SUPPLEMENTAL INFORMATION

Data Description

The California Highway Patrol (CHP) maintains the Statewide Integrated Traffic Records System (SWITRS) which contains all crashes that were reported to CHP from local and governmental agencies. The California raw crash data files for 2019-2021 were collected from the Internet SWITRS application. The database has three components, (i) Collision Records, (ii) Party Records, and (ii) Victim Records, all of which contain information related to crash attributes. In the Collision Records, there is information about each crash, whereas the Party Records contain information from all parties such as drivers, pedestrians, bicyclists, and parked vehicles involved in the crashes. The Victim Records tables hold information about the victims associated with each party.

Data Preparation

In this study, we primarily worked with party-level data. The Party Records file contains the demographic and injury information of each party. The crash time information (collision month and collision year) from the Collision Records table to the Party Records table was included for the respective case ID. The recorded collisions from January 2019 to April 2021 were considered for our study. The records before March 2020 were classified as "Pre-pandemic" whereas the records from March 2020 to April 2021 were classified as "Pandemic." In other words, the total collision records for a period of 14 months were assigned in each category. The season variable is included in our models to minimize bias due to the differences in seasons between pre-pandemic and pandemic periods in the dataset. The "season" variable was categorized into winter, spring, fall, and summer.

Age was transformed into a categorical variable with eight groups of 10-year age intervals for easier interpretability. The descriptive statistics of the dataset with all organized data are presented in Table 1.

Table 1. Descriptive statistics of study variables

Variable	Description	Count	Percentage (%)
Crash injury severity	Injury	382,345	27.1%
	No Injury	1,031,316	72.9%
Driver sex	Male	892,685	63.1%
	Female	520,976	36.9%
Driver age group	18-27	387,202	27.4%

	28-37	337,313	23.9%
	38-47	244,257	17.3%
	48-57	210,900	14.9%
	58-67	146,589	10.4%
	68-77	63,020	4.5%
	78-87	20,797	1.5%
	88-97	3,530	0.25%
Driver race	White	461,933	32.7%
	Black	143,443	10.1%
	Hispanic	598,225	42.3%
	Asian	110,982	7.9%
	Others	99,078	7.0%
Lighting	Daylight	962,362	68.0%
	Dusk-dawn	52,494	3.7%
	Dark streetlights	277,137	19.6%
	Dark-no streetlights	121,668	8.6%
Season	Winter	404,551	28.6%
	Spring	374,598	26.5%
	Summer	302,375	21.4%
	Fall	332,137	23.5%
Pandemic effect	Pre-pandemic	826,190	58.4%
	Pandemic	587,471	41.6%

Modeling Approach Details

Discreet outcome models are generally applied to model crash severity as a function of crash-contributing factors. Adding interaction terms to a regression model can test more specific hypotheses. Interaction terms are used to infer how the effect of one independent variable on the dependent variable depends on the magnitude of another independent variable. The goal of our study was to assess which driver demographic characteristics determined changes in crash injury severity during COVID-19. A binary logistic regression-based approach with interaction terms was employed to determine the effect of the pandemic.

First, a bivariate analysis of each variable was performed to examine the effect of each explanatory variable on motor vehicle crash injury severity. A chi-square test was then conducted for all the categorical predictors. The review of each explanatory variable and its corresponding effect on crash injury severity revealed that all the variables, driver sex, driver age group, driver race, roadway lighting, season, and pandemic effect had the greatest influence at the 95% confidence level ($p < 0.05$).

Six separate logistic regression models were then developed to explore the two-way and three-way interactions of pandemic effect with other independent variables. The likelihood ratio test was performed to determine the significance of the interaction term in each model. The marginal probability and confidence interval were used to quantify the main effects and interaction effects of significant independent variables on the dependent variable. All data analysis and modeling were conducted using R Programming.