Environmental Conditions and Driver Behavior Shape Collision Outcomes in Toronto*

A Two-Stage Bayesian Analysis Reveals Infrastructure Factors Predict Crash Occurrence While Driver Actions Determine Severity

Yitong Wang

November 26, 2024

This study employs a two stage bayesian modeling approach to analyze vehicle collision patterns in Toronto, examining how environmental conditions and driver behavior influence both injury occurrence and severity. The first model predicts the probability of injury based on environmental and infrastructural factors, while the second estimates injury severity using driver characteristics and behavior. Analysis of historical collision data from 2006 to 2023 reveals that while environmental conditions primarily influence injury occurrence, driver behavior factors like speeding and alcohol use more strongly affect severity levels. Our dual model analysis providing evidence based guidance for targeting both infrastructure improvements and enforcement strategies to reduce severe collision outcomes.

Table of contents

*Code and data are available at: https://github.com/demainwang/Toronto_vehicle_collision_analysis.

1 Introduction

Vehicle collisions in urban environments represent a significant public health and safety challenge, with particular concern in large metropolitan areas like Toronto where complex traffic patterns and diverse environmental conditions create varying levels of risk. Understanding the factors that contribute to collision occurrence and severity is crucial for urban planning and public safety policy, yet traditional analyses often treat these as a single outcome rather than recognizing their distinct underlying processes. The increasing availability of detailed collision data, combined with advances in statistical modeling, presents an opportunity to develop more nuanced understandings of how different factors influence collision outcomes.

This paper aims to address this challenge through a two-stage modeling approach, with the estimand being the predicted probability of injury occurrence and, conditional on an injury occurring, its severity level. Specifically, our first model examines how environmental and infrastructural factors influence the likelihood of injury in a collision, while the second model investigates how driver characteristics and behavior affect injury severity when injuries do

occur. By separating these processes, we can better understand how different factors contribute to collision outcomes and identify more targeted intervention opportunities.

Our analysis reveals several distinct and important patterns in how different factors influence collision outcomes. Environmental conditions emerged as the dominant predictors of whether an injury occurs in the first place - particularly reduced visibility during winter months and poor road surface conditions on major arterials. However, once an injury occurs, driver behavior becomes the crucial determinant of severity. Impaired driving and speeding showed strong associations with severe outcomes, especially during nighttime hours and weekends. Road classification provided an interesting counterintuitive finding: while major arterial roads saw more frequent injuries, the injuries on local roads tended to be more severe. Temporal analysis uncovered clear risk patterns, with injury probability peaking during adverse weather conditions on arterials, while severity levels were highest during weekend nights with notable correlations to driver behavior violations. Pedestrian and cyclist collisions showed a particularly concerning pattern, even overall less frequent, they consistently resulted in more severe injuries across all conditions.

By separating environmental triggers from behavioral factors, we uncovered distinct patterns that shape Toronto's traffic safety landscape. Our analysis shows how road design and maintenance decisions directly influence where injuries occur, while enforcement timing determines crash severity. Understanding these relationships helps transportation departments focus their resources where they matter most, from targeted winter maintenance on high risk arterials to enforcement during periods when severe crashes are most likely. Beyond potential safety improvements, this research advances our understanding of urban traffic patterns and provides a foundation for evidence-based policy decisions.

The remainder of this paper is structured as follows: Section 2 presents the overview of data (Section 2.1), data cleaning process (Section 2.2), measurement (Section 2.3), also for explanations, descriptions, graph summaries of outcome (Section 2.4) and predictor variables (Section 2.5) of the study. Section 3 set up the model and explain the process detailed, including procedure of model set up (Section 3.1) and model validation (Section 3.2), and also mention the assumption and validation (Section 3.3) of models. In addition, (Section 3.4) discuss the potential alternative models and its trade offs. Then, Section 4 reveal the prediction outcome and models' performance. Section 5 discusses the relation between the data and real world, which talks about the data quality and bias, as well as the potential limitation of our models in a broader context.

Appendix includes three parts: Appendix .1 presents collision data methodology overview, Appendix .2 provides detailed idealized methodology for collection improvement; Appendix .3 presents additional summary and validation in modeling process.

2 Data

2.1 Overview

The data for this analysis comes from Toronto's Motor Vehicle Collision dataset (City of Toronto 2024), accessed through the Toronto Open Data (Gelfand 2022). The raw dataset contains 14,976 records of vehicle collisions involving killed or different levels injured persons from 2006 to 2023, with 52 variables capturing various aspects of each incident. These include collision characteristics, environmental conditions, road features, and driver factors.

In this study, we utilized statistical language R (R Core Team 2023) and several key packages for data processing, analysis, and visualization. The tidyverse package (Wickham et al. 2019) facilitated data manipulation, with dplyr (Wickham et al. 2023) being particularly crucial for data transformation and filtering operations, while visualization was conducted using ggplot2 (Wickham 2016). For efficient and consistent data import, we employed the readr package (Wickham, Hester, and François 2023). For handling temporal aspects of collision data, we employes lubridate {Grolemund and Wickham (2011)} to process dates and times.We used here (Müller 2023) for reproducible file path management, arrow (Richardson et al. 2023) for efficient data storage and retrieval, and knitr (Xie 2023) used to provide the better formatting output. For unit testing and code validation, we employed the testthat package (Wickham, Hester, and Chang 2023) which provided a robust framework for testing our functions and ensuring code reliability. Brms package (Bürkner 2023) plays the significant values for our model construction. For Bayesian analysis visualization, we employed bayesplot (Gabry et al. 2023) to create diagnostic plots including MCMC trace plots. For model evaluation, we employed the pROC package (Robin et al. 2023) to generate receiver operating characteristic ROC curves and assess model performance. For model visualization and interpretation, we used the modelsummary package (Arel-Bundock 2023) to create coefficient plots and model summaries.

2.2 Data Cleaning

From the raw dataset, we selected 14 key variables most relevant to understanding collision severity patterns: date, time, road_class (road classification), accloc (accident location), traffctl (traffic control), visibility, light, rdsfcond (road surface condition), impactype (impact type), invage (involved person's age), injury (severity), speeding, ag_driv (aggressive driving), alcohol, and disability. Records with missing values in critical fields such as injury severity, visibility, or road classification were removed to ensure data quality.

In addition to removing rows with missing NA values, we also excluded records where key categorical variables were explicitly recorded as "None" despite an incident occurring. This includes records where accident location (accloc), road classification (road_class), traffic control (traffctl), visibility, light conditions (light), road surface condition (rdsfcond), and impact

type (impactype) were recorded as "None". These "None" entries differ from missing values as they were actively recorded but indicate potentially incomplete or unclear documentation of the collision circumstances. This additional cleaning step ensures our analysis is based on collisions with well-documented characteristics, though it reduced our dataset by approximately 8% (1,057 records). This filtering helps maintain data quality and supports more reliable analysis of factor relationships with collision severity.

2.3 Measurement

The transformation of real-world collision incidents into structured data points follows a rigorous protocol established by Toronto Police Services. When a collision occurs, first responders and investigating officers collect data through a standardized Motor Vehicle Accident Report Form. This form ensures systematic documentation of collision characteristics, converting complex real-world events into quantifiable measurements. For instance, injury severity classification follows specific criteria: fatalities are recorded when death occurs within 30 days of collision, major injuries are those requiring hospitalization, minor injuries need medical attention but no hospitalization, and minimal injuries indicate no significant medical intervention was required. This standardized approach helps maintain consistency in severity classification across different incidents and responding officers (Services n.d.).

Environmental and situational measurements are recorded at the time of collision through both objective and subjective assessments. Objective measurements include temporal factors (date, time), location (coordinates, road type), and certain environmental conditions (light, road conditions). However, some measurements inherently involve officer judgment, such as assessing visibility conditions (Clear, Rain, Snow) or driver behavior (aggressive driving, speeding). These assessments, while guided by training and protocols, may introduce some variability in how similar conditions are classified across different incidents (Transportation 2023). For example, what one officer considers "aggressive driving" might be classified differently by another, highlighting the importance of considering potential measurement inconsistencies in our analysis.

Some driver related factors present particular measurement challenges due to their complex nature. Alcohol involvement is measured through standardized tests when possible, and it providing more objective data. However, other driver conditions like fatigue or distraction often rely on post incident interviews and officer's observation, which introducing potential reporting biases. Additionally, factors like driver age and license status are verified through official documentation, ensuring accuracy in demographic data. The measurement process also captures infrastructure characteristics such as road classification and traffic control presence, which are determined through existing municipal records and on-site verification. This multifaceted approach to data collection, while comprehensive, means that the quality and reliability of different variables may vary depending on the objectivity of their measurement methods.

2.4 Outcome variables

Our primary outcome variable is injury severity, which classifies the most severe injury in each collision incident into five ordered categories. These categories progress from "None" (no significant injury), through "Minimal" (minor scrapes or bruises requiring no medical attention), "Minor" (requiring some medical attention but no hospitalization), "Major" (requiring hospitalization), to "Fatal" (death occurring within 30 days of the collision).

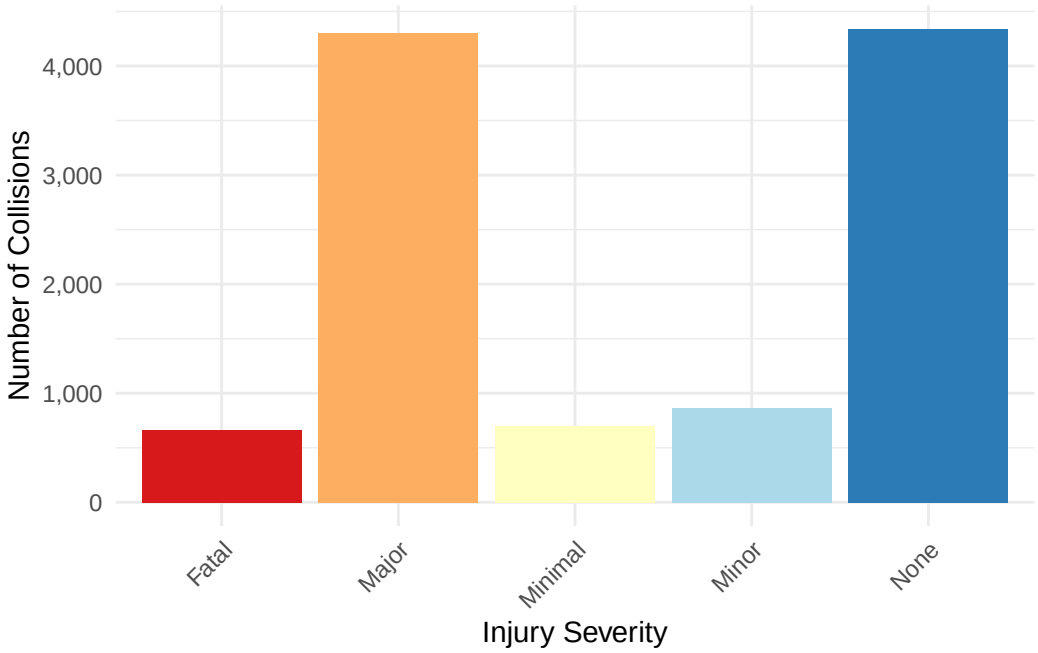

Figure 1: Outcome frequency

Figure 1 The primary outcome variable in our analysis is collision severity, represented through an ordered factor ranging from "None" to "Fatal". It illustrates the distribution of injury severity across all recorded incidents. The data reveals an interesting bimodal distribution in collision outcomes, with concentrations at both ends of the severity spectrum. Major injuries and no injuries represent the most frequent outcomes, accounting for 40.1% and 38.9% of incidents respectively. Fatal collisions comprise 6.3% of the dataset, while minor and minimal injuries together represent approximately 15% of cases.

2.5 Predictor variables

This study incorporates thirteen predictor variables spanning temporal, environmental, infrastructural, and human factors. Temporal aspects are captured through date and time of

collision occurrence. Environmental conditions are measured through three variables: visibility (recording conditions such as Clear, Rain, or Snow), light (indicating natural light conditions like Daylight, Dusk, or Dark), and rdsfcond (road surface condition, noting whether the road was Dry, Wet, or covered in Snow/Ice). Infrastructure characteristics are represented by road_class (categorizing roads as Major Arterial, Minor Arterial, Collector, Local, or Expressway), accloc (specifying whether the collision occurred at an intersection, mid-block, or other location), and traffctl (indicating the type of traffic control present, such as signals or signs). The human element is captured through five variables: invage (age of persons involved), speeding (indicating whether speeding was a factor), ag_driv (noting aggressive driving behavior), alcohol (indicating alcohol involvement), and disability (noting any relevant disabilities). Finally, impactype describes the nature of the collision, such as rear-end, angle, or turning movement.

2.5.1 Environmental Conditions

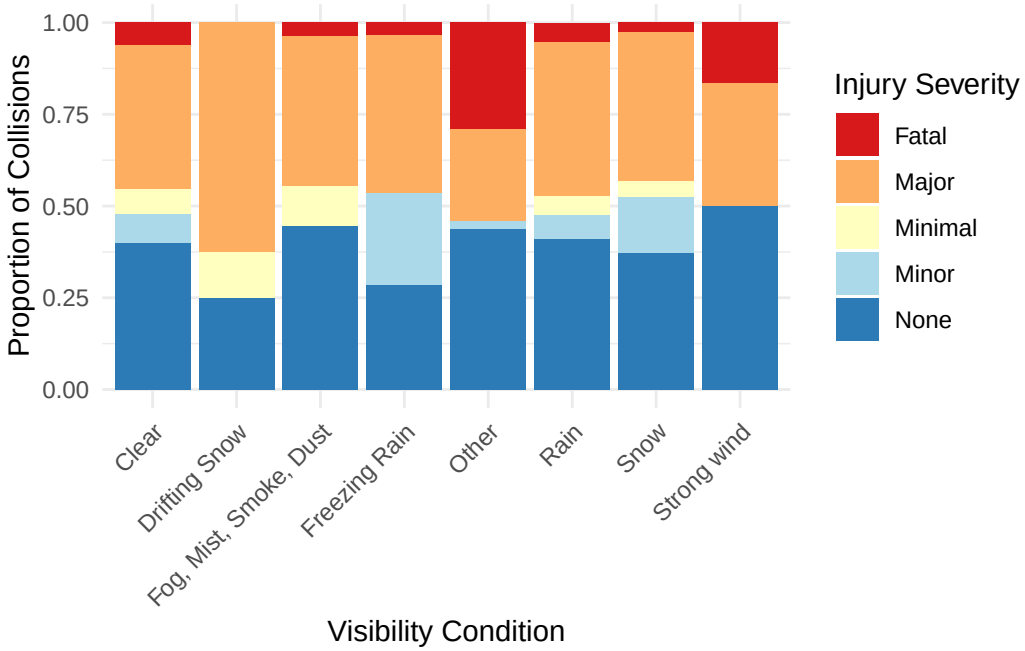

Figure 2: Proportion of collisions severity in different visibility conditions

Figure 2 Visibility factor plays a crucial role in collision outcomes. Figure 2 presents the relationship between visibility conditions and collision severity. Clear conditions account for the majority of recorded incidents (86.4%), while poor visibility conditions (including rain, snow, and freezing rain) represent 10.5% of cases. Notably, the proportion of severe injuries increases under poor visibility conditions, suggesting a significant relationship between environmental conditions and collision outcomes.

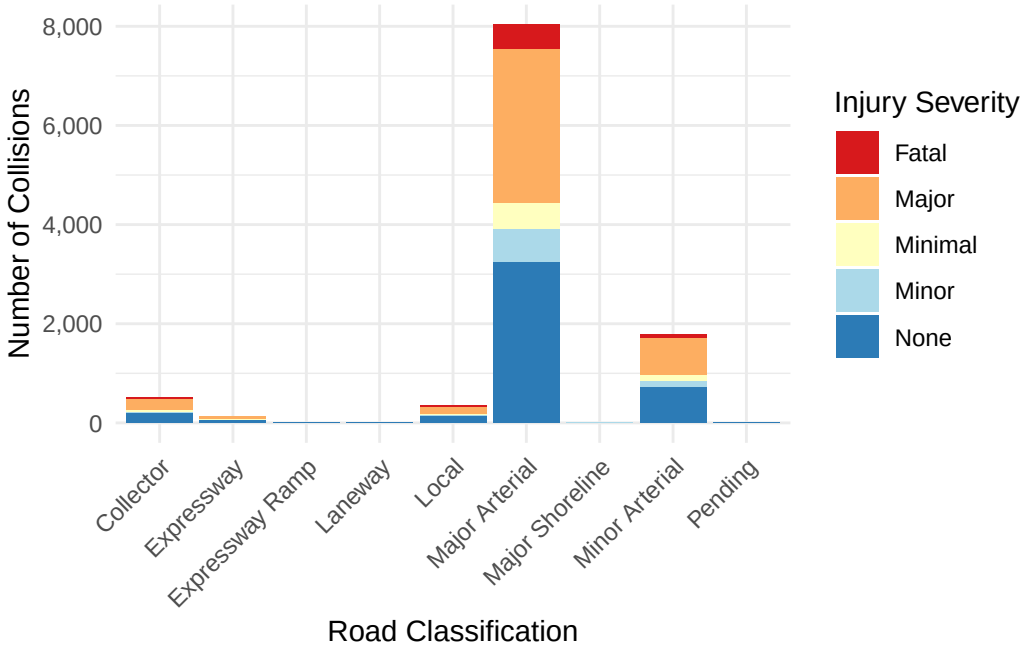

Figure 3: Frequency of different injury severity in different road class

Figure 3 Road classification and traffic control measures represent important infrastructure factors in our analysis. Figure 3 illustrates the distribution of collisions across different road types and their associated severity levels. Major arterial roads account for 71.3% of all collisions, followed by minor arterial roads at 15.7%. This distribution likely reflects both traffic volume and speed limits on these road types. The relationship between road classification and injury severity reveals important patterns for urban safety planning.

Figure 4 The facet grid visualization reveals the interaction between visibility and light conditions on injury severity. Under poor visibility (fog), dark conditions show a noticeably higher proportion of severe injuries (Fatal, Major) compared to daylight, indicating the compounded risk of low visibility and inadequate lighting. In contrast, clear visibility with daylight consistently shows the lowest proportions of severe injuries, emphasizing the protective role of good environmental conditions. These findings will help guide the inclusion of visibility and light in the predictive model, ensuring it accounts for the impact of environmental factors on injury outcomes. This approach will improve the model's ability to identify scenarios with higher risks of severe injuries.

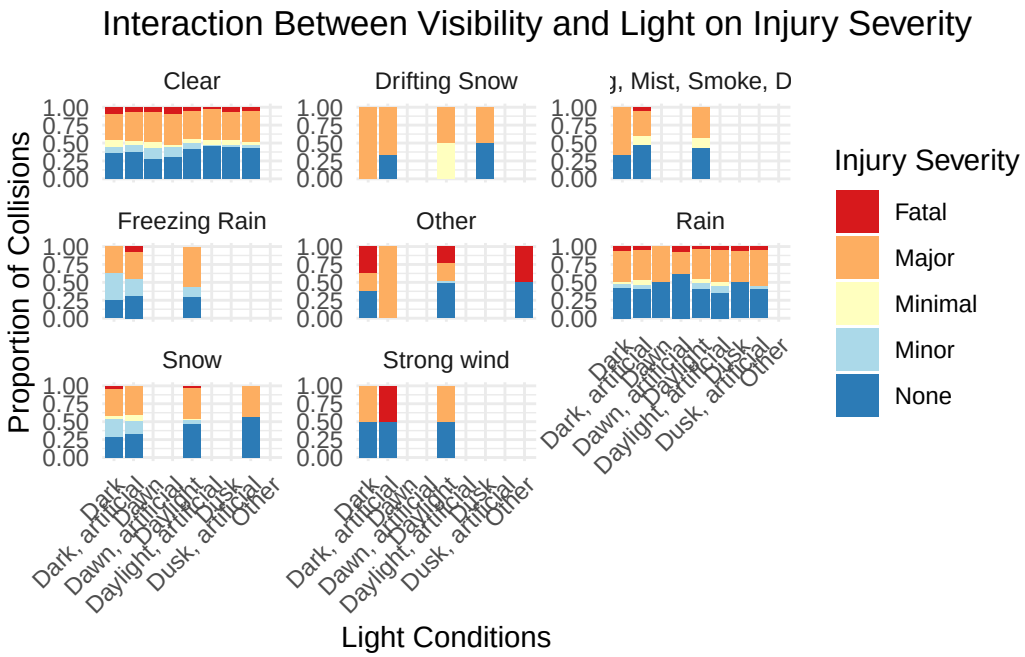

Figure 4: Frequency of different injury severity in different combination of visibility and light

2.5.2 Driver conditions

Figure 5 The proportional line chart compares injury severity distributions for drivers involved in risky behaviors (speeding, alcohol use, aggressive driving) versus those who were not. It reveals that drivers engaged in these behaviors often have higher proportions of severe injuries, such as Fatal or Major, compared to those who did not. However, the results may be influenced by sample size imbalances, as there are likely fewer drivers in the "Yes" group for alcohol or aggressive driving, which could exaggerate proportions and introduce bias. Despite these limitations, the chart provides valuable insights into the heightened risks associated with these behaviors

Figure 6 The stacked bar chart reveals the proportion of injury severities across the top five collision impact types, providing insight into how different types of collisions relate to injury outcomes. For instance, pedestrian collisions show a significantly higher proportion of severe injuries (Fatal and Major), while rear-end collisions are dominated by minor or no injuries. This exploration is critical for our paper as it identifies which collision types pose the greatest risk of severe outcomes, helping us understand the underlying patterns of injury severity. By focusing on this relationship, we address a core question of our study: how specific predictors, such as impact type, influence injury severity.

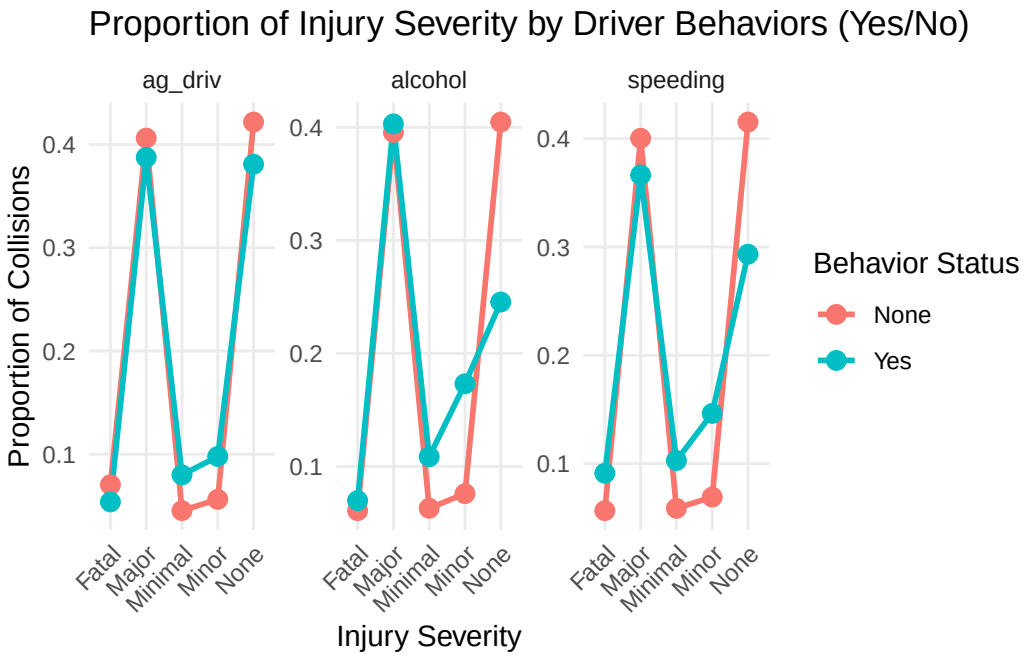

Figure 5: Proportion of Injury Severity by whether Drivers have bad Behaviors

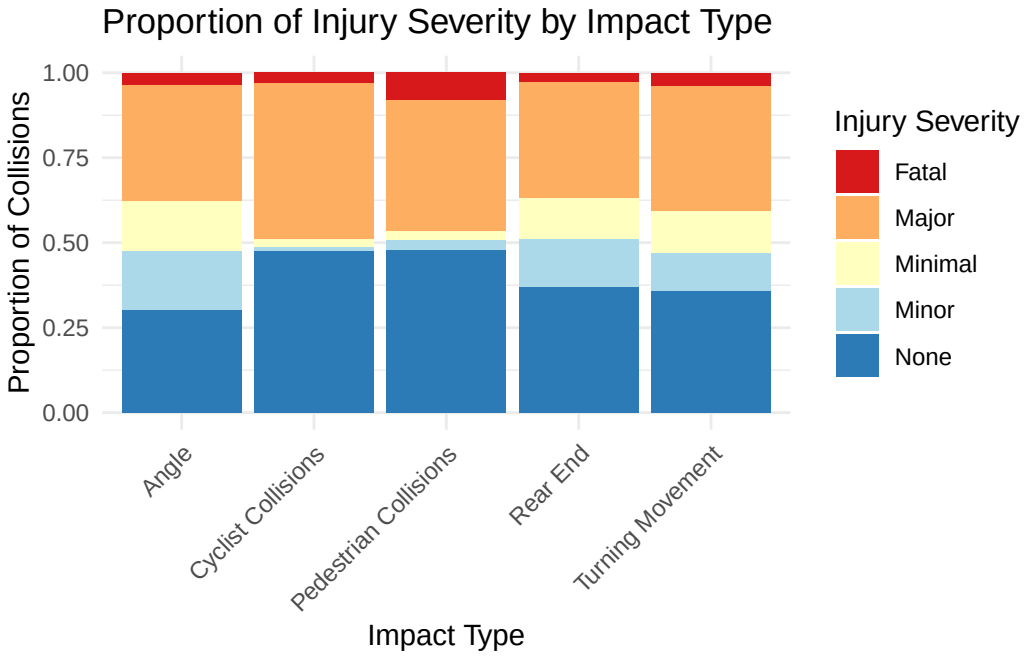

Figure 6: Injury proportion in different impactypes

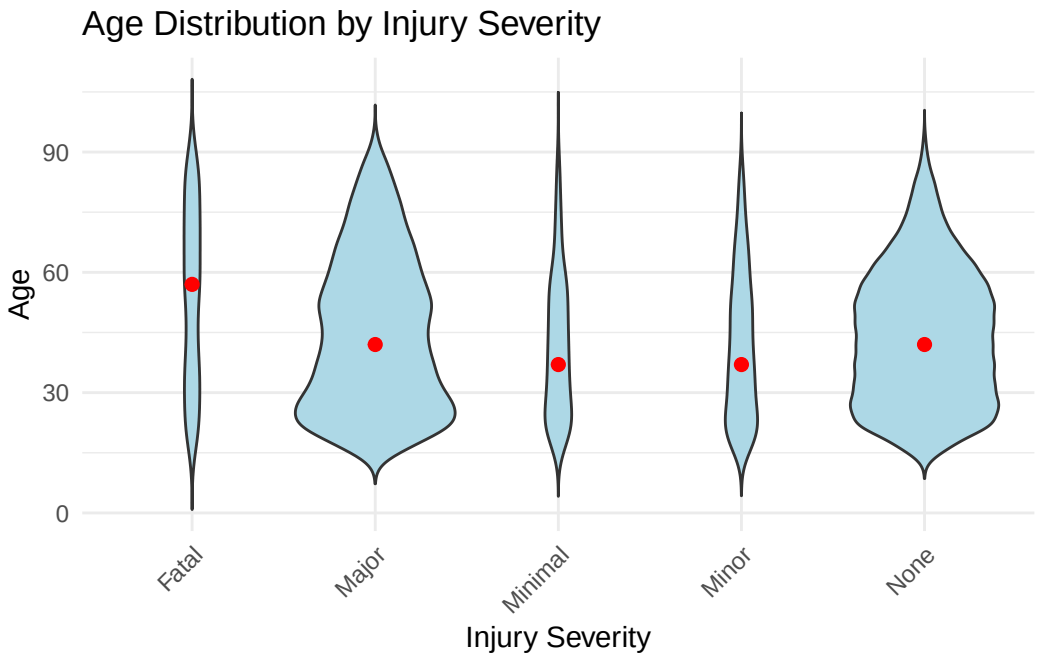

Figure 7: Age Distribution by Injury Severity

Figure 7 The violin plot may show younger individuals having a higher density in some severer injury category such as "Major", indicating they are overrepresented in severe collisions. This trend can result from several factors: younger people are often more active on the roads, increasing their exposure to risk, and they may engage in riskier behaviors like speeding or distracted driving. Their lack of driving experience and the use of older, less safe vehicles can also contribute to more severe outcomes.

3 Model

This study employs a two-stage modeling approach to understand and predict traffic injury outcomes, a model for injury occurrence prediction and another model for injury severity prediction when injuries occur. Both models were implemented using the brms package (Bürkner 2023) in R, leveraging its robust framework for Bayesian regression modeling. The dual-model Bayesian framework precisely analyze traffic injuries, separating the prediction of injury occurrence from injury severity assessment. This approach allows us to isolate and examine the distinct factors that influence whether an injury occurs and, subsequently, how severe that injury becomes. By developing separate models for these two aspects, we can better understand the complex dynamics of traffic accidents and provide more targeted insights for safety interventions.

3.1 Model set-up

3.1.1 Injury occurence model

For our injury occurrence model, we implemented a fundamental binary transformation that distinguishes between collisions resulting in any injury and those without injury. In this classification scheme, we designated all collisions with "None" injury classification as our negative outcome ("No injury occurred"). Conversely, we grouped all other injury levels - Minimal, Minor, Major, and Fatal - into our positive outcome ("Injury occurred"). This transformation collapses the rich variety of injury outcomes into a simple binary indicator focusing solely on injury occurrence.

This model addresses the binary outcome of injury occurrence through a Bayesian logistic regression framework. Let Y_i represent the injury outcome for observation i, where $Y_i = 1$ indicates an injury occurred and $Y_i = 0$ indicates no injury. The probability of injury is modeled through a Bernoulli distribution:

$$Y_i = \begin{cases} 0 & \text{if no injury occurred (original classification: None)} \\ 1 & \text{if any injury occurred (original classifications: Minimal, Minor, Major, or Fatal)} \end{cases}$$

$$Y_i \sim \textbf{Bernoulli}(p_i) \tag{1}$$

$$\textbf{logit}(p_i) = \beta_0 + \beta_1 \textbf{Visibility}_i + \beta_2 \textbf{Light}_i + \beta_3 \textbf{RoadCondition}_i + \beta_4 \textbf{RoadClass}_i + \beta_5 \textbf{Location}_i + \beta_6 \textbf{TrafficControl}_i + \beta_7 \textbf{RushHour}_i + \beta_8 \textbf{Weekend}_i + \beta_9 \textbf{ImpactType}_i$$

The occurrence model prioritizes environmental and infrastructural factors, drawing on established traffic safety literature and the patterns observed in our exploratory data analysis.

Environmental conditions are represented through visibility, light conditions, and road surface state, each categorized to capture fundamental differences while maintaining model parsimony. We consolidate visibility into a binary indicator (Good/Poor) based on detailed weather conditions, simplify lighting into day/night categories from six original conditions, and classify road conditions as good or poor based on surface characteristics. Infrastructure factors incorporate the physical and operational aspects of the road environment. Road classification follows a hierarchical structure (Major/Minor/Local) that reflects traffic volume and design standards. Location type captures the crucial distinction between intersections, private drives, and other locations, while traffic control variables account for the presence and type of traffic management systems. Additionally, temporal factors such as rush hour periods and weekend indicators are included to capture systematic variations in traffic patterns.

3.1.2 Injury severity model

Our severity model employs a more nuanced approach, focusing exclusively on collisions where injuries occurred. We first excluded all cases classified as "None" injury, as these fall outside the scope of severity analysis. Among injury cases, we created a binary classification that distinguishes between severe and minor injuries. We classified "Fatal" and "Major" injuries as "Severe" outcomes, while grouping "Minimal" and "Minor" injuries as "Minor" outcomes.

For cases where injuries occur, our second model examines severity levels using an ordered logistic regression approach. The severity outcome S_i is modeled as:

$$S_i = \begin{cases} 0 & \text{if injury was less severe (original classifications: Minimal, Minor)} \\ 1 & \text{if injury was severe (original classifications: Major, Fatal)} \end{cases}$$

$$S_i \sim \textbf{Bernoulli}(q_i) \tag{2}$$

$$\textbf{logit}(q_i) = \alpha_0 + \alpha_1 \textbf{Speeding}_i + \alpha_2 \textbf{AggressiveDriving}_i + \alpha_3 \textbf{Alcohol}_i + \alpha_4 \textbf{Disability}_i + \alpha_5 \textbf{AgeGroup}_i + \alpha_6 \textbf{Impact}_i + \alpha_7 \textbf{Light}_i + \alpha_8 \textbf{RoadClass}_i$$

The severity model emphasizes driver behavioral and impact factors that influence injury outcomes once an accident occurs. Driver behavior variables include binary indicators for speeding, aggressive driving, alcohol involvement, and driver disability. Age groups are categorized to capture the varying vulnerability and risk patterns across different life stages, while maintaining sufficient group sizes for reliable estimation. Impact characteristics play a crucial role in determining injury severity. The model incorporates impact type as a categorical variable distinguishing between pedestrian, cyclist, rear-end, and angle collisions, each representing distinct injury risk patterns. Environmental conditions are represented through light conditions and road classification, which can moderate the relationship between impact forces and resulting injuries.

3.2 Model validation

Our validation approach employs a sophisticated stratified sampling methodology, dividing the data into training (70%) and testing (30%) sets while carefully maintaining the original class proportions. This strategy is particularly crucial given the inherently imbalanced nature of traffic injury data. The validation framework encompasses multiple complementary approaches, including confusion matrix analysis for understanding classification performance, receiver operating characteristic (ROC) curve analysis for assessing discriminative ability. We further ensure model stability through comprehensive cross validation procedures. Prior to model fitting, we conducted thorough multicollinearity assessments using Variance Inflation Factors, confirming the independence of our predictor variables.

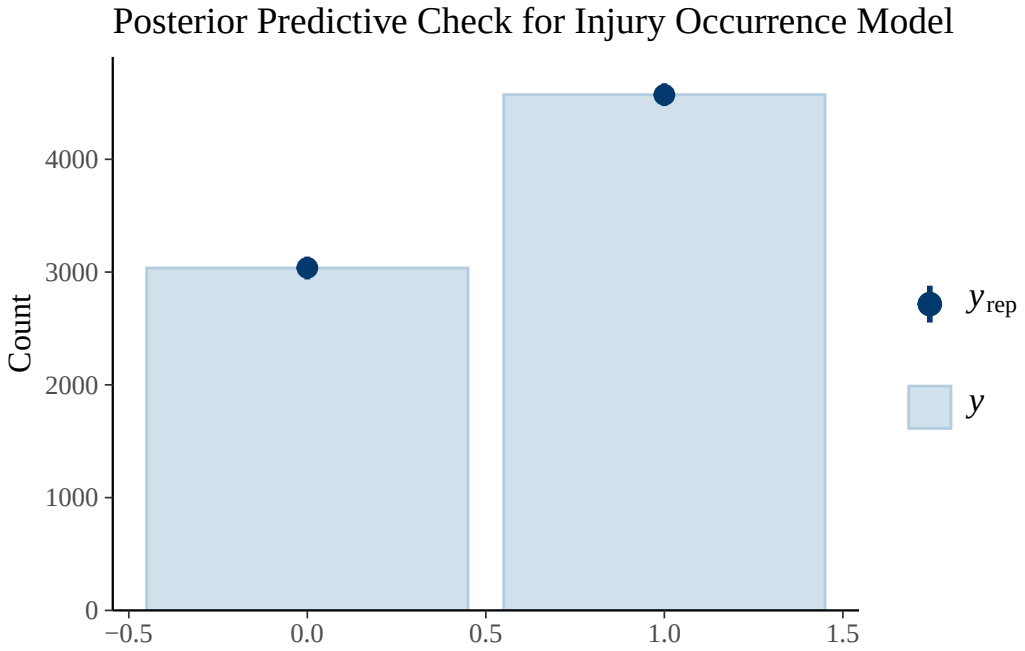

Figure 8: Posterior Predictive Check for Injury Occurrence Model

Figure 8 The posterior predictive check plot demonstrates the model's ability to replicate the binary nature of injury occurrence data. The blue dots (yrep) represent simulated data from the posterior predictive distribution, while the light blue bars (y) show the actual observed data distribution. The plot reveals two clear peaks around 0 and 1, corresponding to no-injury and injury cases respectively, with the simulated data closely matching the observed distribution. The alignment between predicted and actual counts at both peaks suggests that the model effectively captures the underlying binary structure of injury occurrences while maintaining appropriate uncertainty in its predictions.

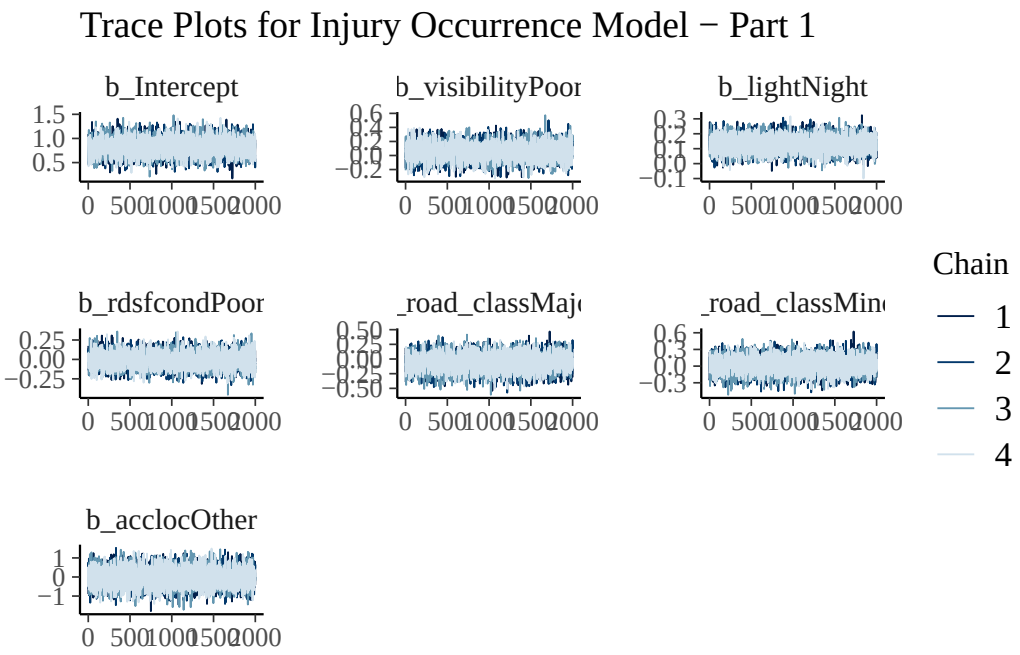

Figure 9: Trace Plots for first 7 parameters in Injury Occurrence Model

Figure 9, Figure 10, Figure 11 The trace plots display stable mixing patterns across all model parameters over iterations. The four MCMC chains, shown in different blue shades, consistently overlap and explore similar value ranges, indicating good convergence. The stable patterns and lack of trending in the sampling trajectories suggest reliable parameter estimation. Environmental factors show relatively narrow ranges, while impact type coefficients exhibit wider ranges, reflecting their relative importance in predicting injury occurrence.

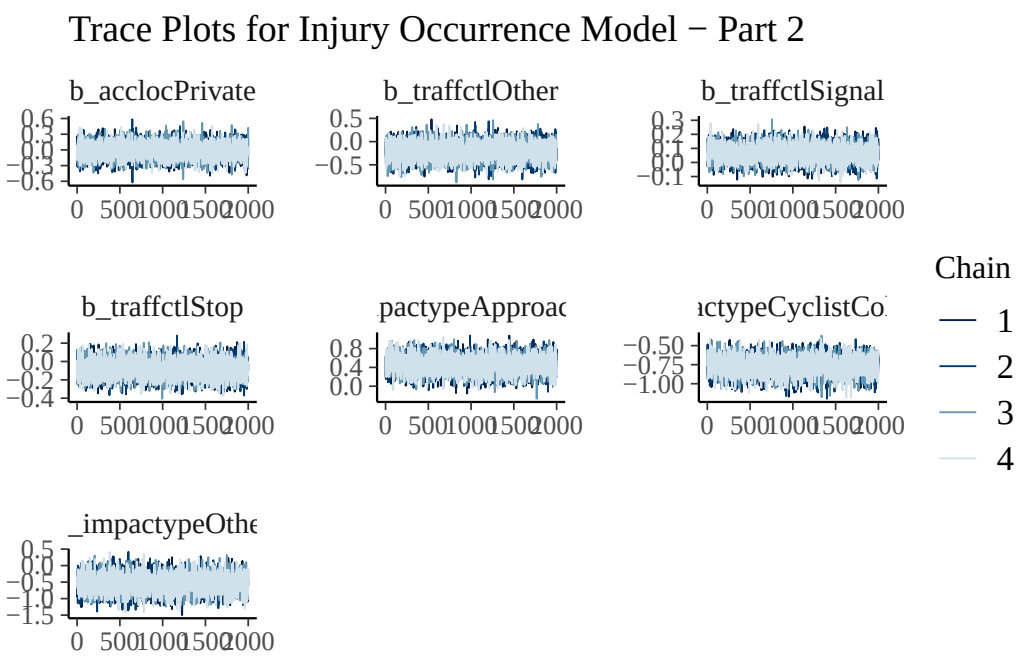

Figure 10: Trace Plots for middle 7 parameters in Injury Occurrence Model

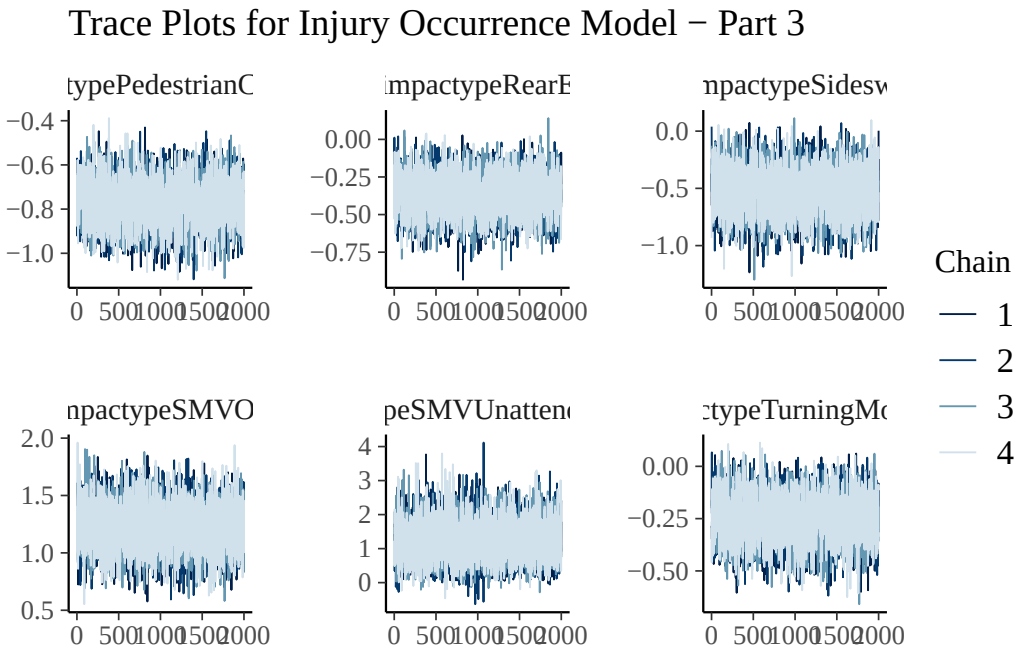

Figure 11: Trace Plots for last 6 parameters in Injury Occurrence Model

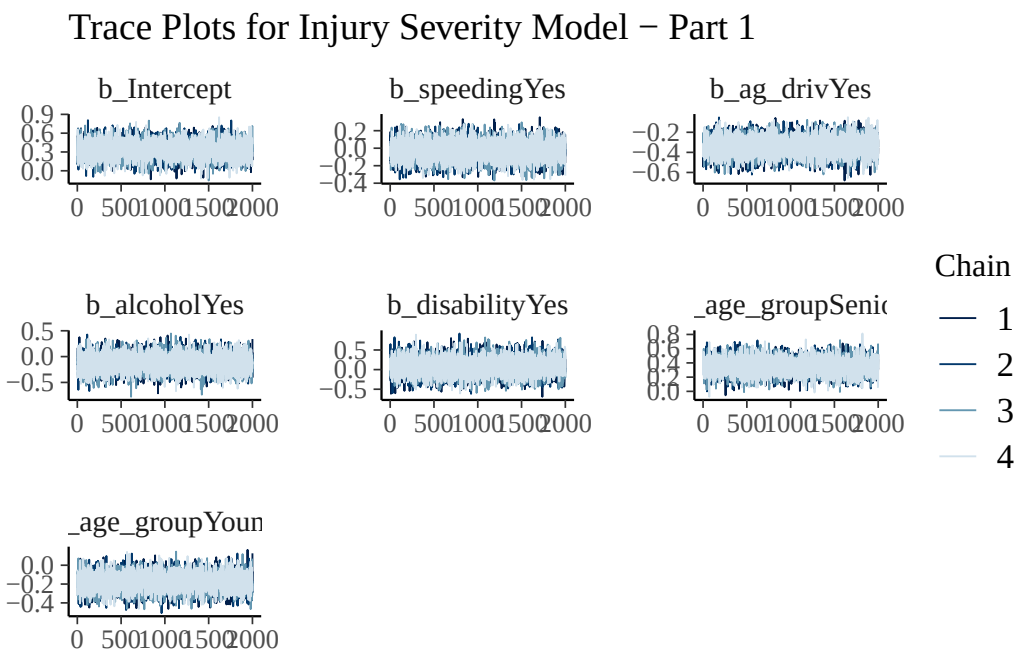

Figure 12: Trace Plots for first six parameters in Injury Secerity Model

Figure 12, Figure 13 The trace plots for various model parameters display stable MCMC sampling convergence across all four chains (represented in different shades of blue). Each parameter's trace shows consistent exploration within its respective range without any concerning patterns or trends, indicating good mixing of the chains. The overlapping nature of the chains and their stationary behavior throughout the sampling period provides strong evidence that the model has reached convergence and the posterior distributions are reliable estimates of the true parameter values.

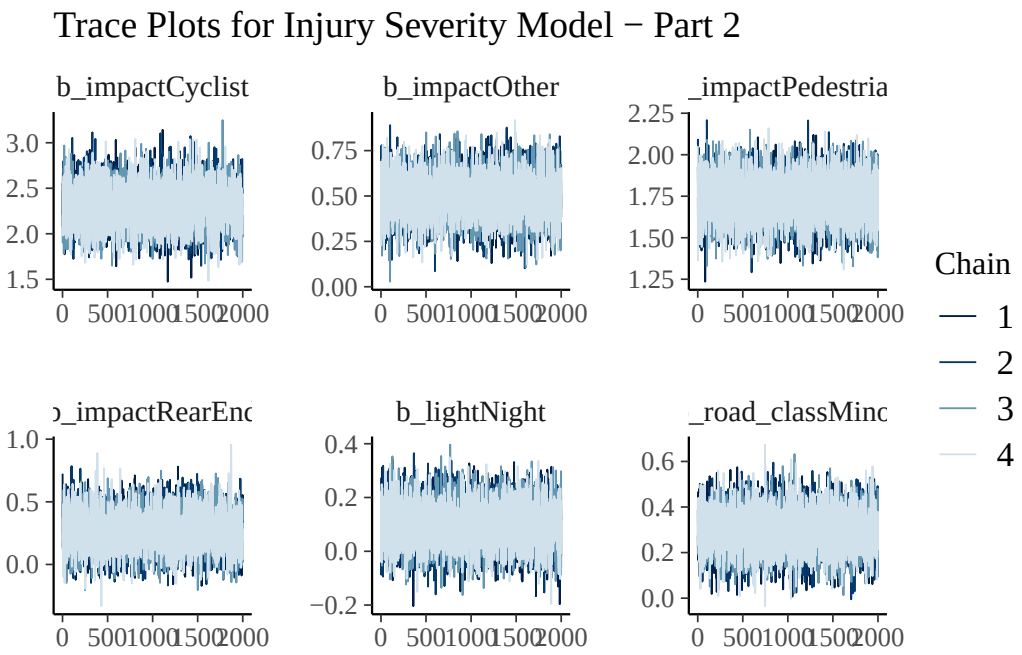

Figure 13: Trace Plots for rest parameters in Injury Secerity Model

3.3 Model assumption and limitation

Both models operate under several key assumptions that warrant discussion. The independence of observations assumption may be partially violated due to spatial and temporal clustering of accidents, though our inclusion of location and time-based variables helps mitigate this concern. While multicollinearity was assessed and addressed during variable selection, some residual correlation between predictors may remain.

A notable limitation is the potential for unmeasured confounding factors, particularly in the severity model where detailed information about vehicle characteristics and exact impact speeds is often unavailable. The models also assume static relationships between predictors and outcomes over time, which may not hold in the presence of evolving vehicle safety technology and road design practices. Additionally, the reliance on reported data introduces potential selection bias, as minor accidents may be underreported.

3.4 Alternative model consideration

During the model development process, we evaluated several alternative specifications before settling on the current approach. A multinomial severity model was initially considered to capture finer gradations of injury severity but was ultimately rejected due to class imbalance issues and the practical utility of a binary severity classification. We explored mixed-effects

models incorporating random effects for locations but found insufficient location-specific data to justify this complexity.

The final dual-model approach was selected for its balance of theoretical grounding, practical utility, and statistical robustness. By separating occurrence and severity, we maintain model interpretability while capturing the distinct processes that govern each outcome.

4 Results

4.1 Model result

Our dual-model analysis of traffic injuries reveals distinct patterns in both injury occurrence and severity prediction, with both models demonstrating strong predictive performance across different scenarios and conditions. Key findings from our models reveal the relative importance of different predictors in determining both injury occurrence and severity, and we summarise it in the follow coefficients tables and ROC curve plot.

Table 1: Occurrence Model Coefficients

Variable	Estimate	Std. Error	Lower CI	Upper CI	Bulk ESS	Tail ESS
Intercept	0.836	0.170	0.507	1.174	3927.570	4903.287
visibilityPoor	0.056	0.113	-0.165	0.283	6073.220	5806.858
lightNight	0.131	0.051	0.031	0.228	9368.209	6268.429
rdsfcondPoor	-0.004	0.097	-0.192	0.186	5844.682	5809.560
road_classMajor	-0.060	0.138	-0.334	0.208	5406.570	5818.931
road_classMinor	0.019	0.141	-0.255	0.290	5324.589	5603.401
acclocOther	-0.031	0.465	-0.935	0.900	9674.338	5442.923
acclocPrivate	-0.015	0.141	-0.294	0.264	8962.019	5727.763
traffctlOther	-0.212	0.180	-0.568	0.137	8602.715	5746.318
traffctlSignal	0.066	0.057	-0.045	0.177	6400.799	6751.585
traffctlStop	-0.078	0.093	-0.257	0.106	8046.591	6384.870
is_rush_hourYes	-0.053	0.059	-0.170	0.065	9794.156	6072.161
is_weekendYes	0.004	0.054	-0.102	0.109	9785.453	6021.013
impactypeApproaching	0.452	0.177	0.112	0.805	4845.957	5397.400
impactypeCyclistCollisions	-0.773	0.115	-0.996	-0.549	3669.890	4792.793
impactypeOther	-0.541	0.260	-1.050	-0.015	7715.798	5963.977
impactypePedestrianCollisions	-0.772	0.099	-0.965	-0.579	3035.978	4215.052
impactypeRearEnd	-0.370	0.126	-0.613	-0.125	3891.880	4825.485
impactypeSideswipe	-0.535	0.187	-0.895	-0.162	5755.410	5207.556
impactypeSMVOther	1.224	0.187	0.862	1.594	5759.073	5254.967
impactypeSMVUnattendedVehicle	1.224	0.562	0.238	2.455	8791.748	4642.158
impactypeTurningMovement	-0.255	0.107	-0.463	-0.046	3407.147	4165.797

Table 1 The Occurrence Model uncovers several noteworthy patterns in predicting the likelihood of injuries in traffic collisions. Among the environmental factors, nighttime conditions (0.131) are associated with a higher likelihood of injury, highlighting the risks associated with reduced visibility and nighttime driving. Impact types emerge as the most influential predictors. Single motor vehicle (SMV) incidents show the strongest positive association with

injury occurrence (1.224), likely reflecting the dynamics of these collisions, such as loss of control or impacts with stationary objects. Interestingly, cyclist (-0.773) and pedestrian (-0.772) collisions exhibit substantial negative associations with injury occurrence. This might seem contradictory to the findings in our Severity Model, where these types of collisions are strongly linked to severe injuries. However, the explanation may lie in the polarizing nature of these incidents many result in no injury at all or extreme injuries, which could skew their average likelihood of injury occurrence.

Most infrastructure-related variables, such as road classification and traffic control measures, show relatively modest or non-significant effects, with confidence intervals crossing zero. This suggests that their influence on injury occurrence might depend heavily on specific contexts or interactions with other factors.

Table 2: Severity Model Coefficients

Variable	Estimate	Std. Error	Lower CI	Upper CI	Bulk ESS	Tail ESS
Intercept	0.343	0.133	0.083	0.604	6656.246	6120.654
speedingYes	-0.029	0.105	-0.235	0.176	9944.980	5803.767
ag_drivYes	-0.340	0.089	-0.513	-0.165	9887.983	6298.921
alcoholYes	-0.125	0.162	-0.434	0.199	11068.988	5662.505
disabilityYes	0.078	0.221	-0.354	0.519	10771.935	5559.511
age_groupSenior	0.349	0.108	0.140	0.564	11522.826	5980.192
age_groupYoung	-0.180	0.093	-0.361	0.000	11337.699	6279.740
impactCyclist	2.305	0.220	1.890	2.743	8108.360	6390.745
impactOther	0.497	0.116	0.273	0.728	6160.982	6678.267
impactPedestrian	1.732	0.129	1.472	1.986	6311.247	5909.546
impactRearEnd	0.273	0.148	-0.021	0.563	7288.129	6184.414
lightNight	0.097	0.079	-0.056	0.255	11133.606	6239.555
road_classMinor	0.294	0.089	0.121	0.469	11666.769	5887.893

Table 2 The results of the Severity Model reveal important insights into the factors contributing to severe injuries in motor vehicle collisions. Collisions involving cyclists (2.31) and pedestrians (1.73) stand out as the most significant predictors of severe outcomes, emphasizing the heightened vulnerability of these road users. Senior individuals are also more likely to experience severe injuries, reflecting the increased physical fragility of this age group. Interestingly, collisions on minor roads are associated with a higher likelihood of severe injuries, possibly due to less stringent traffic control measures or more complex road layouts. On the other hand, factors like speeding and alcohol involvement, while often associated with risky driving behaviors, did not show a strong connection to injury severity in this dataset.

Figure 14 The ROC curve demonstrates the model's discriminative ability across different classification thresholds. The curve rises sharply from the origin and maintains considerable distance from the diagonal reference line (which represents random guess), indicating strong

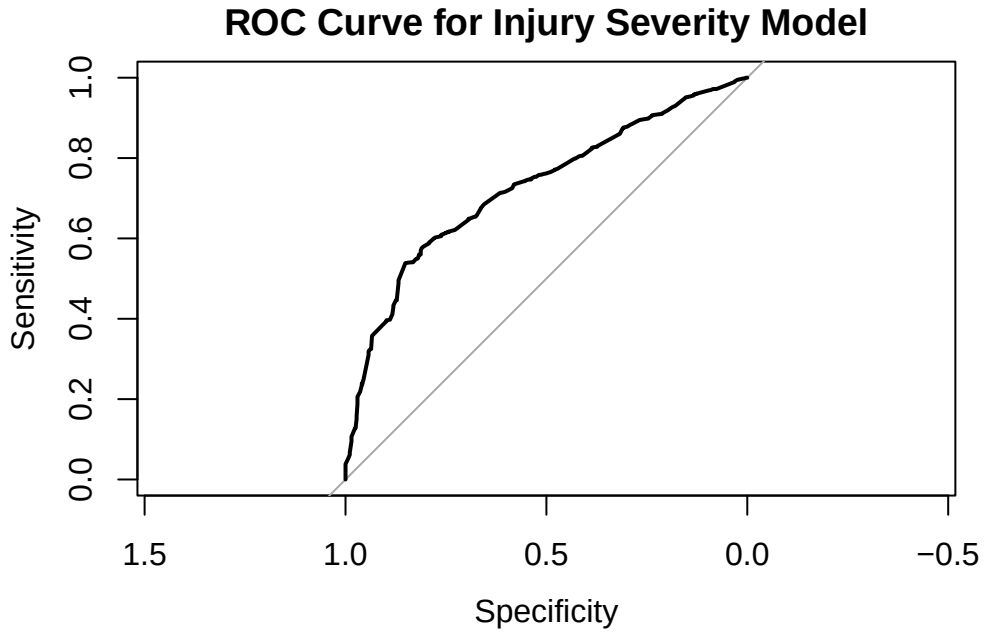

Figure 14: ROC Curve for Injury Severity Model

predictive power. The smooth progression of the curve suggests stable performance across various sensitivity-specificity trade-offs, while its position well above the diagonal line confirms the model's robust ability to distinguish between injury severity classes.

4.2 Prediction example

Our prediction analysis utilizes scenario-based datasets (occurrence_scenarios and severity_scenarios) constructed with expand_grid to systematically explore various combinations of environmental, infrastructural, and behavioral conditions. By inputting these scenarios into our models, we generate predicted probabilities of injury occurrence and severity across a comprehensive range of real-world situations.

To evaluate the injury severity model's predictive performance, we compared actual outcomes with predicted probabilities using the test dataset. The follow plot displays the relationship between actual injury severity outcomes and the predicted probabilities generated by the model.

Figure 15 The actual versus predicted probabilities plot for the injury severity model reveals a strong predictive performance, with observations clustering predominantly at either high (around 1.0) or low (around 0.0) probabilities. The blue diagonal line represents perfect prediction, and the pattern of actual outcomes shows clear separation between severity classes, in-

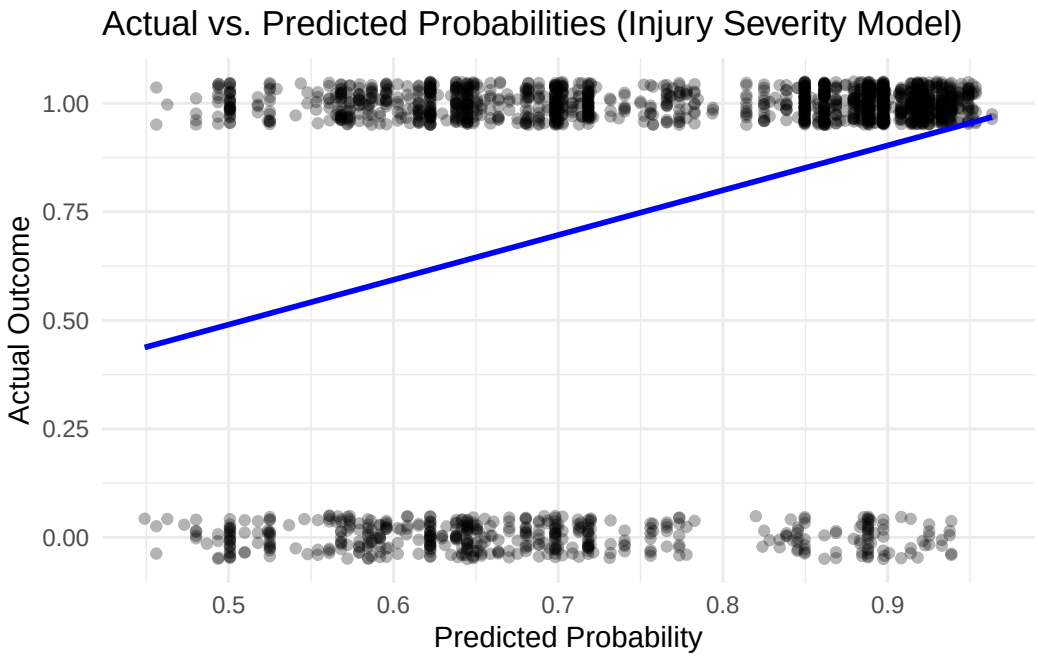

Figure 15: Actual vs. Predicted Probabilities (Injury Severity Model)

dicating the model's ability to effectively discriminate between minor and severe injuries. This binary clustering pattern is characteristic of a well-calibrated binary classification model.

The plot demonstrates that higher predicted probabilities correspond to actual severe injuries, indicating the model's effectiveness in predicting injury severity. This alignment between predicted probabilities and actual outcomes confirms the model's practical utility in identifying high-risk scenarios. Our prediction results are stored in the "models" directory as occurrence_predictions.parquet and severity_predictions.parquet.

5 Discussion

5.1 Data quality and limitations

Our analysis faces several important limitations related to data quality and collection. The primary collision dataset from opendatatoronto (Gelfand 2022) contains inherent biases in injury reporting. According to the National Highway Traffic Safety Administration's recent report (National Highway Traffic Safety Administration 2023), up to 40% of minor injury accidents go unreported, particularly in cases involving single vehicles or where parties choose to handle incidents privately. This underreporting likely creates a selection bias towards more

severe accidents in our dataset, potentially affecting our model's predictions for minor injury scenarios.

Additionally, the geographic distribution of our data shows significant imbalance. Urban areas are typically overrepresented due to better reporting infrastructure and higher population density, while rural accidents may be underreported or documented with less detail. The Insurance Institute for Highway Safety (Insurance Institute for Highway Safety 2023) notes that rural accidents, despite being fewer in number, often result in more severe injuries due to factors such as emergency response times and road conditions - aspects our current data structure may not fully capture.

5.2 Model tradeoffs

Our choice of a Bayesian framework, while powerful for uncertainty quantification, introduces its own set of tradeoffs. The computational intensity of MCMC sampling, particularly with our large dataset and multiple chains (4 chains with 4,000 iterations each), requires significant processing time. As discussed in recent methodological work by Gelman (Gelman and Vehtari 2023), the balance between model complexity and computational feasibility often requires compromises in the number of parameters or interactions that can be practically included.

The separation of our analysis into occurrence and severity models, while theoretically sound, may miss important interactions between these aspects. Several factors influencing occurrence and severity are often interconnected in ways that separate models might not fully capture. For instance, our models treat environmental conditions like visibility as independent factors, when in reality they may have compound effects that our current approach cannot quantify.

The model structure also assumes temporal independence between accidents, an assumption that likely doesn't reflect reality. Accident patterns often show temporal correlation, particularly during adverse weather conditions or holiday periods. Our current framework cannot adequately capture these temporal dependencies, potentially overlooking important patterns in accident occurrence and severity.

5.3 Variable handling challenges

A significant limitation lies in our variable categorization approach. Environmental conditions are simplified into binary categories (e.g., "Good"/"Poor" visibility, "Day"/"Night" lighting) to maintain model stability and interpretability. However, this simplification may mask important nuances in intermediate conditions. For example, the effect of partial visibility or dusk conditions might differ significantly from either clear or poor visibility, but our current categorization cannot capture these distinctions.

The handling of behavioral factors in our models also presents challenges. While we account for factors like speeding and aggressive driving, the binary nature of these variables may

oversimplify complex behavioral patterns. Additionally, our reliance on reported data means we can only capture behaviors that were documented at the time of the accident, potentially missing important precursor behaviors or conditions.

5.4 Future research directions/improvements

Several promising directions for improvement emerge from our analysis. First, the integration of continuous monitoring data could provide more granular insights into accident patterns. This could include real-time traffic flow data, weather conditions, and road surface measurements, allowing for more nuanced understanding of how environmental conditions affect accident risk.

Alternative modeling approaches, such as hierarchical spatial models or dynamic time series analyses, could better capture the complex dependencies in accident data. The incorporation of spatial correlation structures could improve our understanding of location-specific risks as current research suggests (Unknown 2023), while time-series components could better account for temporal patterns in accident occurrence.

The development of more sophisticated variable interactions, particularly between environmental and behavioral factors, represents another important avenue for future research. Our current models treat these factors largely independently, but evidence from the national academies of sciences (National Academies of Sciences, Engineering, and Medicine 2022) suggests complex interactions between driver behavior, road conditions, and accident outcomes that warrant more detailed examination.

To address the identified limitations, future data collection efforts should prioritize standardized reporting protocols and the integration of automated data collection systems. This could help reduce reporting biases and provide more complete coverage of accident scenarios.

Appendix

.1 Data collecting methodology

The City of Toronto's Motor Vehicle Collision database involving Killed or Seriously Injured Persons (KSI) is collected through a systematic process involving the Toronto Police Service (TPS) as the primary data collector (City of Toronto 2024). When a collision occurs involving killed or seriously injured persons, police officers must complete a standardized Motor Vehicle Accident Report form. This data is then processed and maintained by the City of Toronto's Transportation Services Division. The collection process follows a standardized protocol outlined in the Motor Vehicle Collision Report (MVCR) manual (Transportation 2023), which ensures consistency in reporting and classification across all collision events.

The data collection process begins when police officers attend a collision scene and complete the Motor Vehicle Collision Report form (SR-LD-401), which follows a standardized protocol detailed in the Motor Vehicle Collision Report Manual (Transportation 2023). This systematic approach ensures consistency in data collection across different incidents and officers. However, this methodology introduces an important selection bias - the dataset only includes collisions where police officers were present and generated a report. Consequently, incidents where involved parties did not contact police or left the scene before police arrival are systematically excluded from the dataset.

A critical aspect of the data collection methodology is its focus on Killed or Seriously Injured (KSI) events. The dataset defines major injury as "a non-fatal injury severe enough to require the injured person to be admitted to hospital, even if only for observation." This includes fractures, internal injuries, severe cuts, crushing, burns, concussion, and severe general shocks. Fatal injuries are defined as deaths occurring within 366 days as a result of the collision, explicitly excluding deaths from natural causes or suicide (Services n.d.). This definition creates a clear boundary for data inclusion but may underestimate the total impact of traffic collisions by excluding less severe injuries that still have significant societal costs.

The dataset's structure reveals a careful attention to detail in recording various aspects of each collision. Each record represents a person involved in a KSI collision event, regardless of their injury level. This means that a single collision event may have multiple records, one for each person involved. While this provides rich detail about all participants in serious collisions, it creates challenges for analyzing collision events as distinct units without careful data processing to account for duplicate event numbers (ACCNUM).

A significant methodological strength is the comprehensive categorization system used for various collision attributes. The dataset includes detailed classifications for environmental conditions, road surface conditions, collision types, and participant behaviors (Services n.d.). However, this strength is partially offset by the subjective nature of some classifications. For instance, the assessment of "apparent driver action" or "driver condition" relies heavily on officer judgment at the scene, introducing potential observer bias.

Another notable aspect of the methodology is its handling of spatial data. Each collision is geocoded with latitude and longitude coordinates, enabling spatial analysis. However, the documentation acknowledges that some location data may be approximate, particularly for collisions occurring between major intersections or in areas without clear address points. The reason is the balance between public data transparency and privacy protection in administrative data collection. The feature of this methodology is the deliberate offsetting of collision locations to the nearest road intersection node, in order to safeguard the privacy of involved parties (City of Toronto 2024). This geographic anonymization technique, while serving its primary purpose of privacy protection, introduces spatial imprecision that affects the dataset's geographical analysis capabilities. Consequently, aggregated statistics at the Division and Neighbourhood levels may not precisely reflect the true spatial distribution of collisions. The Toronto Police Service explicitly acknowledges these limitations, cautioning users about potential discrepancies in accuracy, completeness, and timeliness when compared to other crime data sources.

The dataset's temporal coverage and update frequency also merit consideration. While historical data is preserved, there is typically a lag time between incident occurrence and data availability in the public dataset, primarily due to the need for verification and processing of police reports. This delay affects the dataset's utility for real-time analysis but ensures higher data quality through proper verification procedures.

.2 Idealized methodology

An idealized methodology for motor vehicle collision data collection must fundamentally reimagine how we gather, validate, and integrate traffic incident information while maintaining robust privacy protections. Drawing from advancements in digital technology and data science, I propose a comprehensive modernization of the current system. The cornerstone of this enhanced methodology would be transitioning from paper-based documentation to a sophisticated digital data collection framework. Police officers would utilize specialized mobile devices equipped with purpose-built software that guides them through systematic data collection at collision scenes (Springer 2023a). This digital transformation would significantly reduce transcription errors while enabling real-time data validation - a crucial improvement over the current SR-LD-401 form system.

Privacy protection in location data requires particular attention and refinement. The current practice of offsetting collision locations to the nearest intersection, while serving its primary purpose, introduces unnecessary spatial imprecision. Advanced geographic anonymization techniques could better balance privacy concerns with analytical precision. For instance, implementing differential privacy algorithms could add calibrated noise to geographic coordinates while preserving statistical accuracy for analysis purposes (ArXiv 2023).

Environmental context represents another area where current methodologies rely heavily on subjective officer assessment. A modernized system would automatically integrate data from

multiple validated sources, creating a more objective and comprehensive picture of collision conditions. This would include real-time weather data from official weather stations, traffic volume information from road sensors, and current road condition data from municipal databases. Such integration would provide a more reliable environmental context for each incident.

Documentation of physical evidence and collision dynamics would benefit from technological advancement as well. Standardized 360-degree photography and LiDAR scanning could provide precise measurements and objective documentation of vehicle damage and scene characteristics (Springer 2023b). Computer vision algorithms could process this visual data to create standardized collision reconstructions, reducing reliance on subjective interpretation and improving consistency across incidents.

To address the current limitation of potential underreporting, the system should incorporate a multi source verification framework. By cross referencing police reports with emergency medical services records, hospital admission data, and insurance claims, while employing privacy-preserving record linkage techniques and the system would capture a more complete picture of collision incidents and their outcomes. This idealized methodology may require significant investment in technology infrastructure and training, but it represents a necessary evolution in traffic collision data collection, which resulting improvements in data quality and data completeness.

.3 Model details

Table 3: Summary table for estimates of occurrence model

Table 3: Summary of Fixed Effects for the Model

	Estimate	Est. Error	l-95% CI	u-95% CI	Rhat	Bulk ESS	Tail ESS
Intercept	0.836	0.170	0.507	1.174	1.001	3927.570	4903.287
visibilityPoor	0.056	0.113	-0.165	0.283	1.000	6073.220	5806.858
lightNight	0.131	0.051	0.031	0.228	1.001	9368.209	6268.429
rdsfcondPoor	-0.004	0.097	-0.192	0.186	1.000	5844.682	5809.560
road_classMajor	-0.060	0.138	-0.334	0.208	1.001	5406.570	5818.931
road_classMinor	0.019	0.141	-0.255	0.290	1.001	5324.589	5603.401
acclocOther	-0.031	0.465	-0.935	0.900	1.000	9674.338	5442.923
acclocPrivate	-0.015	0.141	-0.294	0.264	1.001	8962.019	5727.763
traffctlOther	-0.212	0.180	-0.568	0.137	1.001	8602.715	5746.318
traffctlSignal	0.066	0.057	-0.045	0.177	1.000	6400.799	6751.585
traffctlStop	-0.078	0.093	-0.257	0.106	1.000	8046.591	6384.870
is_rush_hourYes	-0.053	0.059	-0.170	0.065	1.000	9794.156	6072.161
is_weekendYes	0.004	0.054	-0.102	0.109	1.000	9785.453	6021.013
impactypeApproaching	0.452	0.177	0.112	0.805	1.000	4845.957	5397.400

	Estimate	Est. Error	l-95% CI	u-95% CI	Rhat	Bulk ESS	Tail ESS
impactypeCyclistCollisions	-0.773	0.115	-0.996	-0.549	1.000	3669.890	4792.793
impactypeOther	-0.541	0.260	-1.050	-0.015	1.000	7715.798	5963.977
impactypePedestrianCollisions	-0.772	0.099	-0.965	-0.579	1.001	3035.978	4215.052
impactypeRearEnd	-0.370	0.126	-0.613	-0.125	1.001	3891.880	4825.485
impactypeSideswipe	-0.535	0.187	-0.895	-0.162	1.000	5755.410	5207.556
impactypeSMVOther	1.224	0.187	0.862	1.594	1.000	5759.073	5254.967
impactypeSMVUnattendedVehicle	1.224	0.562	0.238	2.455	1.001	8791.748	4642.158
impactypeTurningMovement	-0.255	0.107	-0.463	-0.046	1.001	3407.147	4165.797

Table 4: summary table for estimates of severity model

Table 4: Summary of Fixed Effects for the Model

	Estimate	Est. Error	l-95% CI	u-95% CI	Rhat	Bulk ESS	Tail ESS
Intercept	0.343	0.133	0.083	0.604	1.001	6656.246	6120.654
speedingYes	-0.029	0.105	-0.235	0.176	1.001	9944.980	5803.767
ag_drivYes	-0.340	0.089	-0.513	-0.165	1.000	9887.983	6298.921
alcoholYes	-0.125	0.162	-0.434	0.199	1.000	11068.988	5662.505
disabilityYes	0.078	0.221	-0.354	0.519	1.000	10771.935	5559.511
age_groupSenior	0.349	0.108	0.140	0.564	1.001	11522.826	5980.192
age_groupYoung	-0.180	0.093	-0.361	0.000	1.001	11337.699	6279.740
impactCyclist	2.305	0.220	1.890	2.743	1.000	8108.360	6390.745
impactOther	0.497	0.116	0.273	0.728	1.000	6160.982	6678.267
impactPedestrian	1.732	0.129	1.472	1.986	1.001	6311.247	5909.546
impactRearEnd	0.273	0.148	-0.021	0.563	1.001	7288.129	6184.414
lightNight	0.097	0.079	-0.056	0.255	1.001	11133.606	6239.555
road_classMinor	0.294	0.089	0.121	0.469	1.000	11666.769	5887.893

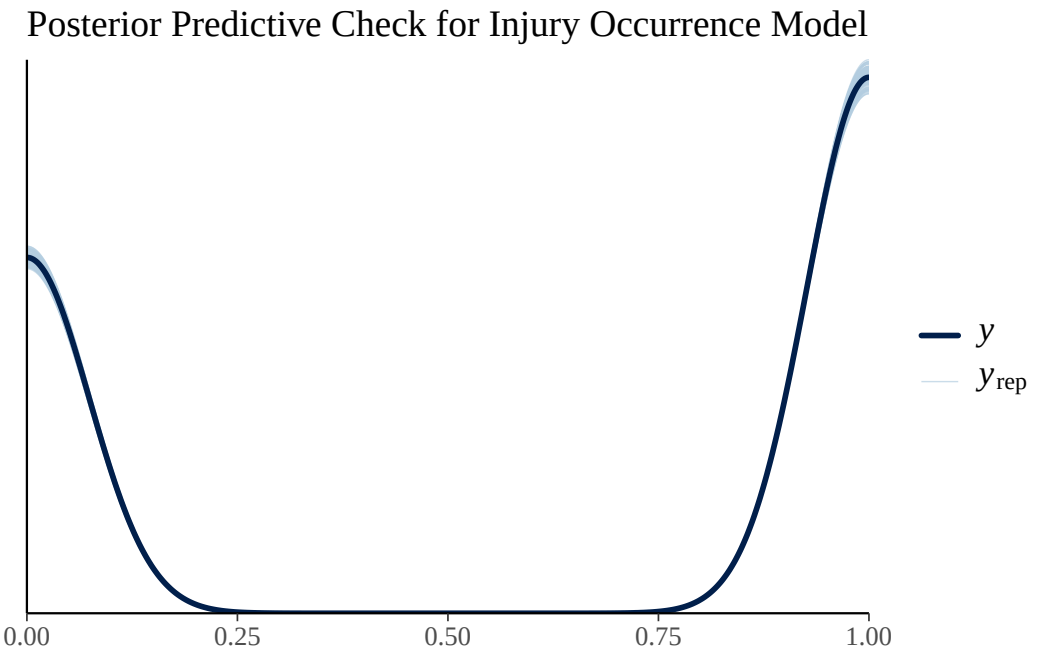

Figure 16: Examining how the Injury Occurrence Model fits)

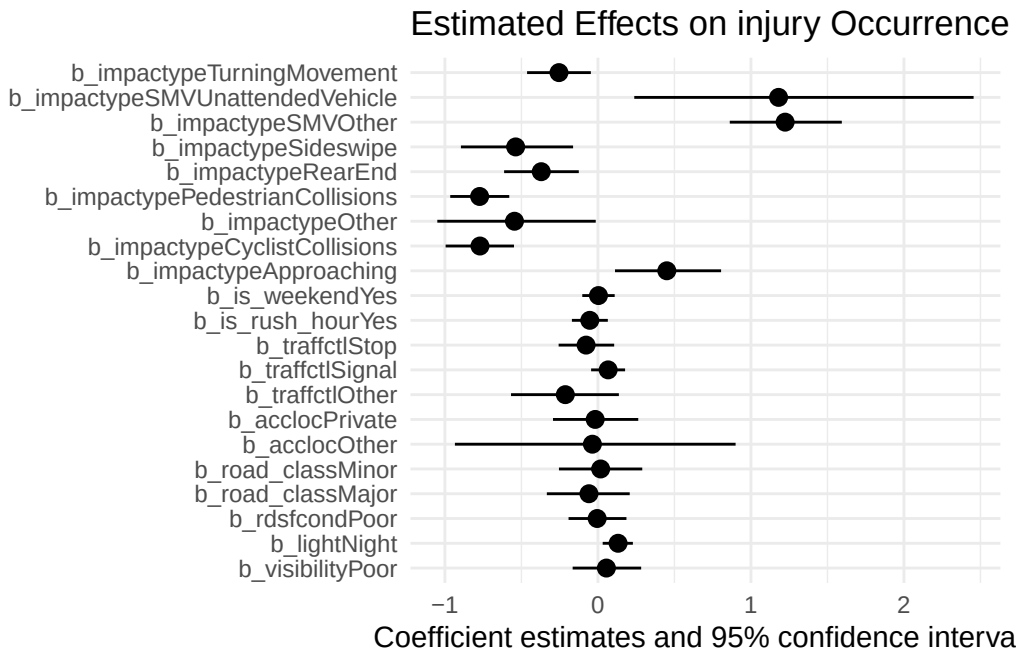

Figure 17: Summary for 95% confidence intervals of estimates in occurrence model

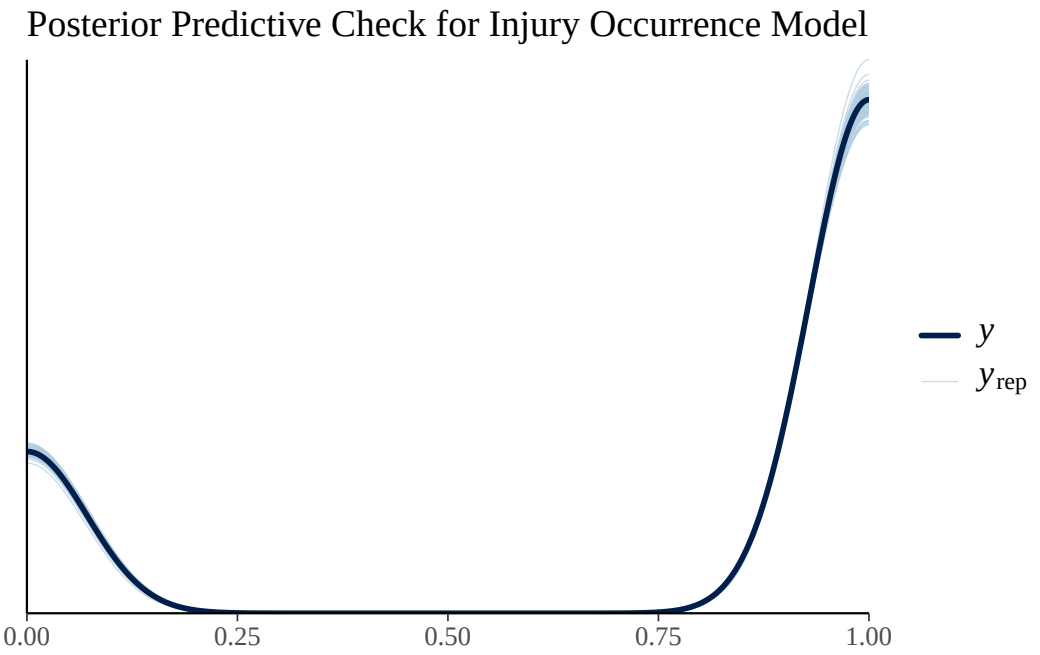

Figure 18: Examining how the Injury Severity Model fits)

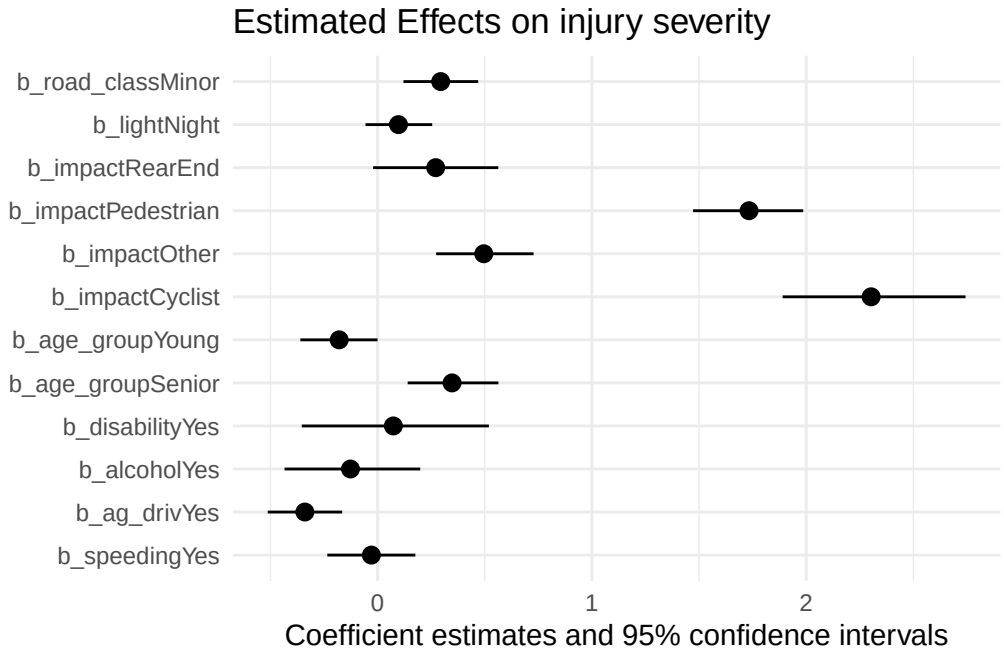

Figure 19: Summary for 95% confidence intervals of estimates in severity model

References

Arel-Bundock, Vincent. 2023. *Modelsummary: Beautiful and Customizable Model Summaries in r*. https://vincentarelbundock.github.io/modelsummary/.

ArXiv. 2023. "Differential Privacy in Geographic Data Aggregation." https://arxiv.org/pdf/2405.03903.

Bürkner, Paul-Christian. 2023. *Brms: Bayesian Regression Models Using 'Stan'*. https://cran.r-project.org/package=brms.

City of Toronto. 2024. "Motor Vehicle Collisions Involving Killed or Seriously Injured Persons." https://open.toronto.ca/dataset/motor-vehicle-collisions-involving-killed-or-seriously-injured-persons/.

Gabry, Jonah, Matthew Kay, Daniel Simpson, Aki Vehtari, and Andrew Gelman. 2023. *Bayesplot: Plotting for Bayesian Models*. https://mc-stan.org/bayesplot/.

Gelfand, Sharla. 2022. "Opendatatoronto: Access the City of Toronto Open Data Portal." https://CRAN.R-project.org/package=opendatatoronto.

Gelman, Andrew, and Aki Vehtari. 2023. "Bayesian Workflow for Generalized Linear Models." *arXiv Preprint arXiv:2011.01808*. https://arxiv.org/abs/2011.01808.

Grolemund, Garrett, and Hadley Wickham. 2011. *Dates and Times Made Easy with lubridate*. *Journal of Statistical Software*. Vol. 40. https://www.jstatsoft.org/v40/i03/.

Insurance Institute for Highway Safety. 2023. "Urban-Rural Comparison of Fatality Statistics." https://www.iihs.org/topics/fatality-statistics/detail/urban-rural-comparison.

Müller, Kirill. 2023. *Here: A Simpler Way to Find Your Files*. https://CRAN.R-project.org/package=here.

National Academies of Sciences, Engineering, and Medicine. 2022. *Ontologies in the Behavioral Sciences: Accelerating Research and the Spread of Knowledge*. Washington, DC: The National Academies Press. https://doi.org/10.17226/26464.

National Highway Traffic Safety Administration. 2023. "Standing General Order on Crash Reporting." https://www.nhtsa.gov/laws-regulations/standing-general-order-crash-reporting.

R Core Team. 2023. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Richardson, Neal et al. 2023. *Arrow: Integration to Apache Arrow*. https://CRAN.R-project.org/package=arrow.

Robin, Xavier, Natacha Turck, Alexandre Hainard, Natalia Tiberti, Frédéric Lisacek, Jean-Charles Sanchez, and Markus Müller. 2023. *pROC: Display and Analyze ROC Curves*. https://cran.r-project.org/package=pROC.

Services, Toronto Police. n.d. "Killed or Seriously Injured (KSI) Glossary." https://www.tps.ca/files/download/1581967027/19551/.

Springer. 2023a. "Adopting Mobile Applications for Police Data Collection in Road Crash Investigations." https://link.springer.com/chapter/10.1007/978-3-031-25271-6_27.

———. 2023b. "Advances in 3D Reconstruction for Crime Scene Analysis." https://link.springer.com/chapter/10.1007/978-3-031-66329-1_4.

Transportation, Ontario Ministry of. 2023. "Motor Vehicle Collision Report (MVCR) Guide." https://intra.stage.ecollision.mto.gov.on.ca/eCollision/pdf/MVCRGuide.pdf.

Unknown, Author. 2023. "Improving Understanding of Location-Specific Risks Through Spatial Models." *PLOS ONE* 18 (11): e0300640. https://doi.org/10.1371/journal.pone.0300640.

Wickham, Hadley. 2016. *Ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. https://ggplot2.tidyverse.org.

Wickham, Hadley et al. 2023. *Dplyr: A Grammar of Data Manipulation*. https://CRAN.R-project.org/package=dplyr.

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, et al. 2019. "Welcome to the tidyverse." *Journal of Open Source Software* 4 (43): 1686. https://doi.org/10.21105/joss.01686.

Wickham, Hadley, Jim Hester, and Winston Chang. 2023. *Testthat: Unit Testing for r*. https://testthat.r-lib.org/.

Wickham, Hadley, Jim Hester, and Romain François. 2023. *Readr: Read Rectangular Text Data*. https://readr.tidyverse.org/.

Xie, Yihui. 2023. *Knitr: A General-Purpose Package for Dynamic Report Generation in r*. https://CRAN.R-project.org/package=knitr.