Risk Factors of Fatalities in Motor Vehicle Crashes Happened in Raleigh, NC in 2015- 2020

Background

Motor vehicle crash deaths is one of the leading causes of death in the United States in these years and a great number of studies were conducted to study the cause of crashes and, therefore, find ways to prevent motor vehicle crash. However, only a handful of studies (https://pubmed.ncbi.nlm.nih.gov/30079869/) were conducted to study the risk factors of fatal cases given a car crash. While it is important to prevent crashes, there are inevitably crashes that cannot be avoided, so it is equally important to know how to reduce fatality when crashes occur.

Scientific Questions

The aim of this study is to identify other general risk factors associated with fatal cases in car crashes. Factors recorded in car crash reports, despite the outcome indicating if the case is fatal or non-fatal, were divided into three categories: crash site demographics, driver-related issues, and vehicle mechanics. Crash site demographics recorded the weather and time when the crash took place; driver-related issues recorded whether the crash involved at least one teen driver, one senior driver, one driver suspected to be intoxicated by the police, and/or one child passenger; vehicle mechanics recorded whether all people involved in the crash had airbags and proper protection, e.g., seatbelt, child restraint etc. The three scientific questions are:

1. Does some particular crash site demographics increase the risk of death, controlled for other factors, among crashes happen in Raleigh in 2015-2020.
2. Does driver-related issues increase the risk of death, controlled for other factors, among crashes happen in Raleigh in 2015-2020.
3. Does lack of proper vehicle mechanics, i.e., lack of airbags and protection belts, increase the risk of death, controlled for other factors, among crashes happen in Raleigh in 2015-2020.

Methods

Study Design and Setting

This retrospective observational cohort study included crash cases happened in Raleigh, North Carolina between January 1, 2015, and December 31, 2020. All data were provided from the standard DMV-349 Crash Form completed by the initial officer at the scene of a crash.

Data Collection

Crash data were extracted from data.raleighnc.gov, an open data portal recording public data about city of Raleigh. Two data sets were extracted, one with crash site location information including: indicator of fatal crash, primary weather condition and time of crash, and another with information of each person involved in crash including: age, person type, suspected alcohol/drug status, airbag existence, and protection existence. Age and time of crash were the only two continuous variables. Since intuitively the association between time of crash and fatal crash would not be linear, time of crash were treated as categorical (Morning: 5am-12pm, Afternoon: 12pm-5pm, Night: 5pm-5am). Categorization of age was described in detail in the Study Size and Variables section.

Study Size and Variables

Since two data sets had different measuring unit, data merging process was introduced before data analysis. Information in person-level data set were transformed into case-level indicator variables with the following criteria:

- Teen driver: record with at least one driver age 16-19 in one case
- Senior driver: record with at least one driver aged 65 or above in one case
- Suspected intoxicated driver: record with at least one driver suspected to be intoxicated in one case
- Child passenger: record with at least one passenger aged 12 or below in one case
- No airbag: record with at least one person in car not equipped with airbag in one case
- No protection: record with at least one person in car not equipped with any kind of protection (seatbelt, shoulder belt, child restraint, etc.) in one case

The age categorization above can be found in CDC website:

- https://www.cdc.gov/injury/features/older-driver-safety/index.html
- https://www.cdc.gov/injury/features/teen-drivers/index.html
- https://www.cdc.gov/injury/features/child-passenger-safety/index.html

Missing data appeared both in original data sets and during the process of merging, including person records with unknown person type, person records without concerning information, and records that cannot match due to merging. Records with missing data in all these categories were less than 1% of the whole study population, and thus were excluded from the analysis. As a result, 98,939 cases with 130 fatal ones (outcome variable) were included in the study. Case characteristics are summarized in Table 1.

Table 1. Characteristics of 98939 cases happened in Raleigh, NC from 2015 to 2020.

	Non-fatal (N=98809)	Fatal (N=130)	Overall (N=98939)
WeatherCondition			
Blowing sand, dirt, snow	6.00 (0.0%)	0 (0%)	6.00 (0.0%)
Clear	76594 (77.5%)	77.0 (59.2%)	76671 (77.5%)
Cloudy	11602 (11.7%)	36.0 (27.7%)	11638 (11.8%)
Fog, smog, smoke	141 (0.1%)	2.00 (1.5%)	143 (0.1%)
Other*	29.0 (0.0%)	0 (0%)	29.0 (0.0%)
Rain	9838 (10.0%)	15.0 (11.5%)	9853 (10.0%)
Severe crosswinds	5.00 (0.0%)	0 (0%)	5.00 (0.0%)
Sleet, hail, freezing rain/drizzle	225 (0.2%)	0 (0%)	225 (0.2%)
Snow	369 (0.4%)	0 (0%)	369 (0.4%)
TimeOfDay			
Night	28563 (28.9%)	72.0 (55.4%)	28635 (28.9%)
Morning	30834 (31.2%)	28.0 (21.5%)	30862 (31.2%)
Afternoon	39412 (39.9%)	30.0 (23.1%)	39442 (39.9%)
TeenDriver			
No	89200 (90.3%)	121 (93.1%)	89321 (90.3%)
Yes	9609 (9.7%)	9.00 (6.9%)	9618 (9.7%)
SeniorDriver			
No	86643 (87.7%)	112 (86.2%)	86755 (87.7%)
Yes	12166 (12.3%)	18.0 (13.8%)	12184 (12.3%)
IntoxicatedSuspected			
No	96643 (97.8%)	111 (85.4%)	96754 (97.8%)
Yes	2166 (2.2%)	19.0 (14.6%)	2185 (2.2%)
ChildPassenger			
No	90095 (91.2%)	121 (93.1%)	90216 (91.2%)
Yes	8714 (8.8%)	9.00 (6.9%)	8723 (8.8%)
NoAirbag			
No	92114 (93.2%)	108 (83.1%)	92222 (93.2%)
Yes	6695 (6.8%)	22.0 (16.9%)	6717 (6.8%)
NoSeatBelt			
No	95819 (97.0%)	96.0 (73.8%)	95915 (96.9%)
Yes	2990 (3.0%)	34.0 (26.2%)	3024 (3.1%)

Above covariates were divided into three categories for further analysis:

- Crash site demographics: weather condition, time of day
- Driver-related issues: teen driver, senior driver, intoxicated suspected, child passenger
- Vehicle mechanics: no airbag, no seatbelt

Statistical Analysis

To test the null hypothesis that there is no association between any factor recorded in the crash report and fatal crashes, i.e., no risk factors associated with a fatal crash, we examined factors with the use of multivariable logistic regression. The outcome was assessed with four models: model 1 to 3 included covariates in each category only, and model 4 included additional covariates in other 2 categories as adjustment for confounding. All model covariates were selected a priori. All tests were conducted using two-sided, 5% significance level, and Benjamini-Hochberg Procedure were used to address multiple comparison adjustment.

Results

Table 2. Odds Ratios for Crash Site Demographics (Model 1).

Variable	Odds Ratio (95% CI)	P-value (after BH)
Weather Condition		
Clear	Reference	
Blowing sand, dirt, snow	0.000 (0.000, 5.54e+174)	0.998
Cloudy	3.363 (2.235, 4.965)	**< 0.001**
Fog, smog, smoke	13.398 (2.181, 43.561)	**< 0.001**
Rain	1.410 (0.779, 2.380)	0.494
Severe crosswinds	0.000 (0.000, 6.15e+207)	0.998
Sleet, hail, freezing rain/drizzle	0.000 (0.000, 2.44e+04)	0.998
Snow	0.000 (0.000, 216.66)	0.998
Other	0.000 (0.000, 1.77e+37)	0.998
Time of Crash		
Afternoon	Reference	
Morning	1.128 (0.669, 1.893)	0.998
Night	3.406 (2.245, 5.297)	**< 0.001**

FACTORS ASSOCIATED WITH CRASH SITE DEMOGRAPHICS

Table 2 shows the unadjusted odds of fatal crash for crash site demographic factors. In the unadjusted analysis, the “Clear” in weather condition and “Afternoon” in time of crash were chosen as the reference since these two categories have the largest number of total crashes, thus presenting the most common situation a car crash could take place in Raleigh.

From the table, cloudy weather was associated with approximately 3.4 times (95% confidence interval: 2.2 to 5.0) and foggy weather was associated with about 13.4 times (95% confidence interval: 2.2 to 43.6) the odds of fatal crash as clear weather. Other weathers show no statistically significant difference in the odds of a fatal crash compared to clear weather. Driving at night (5 pm to 5 am) was associated with approximately 3.4 times (95% confidence interval: 2.2 to 5.3) the odds of fatal crash as driving in the afternoon and driving in the morning shows no statistically significant difference in the odds of a fatal crash compared to driving in the afternoon.

Table 3. **Odds Ratios for Driver-related Issues (Model 2).**

Variable	Odds Ratio (95% CI)	P-value (after BH)
Teen Driver	0.735 (0.346, 1.368)	0.567
Senior Driver	1.211 (0.710, 1.948)	0.567
Intoxicated-suspected Driver	7.616 (4.521, 12.145)	**< 0.001**
Child Passenger	0.830 (0.391, 1.545)	0.591

FACTORS ASSOCIATED WITH DRIVER-RELATED ISSUES

Table 3 shows the unadjusted odds of fatal crash for driver-related issue factors. In the unadjusted analysis, intoxicated-suspected drivers were associated with 7.6 times (95% confidence interval: 4.5 to 12.1) the odds of fatal crash as sober drivers. Teen driver and child passenger were associated with slightly lower odds of fatal cases, and senior driver was associated with slightly higher odds of fatal cases, but none of these factors show statistically significant difference in the odds of a fatal crash.

Table 4. **Odds Ratios for Vehicle mechanics (Model 3).**

Variable	Odds Ratio (95% CI)	P-value (after BH)
No Airbag	1.364 (0.807, 2.217)	0.227
No Belt/Protection	10.278 (6.599, 15.613)	**< 0.001**

FACTORS ASSOCIATED WITH VEHICLE MECHANICS

Table 4 shows the unadjusted odds of fatal crash for vehicle mechanics factors. In the unadjusted analysis, lack of seatbelt or other belt-like protection was associated with 10.3 times (95% confidence interval: 6.6 to 15.6) the odds of fatal cases. Lack of airbags was associated with higher odds of fatal cases but did not show a statistically significant difference.

Table 5. **Adjusted Odds Ratios (Model 4).**

Variable	Odds Ratio (95% CI)	P-value (after BH)
Crash Site Demographics		
Weather Condition		
Clear	Reference	
Blowing sand, dirt, snow	0.000 (0.000, 2.42e+164)	0.998
Cloudy	3.578 (2.373, 5.297)	**< 0.001**
Fog, smog, smoke	10.486 (1.632, 36.727)	**0.0055**
Rain	1.528 (0.843, 2.585)	0.256
Severe crosswinds	0.000 (0.000, 8.63e+209)	0.998
Sleet, hail, freezing rain/drizzle	0.000 (0.000, 1.33e+04)	0.998
Snow	0.000 (0.000, 117.28)	0.998

Other	0.000 (0.000, 4.19e+35)	0.998
Time of Crash		
Afternoon	Reference	
Morning	1.141 (0.678, 1.916)	0.874
Night	2.923 (1.894, 4.611)	**< 0.001**
Driver-related Issue		
Teen Driver	0.702 (0.330, 1.310)	0.477
Senior Driver	1.530 (0.891, 2.482)	0.215
Intoxicated-suspected Driver	3.527 (2.018, 5.866)	**< 0.001**
Child Passenger	0.686 (0.320, 1.291)	0.477
Vehicle Mechanics		
No Airbag	1.652 (0.977, 2.685)	0.123
No Belt/Protection	8.753 (5.587, 13.406)	**< 0.001**

Table 5 shows the adjusted odds of fatal crash for all factors. The adjusted analysis shows very similar results as the unadjusted analyses. Cloudy weather (OR: 3.6, 95% CI: 2.4 to 5.3) and foggy weather (OR: 10.5, 95% CI: 1.6 to 36.7) were associated with significantly higher odds compared to clear weather. Driving at night (OR: 2.9, 95% CI: 1.9 to 4.6) was associated with significantly higher odds compared to driving in the afternoon. Intoxicated drivers (OR: 3.5, 95% CI: 2.0 to 5.9) and lack of belt protections (OR: 8.8, 95% CI: 5.6 to 13.4) were associated with significantly higher odds of fatal crashes.

Discussion

This study explored the risk factors associated with fatal cases when a crash happened using logistic regression. Most of the assumptions are met due to the characteristics of this data set. Since all covariates were categorical, there were no outliers in the data set and linear relationship was not required between the logit of the outcome and each covariate. The analysis was conducted on cases, so each observation was independent.

Multicollinearity can be a problem since some of the covariates might be correlated. One intuitive example is that people are more likely to drink at night than in the morning or in the afternoon. To check if this assumption is violated, we use a correlation heatmap and check the VIF for the model.

Figure 1. **Correlation Heatmap for All Covariates Included in Model.**

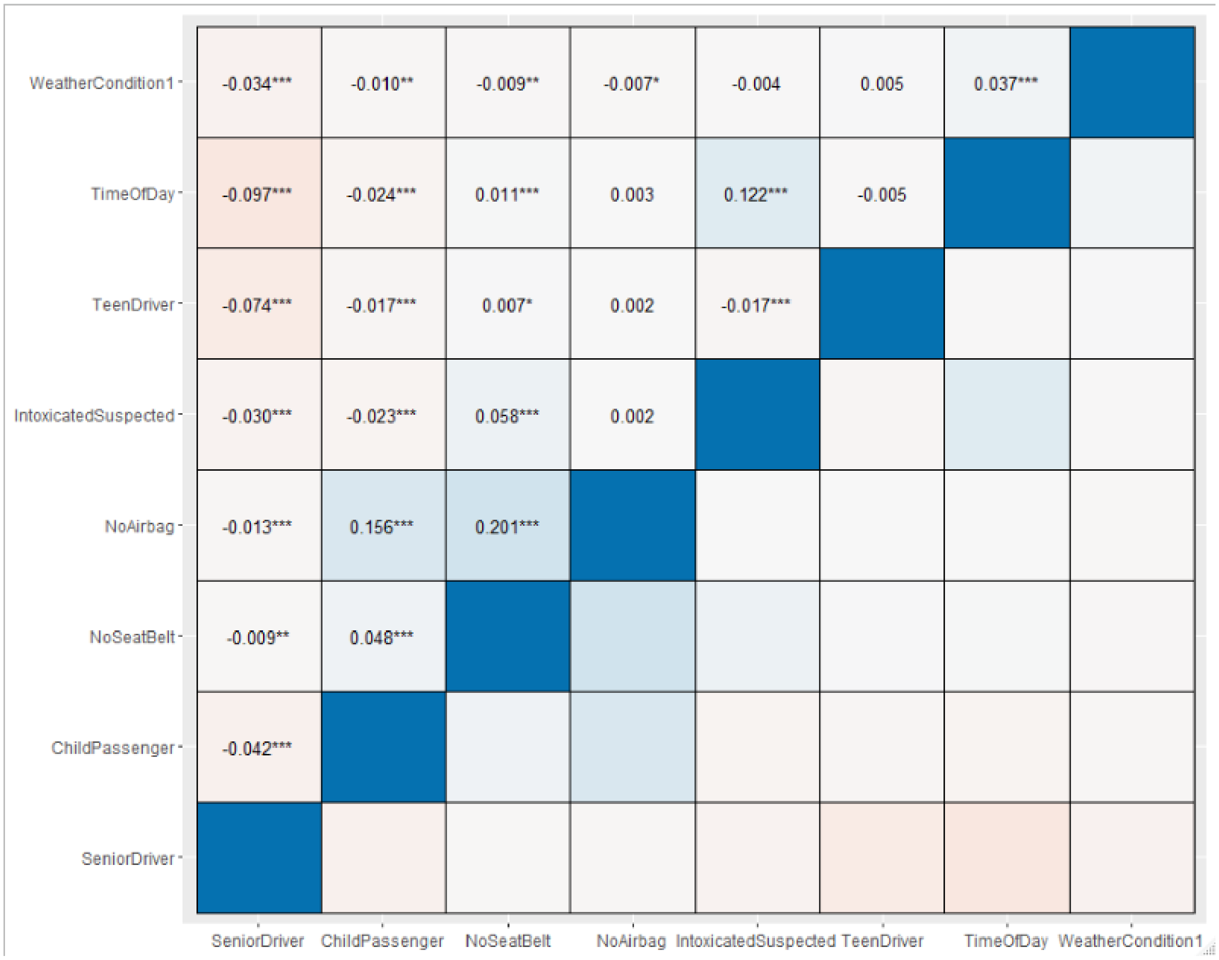

Table 6. **VIF for All Covariates Included in Model.**

Weather Condition	Time of Day	Teen Driver	Senior Driver	Intoxicated Suspected	Child Passenger	No Airbag	No Belt Protection
1.03	1.11	1.01	1.04	1.15	1.03	1.18	1.22

The correlation heatmap (Figure 1) does not show high correlations between any pair of covariates since the absolute values are all below 0.2. The VIFs for all covariates (Table 6) are well below 5, so the assumption of no multicollinearity is not violated.

Furthermore, from Table 1, we can see some categories have a very small event rate (event rate < 1%), which might largely skew the results of the logistic regression. As a result, some of the confidence intervals of Model 1 and Model 4 shown in Table 2 and Table 5 have an unreasonably wide range. Thus, to obtain more meaningful and interpretable estimated coefficients, we can remove the observations within categories with low event rate and fit the models once again.

Table 7. Odds Ratios for Crash Site Demographics Removing Categories with Event Rate < 1% (Model 1 Supplement).

Variable	Odds Ratio (95% CI)	P-value (after BH)
Weather Condition		
Clear	Reference	
Cloudy	3.363 (2.235, 4.965)	**< 0.001**
Fog, smog, smoke	13.398 (2.181, 43.561)	**< 0.001**
Rain	1.410 (0.779, 2.380)	0.270
Time of Crash		
Afternoon	Reference	
Morning	1.128 (0.669, 1.893)	0.649
Night	3.406 (2.245, 5.297)	**< 0.001**

Table 8. Adjusted Odds Ratios Removing Categories with Event Rate < 1% (Model 4 Supplement).

Variable	Odds Ratio (95% CI)	P-value (after BH)
Crash Site Demographics		
Weather Condition		
Clear	Reference	
Cloudy	3.578 (2.373, 5.297)	**< 0.001**
Fog, smog, smoke	10.486 (1.632, 36.727)	**0.0039**
Rain	1.528 (0.843, 2.585)	0.181
Time of Crash		
Afternoon	Reference	
Morning	1.141 (0.678, 1.916)	0.617
Night	2.923 (1.894, 4.611)	**< 0.001**
Driver-related Issue		
Teen Driver	0.702 (0.330, 1.310)	0.337
Senior Driver	1.530 (0.891, 2.482)	0.152
Intoxicated-suspected Driver	3.527 (2.018, 5.866)	**< 0.001**
Child Passenger	0.686 (0.320, 1.291)	0.337
Vehicle Mechanics		
No Airbag	1.652 (0.977, 2.685)	0.087
No Belt/Protection	8.753 (5.587, 13.406)	**< 0.001**

Table 7 and Table 8 demonstrate the results for models without low event rate categories. These results are almost the same as models including all observations. Thus, these categories did not have much effect on the original model, probably due to the huge sample size, but deleting them might make the results look more convincible.

Though the models did not violate any important assumptions, this study has several limitations. The study population of this study was limited to one city in a certain period, and therefore may be biased if we want to generalize this result to other locations. Also, though the sample size is large enough to generate enough power, some categories still have a very small number of observations and even smaller number of events as addressed above, so the results are neither valid nor interpretable for those categories.

The data merging process also yielded some problems that can potentially affect the validity of the results. Since the person-level data set was merged into the case-level data set, the final response variable in the analysis may not reflect the true situation. For example, for two drivers involved in one crash, one sober and one intoxicated, the death of either one of them would generate the same case-level data, but the meaning of death of these two are apparently different and should yield different result. This is a limitation in the statistical analysis plan and should be taken care of if similar studies are to be conducted.

There are also problems with the original data set. Since this study used data collected for administrative use rather than research use, and all data were inputted manually, there are inevitable mistakes with unknown pattern and unknown frequency in the data sets. During data selection by inclusion criteria, some observations obviously misspelled the name of location by adding or losing a couple of letters. These observations were selected based on intuition and did not have any ways of validation. In the case of including locations with a similar spelling as the study location or excluding observations due to spelling mistakes, results can be biased. Errors also occur during data merging process in the form of mismatching information, i.e., the information concluded from the person-level data did not match that in the case-level data. This type of error was somehow minor since the outcome is binary, but it is still problematic when the information shows contradiction. These issues should be addressed and considered in a sensitivity analysis but given limited resources to the public data and difficulties validating past records, this study will not conduct such analysis.

Notwithstanding these limitations, this study provides an overview of possible risk factors associated with fatal crashes when a car crash happens, which therefore provides some insights on how to drive more safely and which situations to avoid driving in the future.