An empirical assessment of strategies to model opponent effects in crash severity analysis

Author 1[a], Author 2[b]

[a] *Department, Street, City, State, Zip*
[b] *Department, Street, City, State, Zip*

Abstract

Road accidents impose an important burden on health and the economy. Numerous efforts to understand the factors that affect road collisions have been undertaken. One stream of research focus on modelling the severity of crashes. Crash severity research is useful to clarify the way different factors can influence the outcome of the event. The objective of this paper is to assess different strategies to model the interactions between participants in a crash, in the context of crashes involving two parties. Towards this objective, a series of models are estimated using data from Canada's National Collision Database. Three levels of crash severity (no injury/injury/fatality) are analyzed using ordered logit models and covariates for the participants in the crash and the conditions of the crash. Modelling strategies include different ways of introducing the covariates (e.g., in a single-level or multi-level form), as well as by subsetting the dataset. The models are assessed using predicted shares and outcomes, and the results highlight the importance of considering opponent effects in crash severity analysis. On the other hand, the study suggests that hierarchical (i.e., multi-level) specifications and subsetting do not perform necessarily better than a relatively simple single-level model with opponent effects.

1. Introduction

Road safety continues to be a concern world-wide. According to a recent report from the World Health Organization (2019), road accidents are the 8th leading cause of death for all ages, and the number one cause of death for children and young people between the ages of 5 to 29. Of all leading causes of death, road accidents are the only cause of death unrelated to disease, disorder, or infection. Road accidents impose a heavy burden on individuals and society as a whole. Gobally, the rate of road collision-related deaths per 100,000 population and 100,000 vehicles have both fallen, even as the number of vehicles has grown (World Health Organization, 2019, Figs. 1 and 2). These gains, although they are to be celebrated, cannot distract from the crushing economic cost of premature death (e.g., Symons et al., 2019; Wijnen et al., 2019), not to mention the long-term consequences for survivors, measured in sometimes crippling emotional and physical pain (e.g., Merlin et al., 2007; Devlin et al., 2019; Pelissier et al., n.d.).

Evidence from across the world suggests that the burden of road accidents is not borne evenly. There are important disparities at the international level, where the odds of death due to road crashes are three times higher in low-income countries compared to high-income countries; in fact, no reductions in road accident-related fatalities were appreciated in low-income countries between 2013 and 2016 (World Health Organization, 2019). In the case of high-income countries, where substantial gains in road safety have been observed for years, said gains have also been unevenly distributed; thus, while fatal crashes involving older adults in the United States and Great Britain declined between 1997 and 2010 (despite the graying of the population), the trend remained stable or increased slightly in Australia in roughly the same period (Thompson et al., 2018). There are also systematic differences in the impact of road accidents. For example, in a study in the United States, Obeng (2011) reported that the impact of covariates of crash severity varied substantially in magnitude by gender. More recently, Regev et al. (2018) used adjusted crash risk to find that the risk of crashes in Great Britain peaked for people 21 to 29 years of age; on the other hand, the risk of fatal injuries for older drivers was constant, irrespective of the seriousness of the crash - which highlights the

Email addresses: a1@example.com (Author 1), a2@example.com (Author 2)

perils of accidents at older ages. Other studies have concentrated on the consequences of road accidents for the young (e.g., Peek-Asa et al., 2010), the old (e.g., Rakotonirainy et al., 2012), as well as pedestrians and cyclists (e.g., Hanson et al., 2013; McArthur et al., 2014).

Given the relevance and cost of this matter, as well as the important variations of the impacts among different population segments, numerous efforts efforts have been conducted to better understand the factors that affect road safety - including the probable consequences of crashes. Consequently, a stream of research in the analysis of road accidents is concerned with the severity of crashes. In particular, multivariate analysis of crash severity is a useful way to clarify the way various factors can affect the outcome of an incident, to discriminate between various levels of injury, from no injury (i.e., property damage only), to different degrees of injury up to and including fatality. This is an active area of research (e.g., Savolainen et al., 2011), and one where methodological developments have aimed at improving the reliability, accuracy, and precision of models.

This paper aspires to contribute to the literature on crash severity by assessing different modelling strategies useful to incorporate opponent effects in crash analysis, in the context of incidents involving two parties. The importance of these interactions has been recognized in the existing literature (e.g., Chiou et al., 2013; Lee and Li, 2014; Li et al., 2017; Tarrao et al., 2014), and a number of different modelling strategies have been proposed. In this paper we present a systematic assessment of several relevant modelling strategies, ranging from the way variables are defined in single-level models, in multi-level models (i.e., hierarchical models), as well as using data subsetting approaches. For the assessment we use data from Canada's National Collision Database, a database that collects all police-reported collisions in the country. Using the most recent version of the dataset (2017), three levels of crash severity (no injury/injury/fatality) are analyzed using ordered logit models and covariates for the participants in the crash and the conditions of the crash. For model assessment, we conduct an in-sample prediction exercise using the estimation sample (i.e., *nowcasting*), and also an out-of-sample prediction exercise using the dataset corresponding to 2016 (i.e., *backcasting*). The models are assessed using predicted shares and predicted individual outcomes using an extensive array of verification statistics. The results highlight the importance of considering opponent effects in crash severity analysis to improve the goodness-of-fit and predictive performance o. On the other hand, the study suggests that hierarchical variable specifications and subsetting do not perform necessarily better than a relatively simple single-level model with opponent effects.

The rest of this paper is structured as follows. In Section 2 we present a concise review of the methods used to analyze crash severity, with a particular focus on techniques that consider the interactions between participants in a crash. Section 3 describes the data requirements, data preprocessing, and the modelling strategy, along with the results of model estimation. The results of assessing the models and the discussion of these results is found in Section 4. We then present some additional thoughts about the applicability of this approach in Section 6 before offering some concluding remarks in Section 7.

2. Methodological approaches in crash severity analysis

2.1. General considerations

Modelling the outcomes of crashes in terms of the severity of injuries to participants has been a preoccupation of transportation researchers, planners, auto insurance companies, governments and the public for decades. One of the earliests studies to investigate the severity of injuries conditional on an accident having occurred was by White and Clayton (1972). Kim et al. (1995) later stated that the "linkages between severity of injury and driver characteristics and behaviors have not been thoroughly investigated" (p. 470). Nowadays, there is a burgeoning literature on this subject, including methodological developments, case studies, and more niche research with a focus on particular situations (e.g., crashes at intersections, Mussone et al., 2017; crashes in rural roads, Gong and Fan, 2017) or special populations (e.g., crashes involving motorcyclists or active travelers; see Shaheed et al., 2013; Salon and McIntyre, 2018).

Crash severity is often modelled using models for discrete outcomes, including classification techniques from machine learning (e.g., Iranitalab and Khattak, 2017; Chang and Wang, 2006; Effati et al., 2015; Khan et al., 2015), Poisson models for counts (e.g., Ma et al., 2008), unordered logit/probit models (e.g., Tay et al., 2011), as well as ordered logit/probit models (e.g., Rifaat and Chin, 2007), with numerous variants, such as random parameters/mixed logit (e.g., Aziz et al., 2013; Haleem and Gan, 2013), partial proportional odds

Table 1: Categories of variables used in the analysis of crash severity with examples

Category	Examples
Person-related	Attributes of participants in the crash, e.g., injury status, age, gender, licensing status, professional driver status
Traffic unit-related	Attributes of the traffic unit, e.g., type of traffic unit (car, motorcycle, etc.), maneouver, etc.
Crash-related	Attributes of the crash, e.g., location, weather conditions, light conditions, number of parties, etc.
Road-related	Attributes of the road, e.g., surface condition, grade, geometry, etc.
Opponent-related	Attributes of the opponent, e.g., age of opponent, gender of opponent, opponent vehicle type, etc.

models (e.g., Mooradian et al., 2013; Sasidharan and Menendez, 2014), and the use of copulas (e.g., Wang et al., 2015). Recent reviews of methods include Savolainen et al. (2011) and Shamsunnahar and Eluru (2013).

Irrespective of the modelling framework employed, models of crash severity often include variables in several categories, as shown with examples in Table 1 (also see Montella et al., 2013). Many crash databases and analyses also account for the multievent nature of many crashes. Participants may have had different roles in a crash depending on their context, with some acting as operators of a vehicle (i.e., drivers, bicyclists), while others were passengers. They also may differ depending on what type of traffic unit they were, for example occupants of a light duty vehicle or a truck, motorcyclists, or pedestrians. The multiplicity of roles makes for complicated modelling choices when trying to understand the severity of injuries; for example, what is the unit of analysis, the person, the traffic unit, or the collision? Not surprisingly, it is possible to find examples of studies that adopt different perspectives. A common simplifying strategy to specify a model is to consider only *drivers* and/or only *single-vehicle* crashes (e.g., Kim et al., 2013; Gong and Fan, 2017; Lee and Li, 2014; Osman et al., 2018). This strategy reduces the dimensions of the event, and it becomes possible, for example, to equate the traffic unit to the person for modelling purposes.

The situation becomes more complex when dealing with events that involve two traffic units (e.g., Tarrao et al., 2014; Wang et al., 2015) and multi-traffic unit crashes (e.g., Wu et al., 2014; Bogue et al., 2017). Different strategies have been developed to study these, more complex cases. A number of studies advocate the estimation of separate models for different participants and/or situations. In this way, Wang and Kockelman (2005) estimated models for single-vehicle and two-vehicle crashes, while Savolainen and Mannering (2007) estimated models for single-vehicle and multi-vehicle crashes. More recently, Duddu et al. (2018) and Penmetsa et al. (2017) presented research that estimated separate models for at-fault and not-at-fault drivers. The strategy of estimating separate models also relies on subsetting the dataset, although it is possible to link the relevant models more tightly by means of a common covariance structure, as is the case of bivariate models (e.g., Chiou et al., 2013; Chen et al., 2019) or models with copulas (e.g., Rana et al., 2010; Shamsunnahar et al., 2014; Wang et al., 2015).

A related strategy to specify a crash severity model is to organize the data in such a way that it is possible to model the influence of the attributes of the opponent in a crash. There are numerous examples of studies that consider at least some characteristics of the oponents in two- or multi-vehicle crashes. For example, Wang and Kockelman (2005) considered the type of the opposing vehicle in their model for two-vehicle collisions. Similarly, Tarrao et al. (2014) included in their analysis the age, wheelbase, weight, and engine size of the opposing vehicle while Bogue et al. (2017) used the body type of the opposing vehicle. Penmetsa et al. (2017) and Duddu et al. (2018) are two of the most comprehensive examples of using opponent's information, with the physical condition, sex, age, and vehicle type of the opponent. The twin strategies of subsetting the sample and using the attributes of the opponent are not mutually exclusive, but neither are they consistently used together, as a scan of the literature reveals.

2.2. Modelling techniques

With respect to model structures, Shamsunnahar and Eluru (2013) conducted an extensive comparison of models for discrete outcomes and found only small differences in the performance of unordered models and

ordered models; however, ordered models are usually more parsimonious since only one latent functions needs to be estimated for all outcomes, as opposed to one for each outcome in unordered modelling mechanisms. Bogue et al. (2017) also compared unordered and ordered models in the form of the mixed multinomial logit and a modified rank ordered logit, respectively, and found that the ordered model performed best. To keep the empirical assessment managable we will consider only the ordinal logit model, and will comment on potential extensions in Section 6.

The ordinal model is a latent-variable approach, whereby the severity of the crash (observed) is linked to an underlying latent variable that is a function of the variables of interest, as follows:

$$y^*_{itk} = \sum_{l=1}^{L} \alpha_l p_{itkl} + \sum_{m=1}^{M} \beta_m u_{tkm} + \sum_{q=1}^{Q} \kappa_q c_{kq} + \epsilon_{itk} \tag{1}$$

The left-hand side of the expression above (y^*_{itk}) is a latent (unobservable) variable that is associated with the severity of crash k ($k = 1, \cdots, K$) for participant i in traffic unit t. The right-hand side of the expression is split in four parts. The first part collects $l = 1, \cdots, L$ individual attributes p for participant i in traffic unit t and crash k; these could relate to the person (e.g., age, gender, and road user class). The second part collects $m = 1, \cdots, M$ attributes u related to traffic unit t in crash k; these could be items such as maneuver or vehicle type. The third part collects $q = 1, \cdots, Q$ attributes c related to the crash k, including crash-related and road-related data, such as weather conditions, road alignment, and type of surface. Lastly, the fourth element is a random term specific to participant i in traffic unit t and crash k. The function consists of a total of $Z = L + M + Q$ covariates and associated parameters.

When opponent-related variables are included, the function is augmented as follows:

$$y^*_{itk} = \sum_{l=1}^{L} \alpha_l p_{itkl} + \sum_{m=1}^{M} \beta_m u_{tkm} + \sum_{q=1}^{Q} \kappa_q c_{kq} + \sum_{r=1}^{R} \delta_r o_{jvkr} + \epsilon_{itk} \tag{2}$$

The additional part collects $r = 1, \cdots, R$ attributes o related to individual j in traffic unit v and crash k that opposed individual i in traffic unit t and crash k. These could be individual characteristics of the opponent (such as age and gender) and/or characteristics of the opposing vehicle (such as vehicle type or weight). To qualify as an opponent, individual j must have been a participant in crash k but operating traffic unit $v \neq t$. Sometimes the person *is* the traffic unit, as is the case of a pedestrian. And we exclude passengers of vehicles as opponents, since they do not operate the traffic unit. In case the opponent attributes include only characteristics of the traffic unit, the condition for the traffic unit to be an opponent is that it participated in crash k and was different from t. After introducing this new set of terms, the latent function now consists of a total of $Z = L + M + Q + R$ covariates and associated parameters.

For conciseness, in what follows we will abbreviate the function as follows:

$$y^*_{itk} = \sum_{z=1}^{Z} \theta_z x_{itkz} + \epsilon_{itk} \tag{3}$$

The latent variable is not observed directly, but it is possible to posit a probabilistic relationship with the outcome y_{ik} (the severity of crash k for participant i). Depending on the characteristics of the data and the assumptions made about the random component of the latent function different models can be obtained. For example, if crash severity is coded as a binary variable (e.g., non-fatal/fatal), we can relate the latent variable to the outcome as follows:

$$y_{itk} = \begin{cases} \text{fatal} & \text{if } y^*_{itk} > 0 \\ \text{non-fatal} & \text{if } y^*_{itk} \leq 0 \end{cases} \tag{4}$$

Due to the stochastic nature of the latent function, the outcome of the crash is not fully determined. However, we can make the following probability statement:

$$P(y_{itk} = \text{fatal}) = P(y^*_{itk} > 0) \tag{5}$$

In other words, the probability that individual i in traffic unit k and crash k was a fatality equals the probability that the latent variable is greater than zero. This implies (see Maddala, 1986, p. 22):

$$\begin{aligned} P(y_{itk} = \text{fatal}) &= P(\textstyle\sum_{z=1} \theta_z p_{itkz} + \epsilon_{itk} > 0) \\ &= P(\epsilon_{itk} > -\textstyle\sum_{z=1} \theta_z p_{itkz}) \end{aligned} \tag{6}$$

If the random terms ϵ_{itk} are assumed to follow the logistic distribution, the the binary logit model is obtained; if they are assumed to follow the normal distribution, the binary probit model is obtained. More often, though, the outcome is recorded using more categories, for example property damage only (PDO)/injury/fatality. A similar approach can be adopted, with a latent variable that relates to the outcome as follows:

$$y_{itk} = \begin{cases} \text{fatality} & \text{if } y^*_{itk} > \mu_2 \\ \text{injury} & \text{if } \mu_1 < y^*_{itk} < \mu_2 \\ \text{PDO} & \text{if } y^*_{itk} < \mu_1 \end{cases} \tag{7}$$

where μ_1 and μ_2 are estimable thresholds. In this case, the associated probability statements are as follows:

$$\begin{aligned} P(y_{itk} = \text{PDO}) &= 1 - P(y_{itk} = \text{injury}) - P(y_{itk} = \text{fatality}) \\ P(y_{itk} = \text{injury}) &= P(\mu_1 - \textstyle\sum_{z=1} \theta_z p_{itkz} < \epsilon_{itk} < \mu_2 - \textstyle\sum_{z=1} \theta_z p_{itkz}) \\ P(y_{itk} = \text{fatality}) &= P(\epsilon_{itk} < \mu_1 - \textstyle\sum_{z=1} \theta_z p_{itkz}) \end{aligned} \tag{8}$$

If the random terms are assumed to follow the logistic distribution, the ordered logit model is obtained; if the normal distribution, then the ordered probit model. Estimation methods for these models are very well-established (e.g., Maddala, 1986; Train, 2009)

There are numerous variations of the basic modelling framework above, including hierarchical models, bivariate models, multinomial models, and Bayesian models, among others (see Savolainen et al., 2011 for a review of methods).

2.3. Model specification strategies

In this paper we consider three general model specification strategies, as follows:

- Strategy 1. Introducing opponent-related variables
- Strategy 2. Single-level model and multi-level (hierarchical) model specifications
- Strategy 3. Full sample and sample subsetting

Introduction of opponent related-variables was explained in the preceding subsection. In this way, a base model is given by Equation 1, whereas Strategy 1 (inclusion of opponent-related variables) is Equation 12. These two equations are also examples of single-level model. Next we describe Strategies 2 and 3.

2.3.1. Hierarchical model specification

We can choose to conceptualize the event leading to the outcome as a hierarchical process. There are a few different ways of doing this. For example, the hierarchy could be based on individuals in traffic units. In this case, we can rewrite the latent function as follows:

$$y^*_{itk} = \sum_{m=1}^{M} \beta_m u_{tkm} + \sum_{q=1}^{Q} \kappa_q c_{kq} + \sum_{r=1}^{R} \delta_r o_{jvkr} + \epsilon_{itk} \tag{9}$$

The coefficients of the traffic unit nest the individual attributes as follows. For any given coefficient q:

$$\beta_m = \sum_{l=1}^{L} \beta_{ml} p_{itkl} \tag{10}$$

Therefore, the corresponding term in the latent function becomes (assuming that $p_{itk1} = 1$, i.e., it is a constant term):

$$\begin{aligned} \beta_m u_{tkm} &= \big(\beta_{m1} + \beta_{m2} p_{itk2} + \cdots + \beta_{mL} p_{itkL}\big) u_{tkm} \\ &= \beta_{m1} u_{tkm} + \beta_{m2} p_{itk2} u_{tkm} + \cdots + \beta_{mL} p_{itkL} u_{tkm} \end{aligned} \tag{11}$$

As an alternative, the nesting unit could be the interaction person-opponent, in which case the opponent-level attributes are nested

$$y^*_{itk} = \sum_{l=1}^{L} \alpha_l p_{itkl} + \sum_{m=1}^{M} \beta_m u_{tkm} + \sum_{q=1}^{Q} \kappa_q c_{kq} + \epsilon_{itk} \tag{12}$$

with any person-level coefficient l that we wish to expand defined as follows:

$$\alpha_l = \sum_{r=1}^{R} \alpha_{lr} o_{jvkr} \tag{13}$$

with the same conditions as before, that $j \neq i$ is the operator of traffic unit $v \neq t$. The corresponding term in the latent function is now (assuming that $o_{jvk1} = 1$, i.e., it is a constant term):

$$\begin{aligned} \alpha_l p_{itkl} &= \big(\alpha_{l1} + \alpha_{l2} o_{jvk2} + \cdots + \alpha_{lR} o_{jvkR}\big) p_{itkl} \\ &= \alpha_{l1} p_{itkl} + \alpha_{l2} o_{jvk2} p_{itkl} + \cdots + \alpha_{lR} o_{jvkR} p_{itkl} \end{aligned} \tag{14}$$

Alterted readers will identify this model specification strategy as Casetti's expansion method (Casetti, 1972). This is a deterministic strategy for modelling contextual effects which, when augmented with random components, becomes the well-known multi-level modelling method (Hedeker and Gibbons, 1994, more on this in Section 6). It is worthwhile to note that higher-order hierarchical effects are possible; for instance, individual attributes nested within traffic units, which in turn are nested within collisions. We do not explore this further in the current paper.

2.3.2. Sample subsetting

The third model specification strategy that we will consider is subsetting the sample. This is applicable in conjunction with any of the other strategies discussed above. In essence, we define the latent function with restrictions as follows. Consider a continuous variable, e.g., age of person, and we wish to concentrate the analysis on older adults (e.g., Dissanayake and Lu, 2002). The latent function is defined as desired (see above), however, the following restriction might be applied to the sample

$$\text{Age of individual } i \text{ in traffic unit } t \text{ in crash } k = \begin{cases} \geq 65 & \text{use record } itk \\ < 65 & \text{do not use record } itk \end{cases} \tag{15}$$

Suppose instead that we are interested in crashes by or against a specific type of traffic unit (e.g., pedestrians, Amoh-Gyimah et al., 2017):

$$\text{Road user class of individual } i \text{ in traffic unit } t \text{ in crash } k = \begin{cases} \text{Pedestrian} & \text{use record } itk \\ \text{Not pedestrian} & \text{do not use record } itk \end{cases} \tag{16}$$

or:

$$\text{Road user class of individual } j \text{ in traffic unit } v \text{ in crash } k = \begin{cases} \text{Pedestrian} & \text{use record } jvk \\ \text{Not pedestrian} & \text{do not use record } jvk \end{cases} \tag{17}$$

More generally, for any variable x of interest:

$$x_{itk} = \begin{cases} \text{Condition: TRUE} & \text{use record } itk \\ \text{Condition: FALSE} & \text{do not use record } itk \end{cases} \tag{18}$$

Several conditions can be imposed to subset the sample in any way that the analyst deems appropriate or suitable.

3. Setting for empirical assessment

In this section we present the setting for the empirical assessment of the modelling strategies discussed in Section 2, namely matters related to data and model estimation.

> **Note:** this paper presents reproducible research. The source file is an R Markdown document. All code and data necessary to reproduce the analysis are available from the following anonymous Drive folder:
>
> https://drive.google.com/open?id=12aJtVBaQ4Zj0xa7mtfqxh0E48hKCb_XV
>
> The source files, code, and data will be publicly available in a GitHub repository upon acceptance of the paper for publication

3.1. Data for empirical assessment

To assess the performance of the various modelling strategies we use data from Canada's National Collision Database (NCDB). This is database contains all motor vehicle collisions on public roads in Canada as reported by a police service. Data are collected by provinces and territories, and shared with the federal government, where data are combined, tracked, and analyzed for reporting of deaths, injuries, and collisions in Canada at the national level. The NCDB is provided by Transport Canada, the agency of the federal government of Canada in charge of transportation policies and programs, under the Open Government License - Canada version 2.0 [https://open.canada.ca/en/open-government-licence-canada].

The NCDB is available from 1999. For the purpose of this paper, we use the data corresponding to 2017, which is the most recent year available as of this writing. Furthermore, for assessment we also use the data corresponding to 2016. Similar to databases in other jurisdictions (see Montella et al., 2013), the NCDB contains information pertaining to the collision, the traffic unit(s), and the person(s) involved in a crash, as shown in Tables 2, 3, and 4. Notice that, compared to Table 1, crash-related variables and road-related variables are collected in a single variable class, namely collision-related, since they are unique for each crash.

Data are organized by person; in other words, there is one record per participant in a collision, be they drivers, passengers, pedestrians, etc. The only variable directly available with respect to opponents in a collision is the number of vehicles involved (see models in Bogue et al., 2017). Therefore, the data needs to be processed to obtain attributes of opponents for each participant in a collision. The protocol to do this is described next.

Table 2: Contents of National Collision Database: Collision-level variables

Variable	Description	Notes
C_CASE	Unique collision identifier	Unique identifier for collisions
C_YEAR	Year	Last two digits of year.
C_MNTH	Month	14 levels: January - December; unknown; not reported by jurisdiction.
C_WDAY	Day of week	9 levels: Monday - Sunday; unknown; not reported by jurisdiction.
C_HOUR	Collision hour	25 levels: hourly intervals; unknown; not reported by jurisdiction.
C_SEV	Collision severity	4 levels: collision producing at least one fatality; collision producing non-fatal injury; unknown; not reported by jurisdiction.
C_VEHS	Number of vehicles involved in collision	Number of vehicles: 1-98 vehicles involved; 99 or more vehicles involved; unknown; not reported by jurisdiction.
C_CONF	Collision configuration	21 levels: SINGLE VEHICLE: Hit a moving object (e.g. a person or an animal); Hit a stationary object (e.g. a tree); Ran off left shoulder; Ran off right shoulder; Rollover on roadway; Any other single vehicle collision configuration; TWO-VEHICLES SAME DIRECTION OF TRAVEL: Rear-end collision; Side swipe; One vehicle passing to the left of the other, or left turn conflict; One vehicle passing to the right of the other, or right turn conflict; Any other two vehicle - same direction of travel configuration; TWO-VEHICLES DIFFERENT DIRECTION OF TRAVEL: Head-on collision; Approaching side-swipe; Left turn across opposing traffic; Right turn, including turning conflicts; Right angle collision; Any other two-vehicle - different direction of travel configuration; TWO-VEHICLES, HIT A PARKED VEHICLE: Hit a parked motor vehicle; Choice is other than the preceding values; unknown;not reported by jurisdiction.

Table 2: Contents of National Collision Database: Collision-level variables *(continued)*

Variable	Description	Notes
C_RCFG	Roadway configuration	15 levels: Non-intersection; At an intersection of at least two public roadways; Intersection with parking lot entrance/exit, private driveway or laneway; Railroad level crossing; Bridge, overpass, viaduct; Tunnel or underpass; Passing or climbing lane; Ramp; Traffic circle; Express lane of a freeway system; Collector lane of a freeway system; Transfer lane of a freeway system; Choice is other than the preceding values; unknown;not reported by jurisdiction.
C_WTHR	Weather condition	10 levels: Clear and sunny; Overcast, cloudy but no precipitation; Raining; Snowing, not including drifting snow; Freezing rain, sleet, hail; Visibility limitation; Strong wind; Choice is other than the preceding values; unknown;not reported by jurisdiction.
C_RSUR	Road surface	12 levels: Dry, normal; Wet; Snow (fresh, loose snow); Slush, wet snow; Icy, packed snow; Debris on road (e.g., sand/gravel/dirt); Muddy; Oil; Flooded; Choice is other than the preceding values; unknown;not reported by jurisdiction.
C_RALN	Road alignment	9 levels: Straight and level; Straight with gradient; Curved and level; Curved with gradient; Top of hill or gradient; Bottom of hill or gradient; Choice is other than the preceding values; unknown;not reported by jurisdiction.
C_TRAF	Traffic control	21 levels: Traffic signals fully operational; Traffic signals in flashing mode; Stop sign; Yield sign; Warning sign; Pedestrian crosswalk; Police officer; School guard, flagman; School crossing; Reduced speed zone; No passing zone sign; Markings on the road; School bus stopped with school bus signal lights flashing; School bus stopped with school bus signal lights not flashing; Railway crossing with signals, or signals and gates; Railway crossing with signs only; Control device not specified; No control present; Choice is other than the preceding values; unknown; not reported by jurisdiction.

Note:
Source NCDB available from https://open.canada.ca/data/en/dataset/1eb9eba7-71d1-4b30-9fb1-30cbdab7e63a
Preprocessed data for analysis available from https://drive.google.com/open?id=12aJtVBaQ4Zj0xa7mtfqxh0E48hKCb_XV

Table 3: Contents of National Collision Database: Traffic unit-level variables

Variable	Description	Notes
V_ID	Vehicle sequence number	Number of vehicles: 1-98; Pedestrian sequence number: 99; unknown.
V_TYPE	Vehicle type	21 levels: Light Duty Vehicle (Passenger car, Passenger van, Light utility vehicles and light duty pick up trucks); Panel/cargo van (<= 4536 KG GVWR Panel or window type of van designed primarily for carrying goods); Other trucks and vans (<= 4536 KG GVWR); Unit trucks (> 4536 KG GVWR); Road tractor; School bus; Smaller school bus (< 25 passengers); Urban and Intercity Bus; Motorcycle and moped; Off road vehicles; Bicycle; Purpose-built motorhome; Farm equipment; Construction equipment; Fire engine; Snowmobile; Street car; Data element is not applicable (e.g. dummy vehicle record created for pedestrian); Choice is other than the preceding values; unknown; not reported by jurisdiction.
V_YEAR	Vehicle model year	Model year; dummy for pedestrians; unknown; not reported by jurisdiction.

Note:
Source NCDB available from https://open.canada.ca/data/en/dataset/1eb9eba7-71d1-4b30-9fb1-30cbdab7e63a
Preprocessed data for analysis available from https://drive.google.com/open?id=12aJtVBaQ4Zj0xa7mtfqxh0E48hKCb_XV

Table 4: Contents of National Collision Database: Personal-level variables

Variable	Description	Notes
P_ID	Person sequence number	Sequence number: 1-99; Not applicable (dummy for parked vehicles); not reported by jurisdiction.
P_SEX	Person sex	5 levels: Male; Female; Not applicable (dummy for parked vehicles); unknown (runaway vehicle); not reported by jurisdiction.

Table 4: Contents of National Collision Database: Personal-level variables *(continued)*

Variable	Description	Notes
P_AGE	Person age	Age: less than 1 year; 1-98 years old; 99 years or older; Not applicable (dummy for parked vehicles); unknown (runaway vehicle); not reported by jurisdiction.
P_PSN	Person position	Person position: Driver; Passenger front row, center; Passenger front row, right outboard (including motorcycle passenger in sidecar); Passenger second row, left outboard, including motorcycle passenger; Passenger second row, center; Passenger second row, right outboard; Passenger third row, left outboard;...; Position unknown, but the person was definitely an occupant; Sitting on someone's lap; Outside passenger compartment; Pedestrian; Not applicable (dummy for parked vehicles); Choice is other than the preceding values; unknown (runaway vehicle); not reported by jurisdiction.
P_ISEV	Medical treatment required	6 levels: No Injury; Injury; Fatality; Not applicable (dummy for parked vehicles); Choice is other than the preceding values; unknown (runaway vehicle); not reported by jurisdiction.
P_SAFE	Safety device used	11 levels: No safety device used; Safety device used; Helmet worn; Reflective clothing worn; Both helmet and reflective clothing used; Other safety device used; No safety device equipped (e.g. buses); Not applicable (dummy for parked vehicles); Choice is other than the preceding values; unknown (runaway vehicle); not reported by jurisdiction.
P_USER	Road user class	6 levels: Motor Vehicle Driver; Motor Vehicle Passenger; Pedestrian; Bicyclist; Motorcyclist; Not stated/Other/Unknown.

Note:
Source NCDB available from https://open.canada.ca/data/en/dataset/1eb9eba7-71d1-4b30-9fb1-30cbdab7e63a
Preprocessed data for analysis available from https://drive.google.com/open?id=12aJtVBaQ4Zj0xa7mtfqxh0E48hKCb_XV

3.1.1. Data preprocessing: initial filter

The database is scanned to remove all records that are not a person (including parked cars and other objects) or that are missing information (e.g., runaway vehicles). Next, records missing at least one of the next variables are removed: P_USER (the road user class), P_SEX (sex), P_AGE (age), and P_ISEV (individual-level severity of crash). This initial filter ensures that all records are complete from the perspective of key information.

3.1.2. Data preprocessing: filter two-vehicle collisions

After the initial filter, the database is summarized to find the number of individual-level records that correspond to each collision (C_CASE). At this point, there are 32,298 collisions, involving only one (known) participant, there are 46,483 collisions involving two participants, 19,433 involving 3 participants, 8,250 involving four participants, 3,783 involving 5 participants, 1,789 involving 6 participants, and 1,491 collisions involving seven or more participants. These participants were possibly occupants in different vehicles or were otherwise their own traffic units, as follows: the sample includes 174,741 drivers, 61,403 passengers, 10,798 pedestrians, 5,286 bicyclists, and 6,564 motorcyclists.

The next step is to remove all collisions involving only one participant. This still leaves numerous cases where multiple participants could have been in a single vehicle, for instance in a collision against a stationary object. Therefore, we proceed to use the vehicle sequence number can to find the number of vehicles involved in each collision. This reveals that there are 20,732 collisions that involve only one vehicle but possibly multiple participants (i.e., driver and one or more passengers). In addition, there are 165,520 collisions involving two vehicles (and possibly multiple participants). Finally, there are 40,242 collisions with three or more vehicles.

Once we have identified the number of vehicles in each collision, we proceed to select all cases that involve only two vehicles. The most common cases in two-vehicle collisions are those that include drivers (40,297 collisions; reflective of the prevalence of single-occupant vehicles), followed by cases with passengers (14,120 collisions), pedestrians (5,204 collisions), byciclists (2,238 collisions), and motorcyclists (1,016 collisions). The distribution of individuals per traffic unit is as follows: 80,382 individuals are coded as being in V_ID = 1, 76,523 individuals are coded as being in V_ID = 2, and 7,932 individuals are coded as pedestrians. In addition, 683 individuals are coded as being in vehicles 3 through 9, despite our earlier filter to retain only

collisions with two vehicles. We therefore proceed to select only individuals assigned to vehicles 1 or 2, as well as pedestrians. As a result of this filter a number of cases with only one known participant need to be removed.

3.1.3. Data preprocessing: extract opponent information and join to individual records

Up to this point, the goal of data preprocessing has been to obtain a complete, workable sample of individual records of participants in two-vehicle collisions. There are two possible cases for the collisions, depending on the traffic units involved: 1) vehicle vs vehicle collisions ("vehicle" includes all motorized vehicles, as well as motorcycles/mopeds and bicycles); and 2) vehicle vs pedestrian collisions. To identify opponents in a collision, it is convenient to classify collisions by pedestrian involvement. Accordingly, we find that the database includes 16,636 collisions are vehicle vs pedestrian (possibly multiple pedestrians), and 147,594 collisions involving two vehicles. After splitting the database in this way, we can now extract relevant information about participants in the collision. This involves renaming the person-level variables depending so that we can distinguish each individual by their role as an individual or an opponent in a given collision. Notice that when working with individuals in vehicles, only the driver is considered an opponent in a collisions (i.e., passengers are never considered opponents).

Once the personal attributes of individuals that can be considered opponents in a given collision have ben extracted, their information can be joined to the individual records by means of the collision unique identifier. As a result of this process, a new set of variables are now available for analysis: the age, sex, use of safety device, and road user class of the opponent, as well as the type of the opponent vehicle. A summary of opponent interactions and outcomes can be found in Table 5. The information there shows that the most commont type of opponent for drivers is other drivers, followed by pedestrians. The only opponent of pedestrians, on the other hand, are drivers. Bicyclists and motorcyclists, on the other hand, are mostly opposed by drivers, but occasionally by other road users as well. In terms of outcomes, we observe that virtually all fatalities occur when the opponent is a driver, and only very rarely when the opponent is a motorcyclist. Injuries are also more common when the opponent is a driver, whereas the likelihood of no injuries is higher when the opponent is a pedestrian or a bicyclist.

3.2. Model estimation

Before model estimation, the variables are prepared as follows. First, age is scaled from years to decades. Secondly, new variables are defined to describe the vehicle type. Three classes of vehicle types are considered: 1) light duty vehicles (which in Canada include passenger cars, passenger vans, light utility vehicles, and light duty pick up trucks); 2) light trucks (all other vehicles $\leq$ 4536 kg in gross vehicle weight rating); and heavy vehicles (all other vehicles $\geq$ 4536 in gross vehicle weight rating). Furthermore, this classification of vehicle types is combined with the road user class to distiguish between drivers and passengers of light duty vehicles, light trucks, and heavy vehicles, in addition to pedestrians, bicyclists, and motorcyclists. This is done for both the individual and the opponent. Variable interactions are calculated to produce hierarchical variables. For example, for a hierarchical definition of traffic unit-level variables, age (and the square of aged) are interacted with gender, road user class, and vehicle type. For hierarchical opponent variables, age (and the square of age) are interacted with age of opponent (and squared aged of opponent). The variables thus obtained are shown in Table 6. As seen in the table, Models 1 and 2 are single-level models, and the difference between them is that Model 2 includes opponent variables. Models 3 and 4, in contrast, are hierarchical models. Model 3 considers the hierarchy on the basis of the traffic unit, while Model 4 considers the hierarchy on the basis of the collision.

Models 1 through 4 are estimated using the full sample. As discussed above, a related modelling strategy is to subset the sample (e.g., Islam et al., 2014; Lee and Li, 2014; Tarrao et al., 2014; Wu et al., 2014). In this case we subset by a combination of traffic unit type of the individual (i.e., light duty vehicle, light truck, heavy vehicle, pedestrian, bicyclist, and motorcyclist) and vehicle type of the oponent (i.e., light duty vehicle, light truck, heavy vehicle). This leads to an ensemble of eighteen models to be estimated using subsets of data (see Table 6). By subsetting the sample, opponent effects are incorporated implicitly. Models 1 and 2 are re-estimated using this strategy, dropping variables as necessary where they are irrelevant (for instance, after filtering for pedestrians, no other traffic unit types are present in the subset of data).

Table 7 collects some key summary statistics of the estimated models. Of interest is the goodness of fit of the models, which in the case is measured with Akaike's Information Criterion (AIC). This criterion is

Table 5: Summary of opponent interactions and outcomes by road user class

	Road User Class of Opponent				Outcome			Proportion by Road User Class		
Road User Class	Driver	Pedestrian	Bicyclist	Motorcyclist	No Injury	Injury	Fatality	No Injury	Injury	Fatality
All opponents										
Driver	97582	7880	3799	2498	59180	52143	436	0.52953	0.46657	0.003901
Passenger	35359	1282	667	818	19308	18667	151	0.50643	0.48961	0.003961
Pedestrian	7880	0	0	0	145	7507	228	0.01840	0.95266	0.028934
Bicyclist	3799	1	0	40	49	3760	31	0.01276	0.97917	0.008073
Motorcyclist	2498	30	40	338	204	2598	104	0.07020	0.89401	0.035788
Opponent: Driver										
Driver	97582	0	0	0	45493	51657	432	0.46620	0.52937	0.004427
Passenger	35359	0	0	0	16672	18536	151	0.47151	0.52422	0.004270
Pedestrian	7880	0	0	0	145	7507	228	0.01840	0.95266	0.028934
Bicyclist	3799	0	0	0	43	3725	31	0.01132	0.98052	0.008160
Motorcyclist	2498	0	0	0	98	2299	101	0.03923	0.92034	0.040432
Opponent: Pedestrian										
Driver	0	7880	0	0	7693	187	0	0.97627	0.02373	0.000000
Passenger	0	1282	0	0	1246	36	0	0.97192	0.02808	0.000000
Pedestrian	0	0	0	0	0	0	0	-	-	-
Bicyclist	0	1	0	0	0	1	0	0.00000	1.00000	0.000000
Motorcyclist	0	30	0	0	11	19	0	0.36667	0.63333	0.000000
Opponent: Bicyclist										
Driver	0	0	3799	0	3706	93	0	0.97552	0.02448	0.000000
Passenger	0	0	667	0	649	18	0	0.97301	0.02699	0.000000
Pedestrian	0	0	0	0	0	0	0	-	-	-
Bicyclist	0	0	0	0	0	0	0	-	-	-
Motorcyclist	0	0	40	0	16	24	0	0.40000	0.60000	0.000000
Opponent: Motorcyclist										
Driver	0	0	0	2498	2288	206	4	0.91593	0.08247	0.001601
Passenger	0	0	0	818	741	77	0	0.90587	0.09413	0.000000
Pedestrian	0	0	0	0	0	0	0	-	-	-
Bicyclist	0	0	0	40	6	34	0	0.15000	0.85000	0.000000
Motorcyclist	0	0	0	338	79	256	3	0.23373	0.75740	0.008876

calculated using the estimated likelihood of the model, penalized by the number of coefficients in the model, so that other things being equal, it gives preference to more parsimonious models. Since smaller values of AIC are better fits, it is possible to see that, compared to the base model without opponent variables (Model 1), there are large and significant improvements in fit to be gained by introducing opponent effects. However, the gains are not as large when hierarchical specifications are used, even when the number of additional coefficients that need to be estimated is not substantially larger (recall that the penalty per parameter in AIC is 2).

The likelihood function of the model, and therefore the value of the AIC, depend on the size of the sample, which is why AIC is not comparable across models estimated with different sample sizes. For this reason, the full sample models cannot be compared directly to the models estimated with subsets of data. The models in the ensembles, however, can be compared, and there we find that introducing opponent variables leads to a better fit in most, but not all cases. One model for pedestrians (when the opponent is a heavy vehicle), the two models for bicyclists, and the three models for motorcyclists are better fits when opponent variables are not used.

These results give some preliminary ideas about the relative performance of the different modelling strategies. In the next section we delve more deeply into this question.

4. Model assessment

In this section we report an in-depth examination of the performance of the models. To this end, we use the models to conduct in-sample predictions (i.e., nowcasting), using the same sample that was used to estimate the models. In addition, we also conduct out-of-sample predictions, using the dataset corresponding to the year 2016, processed in identical way as the estimation sample (i.e., the 2017 dataset). This is an example of backcasting. Broadly, we evaluate the models in two ways: first, we compute the estimated shares of each outcome based on the predicted probabilities; and secondly, we evaluate the predicted outcomes.

Table 6: Summary of variables and model specification

Variable	Notes	Model 1	Model 2	Model 3	Model 4
Individual-level variables					
Age	In decades	✓	✓	✓	✓
Age Squared		✓	✓	✓	✓
Sex	Reference: Female	✓	✓	✓	✓
Use of Safety Devices	7 levels; Reference: No Safety Device	✓	✓	✓	✓
Traffic unit-level variables					
Passenger	Reference: Driver	✓	✓		✓
Pedestrian	Reference: Driver	✓	✓		✓
Bicyclist	Reference: Driver	✓	✓		✓
Motorcyclist	Reference: Driver	✓	✓		✓
Light Truck	Reference: Light Duty Vehicle	✓	✓		✓
Heavy Vehicle	Reference: Light Duty Vehicle	✓	✓		✓
Opponent variables					
Age of Opponent	In decades		✓	✓	
Age of Opponent Squared			✓	✓	
Sex of Opponent	Reference: Female		✓	✓	
Opponent: Light Duty Vehicle	Reference: Pedestrian/Bicyclist/Motorcyclist		✓	✓	✓
Opponent: Light Truck	Reference: Pedestrian/Bicyclist/Motorcyclist		✓	✓	✓
Opponent: Heavy Vehicle	Reference: Pedestrian/Bicyclist/Motorcyclist		✓	✓	✓
Hierarchical traffic unit variables					
Light Truck Driver:Age				✓	
Light Truck Driver:Age Squared				✓	
Heavy Vehicle Driver:Age				✓	
Heavy Vehicle Driver:Age Squared				✓	
Light Truck Passenger:Age				✓	
Light Truck Passenger:Age Squared:				✓	
Heavy Vehicle Passenger:Age				✓	
Heavy Vehicle Passenger:Age Squared				✓	
Pedestrian:Age				✓	
Pedestrian:Age Squared				✓	
Bicyclist:Age				✓	
Bicyclist:Age Squared				✓	
Motorcyclist:Age				✓	
Motorcyclist:Age Squared				✓	
Hierarchical opponent variables					
Age:Age of Opponent					✓
Age:Age of Female Opponent					✓
Age:Age of Male Opponent Squared					✓
Age:Age of Female Opponent Squared					✓
Age Squared:Age of Male Opponent					✓
Age Squared:Age of Female Opponent					✓
Collision-level variables					
Crash Configuration	19 levels; Reference: Hit a moving object	✓	✓	✓	✓
Road Configuration	12 levels; Reference: Non-intersection	✓	✓	✓	✓
Weather	9 levels; Reference: Clear and sunny	✓	✓	✓	✓
Surface	11 levels; Reference: Dry	✓	✓	✓	✓
Road Alignment	8 levels; Reference: Straight and level	✓	✓	✓	✓
Traffic Controls	19 levels; Reference: Operational traffic signals	✓	✓	✓	✓
Month	12 levels; Reference: January	✓	✓	✓	✓

Table 7: Summary of model estimation results

Model	Number of observations	Number of coefficients	AIC
Full sample models			
Model 1	164,511	104	195,215
Model 2	164,511	110	178,943
Model 3	164,511	120	181,333
Model 4	164,511	113	179,018
Model 1 Ensemble (sample subsets by user type vs opponent)			
Light duty vehicle vs light duty vehicle	114,841	96	145,390
Light duty vehicle vs light truck	3,237	81	3,943
Light duty vehicle vs heavy vehicle	5,013	90	5,895
Light truck vs light duty vehicle	3,121	81	3,885
Light truck vs light truck	809	69	1,170
Light truck vs heavy vehicle	198	66	288
Heavy vehicle vs light duty vehicle	4,726	81	4,326
Heavy vehicle vs light truck	180	66	225
Heavy vehicle vs heavy vehicle	779	76	1,147
Pedestrian vs light duty vehicle	7,176	90	2,826
Pedestrian vs light truck	328	64	202
Pedestrian vs heavy vehicle	376	66	409
Bicyclist vs light duty vehicle	3,521	82	654
Bicyclist vs light truck	116	44	84
Bicyclist vs heavy vehicle	NA	NA	NA
Motorcyclist vs light duty vehicle	2,298	80	1,367
Motorcyclist vs light truck	127	58	153
Motorcyclist vs heavy vehicle	62	45	88
Model 2 Ensemble (sample subsets by user type vs opponent)			
Light duty vehicle vs light duty vehicle	114,841	99	143,903
Light duty vehicle vs light truck	3,237	84	3,927
Light duty vehicle vs heavy vehicle	5,013	93	5,878
Light truck vs light duty vehicle	3,121	84	3,877
Light truck vs light truck	809	72	1,156
Light truck vs heavy vehicle	198	69	281
Heavy vehicle vs light duty vehicle	4,732	86	4,283
Heavy vehicle vs light truck	179	67	205
Heavy vehicle vs heavy vehicle	779	79	1,136
Pedestrian vs light duty vehicle	7,176	93	2,821
Pedestrian vs light truck	328	67	200
Pedestrian vs heavy vehicle	376	69	410
Bicyclist vs light duty vehicle	3,521	85	659
Bicyclist vs light truck	145	59	114
Bicyclist vs heavy vehicle	NA	NA	NA
Motorcyclist vs light duty vehicle	2,298	83	1,373
Motorcyclist vs light truck	127	61	153
Motorcyclist vs heavy vehicle	63	45	90

Note:
There are zero cases of Bicyclist vs heavy vehicle in the sample

4.1. Outcome shares based on estimated probabilities

The shares of each outcome are calculated as the sum of the estimated probabilities for each observation:

$$\begin{array}{rcl} \hat{S}_{\text{PDO}} & = & \sum_{itk} \hat{P}(y_{itk} = \text{PDO}) \\ \hat{S}_{\text{injury}} & = & \sum_{itk} \hat{P}(y_{itk} = \text{injury}) \\ \hat{S}_{\text{fatality}} & = & \sum_{itk} \hat{P}(y_{itk} = \text{fatality}) \end{array} \tag{19}$$

where $\hat{P}(y_{itk} = h_w)$ is the estimated probability of outcome h_w for individual i in traffic unit t and crash k. The estimated share of outcome h is $\hat{S}_{h_w}$.

The estimated shares can be used to assess the ability of the model to forecast for the population the total number of cases of each outcome. A summary statistic useful to evaluate the performance is the Average Percentage Error, or APE (see Bogue et al., 2017, p. 31), which is calculated for each outcome as follows:

$$APE_{h_w} = \left| \frac{\hat{S}_{h_w} - S_{h_w}}{S_{h_w}} \right| \times 100 \tag{20}$$

The Weighted Average Percentage Error ($WAPE$) aggregates the APE as follows:

$$WAPE = \frac{\sum_w^W APE_{h_w} \times S_{h_w}}{\sum_w^W S_{h_w}} \tag{21}$$

The results of this exercise are reported in Table 8. Of the four full-sample models (Models 1-4), the APE of Model 2 is lowest in the nowcasting exercise for every outcome, with the exception of Fatality, where Model 4 produces a considerably lower APE. When the results are aggregated by means of the $WAPE$, Model 2 gives marginally better results than Model 4. It is interesting to see that the two Ensemble models have lower APE values across the board in the nowcasting exercise, and much better $WAPE$ than the full sample models. However, when we turn to the results of the backcasting (out-of-sample) predictions, these results do not hold, as the Average Percentage Errors worsen considerably, particularly in the case of Fatality. Excellent in-sample predictions but mediocre out-of-sample predictions could constitute evidence of overfitting by the Ensemble models. In terms of backcasting, again Model 4 is marginally better than Model 2, especially due to more accurate shares of No Injury.

Table 8: Predicted shares and average prediction errors (APE) by model (percentages)

	No Injury			Injury			Fatality			
Model	Observed	Predicted	APE	Observed	Predicted	APE	Observed	Predicted	APE	WAPE
In-sample (nowcasting using 2017 dataset, i.e., estimation dataset)										
Model 1	78886	79029.00	0.18	84675	84533.74	0.17	950	948.26	0.18	0.17
Model 2	78886	78928.98	0.05	84675	84641.94	0.04	950	940.08	1.04	0.05
Model 3	78886	79027.29	0.18	84675	84512.50	0.19	950	971.21	2.23	0.20
Model 4	78886	78939.18	0.07	84675	84622.54	0.06	950	949.28	0.08	0.06
Model 1 Ensemble	62413	62402.78	0.02	83564	83573.58	0.01	931	931.64	0.07	0.01
Model 2 Ensemble	62417	62407.00	0.02	83595	83604.14	0.01	931	931.86	0.09	0.01
Model 4 Ensemble	62405	62395.28	0.02	83578	83586.75	0.01	932	932.97	0.10	0.01
Out-of-sample (backcasting using 2016 dataset)										
Model 1	96860	96364.67	0.51	101605	102002.59	0.39	1109	1206.74	8.81	0.50
Model 2	96860	96361.41	0.51	101605	102112.08	0.50	1109	1100.51	0.77	0.51
Model 3	96860	96354.01	0.52	101605	102086.18	0.47	1109	1133.82	2.24	0.51
Model 4	96860	96325.85	0.55	101605	102136.72	0.52	1109	1111.43	0.22	0.54
Model 1 Ensemble	77457	76822.49	0.82	100013	100580.60	0.57	1072	1138.91	6.24	0.71
Model 2 Ensemble	77459	76799.11	0.85	100049	100630.48	0.58	1071	1149.41	7.32	0.74
Model 4 Ensemble	77461	76766.08	0.90	100029	100630.21	0.60	1070	1163.71	8.76	0.78

4.2. Predicted outcomes

Words go here.

Verification statistics used are summarized in Table 9.

Table 9: Verification statistics

Statistic	Description	Notes
Percent Correct (PC)	Total hits and correct rejections divided by number of cases	Strongly influenced by most common category
Percent Correct by Class (PC_c)	Same as Percent Correct but by category	Strongly influenced by most common category
Bias (B)	Total predicted by category, divided by total observed by category	$B > 1$: class is overpredicted; $B < 1$: class is underpredicted
Critical Success Index (CSI)	Total hits divided by total hits + false alarms + misses	$CSI = 1$: perfect score; $CSI = 0$: no skill
Probability of False Detection (F)	Proportion of no events forecast as yes; sensitive to false alarms but ignores misses	$F = 0$: perfect score
Probability of Detection (POD)	Total hits divided by total observed by class	$POD = 1$: perfect score
False Alarm Ratio (FAR)	Total false alarms divided by total forecast yes by class; measures fraction of predicted yes that did not occur	$FAR = 0$: perfect score
Heidke Skill Score (HSS)	Fraction of correct predictions after removing predictions attributable to chance; measures fractional improvement over random; tends to reward conservative forecasts	$HSS = 1$: perfect score; $HSS = 0$: no skill; $HSS < 0$: random is better
Peirce Skill Score (PSS)	Combines POD and F; measures ability to separate yes events from no events; tends to reward conservative forecasts	$PSS = 1$: perfect score; $PSS = 0$: no skill
Gerrity Score (GS)	Measures accuracy of predicting the correct category, relative to random; tends to reward correct forecasts of less likely category	$GS = 1$: perfect score; $GS = 0$: no skill

We next evaluate the outcomes of the nowcast using an array of verification statistics. See Table 11. We next evaluate the outcomes of the nowcast using an array of verification statistics. See Table 11.

Table 10: Assessment of in-sample outcomes (nowcasting using 2017 dataset, i.e., estimation dataset)

Observed	Predicted Outcome			Verification Statistics									
Outcome	No Injury	Injury	Fatality	Percent Correct	Percent Correct by Class	Bias[1]	Critical Success Index[2]	Probability of False Detection[3]	Probability of Detection[4]	False Alarm Ratio[5]	Heidke Skill Score[6]	Peirce Skill Score[7]	Gerrity Score[8]
Model 1													
No Injury	50652	22503	150		69.068	0.929	0.499	0.265	0.642	0.309			
Injury	28232	62121	797	68.552	68.645	1.076	0.546	0.364	0.734	0.318	0.372	0.370	0.190
Fatality	2	51	3		99.392	0.059	0.003	0.000	0.003	0.946			
Model 2													
No Injury	51531	17137	85		72.903	0.872	0.536	0.201	0.653	0.250			
Injury	27355	67514	864	72.364	72.415	1.131	0.598	0.353	0.797	0.295	0.447	0.443	0.227
Fatality	0	24	1		99.409	0.026	0.001	0.000	0.001	0.960			
Model 3													
No Injury	51101	17296	79		72.549	0.868	0.531	0.203	0.648	0.254			
Injury	27785	67338	868	71.996	72.044	1.134	0.594	0.359	0.795	0.298	0.440	0.436	0.224
Fatality	0	41	3		99.399	0.046	0.003	0.000	0.003	0.932			
Model 4													
No Injury	51575	17316	84		72.822	0.874	0.536	0.203	0.654	0.252			
Injury	27311	67336	863	72.283	72.334	1.128	0.597	0.353	0.795	0.295	0.446	0.441	0.227
Fatality	0	23	3		99.410	0.027	0.003	0.000	0.003	0.885			
Model 1 Ensemble													
No Injury	34664	16434	63		69.882	0.820	0.439	0.195	0.555	0.322			
Injury	27749	67120	829	69.311	69.354	1.145	0.599	0.451	0.803	0.299	0.363	0.352	0.201
Fatality	0	10	39		99.386	0.053	0.041	0.000	0.042	0.204			
Model 2 Ensemble													
No Injury	35443	16145	60		70.615	0.827	0.451	0.192	0.568	0.314			
Injury	26974	67437	829	70.042	70.083	1.139	0.605	0.439	0.807	0.292			0.211
Fatality	0	13	42		99.386	0.059	0.044	0.000	0.045	0.236			
Model 4 Ensemble											0.378	0.368	
No Injury	35553	16297	59		70.590	0.832	0.451	0.194	0.570	0.315			
Injury	26852	67271	827	70.020	70.060	1.136	0.605	0.437	0.805	0.292			0.213
Fatality	0	10	46		99.390	0.060	0.049	0.000	0.049	0.179			

Notes:
[1] $B > 1$: class is overpredicted; $B < 1$: class is underpredicted;
[2] $CSI = 1$: perfect score; $CSI = 0$: no skill;
[3] $F = 0$: perfect score;
[4] $POD = 1$: perfect score;
[5] $FAR = 0$: perfect score;
[6] $HSS = 1$: perfect score; $HSS = 0$: no skill; $HSS < 0$: random is better;
[7] $PSS = 1$: perfect score; $PSS = 0$: no skill;
[8] $GS = 1$: perfect score; $GS = 0$: no skill.

5. Best model: insights

Create a table with the results of Model 2.

6. Further considerations

Here I plan to discuss the applicability of the modelling strategy to advanced modelling techniques (partial proportional odds, heterogeneity, hierarchical models, etc.)

7. Concluding remarks

Different modelling strategies can be used to model complex hierarchical, multievent outcomes such as the severity of injuries following a collision. The objective of this paper was to assess the performance of different strategies to model opponent effects in two-vehicle crashes. In broad terms, three strategies were considered: 1) incorporating opponent-level variables in the model; 2) single- versus multi-level model specifications; and 3) sample subsetting and estimation of separate models for different types of individual-opponent interactions.

The results of the empirical assessment strongly suggest that incorporating opponent effects can greatly improve the fit and predictive performance of a model. There is some evidence that subsetting the sample can

Table 11: Assessment of in-sample outcomes (nowcasting using 2017 dataset, i.e., estimation dataset)

Observed	Predicted Outcome			Verification Statistics									
Outcome	No Injury	Injury	Fatality	Percent Correct	Percent Correct by Class	Bias[1]	Critical Success Index[2]	Probability of False Detection[3]	Probability of Detection[4]	False Alarm Ratio[5]	Heidke Skill Score[6]	Peirce Skill Score[7]	Gerrity Score[8]
Model 1													
No Injury	61684	27447	184		68.529	0.922	0.495	0.269	0.637	0.309			
Injury	35171	74073	915	68.028	68.123	1.084	0.538	0.368	0.729	0.328	0.363	0.360	0.188
Fatality	5	85	10		99.404	0.090	0.008	0.000	0.009	0.900			
Model 2													
No Injury	62735	21013	106		72.319	0.866	0.532	0.206	0.648	0.252			
Injury	34125	80569	996	71.808	71.862	1.139	0.589	0.358	0.793	0.304	0.437	0.434	0.224
Fatality	0	23	7		99.436	0.027	0.006	0.000	0.006	0.767			
Model 3													
No Injury	62248	21133	107		72.014	0.862	0.527	0.207	0.643	0.254			
Injury	34610	80433	996	71.496	71.550	1.142	0.586	0.363	0.792	0.307	0.431	0.427	0.221
Fatality	2	39	6		99.427	0.042	0.005	0.000	0.005	0.872			
Model 4													
No Injury	62788	21247	102		72.230	0.869	0.531	0.208	0.648	0.254			
Injury	34071	80331	1000	71.716	71.767	1.136	0.588	0.358	0.791	0.304	0.435	0.432	0.223
Fatality	1	27	7		99.434	0.032	0.006	0.000	0.006	0.800			
Model 1 Ensemble													
No Injury	42896	20230	95		69.259	0.816	0.439	0.201	0.554	0.321			
Injury	34546	79692	962	68.669	68.731	1.152	0.588	0.452	0.797	0.308	0.354	0.345	0.183
Fatality	15	91	15		99.349	0.113	0.013	0.001	0.014	0.876			
Model 2 Ensemble													
No Injury	43486	19937	95		69.758	0.820	0.446	0.198	0.561	0.315			
Injury	33953	80009	961	69.163	69.227	1.149	0.593	0.445	0.800	0.304	0.364	0.355	0.188
Fatality	20	103	15		99.340	0.129	0.013	0.001	0.014	0.891			
Model 4 Ensemble													
No Injury	43560	20160	94		69.671	0.824	0.446	0.200	0.561	0.317			
Injury	33876	79762	959	69.074	69.141	1.146	0.591	0.444	0.800	0.304	0.363	0.354	0.189
Fatality	25	107	17		99.336	0.139	0.014	0.001	0.014	0.886			

Notes:
[1] $B > 1$: class is overpredicted; $B < 1$: class is underpredicted;
[2] $CSI = 1$: perfect score; $CSI = 0$: no skill;
[3] $F = 0$: perfect score;
[4] $POD = 1$: perfect score;
[5] $FAR = 0$: perfect score;
[6] $HSS = 1$: perfect score; $HSS = 0$: no skill; $HSS < 0$: random is better;
[7] $PSS = 1$: perfect score; $PSS = 0$: no skill;
[8] $GS = 1$: perfect score; $GS = 0$: no skill.

improve the results in some isolated situations (e.g., when modelling the severity of crashes involving active travellers or motorcyclists), possibly at the risk of overfitting. In this paper we did not compare individual models in our Ensemble approach, but we suggest that this is an avenue for future research.

In terms of the full sample models, the evidence was not conclusive in favor of a single-level model with opponent variables (Model 2), or a hierarchical model with individual-opponent interactions (Model 4). On the one hand, the AIC tended to favor the

References

Amoh-Gyimah, R., Aidoo, E.N., Akaateba, M.A., Appiah, S.K., 2017. The effect of natural and built environmental characteristics on pedestrian-vehicle crash severity in ghana. International Journal of Injury Control and Safety Promotion 24, 459–468. doi:10.1080/17457300.2016.1232274

Aziz, H.M.A., Ukkusuri, S.V., Hasan, S., 2013. Exploring the determinants of pedestrian-vehicle crash severity in new york city. Accident Analysis and Prevention 50, 1298–1309. doi:10.1016/j.aap.2012.09.034

Bogue, S., Paleti, R., Balan, L., 2017. A modified rank ordered logit model to analyze injury severity of occupants in multivehicle crashes. Analytic Methods in Accident Research 14, 22–40. doi:10.1016/j.amar.2017.03.001

Casetti, E., 1972. Generating models by the expansion method: Applications to geographic research. Geographical Analysis 4, 81–91.

Chang, L.Y., Wang, H.W., 2006. Analysis of traffic injury severity: An application of non-parametric classification tree techniques. Accident Analysis and Prevention 38, 1019–1027. doi:10.1016/j.aap.2006.04.009

Chen, F., Song, M.T., Ma, X.X., 2019. Investigation on the injury severity of drivers in rear-end collisions between cars using a random parameters bivariate ordered probit model. International Journal of Environmental Research and Public Health 16. doi:10.3390/ijerph16142632

Chiou, Y.C., Hwang, C.C., Chang, C.C., Fu, C., 2013. Modeling two-vehicle crash severity by a bivariate generalized ordered probit approach. Accident Analysis and Prevention 51, 175–184. doi:10.1016/j.aap.2012.11.008

Devlin, A., Beck, B., Simpson, P.M., Ekegren, C.L., Giummarra, M.J., Edwards, E.R., Cameron, P.A., Liew, S., Oppy, A., Richardson, M., Page, R., Gabbe, B.J., 2019. The road to recovery for vulnerable road users hospitalised for orthopaedic injury following an on-road crash. Accident Analysis and Prevention 132, 10. doi:10.1016/j.aap.2019.105279

Dissanayake, S., Lu, J.J., 2002. Factors influential in making an injury severity difference to older drivers involved in fixed object-passenger car crashes. Accident Analysis and Prevention 34, 609–618. doi:10.1016/s0001-4575(01)00060-4

Duddu, V.R., Penmetsa, P., Pulugurtha, S.S., 2018. Modeling and comparing injury severity of at-fault and not at-fault drivers in crashes. Accident Analysis and Prevention 120, 55–63. doi:10.1016/j.aap.2018.07.036

Effati, M., Thill, J.C., Shabani, S., 2015. Geospatial and machine learning techniques for wicked social science problems: Analysis of crash severity on a regional highway corridor. Journal of Geographical Systems 17, 107–135. doi:10.1007/s10109-015-0210-x

Gong, L.F., Fan, W.D., 2017. Modeling single-vehicle run-off-road crash severity in rural areas: Accounting for unobserved heterogeneity and age difference. Accident Analysis and Prevention 101, 124–134. doi:10.1016/j.aap.2017.02.014

Haleem, K., Gan, A., 2013. Effect of driver's age and side of impact on crash severity along urban freeways: A mixed logit approach. Journal of Safety Research 46, 67–76. doi:10.1016/j.jsr.2013.04.002

Hanson, C.S., Noland, R.B., Brown, C., 2013. The severity of pedestrian crashes: An analysis using google street view imagery. Journal of Transport Geography 33, 42–53. doi:10.1016/j.jtrangeo.2013.09.002

Hedeker, D., Gibbons, R.D., 1994. A random-effects ordinal regression-model for multilevel analysis. Biometrics 50, 933–944.

Iranitalab, A., Khattak, A., 2017. Comparison of four statistical and machine learning methods for crash severity prediction. Accident Analysis and Prevention 108, 27–36. doi:10.1016/j.aap.2017.08.008

Islam, S., Jones, S.L., Dye, D., 2014. Comprehensive analysis of single- and multi-vehicle large truck at-fault crashes on rural and urban roadways in alabama. Accident Analysis & Prevention 67, 148–158. doi:https://doi.org/10.1016/j.aap.2014.02.014

Khan, G., Bill, A.R., Noyce, D.A., 2015. Exploring the feasibility of classification trees versus ordinal discrete choice models for analyzing crash severity. Transportation Research Part C-Emerging Technologies 50, 86–96. doi:10.1016/j.trc.2014.10.003

Kim, J.K., Ulfarsson, G.F., Kim, S., Shankar, V.N., 2013. Driver-injury severity in single-vehicle crashes in california: A mixed logit analysis of heterogeneity due to age and gender. Accident Analysis and Prevention 50, 1073–1081. doi:10.1016/j.aap.2012.08.011

Kim, K., Nitz, L., Richardson, J., Li, L., 1995. PERSONAL and behavioral predictors of automobile crash and injury severity. Accident Analysis and Prevention 27, 469–481. doi:10.1016/0001-4575(95)00001-g

Lee, C., Li, X.C., 2014. Analysis of injury severity of drivers involved in single- and two-vehicle crashes on highways in ontario. Accident Analysis and Prevention 71, 286–295. doi:10.1016/j.aap.2014.06.008

Li, L., Hasnine, M.S., Habib, K.M.N., Persaud, B., Shalaby, A., 2017. Investigating the interplay between the attributes of at-fault and not-at-fault drivers and the associated impacts on crash injury occurrence and severity level. Journal of Transportation Safety & Security 9, 439–456. doi:10.1080/19439962.2016.1237602

Ma, J.M., Kockelman, K.M., Damien, P., 2008. A multivariate poisson-lognormal regression model for prediction of crash counts by severity, using bayesian methods. Accident Analysis and Prevention 40, 964–975. doi:10.1016/j.aap.2007.11.002

Maddala, G.S., 1986. Limited-dependent and qualitative variables in econometrics. Cambridge university press.

McArthur, A., Savolainen, P.T., Gates, T.J., 2014. Spatial analysis of child pedestrian and bicycle crashes development of safety performance function for areas adjacent to schools. Transportation Research Record 57–63. doi:10.3141/2465-08

Merlin, E.P.R., Gonzalez-Forteza, C., Lira, L.R., Tapia, J.A.J., 2007. Post-traumatic stress disorder in patients with non intentional injuries caused by road traffic accidents. Salud Mental 30, 43–48.

Montella, A., Andreassen, D., Tarko, A.P., Turner, S., Mauriello, F., Imbriani, L.L., Romero, M.A., 2013. Crash databases in australasia, the european union, and the united states review and prospects for improvement. Transportation Research Record 128–136. doi:10.3141/2386-15

Mooradian, J., Ivan, J.N., Ravishanker, N., Hu, S., 2013. Analysis of driver and passenger crash injury severity using partial proportional odds models. Accident Analysis and Prevention 58, 53–58. doi:10.1016/j.aap.2013.04.022

Mussone, L., Bassani, M., Masci, P., 2017. Analysis of factors affecting the severity of crashes in urban road intersections. Accident Analysis and Prevention 103, 112–122. doi:10.1016/j.aap.2017.04.007

Obeng, K., 2011. Gender differences in injury severity risks in crashes at signalized intersections. Accident Analysis and Prevention 43, 1521–1531. doi:10.1016/j.aap.2011.03.004

Osman, M., Mishra, S., Paleti, R., 2018. Injury severity analysis of commercially-licensed drivers in single-vehicle crashes: Accounting for unobserved heterogeneity and age group differences. Accident Analysis and Prevention 118, 289–300. doi:10.1016/j.aap.2018.05.004

Peek-Asa, C., Britton, C., Young, T., Pawlovich, M., Falb, S., 2010. Teenage driver crash incidence and factors influencing crash injury by rurality. Journal of Safety Research 41, 487–492. doi:10.1016/j.jsr.2010.10.002

Pelissier, C., Fort, E., Fontana, L., Hours, M., n.d. Medical and socio-occupational predictive factors of psychological distress 5 years after a road accident: A prospective study. Social Psychiatry and Psychiatric Epidemiology 13. doi:10.1007/s00127-019-01780-0

Penmetsa, P., Pulugurtha, S.S., Duddu, V.R., 2017. Examining injury severity of not-at-fault drivers in two-vehicle crashes. Transportation Research Record 164–173. doi:10.3141/2659-18

Rakotonirainy, A., Steinhardt, D., Delhomme, P., Darvell, M., Schramm, A., 2012. Older drivers' crashes in queensland, australia. Accident Analysis and Prevention 48, 423–429. doi:10.1016/j.aap.2012.02.016

Rana, T.A., Sikder, S., Pinjari, A.R., 2010. Copula-based method for addressing endogeneity in models of severity of traffic crash injuries application to two-vehicle crashes. Transportation Research Record 75–87. doi:10.3141/2147-10

Regev, S., Rolison, J.J., Moutari, S., 2018. Crash risk by driver age, gender, and time of day using a new exposure methodology. Journal of Safety Research 66, 131–140. doi:10.1016/j.jsr.2018.07.002

Rifaat, S.M., Chin, H.C., 2007. Accident severity analysis using ordered probit model. Journal of Advanced Transportation 41, 91–114. doi:10.1002/atr.5670410107

Salon, D., McIntyre, A., 2018. Determinants of pedestrian and bicyclist crash severity by party at fault in san francisco, ca. Accident Analysis and Prevention 110, 149–160. doi:10.1016/j.aap.2017.11.007

Sasidharan, L., Menendez, M., 2014. Partial proportional odds model-an alternate choice for analyzing pedestrian crash injury severities. Accident Analysis and Prevention 72, 330–340. doi:10.1016/j.aap.2014.07.025

Savolainen, P., Mannering, F., 2007. Probabilistic models of motorcyclists' injury severities in single- and multi-vehicle crashes. Accident Analysis and Prevention 39, 955–963. doi:10.1016/j.aap.2006.12.016

Savolainen, P.T., Mannering, F., Lord, D., Quddus, M.A., 2011. The statistical analysis of highway crash-injury severities: A review and assessment of methodological alternatives. Accident Analysis and Prevention 43, 1666–1676. doi:10.1016/j.aap.2011.03.025

Shaheed, M.S.B., Gkritza, K., Zhang, W., Hans, Z., 2013. A mixed logit analysis of two-vehicle crash seventies involving a motorcycle. Accident Analysis and Prevention 61, 119–128. doi:10.1016/j.aap.2013.05.028

Shamsunnahar, Y., Eluru, N., Pinjari, A.R., Tay, R., 2014. Examining driver injury severity in two vehicle crashes - a copula based approach. Accident Analysis and Prevention 66, 120–135. doi:10.1016/j.aap.2014.01.018

Symons, J., Howard, E., Sweeny, K., Kumnick, M., Sheehan, P., 2019. Reduced road traffic injuries for young people: A preliminary investment analysis. Journal of Adolescent Health 65, S34–S43. doi:10.1016/j.jadohealth.2019.01.009

Tarrao, G.A., Coelho, M.C., Rouphail, N.M., 2014. Modeling the impact of subject and opponent vehicles on crash severity in two-vehicle collisions. Transportation Research Record 53–64. doi:10.3141/2432-07

Tay, R., Choi, J., Kattan, L., Khan, A., 2011. A multinomial logit model of pedestrian-vehicle crash severity. International Journal of Sustainable Transportation 5, 233–249. doi:10.1080/15568318.2010.497547

Thompson, J.P., Baldock, M.R.J., Dutschke, J.K., 2018. Trends in the crash involvement of older drivers in australia. Accident Analysis and Prevention 117, 262–269. doi:10.1016/j.aap.2018.04.027

Train, K., 2009. Discrete choice methods with simulation, 2nd Edition. ed. Cambridge University Press, Cambridge.

Wang, K., Yasmin, S., Konduri, K.C., Eluru, N., Ivan, J.N., 2015. Copula-based joint model of injury severity and vehicle damage in two-vehicle crashes. Transportation Research Record 158–166. doi:10.3141/2514-17

Wang, X.K., Kockelman, K.M., 2005. Use of heteroscedastic ordered logit model to study severity of occupant injury - distinguishing effects of vehicle weight and type, in: Statistical Methods; Highway Safety Data, Analysis, and Evaluation; Occupant Protection; Systematic Reviews and Meta-Analysis, Transportation Research Record. pp. 195–204.

White, S., Clayton, S., 1972. Some effects of alcohol, age of driver, and estimated speed on the likelihood of driver injury. Accident Analysis & Prevention 4.

Wijnen, W., Weijermars, W., Schoeters, A., Berghe, W. van den, Bauer, R., Carnis, L., Elvik, R., Martensen, H., 2019. An analysis of official road crash cost estimates in european countries. Safety Science 113, 318–327. doi:10.1016/j.ssci.2018.12.004

World Health Organization, 2019. Global status report on road safety 2018 (2018). Geneva.

Wu, Q., Chen, F., Zhang, G.H., Liu, X.Y.C., Wang, H., Bogus, S.M., 2014. Mixed logit model-based driver injury severity investigations in single- and multi-vehicle crashes on rural two-lane highways. Accident Analysis and Prevention 72, 105–115. doi:10.1016/j.aap.2014.06.014

Yasmin, S., Eluru, N., 2013. Evaluating alternate discrete outcome frameworks for modeling crash injury severity. Accident Analysis & Prevention 59, 506–521. doi:https://doi.org/10.1016/j.aap.2013.06.040