the hyper-parameter that assesses the strength of the ℓ_2 regularization term $\parallel \beta \parallel_2^2$ and contributes to improve the prediction ability of the logistic model. Its value is set during the so-called hyperparameter optimization phase. From an optimization point of view, the addition of the regularization term does not change the convexity of the problem. Once the vector of coefficients β has been estimated from the training data, it is possible to compute the associated conditional probability $\Pr(Y = 1|x)$ for a generic new input x by means of the sigmoid function (Eq. (10)): by this function, the binary predictions can be straightforwardly derived.

2.3.3. Mixed effects logistic regression

Mixed effects logistic regression (also known as multilevel logistic or random parameters logistic) belongs to the class of Generalized Linear Mixed Model (GLMM) (Goldstein, 2011) and provides an extension of the classical logistic regression to include both fixed and random effects. There is no unanimity in defining precisely what a random effect is (Gelman et al., 2005). Bolker et al. (2009) define random effects as factors whose levels are sampled from a larger population, or whose interest lies in the variation among them rather than the specific effects of each level. Statistical units which share these same levels are interdependent and the presence of the random effects is necessary to take care of the correlation between observations. For this reason, the scenario in which these models are employed is the one of a population with a prior known clustered structure. When it is possible to associate a hierarchy with these clusters or groups, such models are also indicated as multilevel models. This hierarchical structure is often encountered in accident analysis applications, where the unknown relationship among the observations in a single crash is referred to as unobserved heterogeneity (Fountas and Anastasopoulos, 2017; Mannering et al., 2016; Fountas et al., 2019; Behnood et al., 2014). The mixed effects logistic regression allows weakening the heterogeneity bias on the observation by introducing the random effects inside the classical logistic regression model; this is the reason why the mixed effects logistic regression is nowadays becoming a widespread tool in accident-related research (Alogaili and Mannering, 2020; Fountas et al., 2018; Truong et al., 2020).

Random intercept logistic regression is a peculiar type of mixed effects logistic regression: it includes the presence of a random intercept term $u_j \sim \mathbb{N}(0, \ \sigma^2)$ for each j-th cluster constituting the second level of the population, composed of a total of n observations allocated in J clusters. The random intercept term u_j represents the contribution of belonging to the j-th cluster and the unobserved heterogeneity. Differently from the classical logistic regression, probabilities π_{ij} vary over observations and clusters as a sigmoid function of a linear predictor η_{ij} which includes the random intercept term:

$$u_j \sim \mathbb{N}(0, \sigma^2), \tag{14}$$

$$Y_{ij}|x_{ij}, u_j \sim \text{Bernoulli}(\pi_{ij}), \tag{15}$$

$$\pi_{ij} = \frac{1}{1 + exp(-\eta_{ij})}, \tag{16}$$

$$\eta_{ij} = ln\left(\frac{\pi_{ij}}{1-\pi_{ij}}\right) = \beta_0 + \beta_1 x_{i1}^{(j)} + \cdots + \beta_k x_{ik}^{(j)} + u_j. \tag{17}$$

Differently from the classical logistic regression, it is substantial to note that the distribution of the outcome random variable Y_{ij} is conditioned to the random effect u_j, which in turn comes from a zero-mean normal distribution whose variance parameter σ^2 is the level 2 (residual) variance, or the between-group variance. The aim of the inference for the random intercept logistic model involves both the estimation of the vector of parameters $\beta = (\beta_0, \ldots, \beta_k)^\top$ and the variance parameter σ^2 as well as the prediction of the random effects u_j. Inference is based on the maximum likelihood estimation criterion; however the likelihood function does not have a closed form and therefore needs to be approximated. Although different likelihood approximations exist, the one based on the Adaptive Gaussian Quadrature (AGQ) method seems to be efficient and widespread (Pinheiro and Chao, 2006).

For what concerns the prediction task, caution is suggested when prediction is required for a unit in a new cluster (Skrondal and Rabe-Hesketh, 2009). The prediction must be evaluated according to the population parameters $\beta^{(pa)} = (\beta_0^{(pa)}, \ldots, \beta_k^{(pa)})^\top$, representing the unconditional effects of the variables, differently from the parameters $\beta = (\beta_0, \ldots, \beta_k)^\top$ which represent effects of the variables controlling for, or holding constant, the random effects. Hedeker et al. (2018) computed the population parameters $\beta^{(pa)}$ computed from the marginal probabilities, which in turn are calculated using Monte Carlo integration over the random effects.

2.3.4. Support Vector Machines

The Support Vector Machine (SVM) model (Cortes and Vapnik, 1995; Boser et al., 1992) is widely used as a simple and efficient tool for binary linear and nonlinear classification. In the accident research field, SVM is primarily devoted to the derivation of crash risk models (Wang et al., 2019; Iranitalab and Khattak, 2017). The idea behind SVM is to construct the maximum-margin hyperplane, i.e., the hyperplane that has the maximum distance between data points of both classes. The main advantage of SVM if compared to the logistic regression is given by the possibility to perform nonlinear classification through the introduction of a kernel function (Rasmussen, 2003). Given two generic instances x_i, x_j, the kernel function $k_e(x_i, x_j)$ provides a measure of similarity between the two instances. Training the model on a dataset of n independent samples $\{x_i, y_i\}^n$, it is possible to obtain a binary prediction function $f(x)$ for a generic new input x. SVM predicts the class of the instance x depending on sign $(f(x))$. Two equivalent prediction functions are involved within the SVM linear framework: the primal linear SVM prediction function and the dual linear SVM prediction function; the latter can be expressed as follows:

$$f(x) = \sum_{i=1}^{n} w_i (x^\top x_i) + \beta_0, \tag{18}$$

$w = (w_1, \ldots, w_n)^\top$ in Eq. (18) is a vector of weights obtained by solving a convex linearly constrained optimization problem. Eq. (18) refers to the prediction function derived from the dual formulation of the SVM training problem. The prediction function derived from the primal formulation of the SVM problem is equivalent to Eq. (18). In the primal case, the prediction function $f(x)$ for a generic input x has the following

form:

$$f(x) = x^{\top}\beta + \beta_0. \tag{19}$$

Nonlinear classification is performed by replacing the dot products in Eq. (18) with the kernel evaluations $k(x, x_i)$. The prediction function used by the nonlinear SVM depends on the choice of kernel function and is given by Eq. (20):

$$f(x) = \sum_{i=1}^{n} w_i k_e(x, x_i). \tag{20}$$

Although a high number of kernel functions exists which can be potentially employed, the most commonly used kernel function is the Gaussian one (Goodfellow et al., 2016).

Weighted SVM (Pedregosa et al., 2011; Boughorbel et al., 2017) represents an extension of the classical linear and nonlinear SVM framework to handle the unbalancing of data. Specifically, this is possible by weighting the minority and the majority class instances differently. Each instance of a given class has an associated cost, computed as the product of the soft-margin hyperparameter C and the importance of the corresponding class (i.e., a weight inversely proportional to class frequency). This formulation can be used both in the linear case and the non-linear one. It is important to note that, when performing non-linear SVM, additional hyperparameters must be set. In the present analysis, the `scikit-learn` library (Pedregosa et al., 2011) has been employed to fit the linear SVM in the dual formulation and the non-linear SVM models; both models solve the unbalancing problem by differently weighting the instances.

2.3.5. Random forests

Random forests employment in the accident-related literature is typically reported in crash risk analyses (Iranitalab and Khattak, 2017) and studies on IR (Kusano and Gabler, 2014). A random forest is a classifier consisting of a collection of decision tree classifiers (Breiman, 2001). The goal of each tree is to create a model that predicts the value of a target variable by learning simple decision rules inferred from the data features. During the training phase, a number of decision trees classifiers are fitted independently on various sub-samples of the dataset. When the random forest is asked to predict the class of a generic input x, the prediction function $f(x)$ is calculated by the majority vote from the predicted classes by each tree in the forest. The prediction function is defined as in Eq. (21), where $f_t(x)$ is the predicted class by the t-th tree of the forest (Hastie et al., 2009):

$$f(x) = \text{majority vote}\{f_t(x)\}_1^{\top} \tag{21}$$

In the presence of unbalanced datasets, a random forest classifier tends to be biased on the majority class. In order to alleviate this problem, two main approaches are available: balanced and weighted random forests. The first one consists in using a stratified bootstrap approach to generate each sub-sample; the latter consists in assigning a weight to each class, with the larger weight given to the minority class. These weights are used directly in the criterion for finding splits of the tree and the final prediction of each terminal node is given by a weighted majority vote. This second approach is the one employed to deal with the unbalancing problem, whereas the Gini index represents the criterion for splitting.[4]

Based on these considerations, it can be asserted that the key difference between the random forest and SVM classifiers practically lies in the way positive and negative cases are separated: the random forest predictive model is represented by a piecewise-defined function, representing a partitioning of the space, for which the number of splitting rules derived from the modeling (i.e., the leaves of the resulting decision tree) defines the width of the continuity intervals; this implies that IR does not change if the value of a feature remains in a specific range, the value of the remaining features being the same. The model resulting from SVM is conversely a hyperplane, which can best separate MAIS 3+ injuries from the others based on the values of the involved features.

[4] Established guidelines are presented in the file available at https://statistics.berkeley.edu/sites/default/files/tech-reports/666.pdf, Sections 2.2 and 2.3.

2.4. Evaluation metrics

In general, the choice of an appropriate evaluation metric for the predictive performance of a classification model is not trivial. The accuracy metric, in the presence of unbalancing, suffers of an evident drawback: for a dataset that has 10% positive class and 90% negative class, a Naïve classifier that always outputs the majority class label will have a high accuracy of 90% (Boughorbel et al., 2017). AUC-ROC represents the most popular predictive metric and is commonly used in situations of unbalanced data. Boughorbel et al. (2017) report that AUC-ROC is robust to data unbalancing; its main limitation is that there is no explicit closed formula to compute it. Another evaluation metric, also suitable in situations of unbalanced classes, is given by the Matthews Correlation Coefficient (MCC), frequently used in biomedical research. As explained by Boughorbel et al. (2017), in addition to being robust to data unbalance, this metric has the advantage of being calculable in closed form if compared to the AUC-ROC metric. MCC is defined as in Eq. (22):

$$\text{MCC} = \frac{\text{TP} \times \text{TN} - \text{FP} \times \text{FN}}{\sqrt{(\text{TP}+\text{FN})(\text{TP}+\text{FP})(\text{TN}+\text{FN})(\text{TN}+\text{FN})}}, \tag{22}$$

TP, TN, FN, FP in Eq. (22) are respectively the number of True Positives (predicted “1”, observed “1”), True Negatives (predicted “0”, observed “0”), False Negatives (predicted “0”, observed “1”) and False Positives (predicted “1”, observed “0”), where “1” represents occurrence of a MAIS 3+ injury and “0” otherwise. Following the definition provided in Eq. (22), MCC for a model varies between -1 (total disagreement between prediction and observation) and $+1$ (perfect prediction); 0 indicates no better than random prediction. Considering the lack of a balanced dataset, evaluation on the fitted models is performed by means of both the AUC-ROC and MCC metrics.

3. Results

The objective of the present section is to thoroughly highlight the influence of V_r and CMI on the MAIS 3+ outcome of a crash. First, the ReliefF algorithm is run to objectively rank the variables based on their importance in the prediction task: this solution allows determining the importance of V_r and CMI if compared to the remaining considered variables, i.e., vehicle-related, occupant-related and impact-related variables. Afterwards, predictive models provide complementary and more complete information with respect to the ReliefF results. To enrich the analysis from a statistical standpoint, results regarding significance of the considered variables are reported in Appendix B.

Table 2
Ranking results for 100 runs of ReliefF.

Rank	Variable	Average score	Standard deviation
#1	V_r	0.043	0.002
#2	Impact side	0.027	0.008
#3	CMI	0.019	0.003
#4	Occupant age	0.014	0.003
#5	Registration year	0.008	0.002
#6	Vehicle category	0.002	0.003

Table 3
Results for logistic regression (numbers in bold characters refer to models with the highest predictive power).

	V_r	Impact side	CMI	Occupant age	Registration year	Vehicle category	AUC-ROC	MCC
M_1	Yes	Yes	Yes	Yes	Yes	Yes	0.878	0.425
M_2	Yes	Yes	No	Yes	Yes	Yes	0.869	0.422
M_3	Yes	Yes	Yes	Yes	Yes	No	**0.888**	**0.455**
M_4	Yes	Yes	No	Yes	Yes	No	0.871	0.435
M_5	Yes	Yes	Yes	Yes	No	No	0.882	0.464
M_6	Yes	Yes	No	Yes	No	No	0.864	0.423
M_7	Yes	Yes	Yes	No	No	No	0.867	0.398
M_8	Yes	Yes	No	No	No	No	0.846	0.427

3.1. ReliefF

The performances of ReliefF can be affected by the unbalancing of the dataset (Xie et al., 2017) (about 11% of the observations is labeled as MAIS 3+, referring to Table 1). To overcome such an issue, random undersampling technique is employed: instances are randomly removed from the majority set until the desired balance is achieved, while all of the data points from the minority class are used (Dubey et al., 2014; He and Garcia, 2008). Repeating 100 times random undersampling from the original dataset (Liu et al., 2008), 100 balanced datasets are generated. For each of them, the ReliefF algorithm is run (Algorithm 2).

Algorithm 2. Repeated ReliefF

Input: the unbalanced dataset;
Output: the mean vector $\bar{W} \in \mathbb{R}^{a+1}$ of the scores and the corresponding standard deviation vector $\bar{S} \in \mathbb{R}^{a+1}$;
for $i = 0$ **to** T **do**
  generate a balanced dataset D_i by the undersampling technique;
  run ReliefF on D_i and obtain the vector $W_i \in \mathbb{R}^{a+1}$;
Compute $\bar{W}$ and $\bar{S}$;

Following Algorithm 2, the average and the standard deviation values of the prediction scores have been derived for each variable of the dataset. Table 2 reports the results obtained from Algorithm 2 and variables are ordered according to their average score of predictive importance. In terms of average score solely, V_r results as the variable which primarily affects the injury outcome: the higher the closing velocity, the higher the potential acceleration experienced by the occupant. CMI ranks in third position, thus providing a first evidence of eccentricity influence on the risk of severe injury. Further, it must be noted that all standard deviations of the quality predictive scores are between 0.002 and 0.003, except for the one related to the "impact side" variable (0.008): differing from the remaining ranking positions, the second and third ones are not clearly identified, with a possible higher importance of CMI if compared to "impact side". In the end, the ranking in Table 2 clearly suggests that impact-related variables (first three positions) are more influential on the MAIS 3+ injury outcome with respect to occupant-related (fourth position) and vehicle-related variables (last two positions).

Since Table 2 reflects contribution of each variable to the injury outcome, the results are particularly relevant for IR reduction purposes. From the obtained feature ranking, it can be asserted that the most appropriate intervention to be applied consists in decreasing the closing velocity at collision: this can be achieved, for instance, by lowering the speed limit of the roads or by introducing devices as the Autonomous Emergency Braking (AEB) on the circulating fleet. Modifications to impact side should be subsequently proposed to allow for a lower IR; nevertheless, limited actions can be performed to vary such an impact-related parameter before the collision, since it also depends on occupants' position inside the compartment. Third, more eccentric impacts should be sought through changes to both the infrastructure and the vehicles: eccentricity can be increased by introducing roundabouts at intersections, as well as allowing ADAS devices to modify the vehicle steering angle before the collision. Conversely, the remaining variables (occupant age, registration year and vehicle category) are set, and cannot be thence directly modified to reduce impact severity. Overall, among numerous pre-crash variables, only closing velocity and eccentricity (CMI) can be efficiently employed to decrease IR associated with potential collisions. For a proper quantification of the benefits which can be obtained by specific changes to pre-crash conditions, appropriate IR models must be derived (Section 3.2) and employed.

3.2. Prediction task

Two main parallel objectives are sought in the present section: (1) to obtain a powerful and parsimonious predictive model which can be advantageously employed to perform IR estimates in a pre-crash scenario; (2) to quantitatively assess the effect of the CMI variable on IR by the difference in terms of prediction scores between models which include CMI as a predictor and models which neglect CMI. To this regard, the predictive performances of statistical and machine learning models are compared. All the models involved in the present analysis are trained using only predictors which can be evaluated before the crash; the predictive models are useful to assess the potential benefits of strategies which aim at changing pre-crash conditions, in terms of IR reduction (Section 3.1). The analysis is focused on eight sets of variables, i.e., different subsets of the predictors to each of which four different machine learning and statistical models (Section 2.3) have been fitted:

- $M_1 = \{V_r, \text{ImpactSide}, \text{CMI}, \text{Age}, \text{Year}, \text{Type}\}$
- $M_2 = \{V_r, \text{ImpactSide}, \text{Age}, \text{Year}, \text{Type}\}$
- $M_3 = \{V_r, \text{ImpactSide}, \text{CMI}, \text{Age}, \text{Year}\}$
- $M_4 = \{V_r, \text{ImpactSide}, \text{Age}, \text{Year}\}$
- $M_5 = \{V_r, \text{ImpactSide}, \text{CMI}, \text{Year}\}$
- $M_6 = \{V_r, \text{ImpactSide}, \text{Year}\}$
- $M_7 = \{V_r, \text{ImpactSide}, \text{CMI}\}$
- $M_8 = \{V_r, \text{ImpactSide}\}$

The set M_1 corresponds to the case in which all the variables are considered. The remaining sets consist of a gradually reduced number of

Table 4
Results for random intercept logistic (numbers in bold characters refer to models with the highest predictive power).

	V_r	Impact side	CMI	Occupant age	Registration year	Vehicle category	AUC-ROC	MCC
M_1	Yes	Yes	Yes	Yes	Yes	Yes	0.881	0.41
M_2	Yes	Yes	No	Yes	Yes	Yes	0.86	0.398
M_3	Yes	Yes	Yes	Yes	Yes	No	**0.884**	**0.404**
M_4	Yes	Yes	No	Yes	Yes	No	0.863	0.385
M_5	Yes	Yes	Yes	Yes	No	No	**0.884**	0.379
M_6	Yes	Yes	No	Yes	No	No	0.858	0.278
M_7	Yes	Yes	Yes	No	No	No	0.866	0.349
M_8	Yes	Yes	No	No	No	No	0.839	0.266

Table 5
Results for SVM classifier with Gaussian kernel (numbers in bold characters refer to models with the highest predictive power).

	V_r	Impact side	CMI	Occupant age	Registration year	Vehicle category	AUC-ROC	MCC
M_1	Yes	Yes	Yes	Yes	Yes	Yes	0.887	0.462
M_2	Yes	Yes	No	Yes	Yes	Yes	0.865	0.428
M_3	Yes	Yes	Yes	Yes	Yes	No	**0.891**	**0.471**
M_4	Yes	Yes	No	Yes	Yes	No	0.873	0.444
M_5	Yes	Yes	Yes	Yes	No	No	0.884	0.471
M_6	Yes	Yes	No	Yes	No	No	0.866	0.423
M_7	Yes	Yes	Yes	No	No	No	0.869	0.409
M_8	Yes	Yes	No	No	No	No	0.845	0.397

Table 6
Results for SVM classifier with Linear kernel (numbers in bold characters refer to models with the highest predictive power).

	V_r	Impact side	CMI	Occupant age	Registration year	Vehicle category	AUC-ROC	MCC
M_1	Yes	Yes	Yes	Yes	Yes	Yes	0.877	0.423
M_2	Yes	Yes	No	Yes	Yes	Yes	0.866	0.437
M_3	Yes	Yes	Yes	Yes	Yes	No	**0.883**	**0.451**
M_4	Yes	Yes	No	Yes	Yes	No	0.871	0.431
M_5	Yes	Yes	Yes	Yes	No	No	0.878	0.471
M_6	Yes	Yes	No	Yes	No	No	0.865	0.413
M_7	Yes	Yes	Yes	No	No	No	0.867	0.405
M_8	Yes	Yes	No	No	No	No	0.846	0.404

Table 7
Results for random forest classifier (numbers in bold characters refer to models with the highest predictive power).

	V_r	Impact side	CMI	Occupant age	Registration year	Vehicle category	AUC-ROC	MCC
M_1	Yes	Yes	Yes	Yes	Yes	Yes	0.834	0.268
M_2	Yes	Yes	No	Yes	Yes	Yes	0.815	0.273
M_3	Yes	Yes	Yes	Yes	Yes	No	**0.853**	**0.324**
M_4	Yes	Yes	No	Yes	Yes	No	0.815	0.211
M_5	Yes	Yes	Yes	Yes	No	No	0.778	0.249
M_6	Yes	Yes	No	Yes	No	No	0.741	0.292
M_7	Yes	Yes	Yes	No	No	No	0.749	0.274
M_8	Yes	Yes	No	No	No	No	0.766	0.218

variables, based on the ranking results provided by the ReliefF algorithm. Only the odd indexed sets contain the CMI predictor; in this way, the predictive performances of the fitted models are evaluated both with presence and absence of the CMI variable. This allows to quantify the predictive performances of the CMI predictor varying the models and the subsets of variables. Conversely, V_r cannot be excluded from the models to study its influence on IR: in fact, models not accounting for vehicles' speed would be characterized by limited physical relevance (as demonstrated by the first position of the ReliefF ranking in Table 2).

For what concerns the dataset, about 80% of the crashes with all the associated observations constitutes the full training set, while the remaining 20% represents the testing set. Since the presence of correlation between observations of the dataset is evident, implying the violation of the standard assumption of independent and identically distributed (*i.i.d.*) observations, such a phenomenon must be taken into account during the training phase of the models. In fact, among all the models included in this work, only the mixed logistic regression takes directly account of the presence of correlation between observations. Appropriate log-likelihood ratio tests have been performed, which highlighted the significance of including a random intercept term varying from one vehicle to another vehicle (level 2 variation, referring to Fig. 3). This roughly means that the presence of correlations in the dataset significantly involves observations belonging to the same vehicle. The practical consequence is that during the training phase only the logistic random intercept models have been fitted by means of the full training set of examples, while all the other models have been trained on a subset of *i.i.d.* training examples. The latter is created considering all the crashes in the full training set and uniformly sampling one observation for each crash.

Independently from the model, a grid search strategy has been used to select the appropriate hyperparameters; the best hyperparameter configuration is obtained through a stratified 5-fold cross-validation. Finally, the model corresponding to this optimal choice of the hyperparameters is trained once again on the whole training set.

To compare the predictions provided by the derived IR models, the AUC-ROC and MCC metrics are both employed. The scores of predictions are reported in Tables 3–7 with respect to four different classifiers. Based on these results, several conclusions can be drawn from a predictive standpoint. First, the set of variables which gives the best results in terms of prediction corresponds to M_3, i.e., the set of variables in which CMI is included and "vehicle category" is neglected; since M_3 proves to be the best subset of variables independently of the considered classifier, the influence of CMI on the injury outcome is demonstrated. Nonetheless, comparison between the M_3 and the M_4 subsets allows assessing such an influence: the most evident difference between the subsets is highlighted for the random forest classifier, i.e., about 4% in terms of AUC-ROC and more than 11% considering the MCC metric. For what concerns the remaining classifiers, this difference reaches almost 2% in terms of AUC-ROC and 3% in terms of MCC. As a whole, the analysis highlights that CMI (and eccentricity of the impact as a consequence) is an influential variable on the MAIS 3+ occurrence; nevertheless, the joint use of V_r and "impact side" can be sufficient to obtain a good accuracy model, as evidenced by the high values of AUC-ROC and MCC of M_8. In general terms, the Gaussian SVM outperforms the other classifiers; the predictive performances in terms of AUC-ROC and MCC are however similar to the ones of logistic regression which, among several other classifiers, has proven to be one of the most efficient tool in the injury-related literature (Kusano and Gabler, 2014). Another relevant finding is that the logistic regression exhibits a slightly higher predictive power than the random intercept logistic (Tables 3–4) for the best model (M_3), both in terms of AUC-ROC and MCC; nevertheless, considering all the analyzed models, the results associated with the logistic regression classifier are in line with the ones from the random intercept logistic: despite the capability of the random intercept to compensate for the unobserved heterogeneity bias, the performance of one method over the other depends both on the considered features and the data to be fitted (Fountas et al., 2018).

It is worth noting that results in Tables 3–7 are coherent with the ranking of predictors provided by the ReliefF algorithm, where the CMI ranks in third position. In fact, considering the M_3 and the M_5 subsets in which the "registration year" variable is included and neglected respectively, differences in AUC-ROC and MCC lower than the ones associated with the M_3 and the M_4 models (influence of CMI) is derived; this observation is independent of the considered classifier. The same applies to the M_5 and the M_7 subsets, from which the influence of the "occupant age" variable is obtained. Exclusion of the "vehicle category" variable, conversely, enhances the predictive model with an increase in both the AUC-ROC and the MCC (from M_1 to M_3): this is in line with a low value of average score and the consequent last position in the ReliefF ranking (Table 2). Finally, the analysis evidences the main advantage linked to the use of ReliefF algorithm, i.e., it provides a ranking of variables which is independent of the classifier considered: although the same information can be obtained comparing the results of several classifiers, ReliefF represents an immediate and powerful tool to classify the relative importance of variables on the considered outcome.

4. Discussion

From the performed analysis whose results are collected in Tables 3–7, CMI is evidenced as a variable which significantly influences IR regardless of the classifier employed. This implies that CMI is not only functional to *a priori* determine the injury outcome (Eq. (2)), but that CMI actively contributes to the determination of the injury outcome as an indicator of impact eccentricity. Consistently with the study by Kusano and Gabler (2014), the random forest technique is less suitable for IR prediction than classical methodologies such as logistic regression; overall, logistic regression and SVM exhibit similar performances, with a slightly higher quality in the case of SVM with Gaussian Kernel.

Let us refer once more to the study by Kusano and Gabler (2014), characterized by the best fitting quality for IR models available from the literature; the M_3 subset with SVM classifier with Gaussian kernel derived in this work is associated with a higher, albeit limited, fitting quality in terms of AUC-ROC (89.1% versus 88.8%). The reason is to be found in the greater reliability of the data employed, for which the condition in Eq. (6) is fulfilled; the model is thus associated with a quality which is close to excellence (AUC-ROC 90%). The injury index is the key difference between this study (MAIS 3+) and the analysis by Kusano and Gabler (Injury Severity Score higher or equal to 15, ISS 15+). Nevertheless, the *a priori* approach proposed in this work applies to any ΔV-based IR model, as ΔV can always be disaggregated into CMI and V_r: this property allows to predictively assess the risk to sustain an injury of any degree, whichever the index employed to represent the injury (AIS, MAIS, ISS, etc.); this feature is extremely interesting in all those areas (e.g., infrastructure design, active safety of vehicles (Sander and Lubbe, 2018; Vangi et al., 2019d)) in which changes to pre-crash conditions can be applied to allow for IR reduction.

5. Limitations

Notwithstanding the interesting highlights obtained, it is worth evidencing that the analysis is focused on the determination of IR by the MAIS 3+ index (serious injuries). The soundness of the results in terms of V_r and CMI influence on the injury outcome can be extended by studies focused on:

- a higher number of injury degrees (different MAIS);
- different injury severity indexes (e.g., ISS);
- an injury index which refers to single body segments (AIS).

This would additionally allow for a broader-spectrum study, employing the models classified as most promising by the performed analysis.

The IR models which have been derived are valid for all collision types, because the value of V_r and CMI is based on *a posteriori* data (by ΔV value and Eq. (3)); conversely, calculation of ΔV by Eq. (2) may be inappropriate if sliding occurs during the collision. However, in this latter case, only a limited part of V_r is converted into ΔV (low V_{rPDOF}): a low IR value associated with the collision is obtained and, as a consequence, a lower relevance of sliding impacts in *a priori* analyses can be highlighted if compared to full impacts.

It must be mentioned that the check applied to increase the quality of the analysis (Eq. (6)) exerts a significant effect on the remaining number of cases to be processed; however, this check can be modified or excluded to include a higher number of accidents: even if more cases can improve the fitting quality in practice, the use of partially incorrect data can significantly affect the quality of the resulting model (Vangi et al., 2019a). For what specifically concerns the derived models, it is worth noting that they are obtained using American data only; given the differences between distinct countries in terms of both data collection (in procedure and accuracy (Fildes et al., 2013)) and crashworthiness of vehicles (Flannagan et al., 2018), the proposed IR models need to be modified and contextualized based on the specific scope. The type-approval year of the vehicle model also affects the passive safety datum: the registration year is considered as an indicator of the vehicle generation (and crashworthiness) in the derived IR models because it is the sole element available from the NASS/CDS database; considering the type-approval year, rather than the registration year, could be beneficial to enhance the overall fitting quality.

6. Conclusions

The present work analyses influence of closing velocity at collision (V_r) and eccentricity on the injury risk (IR) associated with road impacts. Specifically, the Crash Momentum Index (CMI) is employed to disclose the role of impact eccentricity in determining the injury outcome: thanks to both its *a priori* and *a posteriori* formulations, CMI allows determining the vehicle velocity change (ΔV) and IR for the occupants when jointly

employed with V_r.

Differing from the eccentricity indicators available at the state-of-the-art, indications on ΔV are intrinsically contained inside the CMI; this allows isolating and studying the single contribution of eccentricity without disturbance from the further parameters considered to derive the IR model.

Through the ReliefF algorithm for feature ranking, it has been objectively highlighted that V_r and CMI have a greater relevance on IR than occupant-related and vehicle-related parameters. Considering impact-related variables only, V_r is decisive in determining the injury outcome; the influence of intrusion area and CMI is instead comparable. The quality of the best IR model derived, in which V_r and CMI are both included among the modeling parameters, is close to a value of 90% in terms of AUC-ROC: this model, therefore, demonstrates as a useful tool for applications in the road safety field.

The use of CMI and V_r provides useful insights for *a priori* studies on IR; this property is crucial in infrastructure design: quantification is enabled of the potential IR reduction which can be obtained allowing for more eccentric, lower speed impacts (e.g., replacing a four way intersection with a roundabout). The same applies to performance enhancement of Advanced Driver Assistance Systems (ADAS): in particular, the CMI-based approach allows IR to be derived based solely on pre-impact parameters, in all those critical scenarios where ADAS intervention does not prevent the impact. The advantage is twofold, translating into the possibility to: (a) quickly assess how a higher market penetration degree of an existing ADAS affects road safety as a whole, in terms of IR decrease; (b) tune the characteristics of an ADAS to be developed so that intervention, when the impact cannot be averted, tends to minimize IR for the occupants of the involved vehicles by acting directly on pre-impact parameters (i.e., CMI and V_r). In the end, the joint use of CMI and V_r offer a broad range of possibilities to direct strategies toward the enhancement of road safety as a whole.

Authors' contributions

Michelangelo-Santo Gulino: Conceptualization, Methodology, Investigation, Writing – Original Draft, Writing – Review & Editing.

Leonardo Di Gangi: Methodology, Software, Data Curation, Writing – Original Draft, Writing – Review & Editing, Visualization.

Alessio Sortino: Methodology, Software, Data Curation, Writing – Original Draft, Writing – Review & Editing, Visualization.

Dario Vangi: Conceptualization, Methodology, Writing – Original Draft, Writing – Review & Editing, Supervision.

Funding sources

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Conflicts of interest

None declared.

Appendix A. Detailed description of the considered dataset

The present appendix provides a more comprehensive description of the considered dataset, making use of Figs. A.1–A.2 to highlight the data distribution.

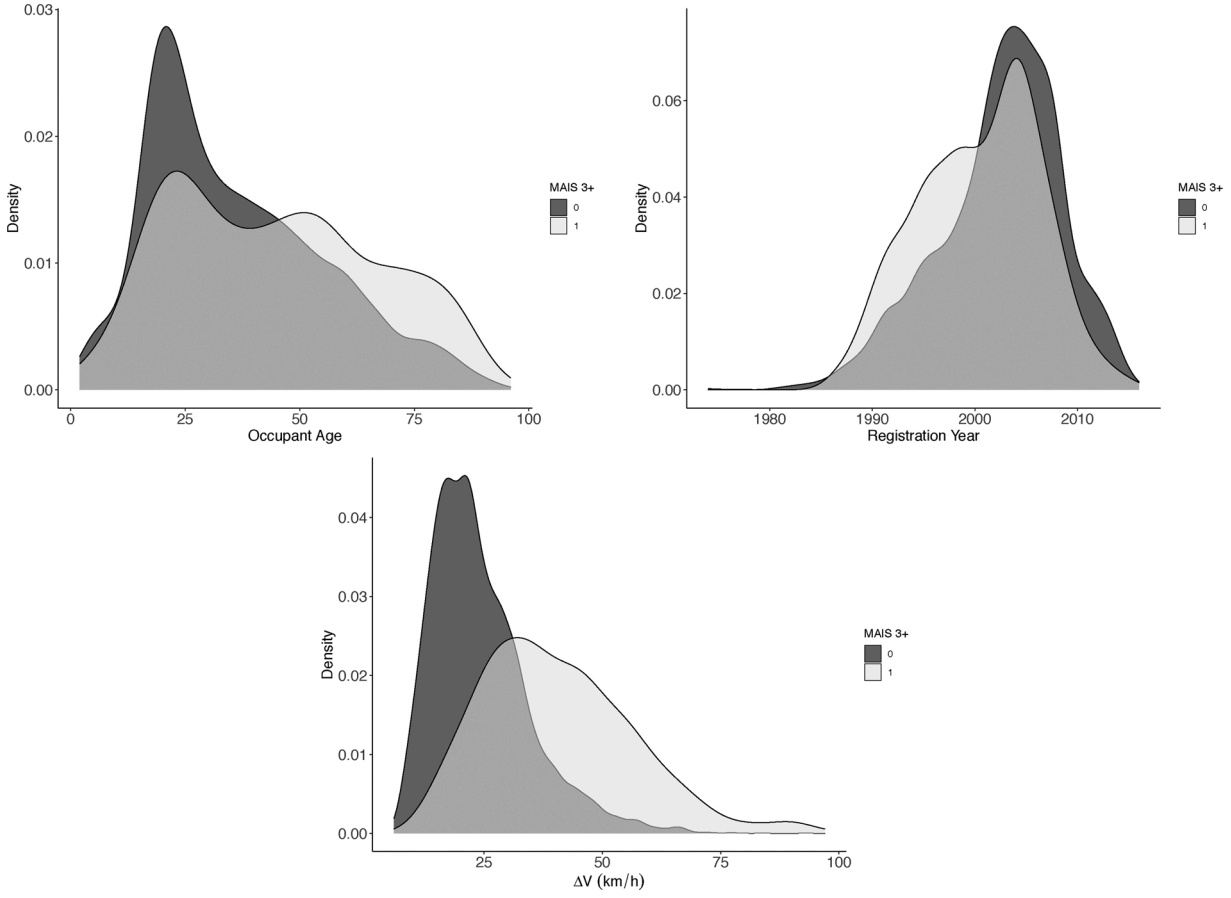

Fig. A.1. Density plots of occupant age, registration year, and ΔV variables, grouped by MAIS 3+ outcome.

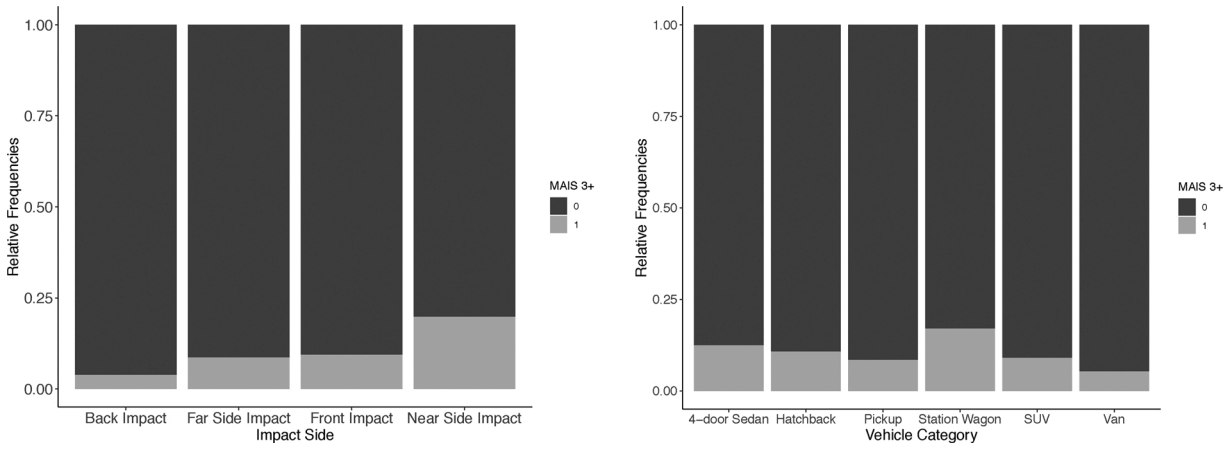

Fig. A.2. Barplots of categorical variables, grouped by MAIS 3+ outcome.

Appendix B. Summary of mixed effects logistic regressions results

In the following, summaries of the fitted random intercept logistic models are provided (Table B.1B.16). Each summary reports the marginal coefficients together with their statistical significance. Both the marginal coefficients and the standard errors are computed by means of a numerical approach, based on Monte Carlo integration, starting from their conditional coefficients counterpart (Hedeker et al., 2018; Rizopoulos, 2019).

Table B.1
M_1 Logistic regression with random intercepts: marginal effects.

	Estimate	Std. error	z value	Pr(> |z|)
(Intercept)	− 5.4575	0.7618	− 7.1642	<1e−04
CMI	0.8267	0.1031	8.0213	<1e−04
V_r	1.1736	0.0839	13.9956	<1e−04
Age	0.7050	0.0808	8.7258	<1e−04
Far side impact	2.1663	0.7536	2.8746	0.0040452
Front impact	2.2559	0.7715	2.9241	0.0034543
Near side impact	3.5757	0.7548	4.7374	<1e−04
Registration year	− 0.5117	0.0803	− 6.3700	<1e−04
Hatchback	− 0.7110	0.4250	− 1.6727	0.0943793
Pickup	− 0.4384	0.3537	− 1.2395	0.2151767
Station Wagon	0.5084	0.3786	1.3427	0.1793552
Suv	− 0.1277	0.2415	− 0.5286	0.5971069
Van	− 0.7436	0.5008	− 1.4848	0.1376014

Table B.2
M_1 Logistic regression with random intercepts: statistics and standard deviation of random intercept.

Log-likelihood	AIC	BIC	SD random intercept
− 708.491	1444.982	1524.157	2.453486

Table B.3
M_2 logistic regression with random intercepts: marginal effects.

	Estimate	Std. error	z value	Pr(> |z|)
(Intercept)	− 4.4852	0.7036	− 6.3746	<1e−04
V_r	1.1866	0.0760	15.6185	<1e−04
Age	0.5956	0.0807	7.3804	<1e−04
Far side impact	1.6564	0.7307	2.2667	0.0234094
Front impact	1.7226	0.7048	2.4439	0.0145296
Near side impact	2.9125	0.7337	3.9696	<1e−04
Year	− 0.4686	0.0838	− 5.5937	<1e−04
Hatchback	− 0.3159	0.4141	− 0.7629	0.4455279
Pickup	− 0.9555	0.2932	− 3.2591	0.0011177
Station Wagon	0.4957	0.4007	1.2370	0.2160932
Suv	− 0.5556	0.2210	− 2.5140	0.0119354
Van	− 1.2479	0.4192	− 2.9769	0.0029120

Table B.4
M_2 logistic regression with random intercepts: statistics and standard deviation of random intercept.

Log-likelihood	AIC	BIC	SD random intercept
− 743.5059	1513.012	1586.532	2.453486

Table B.5
M_3 logistic regression with random intercepts: marginal effects.

	Estimate	Std. error	z value	Pr(> |z|)
(Intercept)	− 8.1770	1.1291	− 7.2419	<1e−04
CMI	1.1747	0.1669	7.0394	<1e−04
V_r	1.7310	0.1930	8.9708	<1e−04
Age	0.9972	0.1437	6.9386	<1e−04
Far side impact	2.7647	0.9375	2.9490	0.0031881
Front impact	2.8567	0.8849	3.2281	0.0012463
Near side impact	4.7936	0.9646	4.9695	<1e−04
Year	− 0.6983	0.1318	− 5.2970	<1e−04

Table B.6
M_3 logistic regression with random intercepts: statistics and standard deviation of random intercept

Log-likelihood	AIC	BIC	SD random intercept
− 713.1024	1444.205	1495.103	2.444645

Table B.7
M_4 logistic regression with random intercepts: marginal effects.

	Estimate	Std. error	z value	Pr(> |z|)
(Intercept)	− 7.6593	1.1896	− 6.4388	<1e−04
V_r	1.8866	0.2255	8.3674	<1e−04
Age	0.9286	0.1479	6.2797	<1e−04
Far side impact	2.3413	0.9803	2.3882	0.016930
Front impact	2.3386	0.9256	2.5266	0.011517
Near side impact	4.2802	1.0081	4.2457	<1e−04
Year	− 0.7182	0.1415	− 5.0762	<1e−04

Table B.8
M_4 logistic regression with random intercepts: statistics and standard deviation of random intercept.

Log-likelihood	AIC	BIC	SD random intercept
− 756.3428	1528.686	1573.929	2.700362

Table B.9
M_5 logistic regression with random intercepts: marginal effects.

	Estimate	Std. Error	z value	Pr(> |z|)
(Intercept)	− 8.0909	1.1547	− 7.0068	<1e−04
CMI	1.1899	0.1723	6.9048	<1e−04
V_r	1.7594	0.2013	8.7386	<1e−04
Age	0.9558	0.1447	6.6069	<1e−04
Far side impact	2.6640	0.9564	2.7854	0.0053456
Front impact	2.7379	0.9033	3.0311	0.0024363
Near side impact	4.6401	0.9820	4.7252	<1e−04

Table B.10
M_5 logistic regression with random intercepts: statistics and standard deviation of random intercept.

Log-likelihood	AIC	BIC	SD random intercept
− 730.8711	1477.742	1522.985	2.579078

Table B.11
M_6 logistic regression with random intercepts: marginal effects.

	Estimate	Std. error	z value	Pr(> |z|)
(Intercept)	− 4.5220	0.6895	− 6.5581	<1e−04
V_r	1.1383	0.0767	14.8484	<1e−04
Age	0.5495	0.0828	6.6340	<1e−04
Far side impact	1.6016	0.7210	2.2214	0.026323
Front impact	1.5838	0.6708	2.3611	0.018223
Near side impact	2.7418	0.6994	3.9201	<1e−04

Table B.12
M_6 logistic regression with random intercepts: statistics and standard deviation of random intercept.

Log-likelihood	AIC	BIC	SD random intercept
− 773.3964	1560.793	1600.38	2.812531

Table B.13
M_7 logistic regression with random intercepts: marginal effects.

	Estimate	Std. error	z value	Pr(> |z|)
(Intercept)	− 5.0038	0.7214	− 6.9362	<1e−04
CMI	0.7212	0.0980	7.3610	<1e−04
V_r	1.0513	0.0748	14.0535	<1e−04
Far side impact	2.0434	0.7484	2.7303	0.0063274
Front impact	2.0094	0.6987	2.8758	0.0040298
Near side impact	3.2951	0.7319	4.5022	<1e−04

Table B.14
M_7 logistic regression with random intercepts: statistics and standard deviation of random intercept.

Log-likelihood	AIC	BIC	SD random intercept
− 765.1173	1544.235	1583.822	2.664012

Table B.15
M_8 logistic regression with random intercepts: marginal effects.

	Estimate	Std. error	z value	Pr(> |z|)
(Intercept)	− 4.3813	0.6623	− 6.6149	<1e−04
V_r	1.0687	0.0787	13.5725	<1e−04
Far side impact	1.6851	0.7029	2.3973	0.016516
Front impact	1.5898	0.6515	2.4403	0.014674
Near side impact	2.7934	0.6808	4.1033	<1e−04

Table B.16
M_8 logistic regression with random intercepts: statistics and standard deviation of random intercept.

Log-likelihood	AIC	BIC	SD random intercept
− 801.5732	1615.146	1649.079	2.846981

References

Abdelwahab, H.T., Abdel-Aty, M.A., 2001. Development of artificial neural network models to predict driver injury severity in traffic accidents at signalized intersections. Transp. Res. Record 1746 (1), 6–13.

Agresti, A., Kateri, M., 2011. Categorical Data Analysis. Springer.

Alogaili, A., Mannering, F., 2020. Unobserved heterogeneity and the effects of driver nationality on crash injury severities in saudi arabia. Accid. Anal. Prev. 144, 105618.

Anastasopoulos, P.C., Mannering, F.L., 2011. An empirical assessment of fixed and random parameter logit models using crash-and non-crash-specific injury data. Accid. Anal. Prev. 43 (3), 1140–1147.

Andricevic, N., Junge, M., Krampe, J., 2018. Injury risk functions for frontal oblique collisions. Traffic Injury Prev. 19 (5), 518–522.

Antonetti, V.W., 1998. Estimating the Coefficient of Restitution of Vehicle-to-Vehicle Bumper Impacts. Technical Report. SAE Technical Paper.

Atkinson, T., Gawarecki, L., Tavakoli, M., 2016. Paired vehicle occupant analysis indicates age and crash severity moderate likelihood of higher severity injury in second row seated adults in frontal crashes. Accid. Anal. Prev. 89, 88–94.

Bagdadi, O., 2013. Estimation of the severity of safety critical events. Accid. Anal. Prev. 50, 167–174.

Bareiss, M., David, M., Gabler, H.C., 2018. Preliminary Estimates of Near Side Crash Injury Risk in Best Performing Passenger Vehicles. Technical Report. SAE Technical Paper.

Behnood, A., Roshandeh, A.M., Mannering, F.L., 2014. Latent class analysis of the effects of age, gender, and alcohol consumption on driver-injury severities. Anal. Methods Accid. Res. 3, 56–91.

Boakye, K.F., Khattak, A., Everett, J., Nambisan, S., 2019. Correlates of front-seat passengers' non-use of seatbelts at night. Accid. Anal. Prev. 130, 30–37.

Bolker, B.M., Brooks, M.E., Clark, C.J., Geange, S.W., Poulsen, J.R., M Henry H Stevens, White, J.-S.S., 2009. Generalized linear mixed models: a practical guide for ecology and evolution. Trends Ecol. Evol. 24 (3), 127–135.

Boser, B.E., Guyon, I.M., Vapnik, V.N., 1992. A training algorithm for optimal margin classifiers. Proceedings of the Fifth Annual Workshop on Computational Learning Theory 144–152.

Boughorbel, S., Jarray, F., El-Anbari, M., 2017. Optimal classifier for imbalanced data using matthews correlation coefficient metric. PLOS ONE 12 (6), e0177678.
Boyd, S., Vandenberghe, L., 2004. Convex Optimization. Cambridge University Press.
Breiman, L., 2001. Random forests. Mach. Learn. 45 (1), 5–32.
Carter, P.M., Flannagan, C.A.C., Reed, M.P., Cunningham, R.M., Rupp, J.D., 2014. Comparing the effects of age, bmi and gender on severe injury (ais 3+) in motor-vehicle crashes. Accid. Anal. Prev. 72, 146–160.
Chang, F., Xu, P., Zhou, H., Chan, A.H.S., Huang, H., 2019. Investigating injury severities of motorcycle riders: a two-step method integrating latent class cluster analysis and random parameters logit model. Accid. Anal. Prev. 131, 316–326.
Chen, C., Zhang, G., Qian, Z., Tarefder, R.A., Tian, Z., 2016. Investigating driver injury severity patterns in rollover crashes using support vector machine models. Accid. Anal. Prev. 90, 128–139.
Chiou, Y.-C., Fu, C., 2013. Modeling crash frequency and severity using multinomial-generalized poisson model with error components. Accid. Anal. Prev. 50, 73–82.
Cortes, C., Vapnik, V., 1995. Support-vector networks. Mach. Learn. 20 (3), 273–297.
Ding, C., Rizzi, M., Strandroth, J., Sander, U., Lubbe, N., 2019. Motorcyclist injury risk as a function of real-life crash speed and other contributing factors. Accid. Anal. Prev. 123, 374–386.
Dubey, R., Zhou, J., Wang, Y., Thompson, P.M., Ye, J., Alzheimer's Disease Neuroimaging Initiative, et al., 2014. Analysis of sampling techniques for imbalanced data: An n = 648 adni study. NeuroImage 87, 220–241.
Fay, R.J., Raney, A.U., Robinette, R.D., 1996. The Effect of Vehicle Rotation on the Occupants' Delta V. Technical Report. SAE Technical Paper.
Fildes, B., Keall, M., Thomas, P., Parkkari, K., Pennisi, L., Tingvall, C., 2013. Evaluation of the benefits of vehicle safety technology: the munds study. Accid. Anal. Prev. 55, 274–281.
Flannagan, C.A.C., Bálint, A., Klinich, K.D., Sander, U., Manary, M.A., Cuny, S., McCarthy, M., Phan, V., Wallbank, C., Green, P.E., et al., 2018. Comparing motor-vehicle crash risk of eu and us vehicles. Accid. Anal. Prev. 117, 392–397.
Forman, J.L., McMurry, T.L., 2018. Nonlinear models of injury risk and implications in intervention targeting for thoracic injury mitigation. Traffic Injury Prev. 19 (Suppl. 2), S103–S108.
Fountas, G., Anastasopoulos, P.C., 2017. A random thresholds random parameters hierarchical ordered probit analysis of highway accident injury-severities. Anal. Methods Accid. Res. 15, 1–16.
Fountas, G., Sarwar, M.T., Anastasopoulos, P.C., Blatt, A., Majka, K., 2018. Analysis of stationary and dynamic factors affecting highway accident occurrence: a dynamic correlated grouped random parameters binary logit approach. Accid. Anal. Prev. 113, 330–340.
Fountas, G., Pantangi, S.S., Hulme, K.F., Anastasopoulos, P.C., 2019. The effects of driver fatigue, gender, and distracted driving on perceived and observed aggressive driving behavior: a correlated grouped random parameters bivariate probit approach. Anal. Methods Accid. Res. 22, 100091.
Fountas, G., Fonzone, A., Gharavi, N., Rye, T., 2020. The joint effect of weather and lighting conditions on injury severities of single-vehicle accidents. Anal. Methods Accid. Res. 100124.
Gelman, A., et al., 2005. Analysis of variance-why it is more important than ever. Ann. Stat. 33 (1), 1–53.
Goldstein, H., 2011. Multilevel Statistical Models, vol. 922. John Wiley & Sons.
Goodfellow, I., Bengio, Y., Courville, A., 2016. Deep Learning. MIT Press.
Høye, A., 2016. How would increasing seat belt use affect the number of killed or seriously injured light vehicle occupants? Accid. Anal. Prev. 88, 175–186.
Hastie, T., Tibshirani, R., Friedman, J., 2009. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer Science & Business Media.
He, H., Garcia, E.A., 2008. Learning from imbalanced data. IEEE Trans. Knowl. Data Eng. (9), 1263–1284.
Hedeker, D., du Toit, S.H.C., Demirtas, H., Gibbons, R.D., 2018. A note on marginalization of regression parameters from mixed models of binary outcomes. Biometrics 74 (1), 354–361.
Hosmer Jr., D.W., Lemeshow, S., Sturdivant, R.X., 2013. Applied Logistic Regression, vol. 398. John Wiley & Sons.
Huang, M., 2002. Vehicle Crash Mechanics. CRC Press.
Imler, S.M., Heller, M.F., Corrigan, C.F., Zhao, K., Watson, H.N., 2010. The Effect of Side Impact Collision Delta-V, Restraint Status, and Occupant Position on Injury Outcome, Technical Report. SAE Technical Paper.
Iranitalab, A., Khattak, A., 2017. Comparison of four statistical and machine learning methods for crash severity prediction. Accid. Anal. Prev. 108, 27–36.
Jeong, H., Jang, Y., Bowman, P.J., Masoud, N., 2018. Classification of motor vehicle crash injury severity: a hybrid approach for imbalanced data. Accid. Anal. Prev. 120, 250–261.
Jianfeng, X., Hongyu, G., Jian, T., Liu, L., Haizhu, L., 2019. A classification and recognition model for the severity of road traffic accident. Adv. Mech. Eng. 11 (5), 1687814019851893.
Jurewicz, C., Sobhani, A., Woolley, J., Dutschke, J., Corben, B., 2016. Exploration of vehicle impact speed-injury severity relationships for application in safer road design. Transp. Res. Proc. 14, 4247–4256.
Khattak, A.J., Rocha, M., 2003. Are suvs "supremely unsafe vehicles"? Analysis of rollovers and injuries with sport utility vehicles. Transp. Res. Record 1840 (1), 167–177.
Kononenko, I., 1994. Estimating attributes: analysis and extensions of relief. European Conference on Machine Learning 171–182.
Kullgren, A., 2008. Dose–response models and edr data for assessment of injury risk and effectiveness of safety systems. In: IRCOBI Conference. Bern, Switzerland, pp. 3–14.
Kullgren, A., Lie, A., Tingvall, C., 2010. Comparison between euro ncap test results and real-world crash data. Traffic Injury Prev. 11 (6), 587–593.
Kusano, K., Gabler, H.C., 2014. Comparison and validation of injury risk classifiers for advanced automated crash notification systems. Traffic Injury Prev. 15 (Suppl. 1), S126–S133.
Kwon, O.H., Rhee, W., Yoon, Y., 2015. Application of classification algorithms for analysis of road safety risk factor dependencies. Accid. Anal. Prev. 75, 1–15.
Lai, X., Ma, C., Hu, J., Zhou, Q., 2012. Impact direction effect on serious-to-fatal injuries among drivers in near-side collisions according to impact location: focus on thoracic injuries. Accid. Anal. Prev. 48, 442–450.
Li, Z., Liu, P., Wang, W., Xu, C., 2012. Using support vector machine models for crash injury severity analysis. Accid. Anal. Prev. 45, 478–486.
Li, Y., Ma, D., Zhu, M., Zeng, Z., Wang, Y., 2018. Identification of significant factors in fatal-injury highway crashes using genetic algorithm and neural network. Accid. Anal. Prev. 111, 354–363.
Liu, X.-Y., Wu, J., Zhou, Z.-H., 2008. Exploratory undersampling for class-imbalance learning. IEEE Trans. Syst. Man Cybern. Part B (Cybern.) 39 (2), 539–550.
Mannering, F.L., Shankar, V., Bhat, C.R., 2016. Unobserved heterogeneity and the statistical analysis of highway accident data. Anal. Methods Accid. Res. 11, 1–16.
McCullagh, P., 2019. Generalized Linear Models. Routledge.
Metzger, K.B., Gruschow, S., Durbin, D.R., Curry, A.E., 2015. Association between ncap ratings and real-world rear seat occupant risk of injury. Traffic Injury Prev. 16 (Suppl. 2), S146–S152.
Mitchell, R.J., Bambach, M.R., Toson, B., 2015. Injury risk for matched front and rear seat car passengers by injury severity and crash type: an exploratory study. Accid. Anal. Prev. 82, 171–179.
Newgard, C.D., 2008. Defining the "older" crash victim: the relationship between age and serious injury in motor vehicle crashes. Accid. Anal. Prev. 40 (4), 1498–1505.
Nishimoto, T., Mukaigawa, K., Tominaga, S., Lubbe, N., Kiuchi, T., Motomura, T., Matsumoto, H., 2017. Serious injury prediction algorithm based on large-scale data and under-triage control. Accid. Anal. Prev. 98, 266–276.
Nishimoto, T., Kubota, K., Ponte, G., 2019. A pedestrian serious injury risk prediction method based on posted speed limit. Accid. Anal. Prev. 129, 84–93.
Pal, C., Narahari, S., Vimalathithan, K., Manoharan, J., Hirayama, S., Hayashi, S., Combest, J., 2018. Real World Accident Analysis of Driver Car-to-Car Intersection Near-Side Impacts: Focus on Impact Location, Impact Angle and Lateral Delta-V, Technical Report. SAE Technical Paper.
Papadimitriou, E., Filtness, A., Theofilatos, A., Ziakopoulos, A., Quigley, C., Yannis, G., 2019. Review and ranking of crash risk factors related to the road infrastructure. Accid. Anal. Prev. 125, 85–97.
Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E., 2011. Scikit-learn: machine learning in Python. J. Mach. Learn. Res. 12, 2825–2830.
Pinheiro, J.C., Chao, E.C., 2006. Efficient laplacian and adaptive gaussian quadrature algorithms for multilevel generalized linear mixed models. J. Comput. Graph. Stat. 15 (1), 58–81.
Rahman Shaon, M.R., Qin, X., Afghari, A.P., Washington, S., Haque, M.M., 2019. Incorporating behavioral variables into crash count prediction by severity: a multivariate multiple risk source approach. Accid. Anal. Prev. 129, 277–288.
Rasmussen, C.E., 2003. Gaussian processes in machine learning. Summer School on Machine Learning. Springer, pp. 63–71.
Rizopoulos, D., 2019. GLMMadaptive: Generalized Linear Mixed Models Using Adaptive Gaussian Quadrature. R package version 0.6-5.
Robnik-Šikonja, M., Kononenko, I., 2003. Theoretical and empirical analysis of relieff and rrelieff. Mach. Learn. 53 (1–2), 23–69.
Rosén, E., Källhammer, J.-E., Eriksson, D., Nentwich, M., Fredriksson, R., Smith, K., 2010. Pedestrian injury mitigation by autonomous braking. Accid. Anal. Prev. 42 (6), 1949–1957.
Sander, U., Lubbe, N., 2018. Market penetration of intersection aeb: characterizing avoided and residual straight crossing path accidents. Accid. Anal. Prev. 115, 178–188.
Savolainen, P.T., Mannering, F.L., Lord, D., Quddus, M.A., 2011. The statistical analysis of highway crash-injury severities: a review and assessment of methodological alternatives. Accid. Anal. Prev. 43 (5), 1666–1676.
Schlögl, M., Stütz, R., 2017. Methodological considerations with data uncertainty in road safety analysis. Accid. Anal. Prev.
Shannon, D., Murphy, F., Mullins, M., Rizzi, L., 2020. Exploring the role of Delta-V in influencing occupant injury severities-a mediation analysis approach to motor vehicle collisions. Accid. Anal. Prev. 142, 105577.
Skrondal, A., Rabe-Hesketh, S., 2009. Prediction in multilevel generalized linear models. J. R. Stat. Soc.: Ser. A (Stat. Soc.) 172 (3), 659–687.
Sobhani, A., Young, W., Logan, D., Bahroloolom, S., 2011. A kinetic energy model of two-vehicle crash injury severity. Accid. Anal. Prev. 43 (3), 741–754.
Truong, L.T., Nguyen, H.T.T., Tay, R., 2020. A random parameter logistic model of fatigue-related motorcycle crash involvement in Hanoi, Vietnam. Accid. Anal. Prev. 144, 105627.
Urbanowicz, R.J., Meeker, M., La Cava, W., Olson, R.S., Moore, J.H., 2018. Relief-based feature selection: introduction and review. J. Biomed. Inform.
Vadeby, A.M., 2004. Modeling of relative collision safety including driver characteristics. Accid. Anal. Prev. 36 (5), 909–917.
Vangi, D., 2014. Impact severity assessment in vehicle accidents. Int. J. Crashworth. 19 (6), 576–587.
Vangi, D., 2020. Vehicle Collision Dynamics: Analysis and Reconstruction. Butterworth-Heinemann.
Vangi, D., Gulino, M.-S., Cialdai, C., 2019a. Coherence assessment of accident database kinematic data. Accid. Anal. Prev. 123, 356–364.

Vangi, D., Gulino, M.-S., Fiorentino, A., Virga, A., 2019b. Crash momentum index and closing velocity as crash severity index. Proc. Inst. Mech. Eng. Part D: J. Automob. Eng. page 0954407018823658.
Vangi, D., Begani, F., Spitzhüttl, F., Gulino, M.-S., 2019c. Vehicle accident reconstruction by a reduced order impact model. Forensic Sci. Int. 298, 426-e1.
Vangi, D., Virga, A., Gulino, M.-S., 2019d. Combined activation of braking and steering for automated driving systems: adaptive intervention by injury risk-based criteria. Proc. Struct. Integr. 24, 423–436.
Viano, D.C., Parenteau, C.S., 2008. Fatalities by Seating Position and Principal Direction of Force (pdof) for 1st, 2nd and 3rd row Occupants. Technical Report. SAE Technical Paper.
Wang, L., Abdel-Aty, M., Lee, J., Shi, Q., 2019. Analysis of real-time crash risk for expressway ramps using traffic, geometric, trip generation, and socio-demographic predictors. Accid. Anal. Prev. 122, 378–384.
Wenzel, T.P., Ross, M., 2005. The effects of vehicle model and driver behavior on risk. Accid. Anal. Prev. 37 (3), 479–494.
Wood, D.P., Simms, C.K., 2002. Car size and injury risk: a model for injury risk in frontal collisions. Accid. Anal. Prev. 34 (1), 93–99.
Xie, Y., Li, D., Zhang, D., Shuang, H., 2017. An improved multi-label relief feature selection algorithm for unbalanced datasets. International Conference on Intelligent and Interactive Systems and Applications 141–151.
Ye, X., Pendyala, R.M., Shankar, V., Konduri, K.C., 2013. A simultaneous equations model of crash frequency by severity level for freeway sections. Accid. Anal. Prev. 57, 140–149.