

2014

## Identifying and Quantifying Factors Affecting Injury Severity of Young Drivers Involved in Single Vehicle Crashes Occurring within Curves on Rural Two-Lane Roads in Louisiana

Cory Hutchinson

*Louisiana State University and Agricultural and Mechanical College*

Follow this and additional works at: [https://digitalcommons.lsu.edu/gradschool\\_dissertations](https://digitalcommons.lsu.edu/gradschool_dissertations)



Part of the [Human Resources Management Commons](#)

---

### Recommended Citation

Hutchinson, Cory, "Identifying and Quantifying Factors Affecting Injury Severity of Young Drivers Involved in Single Vehicle Crashes Occurring within Curves on Rural Two-Lane Roads in Louisiana" (2014). *LSU Doctoral Dissertations*. 1178.

[https://digitalcommons.lsu.edu/gradschool\\_dissertations/1178](https://digitalcommons.lsu.edu/gradschool_dissertations/1178)

This Dissertation is brought to you for free and open access by the Graduate School at LSU Digital Commons. It has been accepted for inclusion in LSU Doctoral Dissertations by an authorized graduate school editor of LSU Digital Commons. For more information, please contact [gradetd@lsu.edu](mailto:gradetd@lsu.edu).

IDENTIFYING AND QUANTIFYING FACTORS AFFECTING  
INJURY SEVERITY OF YOUNG DRIVERS INVOLVED IN SINGLE  
VEHICLE CRASHES OCCURRING WITHIN CURVES ON RURAL TWO-  
LANE ROADS IN LOUISIANA

A Dissertation

Submitted to the Graduate Faculty of the  
Louisiana State University and  
Agriculture and Mechanical College  
in partial fulfillment of the  
requirements for the degree of  
Doctor of Philosophy

in

The School of Human Resource Education  
and Workforce Development

by

Cory Hutchinson

B.S., Louisiana State University, 1991

M.S., Louisiana State University, 1993

MBA, Louisiana State University, 1998

M.S., Louisiana State University, 2014

December 2014

## ACKNOWLEDGEMENTS

I would like to thank Dr. Michael Burnett who served as my committee chair. He has taught me patience and has supported my research with important advice and helpful suggestions for improvement. In addition, his encouragement throughout my program of study has helped me to stay focused.

I am also deeply grateful to Dr. Helmut Schneider, my boss, committee member, and mentor. This study would not have been possible without his guidance, expertise, and assistance. Dr. Schneider always has my best interest in mind and constantly supports me in all my endeavors. I will forever be thankful for his encouragement and hope we will always have such a great working and professional relationship.

I feel honored to have Dr. Satish Verma and Dr. William Black serve on my dissertation committee. I am thankful for all your time, feedback, thoughts, and support during this endeavor.

I would like to thank my parents Barry and Earline Hutchinson for their love, encouragement, and support throughout my life. I especially want to thank them for instilling in me a deep appreciation for higher education.

Finally, special thanks go out to my family. My two children Nicholaus and Corrine have shown tremendous patience and understanding throughout my PhD journey. Last, but not least, my wife Laurene Hutchinson has truly been there for me and without her love and support I would have never been able to accomplish this goal. Her sacrifices and hard work during these past seven years will never be forgotten.

## TABLE OF CONTENTS

ACKNOWLEDGEMENTS .....	ii
LIST OF TABLES .....	vi
LIST OF FIGURES .....	vii
ABSTRACT .....	ix
CHAPTER 1. INTRODUCTION .....	1
Motivation .....	1
Objectives .....	4
CHAPTER 2. LITERATURE REVIEW .....	6
Injury Contributing Factors Overview .....	6
Young Drivers .....	6
Risk Taking.....	7
Gender .....	8
Urban versus Rural .....	8
Curves .....	9
Roadway Departures.....	10
Environmental .....	10
Vehicle.....	11
Injury Severity Prediction Modeling .....	11
Logistic Regression .....	11
Bayesian Networks .....	15
CHAPTER 3. RESEARCH METHODOLOGY .....	18
Injury Severity Prediction Model Selection .....	18
Dependent Variable: Driver Severity .....	18
Independent Variables .....	19
Binary Logistic Regression Explanatory Variables.....	20
Bayesian Network Explanatory Variables.....	23
Model Description .....	26
Logistic Regression Model.....	26
Bayesian Network Model .....	28
CHAPTER 4. DATA .....	33
Crash Data .....	33
Location Data .....	33
Roadway Data .....	34
Crash Trend .....	35
CHAPTER 5. MODEL ESTIMATION.....	37
Binary Logistic Regression .....	37
Correlation .....	37
Binary Logistic Regression Modeling .....	39

Logistic Regression Results.....	43
Bayesian Network Modeling .....	46
Bayesian Network Results.....	52
Jouffe's Likelihood Matching .....	54
Driver Factors .....	57
Environmental Factors.....	58
Roadway Factors.....	58
Vehicle Factors .....	58
Identify Factors Affecting Driver Injury Level .....	58
<b>CHAPTER 6. ANALYSIS AND DISCUSSION .....</b>	<b>60</b>
Impact of Identified Contributing Factors.....	60
Protection System .....	60
Protection System and Driver Ejection .....	60
Protection System and Airbags.....	61
Substance Suspected.....	62
Substance Suspected and Protection System Usage.....	62
Violations.....	63
Violations and Protection System Usage.....	64
Violations and Substance Suspected .....	65
Distraction .....	66
Gender .....	67
Most Harmful Event .....	68
Time of Day.....	69
Vehicle Type.....	70
Vehicle Year.....	70
<b>CHAPTER 7. CONCLUSIONS AND FUTURE RESEARCH DIRECTIONS .....</b>	<b>72</b>
Driver Factors.....	72
Ejection and Protection System .....	73
Substance Suspected.....	74
Violation (Careless Operation and Speeding) .....	74
Youth Drivers .....	75
Environmental Factors .....	75
Harmful Events (Rollover and Hitting a Pole or Tree) .....	75
Other Environmental Factors.....	76
Roadway Factors .....	77
Vehicle Factors.....	77
Benefits of Bayesian Networks .....	78
Causal Inference .....	78
Directed Acyclic Graphs .....	79
Investigation of Multiple Variable Interactions.....	80
Variables can Support Multiple Outcome Values .....	80
Direction of Future Research .....	80
Data.....	80
Creating a Casual Network.....	81
Establishing a Quantitative Relationship between Driver Behavior and Crashes.....	81

REFERENCES .....	83
APPENDIX: HSRG Predicted Alcohol Formula .....	89
VITA .....	90

## LIST OF TABLES

Table 1	Explanatory Variables used in Logistic Regression Model .....	20
Table 2	Explanatory Variables used in Bayesian Network Model .....	23
Table 3	Overview of Single Vehicle Young Driver Curve Crashes onTwo-Lane Rural Highways .....	35
Table 4	Summary Data for all Crashes in Louisiana .....	36
Table 5	Correlated Variables .....	38
Table 6	Logistic Regression Coefficient Table for Driver Injury as a Function of 31 Predictors .....	40
Table 7	Significant Variables within the 31 Predictors Model.....	41
Table 8	Logistic Regression Coefficient Information for Driver Injury.....	42
Table 9	Node Significance with Injury Level.....	53
Table 10	Direct Effect of Driver Factors on Driver Injury .....	57
Table 11	Direct Effect of Vehicle Factors on Driver Injury .....	58
Table 12	Driver Injury Contributing Factors .....	59
Table 13	LA Alcohol Related Crash Information for Young Drivers .....	68

## LIST OF FIGURES

Figure 1	Sample Bayesian Network Model .....	30
Figure 2	ROC Curve Information .....	42
Figure 3	Initial Bayesian Network as Unconnected Nodes .....	47
Figure 4	Bayesian Network for Potential Factors .....	48
Figure 5	Mutual Information Shared Between Nodes .....	49
Figure 6	Bayesian Network without Highly Correlated Variables .....	50
Figure 7	Bayesian Network with Clustering of Factors.....	51
Figure 8	Final Bayesian Network Model.....	52
Figure 9	Relationship between Protection System and Substance Suspected in the BN.....	53
Figure 10	Evaluating Driver Injury Based on Gender .....	55
Figure 11	Direct Effect of Gender on Driver Injury .....	55
Figure 12	Youth Driver Injury Levels .....	60
Figure 13	Direct Effect of Seatbelt Use on Youth Driver Injury.....	61
Figure 14	Youth Driver and Ejection/Seatbelt Information .....	61
Figure 15	Youth Driver and SeatBelt/Airbag Information .....	62
Figure 16	Direct Effect of Substance Suspected On Youth Driver Injury .....	63
Figure 17	Youth Drivers Protection System Usage and Substance Suspected Information.....	63
Figure 18	Youth Driver Injury Percentages for Violation Information.....	64
Figure 19	Protection System Usage When No Violation .....	64
Figure 20	Protection System Usage With Violation.....	65
Figure 21	Substance Suspected and Violation Information.....	66
Figure 22	Youth Driver Injury and Distraction Information .....	66



Figure 23 Youth Driver and Distraction Information .....	66
Figure 24 Youth Driver, Distraction, and Substance Suspected Information.....	66
Figure 25 Youth Driver Gender Information.....	67
Figure 26 Youth Driver and Most Harmful Event with No Violation Information.....	68
Figure 27 Youth Driver and Most Harmful Event with Violation Information .....	69
Figure 28 Youth Driver Injury and Substance Suspected/Time of Day Information .....	70
Figure 29 Driver and Vehicle Type Information .....	70

## ABSTRACT

This study investigates factors affecting young driver injury levels for single vehicle crashes occurring within curves on rural two-lane roads in Louisiana. Although the number of fatal and serious injury crashes involving young drivers is declining, young drivers are still overrepresented in crashes and crashes are still the leading cause of death for young drivers.

Driver injury prediction models are formulated using binary logistic regression and Bayesian Network (BN) modeling. Binary logistic regression models have commonly been used in safety studies to analyze injury levels of occupants involved in crashes over the past few decades. More recently, a few safety studies have begun to use BN models to evaluate injury levels.

This study identifies eight significant factors affecting youth driver injury levels: air bag, distracted, ejected, gender, protection system, substance suspected, violation, and most harmful event. Of these factors distracted, protection system, substance suspected, and violation are human factors which can be modified through educational programs.

While both models are able to identify statistical significant variables, more insight is gained from the BN model. For instance, both models found gender to be statistically significant. While the logistical regression model finds males are 0.751 times less likely to be injured than female, the BN finds gender only has a 0.02% direct effect on injury. The BN shows that it is not gender itself that affects driver injury level, but the different behavior characteristics of males versus females which affect injury levels. Males are less likely to wear seatbelts and more likely to be suspected of alcohol in crashes. It is these driver behaviors, not the gender of the driver, which affects injuries.

This study also has a number of theoretical and practical implications. As the first study to utilize BN modeling in evaluating driver injury levels in Louisiana, it expands the literature of BN models being used for analyzing injury levels in car crashes. The findings are also important to driver educational and safety professionals. By identifying factors affecting young driver injury levels, educational and training programs can be enhanced to target specific human behaviors to save more lives.

## CHAPTER 1. INTRODUCTION

### **Motivation**

The motivation for this study is to gain insight and understanding of driver, environmental, roadway, and vehicle characteristics in single vehicle traffic crashes occurring within rural two-lane curves resulting in a young driver fatality or serious/moderate injury in Louisiana. Identifying and quantifying these characteristics can lead to potential countermeasures, including education and training programs to save lives.

According to the U.S. Department of Transportation's National Highway Traffic Safety Administration (NHTSA), 33,561 people were killed in traffic crashes in 2012 (NHTSA, 2012). This is an average of nearly 92 people a day, or one death nearly every 16 minutes. While this is the first increase in fatalities since 2005, the United States averages more than 30,000 lives lost in traffic crashes on a yearly basis. Within Louisiana, 772 people were killed in traffic crashes in 2012 (HSRG, 2014). This is the state's first increase since 2007 and equals Louisiana average of traffic fatalities over the past five years (HSRG, 2014).

To help save more lives, Louisiana created a comprehensive, multidisciplinary Strategic Highway Safety Plan (SHSP) to reduce motor vehicle-related fatalities and serious/moderate injuries. This ambitious plan aims to have zero deaths with an interim goal of reducing traffic fatalities and serious/moderate injuries by 50% before 2030. In order to achieve these goals, effective crash countermeasures must be put in place, including reducing crashes involving young drivers and roadway departures.

Young drivers lack experience in driving and are more willing to engage in risk taking behaviors. They lack proper skills and judgment one can only obtain from years of driving, making them more susceptible to being involved in a crash. Young drivers are also more likely to not wear their safety belt, speed, drive impaired, and become distracted, all of which increases

their chances of being seriously/moderately injured, if not killed, in automobile crashes (Beirness et al., 2004; Porter and Whitton, 2002; Boyce & Geller, 2002). From 2005 through 2012, young drivers aged 15 – 24 represented 17.23% of all license drivers in Louisiana but accounted for 31.99% of fatal crashes and 40.51% of serious/moderate injury crashes (HSRG, 2014).

Crashes within curves are more likely to result in severe injuries compared to straight roadway sections and a larger portion of single vehicle crashes occur within curves (Hung, 2002). Roadway departure is a major concern within curves, research has shown injury severity levels are higher when drivers leave the roadway and strike a fixed object (Chen, 2010; Hummer, 2010; Torbic et al., 2004). Curve crashes, particularly on rural two-lane roads, have long been a safety concern for transportation professionals (AASHTO, 2010; AASHTO, 2005). Between 2005 and 2012, over one-third of all single vehicle crashes on Louisiana's two-lane rural routes occurred within curves (HSRG, 2014). Of these crashes, 2.36% were fatal and 1.14% involved serious/moderate injuries for the drivers, compared to only 0.60% fatal and 0.86% serious/moderate injuries occurring in curves for all other road types in the state (HSRG, 2014).

When evaluating driver characteristics to reduce injuries, it is important to identify driver, environment, roadway, and vehicle factors which directly influence the driver's injury severity level. Research studies concerning injury severity level of crashes are increasing, especially within the past six years (Mujalli & de Oña, 2011b). Logistic regression, also referred to as logit modeling, is widely used in research with binary logit modeling being the most-used (Mujalli & de Oña, 2011b). The frequent use of these models can be attributed to their ease of use, widespread acceptability, and incorporation into popular software packages (Jones & Jørgensen, 2003).

While logistic regression models are commonly used to analyze injury severity levels resulting from crashes, research utilizing logistic regression within the area of curves, young

drivers, and single vehicles is limited. Most of the research using logistic regression evaluates older drivers (Dissanayake & Lu, 2002; Robertson & Vanlaar, 2008), pedestrian or bicyclist (Eluru et al., 2008), crash types (Gabauer & Gabler, 2008; Donnell & Mason, 2004; Yan et al., 2005; Tay et al., 2008; Chen et al, 2012), and vehicle types involved in crashes (Becker et al., 2003; Pai, 2009)

Logistic regression models have their assumptions and when these assumptions are violated, erroneous estimates of injury severity can occur (Chang & Wang, 2006). Logistic regression assumes a linear relationship between the dependent and independent variables and this assumption does not always hold when analyzing crash data. For instance, the relationship between driver injury and seatbelt use is not linear in nature (HSRG, 2014). Also, logistic regression is sensitive to high correlation among independent variables. When gender and seatbelt use are used as predictor variables to analyze driver injury level, collinearity exists due to the fact that females more often wear their seatbelts in fatal and severe crashes compared to males (HSRG, 2014).

An alternative modeling technique and one that is being used more frequently in other fields, but can be applied to crash data analysis, is Bayesian Network (BN) modeling. One major benefit of BNs over logistic regression models is that BNs do not need to know any pre-defined relationships between predictor variables and the outcome variable of interest. Bayesian Networks also offer the advantages of easily identifying underlying patterns in the data, investigating relationships between variables of interest, and making predictions based on those relationships (de Oña et al., 2011).

## **Objectives**

Given the limited research in young drivers' injury severity levels in single vehicle crashes occurring within curves on rural two-lane roads, this research effort investigates the following research question:

1. What driver, environmental, roadway, and/or vehicle characteristics influence injury severity levels of young drivers involved in single vehicle crashes within Louisiana's rural two-lane curves?

To answer this question, the traditional research methodology of binary logistic regression modeling will be used. However, recent research in driver injury modeling has begun using Bayesian Networks (Conrady & Jouffe, 2013b; de Oña et al, 2013; de Oña et al., 2011; Mujalli & de Oña (2011); Simoncic, 2004). Therefore, this research effort will also address the following sub-questions:

1. Can a Bayesian Network model be developed to identify driver, environmental, roadway, and/or vehicle characteristics influencing injury severity levels of young drivers involved in single vehicle crashes within Louisiana's rural two-lane curves?
2. What benefits, if any, exists using a Bayesian Network model over the traditional binary logistic model?

The original contribution of this study is developing a binary logistic regression model to identify factors which directly affect young driver injury levels occurring in single vehicle crashes within curves on rural two-lanes in Louisiana. This information will help safety and educational professional develop training and educational material to help save more lives. These materials can also be used as countermeasures within LA's SHSP to help reduce the number of young drivers and roadway departures fatal and serious/moderate injury crashes. The

second part of the study involves developing a BN model to answer the same question, compare the results of the two models, and identify any advantages using a BN model.



## CHAPTER 2. LITERATURE REVIEW

### **Injury Contributing Factors Overview**

Over the past few decades, considerable research has been conducted on factors contributing to crash injury levels. Shinar (2007) states 90% of crashes are due to driver errors. In 2003, the Government Accountability Office (GAO, 2003) identified human factors as the most prevalent, followed by roadway and then vehicle factors, when analyzing factors contributing to motor vehicle crashes. Veridian Engineering (Hendericks et al., 1999), when studying driver behaviors and unsafe driving acts, also concluded human factors are the most prominent factors influencing injury levels. The Tri-Level Study (Treat et al., 1979) conducted in Indiana in the late 1970s further identified human factors as most important, while vehicle factors are least important.

#### Young Drivers

Young drivers are overrepresented in automobile crashes. According to teen driver facts sheet produced by the Centers for Disease Control and Prevention (CDC, 2010), young adults aged 15-24 represent 14% of the US population, but account for 58% of the total cost of motor vehicle injuries. Data from the CDC's Web-based Injury Statistics Query and Reporting System (WISQARS, 2010) showed unintentional injury as the leading cause of death for people aged 15 - 24 from 2005 through 2010. Hendrick (2010) analyzed data from the CDC and reported motor vehicle crashes are the top cause of unintentional deaths for 12 – 19 year olds and accounts for 73% of their fatalities.

These trends are seen worldwide. The Organization of Economic Cooperation and Development (OECD, 2006) reported traffic crashes account for the greatest number of deaths of people aged 15 – 24 in 23 industrialized countries. This report also stated drivers below the age of 24 are two times more likely than other drivers to be killed in car crashes in the United States

and are over-represented in single-vehicle motor vehicle crashes, which are closely associated in risk taking behaviors.

While National Highway Transportation Safety Administration (NHTSA) reported the number of crashes and fatalities are declining for young drivers over the past few years, according to the National Center for Health Statistics latest data in 2007, motor vehicle crashes were still the leading cause of death for 15 – 20 year olds (CDC, 2010). This trend can be seen in 2011 where young adults aged 16 – 20 and 21 – 24 had the highest two fatality rates per 100,000 population in the US at 13.98 and 16.61 respectively. In the same year, these two groups also had the highest injury rate at 1,252 (NHTSA, 2011).

### Risk Taking

Young drivers are more willing to engage in risk taking behaviors such as not wearing their safety belts, speeding, driving impaired, and easily becoming distracted, all of which increases their chances of being killed or seriously injured in automobile crashes (Beirness et al. 2004). Porter and Whitton (2002) used GPS and video technology to study driver behaviors of young (20 to 29), middle-aged (30 to 64), and older (65 years of age or older) drivers. They found young drivers drove faster, had shorter deceleration distance, and smaller acceleration times as compared to middle-aged and older drivers. Younger drivers also received a substantially higher number of violation infractions for speeding, not stopping fully at stop signs, and following too close (Porter and Whitton, 2002). Boyce and Geller (2002) using an instrumented vehicle to obtain behavioral data from drivers aged 18 to 82, found younger drivers are more likely to speed, follow too close, and engage in in-vehicle behavior not relevant to the driving task.

## Gender

While young drivers in general were more likely to take additional risk while driving, young male drivers have higher crash rates and are more likely to violate traffic laws and engage in risk taking behaviors (Yagil 1998). Yagil also concluded young male drivers evaluated traffic laws as less important than other laws and are less likely to comply with traffic laws.

Clarke et al. (2006) found that both male and female 17-19 year olds are over-represented in crashes occurring in curves within rural areas. Males are over-represented in crashes occurring at night, with or without street lights. Both males and females demonstrated a decline in curve crashes within rural areas as their age increases from 17-19 to 20-22 and then again from 20-22 to 23-25. Maycock (2002) further identified young males as having higher crash involvement rates than their female counterparts.

## Urban versus Rural

Peek-Asa et al. (2010) studied crash data in Iowa from 1995 – 2004 for drivers between the ages of 10 through 18 to examine their characteristics on rural versus urban roads. In Iowa, teenagers cannot obtain a driver license until the age of 14; however, the Iowa crash database contained a large enough number of drivers under the age of 14 to include them. The study found rural teen crashes are 4.7 times more likely to result in a fatal or severe injury than urban crashes. The study also identified young males have a 30% increased odds for a severe crash than young women in rural crashes, and single vehicle crashes are far more frequent in rural areas (65%) as compared to urban areas (10%). Running off the road was the second leading contributing cause of crashes in urban areas and third for rural areas.

Torbic et al. (2004) reported each year nearly 25% of people killed in automobile crashes in the United States are killed in crashes occurring within curves. Of these crashes, 75% of fatal crashes occur in rural areas and more than 70% are on two-lane secondary roads.

## Curves

Young drivers lack the proper skills and judgment obtained from years of experience in identifying and maneuvering around hazardous situations which makes them more likely to be involved in a car crash (Beirness et al. 2004). One such potential hazardous situation for young drivers is safely driving through curvatures in the roadway.

The influence of horizontal curve crashes on the frequency and severity of crashes has long been a concern for transportation safety professionals. The American Association of State Highway and Transportation Officials' Strategic Highway Safety Plan (AASHTO, 2005) and Highway Safety Manual (AASHTO, 2010) both address the influence of horizontal curves on highway safety.

Huang et al. (2002), from the University of North Carolina's Highway Safety Research Center, conducted a study for the North Carolina Department of Transportation to identify factors and countermeasures for severe crashes in North Carolina. Two of the main conclusions from the study when looking at curves are that crashes within curves are more likely to be more severe than on straight roadway sections and a larger portion of single vehicle crashes occur within curves.

Hummer et al. (2010) also studied crash data in North Carolina to obtain a better understanding of crashes within curves. Roadway data was collected from the North Carolina Department of Transportation and analysis was performed to evaluate curve crashes on two-lane roads, all crashes on two-lane roads, and all crashes on all roads. Evaluating crashes within curves revealed 21% of all two-lane crashes occur in curves compared to only 14% among all roads. Of the two-lane curve crashes, 70% occur in rural areas compared to only 45% for all crashes statewide, demonstrating rural two-lane curves are overrepresented when evaluating crashes.

### Roadway Departures

Hummer et al. (2010) analyzed crashes based on severity levels and found two-lane curve crashes result in nearly twice the percent of fatal and disabling injuries as compared to crashes on all two-lane roads and all roads statewide. Two-lane curve crashes make up the majority (52%) of collision with fixed objects, over twice the percent as compared to all two-lane road crashes. For most harmful events; rollover, collision with trees, and collision with ditches were identified as the main concerns for single vehicle crashes on two-lane curves. This finding was similar to Queensland Transport (2006), which reported that more severe curve related crashes in Australia involve run-off-road, head-on, rollover, and hitting roadside objects.

Torbic et al. (2004) reported each year nearly 25% of people killed in automobile crashes in the United States are killed in crashes occurring within curves. Within fatal curve crashes, 76% involve a single vehicle leaving the road and striking a fixed object. Chen (2010) found a significant relationship existed between crash severity levels of crashes occurring within curves and striking a tree. Huang et al. (2002) identified run-off-the-road crashes occur mostly within curves on rural two-lane roads, and accounted for the largest number of fatal and serious injury crashes.

### Environmental

Hummer et al. (2010) identified two-lane curve crashes to be more evenly dispersed throughout the time of day and day of week than all two-lane crashes and all road crashes. The study also found the majority of crashes occur during the day (lighting present) and on dry surfaces (clear weather). Chen (2010) found a significant relationship existed between the crash severity levels of crashes occurring within curves and time of the crash.

## Vehicle

Chen (2010) found a significant relationship existed between the crash severity levels of crashes occurring within curves and the manufacturing year of the vehicle.

### **Injury Severity Prediction Modeling**

Studies about traffic crash injury severities have been increasing over time, with the largest number of studies being performed within the past five years (Mujalli and de Oña, 2011b). Injury severity studies focus on factors affecting the severity of the crash outcome. These studies are particularly useful for analyzing severity levels for different driver groups (Hauer, 2006). Regression analysis is widely used in crash severity studies and logistic regression is one of the most commonly used models (Chang & Wang, 2006; Savolainen et al., 2011; de Oña et al. 2011).

## Logistic Regression

Dissanayake (2003) used logistic regression modeling to identify roadway, driver, environmental, and vehicle related factors influencing the injury severity of young drivers involved in run-off-the-road crashes. The study used crash data from 1997 – 1998 from the Florida Traffic Crash Database and created a separate model for each severity level. For fatal crashes the following factors were influential; driver under influence of alcohol or drugs, driver ejected in crash, driver was at fault, restraint device was not used, and impact point was side of vehicle. The models for severe crashes had the following influential factors; driver ejected in crash, restraint device was not used, crashes occurred in a rural area, and driver was male.

Dissanayake & Lu (2002) used crash data from the National Center for Statistics and Analysis for years 1994 to 1996 and identified elements which are more likely to produce severe injuries to older drivers involved in passenger car crashes with fixed objects. They evaluated driver, vehicle, roadway, and environmental elements and treated each as dichotomous variables

(0 or 1). The study identified travel speed and restraint device usage as important parameters in making a difference in injury levels. Higher speed and lack of seatbelt usage increase the chances for more severe injuries. Other significant variables were impact point, alcohol and drug use while driving, driver condition, gender, at fault, rural, and curves.

Mercier et al. (1997) utilized logistic regression modeling to evaluate age and gender as predictors of injury severity levels for individuals involved in head-on highway crashes. Crash data was analyzed from the Iowa Department of Transportation from 1986 through 1993. They controlled for speed by examining only crashes on interstates, freeways, and state highways where the speed limit ranged from 55 to 65 miles per hour. Occupant positions were controlled for by including only drivers and right-front-seat passengers. Possible injury and no injury crashes were excluded in the study. Possible injury was also excluded since “possible” may not prove to be an actual injury. Out of fourteen potential independent variables, only age and safety restraint were found to be significant.

Al-Ghamdi (2002) studied 560 injury crashes occurring on urban roads between 1997 and 1998 in Riyadh, the capital of Saudi Arabi. The outcome variable accident level was captured as either fatal or non-fatal, where non-fatal only included injury crashes. Nine independent variables; location, crash type, collision type, time, cause, at fault, driver age, nationality, vehicle type, and license status, were used in the model. Many of the independent variables were categorical in nature. While the variables location and cause were found to be significant, their interaction effect was not significant. Crashes happening at non-intersections (location) and crashes that occurred because of running a red light (cause) are more likely to result in a serious injury. Also, age is significant showing younger and older drivers are more at risk of sustaining a serious injury.

Using crash data in Alberta, Canada from 2003 to 2005, Barua et al. (2010) studied fatality risk of intersection crashes on rural undivided highways. A response variable of fatal or nonfatal crash was used along with eighteen vehicle, roadway, crash, driver, and traffic related independent factors. Vehicle variables in the final model included truck-tractors and motorcycles. Roadway variables shown to be significant included intersection, traffic operations, and vertical alignment. The season, crash time, collision type, intersection type, and roadway surface condition were significant variables within crash factors. Driver age, gender, fatigue, and impairment were all driver variables included in the final model. Traffic volume was the only traffic variable included in the final model.

Zhu et al. (2010) studied fatal crashes on rural two-lane highways in 1997 and 1998 which occurred in Alabama, Georgia, Mississippi, and South Carolina to determine their impact on crash conditions and potential contributing factors. Using logistic regression, the study developed two crash-type prediction models: single-vehicle versus multiple-vehicle fatal crashes and head-on versus other fatal crashes. The crash type of interest in the logistic regression model was single-vehicle run-off the road fatal crashes where the vehicle overturned or struck a fixed object. The ultimate goal of the study was to identify valuable information and quantify relationships between highway design characteristics and associated performance measures.

Zhu et al. (2010) first created a model using four states. The model generated predictor variables for Georgia, Alabama, and South Carolina which were all sufficient in explaining crash differences across all states, except for Mississippi. Since the objective of the study was to identify rural two-lane-highway fatal crashes models to better understand crash trend in Georgia and other states, the researchers created a three-state model (Alabama, Georgia, and South Carolina). This new model produced results which were suitable for predicting crashes in Georgia-specific conditions. The variables of interest in the three-state model were intersection,



curve to left, crest vertical curve, near commercial driveways, dark without supplement lights, and crash between 1:00 am and 3:00 am. Next, the researchers created four separate models, one for each state. These individual models did not contain the same set of significant independent variables, suggesting the three-state model comprises many of the primary, but possibly not all, factors associated with fatal crashes.

Horizontal curves were identified as a target area for safety improvement on rural two-lane roads in Texas. Schneider IV et al. (2009) used multinomial logistic modeling to assess driver injury severity levels resulting from 10,029 single vehicle crashes on rural two-lane roads between 1997 and 2001. The study examined driver, vehicle, roadway, and environmental factors of crashes to assess their effect on driver injury. Four models were developed; all crashes, crashes within small radius curves (less than 500 feet), crashes within medium radius curves (between 500 and 2,800 feet), and crashes within large radius curves (greater than 2,800 feet). The three curve models identified driver injury levels are more likely to occur in curves with a medium radius, followed by the small radius, and then large radius. While the degree of injury is not significantly different between the groups, driver fatalities were slightly less in small radius groups compared to the medium and large radius groups.

The study also identified drivers' injury levels significantly increase in run-off-the-road crashes where the vehicle collided with a roadside object. Crashes occurring during daylight hours with clear weather also tend to be more severe. Gender of the driver was a factor, as females were 23% to 31% more likely to sustain an injury than males, and driver injury was found to increase with driver age, especially as the curve radius decreases.

High-risk was also found to influence higher injury severities. Alcohol and drug use increased the probability of driver injury by 18% - 40% and fatalities by 243% - 549%. Seat belt usage increased the likelihood of no injury by 415% for serious injuries and 1,012% for

fatalities. Crashes that cause the air bag to deploy, increased the risk of injury to drivers for all curve groups.

These results are similar to Zhang (2010) who analyzed driver, vehicle, roadway, and environmental factors affecting crash severity in Louisiana for crashes occurring between 1999 and 2004. This study used multinomial logit, ordered logit, and ordered mix logit models to relate crash severity to ten possible independent variables. All three models found the curve variable to not have a significant impact on crash severity. While, this study did not find the presence of a curve to be significantly important, it also did not examine whether crashes occurring within curves tend to produce more severe injuries.

### Bayesian Networks

While using BNs to analyze crash data is scarce, BNs are being used more frequently in other fields of study and can easily be applied to the area of crash data analysis.

Simoncic (2004) performed a study to show the potential of BNs when modeling road accidents. He analyzed accident outcome evaluating road characteristics, traffic flow characteristics, time/season factors, characteristics of people within the crash, protection system device usage, vehicle types, and speed of vehicles. When generating the BN, external crash variables (weather, day of week, time of day) and variables related to the driver (age, gender, driving experience, use of safety device, and alcohol usage) were used as root nodes. Variables relating to injury level of the drivers and overall crash were used as leaf nodes.

Evaluating inference results by accident type (fatal/serious injury versus other), speed had an odds ratio of 2.1. A slightly smaller odds ratio was found for wrong side/direction and settlement. When evaluating inference results from the intoxication variables (yes versus no), nighttime had an odds ratio of 3.7. High odds ratios was also found for gender, at-fault and

cause. Based on his research, Simoncic concluded BNs can be utilized within the domain of road-accident modeling.

de Oña et al. (2011) utilized BNs to classify crash injuries for crashes on rural highways in Spain. Eighteen variables describing injury levels, roadway information, weather, crash, and driver information were analyzed for 1,536 crashes to classify injury severity level. Results from the BNs showed the following factors to be more significant in fatal and serious injury crashes; head on collisions and rollover crashes, young driver 18 -25 years of age (especially male drivers), hours of darkness, and crashes resulting in at least one injury.

Mujalli and de Oña (2011) used the same data and evaluation methods to analyze BNs using only the most significant variables compared to using all variables in the dataset. After evaluating different possible combinations, the variables accident type, atmospheric factors, lighting, and number of injuries, were identified as most relevant.

In 2013, de Oña et al. (2013) analyzed accident severity for rural highway crashes in the province of Granada (South of Spain). In this study, the same 18 independent variables used in the previous study were analyzed with injury severity as the dependent variable. The following independent variables contributed the most to severity; accident type, sight distance, time, occupants involved, age, lighting, number of vehicles, number of injuries, atmospheric factors, pavement markings, and pavement width. The study also identified teenagers as having higher probability of injury accidents.

Conrady and Jouffe (2013b) used BNs to provide a robust framework for evaluating the impact of regulatory interventions. This study was conducted to evaluate if occupants within smaller vehicles, which obtain greater fuel economy, are placed at greater risk for injury or death. In the study, crash injury severity was used as the dependent variable and only crashes involving two vehicles with no passengers (only drivers) were evaluated. A BN was created

with the following sixteen variables of interest: driver age, driver sex, crash injury severity, air bag deployed, vehicle curb weight, crash angle, total delta V, energy absorption, footprint of vehicle, number of lanes, use of seatbelt, vehicle model year, speed limit, track width, vehicle type, and vehicle wheel base. It was found that seat belt usage, air bag deployment, and vehicle curb weight all had a major effect on the driver's injury.

## CHAPTER 3. RESEARCH METHODOLOGY

### **Injury Severity Prediction Model Selection**

A driver injury prediction model is needed to establish the relationship between the driver's injury level and contributing factors of the crash. Since the outcome of driver injury level studies is discrete in nature, discrete prediction models are selected as the most appropriate choice. The logistic regression model is chosen since it is the most widely discrete model used when evaluating traffic crash injury levels.

However, traditional logistic regression models have certain limitations which can be overcome when using multilevel models such as BNs. By simulating an environment using a set of variables and their conditional dependencies, BNs are highly flexible models. Furthermore, unlike logistic regression models, BNs are not restricted to assumptions of linear relationships and multicollinearity among variables.

### **Dependent Variable: Driver Severity**

The main focus of this study is identifying and quantifying contributing factors leading to the driver's injury level, therefore the injury level of the driver is the dependent variable in the injury prediction models. Driver injury severity will be measured using the driver injury code reported by the officer and collected on the crash report. Louisiana's crash reports closely follow the Model Minimum Uniform Crash Criteria Guidelines (MMUCC) established as a collaborative effort involving the Governors Highway Safety Association (GHSA), the Federal Highway Administration (FHWA), the Federal Motor Carrier Safety Administration (FMCSA), and the National Highway Traffic Safety Administration (NHTSA).

The five MMUC levels of injury status are fatal, suspected serious injury, suspected minor injury, possible injury, and no apparent injury. This is very similar to Louisiana where the

driver's injury severity is defined as fatal, serious (incapacitating) injury, moderate (non-incapacitating) injury, possible (compliant) injury, and no injury (property damage only).

The response variable for this study is injury level and is coded as binary (dichotomous). The two levels of injured are 1 if the driver is injured (fatal, serious injury, moderate injury) and 0 if the driver is not injured or possibly injured. Possible injury is not considered an injury in this study for two reasons. First, this study will be used to assist Louisiana with their Strategic Highway Safety Plan (LA SHSP) which only evaluates crashes that result in a fatality or serious/moderate injury to anyone involved in the crash. Second, a possible injury is not proven to be an actual injury. As such, only an actual injury or fatality, identified by the officer, is considered as an injury.

For any given year, 22% of all crashes in Louisiana are classified as possible injury and 70% as no injury. Of the remaining 8% of crashes; 0.5% are fatal, 1% are serious injury and 6.5% are moderate injury (HSRG, 2014).

### **Independent Variables**

The selection of independent variables includes consideration from the identified literature review and years of experience in analyzing crash data. To meet the objectives of this study, data concerning the driver, environment, roadway, and vehicle will be used as independent variables.

To analyze human factors, driver information is required. Driver data includes the following characteristics: air bag, distracted, ejection, gender, inattentive, predicted alcohol, protection system, race, substance suspected, age, and violation.

Besides human factors, environmental, roadway, and vehicle factors can contribute to the driver's injury level. Environmental characteristics include day of the week, most harmful event,

lighting, time, and weather. Roadway features include average daily traffic (ADT), curve crash modification factor (CMF), curve length, curve radius, lane width, and shoulder width. Vehicle factors are vehicle type, and vehicle year. A curve's CMF, as defined by the Highway Safety Manual (AASHTO, 2010b), is calculated as:

$$CMF = ABS(((1.55 * Curve Length) + \frac{80.2}{Curve Radius}) / (1.55 * Curve Length)) \quad (3.1)$$

### Binary Logistic Regression Explanatory Variables

For the binary logistic regression model, all explanatory variables are treated as dichotomous variables (0 and 1). Dummy variables are created for those independent variables that are continuous or categorical in nature. For example, curve radius is divided into three dummy variables; small, medium, and large, representing the size of the curve. The creation of dummy variables leads to thirty-four potential independent variables; twelve driver variables, eight environmental variables, ten roadway variables, and four vehicle variables. Summary descriptions and characteristics of the factors and variables used in the logistic regression model are shown in Table 1.

Table 1 Explanatory Variables used in Logistic Regression Model

Explanatory Variable	Description		Percentage
Driver			
Airbag Non-Deployed	1	Non-Deployed or Non-Deployed/Switch Off	61.33
	0	Deployed	24.05
Distracted	1	Distracted	45.21
	0	Not Distracted	54.79
Ejected	1	Partially or Totally Ejected	5.87
	0	Not Ejected	93.37
Male	1	Male	66.12
	0	Female	33.78
Inattentive	1	Inattentive	36.70
	0	Not Inattentive	63.30
Predicted Alcohol	1	Predicted Alcohol	18.62
	0	Not Predicted Alcohol	81.38

(Table 1 continued)

Explanatory Variable		Description	Percentage
Driver			
No Protection System	1	No or Improper Seatbelt Usage	14.09
	0	Shoulder and Lap Belt Used	76.24
African American	1	African American	22.36
	0	Caucasian	75.15
Substance Suspected	1	Alcohol and/or Drugs Suspected	18.41
	0	Neither Alcohol nor Drugs Suspected	76.18
Youth Driver	1	Driver Age Between 15 – 24	38.01
	0	Driver Age Between 25 – 54	61.99
Violation			
Careless Operation	1	Careless Operation	63.03
	0	Not Careless Operation	36.97
Speeding	1	Speeding	3.35
	0	Not Speeding	96.65
Environmental			
Weekend	1	Friday, Saturday, or Sunday	50.15
	0	Monday, Tuesday, Wednesday, or Thursday	49.85
Most Harmful Event			
Culvert or Ditch	1	Culvert or Ditch	19.01
	0	Not a Culvert or Ditch	80.99
Other Fixed Object	1	Other Fixed Object Beside Culvert, Ditch, Pole, or Tree	13.19
	0	Not a Fixed Object	86.81
Pole or Tree	1	Pole or Tree	24.97
	0	Not a Pole or Tree	75.03
Rollover	1	Rollover	11.96
	0	Not a Rollover	88.04
Dark	1	Dark - No Street Lights	47.25
	0	Daylight, Dark – Continuous Street Light, Dark- Street Light Intersection Only, Dusk, or Dawn	52.22
6:00 – 19:00	1	Between 6:00 AM and 7:59 PM	54.01
	0	Between 8:00 PM and 5:59 AM	45.99
Non-Clear Weather	1	Cloudy, Rain, Fog/Smoke, Sleet/Hail, Snow, Severe Crosswind, Blowing, Sand/Soil/Dirt/Snow, or Other	35.39
	0	Clear	64.29



(Table 1 continued)

Explanatory Variable		Description	Percentage
Roadway			
ADT GT 3000		Greater Than Equal To 3,000 Less Than 3,000	63.76 36.24
Curve CMF LT .5	1 0	Greater Than Equal To .5 Less Than .5	63.47 36.53
Curve Length Small	1 0	Less Than .15 Greater Than Equal To .15	42.45 57.55
Medium	1 0	Between .15 and .2999 Not Between .15 and .2999	46.33 53.67
Large	1 0	Greater Than Equal to .3 Less than .3	11.23 88.77
Curve Radius Small	1 0	Less Than 500 Greater Than Equal To 500	8.60 91.40
Medium	1 0	Between 500 and 2,799 Not Between 500 and 2,799	56.61 43.39
Large	1 0	Greater Than Equal to 2,800 Less than 2,800	34.79 65.21
Lane Width LT 12	1 0	Less Than 12 Greater Than Equal To 12	19.01 80.99
Shoulder Width LT 4	1 0	Less Than 4 Greater Than Equal To 4	65.74 34.26
Vehicle			
Vehicle Type Passenger Car	1 0	Passenger Car Not a Passenger Car	45.07 54.93
Light Truck	1 0	Light Truck Not a Light Truck	32.29 67.71
SUV	1 0	SUV Not a SUV	11.06 88.94
Vehicle Year LT 2000	1 0	Less than 2000 Greater than or equal to 2000	36.49 64.43

### Bayesian Network Explanatory Variables

Within the BN model, all explanatory variables are treated as categorical variables. Since there is no need to create dummy variables within BNs, there are twenty-four independent variables; eleven driver variables, five environmental variables, six roadway variables, and two vehicle variables. Summary descriptions and characteristics of the factors and variables used in the BN model are shown in Table 2.

Table 2 Explanatory Variables used in Bayesian Network Model

Explanatory Variable	Description	Percentage
Driver		
Airbag	Deployed	24.05
	Non-Deployed	61.15
	Non-Deployed/Switch Off	0.18
	Not Applicable	13.42
	Not Reported	0.28
	Unknown	0.92
Distracted	Not Distracted	54.79
	Distracted	45.21
Ejected	Not Ejected	93.37
	Not Reported	0.20
	Partially	0.66
	Totally Ejected	5.21
	Unknown	0.56
Gender	Female	33.78
	Male	66.12
Inattentive	Not Inattentive	63.30
	Inattentive	36.70
Predicted Alcohol	Not Predicted Alcohol	81.38
	Predicted Alcohol	18.62
Protection System	Lab Belt	0.28
	None Used	13.16
	Not Reported	0.18
	Shoulder and Lap Belt Used	76.24
	Shoulder Belt Only	0.65
	Unknown	9.49
Driver		
Race	African American	22.36
	American Indian	0.13
	Caucasian	75.15
	Not Reported	0.34
	Other	2.02

(Table 2 continued)

Explanatory Variable	Description	Percentage
Substance Suspected	Alcohol	14.57
	Alcohol and Drugs	2.10
	Drugs	1.75
	Neither Alcohol nor Drugs	76.18
	Not Reported	0.74
	Unknown	4.67
Youth Driver	Yes (Driver Age Between 15 – 54)	38.01
	No (Driver Age Between 25 – 24)	61.99
Violation	Careless Operation	63.03
	No Violation	17.68
	Other	14.02
	Speeding (Exceeding Stated Speed Limit or Exceeding Safe Speed)	3.35
	Unknown	1.91
Environmental		
Day of the Week	Monday, Tuesday, Wednesday, or Thursday	49.85
	Friday, Saturday, or Sunday	50.15
Most Harmful Event	Culvert or Ditch	19.01
	Other	30.87
	Other Fixed Object	13.19
	Pole or Tree	24.97
	Rollover	11.96
Lighting	Dark - Street Lights	3.36
	Dark – No Street Lights	47.25
	Dark- Street Light Intersection	2.13
	Dawn	1.90
	Daylight	43.43
	Dusk	1.39
	Not Reported	0.17
	Unknown	0.26
Time of Day	12 AM – 6 AM	25.92
	6 AM – 12 PM	20.79
	12 PM – 6 PM	23.78
	6 PM – 12 AM	29.51
Environmental		
Weather	Blowing Sand, Soil, or Dirt	0.01
	Cloudy	17.64
	Fog/Smoke	2.55
	Not Reported	0.02
	Other	0.07
	Rain	14.68
	Severe Crosswind	0.09
	Sleet/Hail	0.18

(Table 2 continued)

Explanatory Variable	Description	Percentage
Environmental		
	Snow	0.18
	Unknown	0.30
Roadway		
ADT	1 to 1,000	19.83
	1,001 to 3,000	45.19
	3,001 to 6,000	23.22
	6,001 to 10,000	9.14
	Greater than 10,000	2.61
Curve CMF	0 to 0.249	24.97
	0.250 to 0.499	38.49
	0.500 to 0.749	20.28
	0.750 to 0.999	5.47
	Greater than 0.999	10.78
Curve Length	Large	11.23
	Medium	46.33
	Small	42.45
Curve Radius	Large	34.79
	Medium	56.61
	Small	8.60
Lane Width	Less Than 10	2.03
	10	27.56
	11	26.21
	12	38.65
	Greater Than 12	5.55
Shoulder Width	None	0.71
	1 to 3	33.54
	4 to 6	45.87
	7 to 9	15.53
	Greater Than 9	4.35
Vehicle		
Vehicle Type	Light Truck	32.29
	Other	11.58
	Passenger Car	45.07
	SUV	11.06
Vehicle Year 2000	No (Less than 2000)	36.49
	Yes (Greater than or equal to 2000)	64.43
	Unknown	0.09

## Model Description

### Logistic Regression Model

Logistic regression is widely used in automobile safety studies where the dependent variable measures injury level in a binary format (Dissanayake, 2003; Dissanayake & Lu, 2002, Schneider IV et al. 2009; Zhang, 2010; Chang & Wang, 2006; Tay et al., 2008, Al-Ghamdi, 2002, Qin et al., 2013; Mercier et al., 1997). Logistic regression models are linear regression models where the dependent variable is categorical. For example, if the dependent variable denotes serious, moderate, or possible injury, there would be three categories. Logistic regression models can be used to classify/predict cases based on values of the independent/predictor variables.

Binary Logistic Models (BLM) are logistic regression models where the dependent variable of interest is binary, having one of two possible outcomes. Within crash injury severity studies, a binary outcome may be fatal/non-fatal or injury/no injury. Mujalli and de Oña, (2011b) found BLMs or some extension of it are the most commonly used modeling technique when performing studies evaluating crash injury severity levels.

Shmueli et al. (2010) explains the logistic regression model and odds in the following way. Linear regression uses  $Y$  as a dependent variable, however logistic regression uses a function of  $Y$  called the logit. This logit can then be used to model a linear function of the predictors. Whereas  $Y$  can only take the form of 0 or 1 (category identification),  $p$  can have any interval value between 0 and 1. When expressed as a linear function of  $n$  predictors:

$$p = A + B_1X_1 + B_2X_2 + \dots + B_nX_n \quad (3.2)$$

$p$  is not guaranteed to fall within 0 and 1. However, using the logistic response function guarantees  $p$  is in the interval  $[0,1]$ :

$$p = \frac{1}{1 + e^{-(A + B_1X_1 + B_2X_2 + \dots + B_nX_n)}} \quad (3.3)$$

The odds of the dependent variable being in one group as opposed to the other, is defined as the odds ratio:

$$\text{Odds} = \frac{p}{1 - p} \quad (3.4)$$

The probability can then be computed given the odds of an event:

$$p = \frac{\text{odds}}{1 + \text{odds}} \quad (3.5)$$

Substituting (3.3) into (3.5), the relationship between the odds and the predictors is:

$$\text{Odds} = e^{(A + B_1X_1 + B_2X_2 + \dots + B_nX_n)} \quad (3.6)$$

Taking the log of both sides produces the standard logistic regression model:

$$\log(\text{odds}) \text{ or } \text{logit} = A + B_1X_1 + B_2X_2 + \dots + B_nX_n \quad (3.7)$$

Using the model above, the odds ratio represents the dependent variable being in one group as opposed to the other. When independent variable  $X_j$  increases by one unit and everything else remains the same,  $B_j$  is the multiplication factor by which the odds change. When  $B_j < 0$ , an increase in the variable  $X_j$  decreases the odds of belonging to class 1. Likewise, when  $B_j > 0$ , a decrease in the variable  $X_j$  increases the odds of belonging to class 1.

Logistic regression allows for varying predictor variables (continuous, discrete, and dichotomous), are easily used, and are relatively flexible; however, they have their own modeling assumptions. One such assumption is the pre-defined underlying relationships between dependent and independent variables. Logistic regression assumes a linear relationship between the predictor variables and the logit transform of the outcome variable. While there is no assumption concerning the distribution of predictors, having linearity among the predictors

may enhance the power (Tabachnick & Fidell 2006). Another assumption relates to the absence of multicollinearity. Logistic regression is sensitive to extremely high correlations among predictor variables. Lastly, logistic regression assumes responses from different cases are independent of one another. When these assumptions are violated, erroneous estimates of injury severity can occur (Chang & Wang, 2006).

Binary logistic regression techniques model crash severity as a dichotomous response. These models presume each record examined in the estimation procedure corresponds to an individual injury and assume the residual resulting from the models exhibit independence (Jones & Jørgensen, 2003). However, the assumption of independence may often not hold. For example, different vehicles are equipped with different safety features which can influence their occupants' injury levels. This would tend to show injury levels within the same vehicles as having more similar injury levels than from different vehicles.

An alternative strategy which addresses the issues outlined above, is utilizing multilevel models for analyzing injury severity (Jones & Jørgensen, 2003; Lenguerrand & Laumon, 2006). Multilevel models, such as BNs are gaining popularity in recent years. BNs also offer the advantages of bi-directional induction and probabilistic inference (de Oña et al., 2011).

### Bayesian Network Model

Charniak (1991) states, "The best way to understand Bayesian networks is to imagine trying to model a situation in which causality plays a role but where our understanding of what is actually going on is incomplete, so we need to describe things probabilistically." BNs represent a particular situation as a coherent whole and are comprised of two components, qualitative and quantitative. The qualitative portion consists of the directed acyclic graph (DAG), also known as the structure, which represents variables and their dependencies using nodes and links. Whereas,

the quantitative part, captures the probabilities that quantifies the relationships between variables and their parents.

Variables are represented as nodes and their interactive dependencies as links between related nodes. A node symbolizes a variable and captures that variable's current state. More often, variables are discrete in nature having one of two values. However, this is not always true and variables may have multiple values.

The links within a BN specify the independence assumptions between the two variables. This information is used to determine the probability distribution among the variables in the network. Each node is associated with a probability function which uses a set of values from the node's parent variables to form the probability of the variable represented by the node. That is, BNs allow the user to calculate the conditional probability of a node being in a particular state given the states of that node's parents.

Conrady and Jouffe (2013a) present BNs from the perspective of an applied researcher. BNs, named after Rev. Thomas Bayes (1702-1761), relate conditional and marginal probabilities of two events, A and B, given the probability of event B does not equal to zero:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

$P(A)$  is referred to as the prior probability of event A, since it is not influenced by event B. In fact event B does not have to occur after event A.

$P(A|B)$  is the conditional probability of event A given event B. It is called the posterior probability since it depends on the specified value of event B.

$P(B|A)$  is the conditional probability of event B given event A and is referenced as the likelihood.

$P(B)$  is called the marginal probability of event B and is used as a normalizing constant.



Using this formula, Bayes theorem may be used to represent how the conditional probability of event A given B is related to the converse conditional probability of event B given A.

To fully quantify the relationships between all variables and their parents, a complete probabilistic model of the network must exist. Within a BN, a joint probability distribution is created. A joint probability distribution is the probability distribution representing the probability of every possible scenario within the model. Stated differently, it gives the probability for each combination of values for all variables identified within the model. For a model with n dichotomous variables, the joint distribution would contain  $2^n$  values and is represented by

$$P(v_1, \dots, v_n) = P(v_1)P(v_2|v_1) \dots P(v_n|v_1, \dots, v_{n-1})$$

Bayesian networks factor the joint distribution into local conditional distributions for each variable given its parents to compress the overall distribution list (Conrady and Jouffe 2013). This is demonstrated in Figure 1:

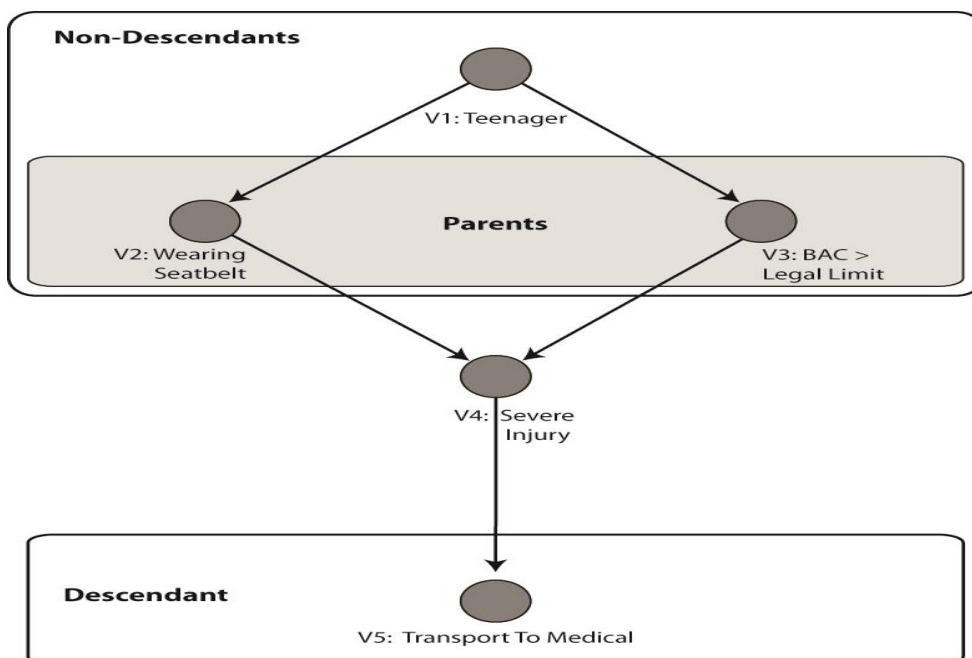


Figure 1 Sample Bayesian Network Model

The joint distribution for the above example would be

$$P(v_1, v_2, v_3, v_4, v_5) = P(v_1)P(v_2|v_1)P(v_3|v_1)P(v_4|v_2, v_3)P(v_5|v_4).$$

Within BNs, using the local conditional distributions, each variable can be treated independently of its non-descendants in the network given the state(s) of its parent(s).

Continuing with the example above,  $v_2$  and  $v_3$  are the parents of  $v_4$  and render  $v_4$  independent of  $v_1$ . This can be seen in the following equal equation:

$$P(v_4|v_1, v_2, v_3) = P(v_4|v_2, v_3)$$

The probability of any variables' state in terms of the conditional probabilities specified in the network can easily be expressed. To determine the probability that a teenage driver was not wearing their seatbelt given that they were transported to a medical facility, can be evaluated as:

$$\begin{aligned} P(v_2 = \text{no} | v_5 = \text{yes}) &= \frac{P(v_2 = \text{no} | v_5 = \text{yes})}{P(v_5 = \text{yes})} \\ &= \frac{\sum_{v_1, v_3, v_4} P(v_1, V_2=\text{no}, v_3, v_4, V_5=\text{yes})}{\sum_{v_1, v_2, v_3, v_4} P(v_1, v_2, v_3, v_4, V_5=\text{yes})} \\ &= \frac{\sum_{v_1, v_3, v_4} P(v_1) P(V_2=\text{no} | v_1) P(v_3 | v_1) P(v_4 | V_2=\text{no}, v_3) P(V_5=\text{yes} | v_4)}{\sum_{v_1, v_2, v_3, v_4} P(v_1) P(v_2 | v_1) P(v_3 | v_1) P(v_4 | v_2, v_3) P(V_5=\text{yes} | v_4)} \end{aligned}$$

According to Darwiche (2009) BNs are attractive for three reasons. First, they offer a complete representation of a particular situation and give a unique probability distribution for the network variables. Second, the network ensures consistency and completeness by utilizing

evaluation that is performed using only variables and their direct causes. Third, they give a compact representative since only an exponentially sized probability distribution is utilized.

When modeling BNs, three main methods can be used for constructing the network (Darwiche 2009). First, the designer uses his own knowledge. Second, the designer uses information gathered from some other type of formal knowledge. These two types of model construction are referred to as knowledge representation. The third method is based on machine learning where the designer allows the network to be built based on learning from the data.

The BN designed for this study will be based on machine learning where the information is learned from the crash data. With the collection of large data sets and the advancement made in machine learning and data mining, models utilizing machine learning techniques are becoming more popular. These models utilize the decreased cost of storage, increased machine power, and advancement in software to analyze large quantities of data.

Traditional statistical techniques utilize sampling methods to draw conclusions about the population. Using designed and controlled experiments, researchers manipulate the variable of interest and measure its effect on the dependent variable. Traditional statistical techniques allow the research to establish cause and effect and ensure outcomes were not attributed to pure random occurrence.

With the advancement in machine learning techniques, researchers can now analyze the entire population data. However, with large data sets, it becomes much more difficult to interpret the results in terms of their structural meaning. While machine learning techniques can create great predictive models (correlation), they often offer little explanatory insight (causation).

## CHAPTER 4. DATA

### **Crash Data**

The crash data used in this study was collected from the Highway Safety Research Group (HSRG) at Louisiana State University (LSU). The HSRG, since 1998, is grant funded by Louisiana's Department of Transportation and Development (LA DOTD) to collect, maintain, analyze, and disseminate crash data

Louisiana's law enforcement agencies utilize a uniform crash report which was approved by the state in 2005. This standardized crash report serves as the basis of the design of the state's crash database. All crash reports submitted to the state, through the HSRG, use the same standard data items and data definitions outlined in the 2005 LA Uniform Crash Report.

### **Location Data**

LA DOTD also maintains a crash database which is updated every two weeks from the HSRG crash database. Using Geographical Information System (GIS) programs, LA DOTD verifies the submitted location information on the crash report against the state's roadway database to determine the accuracy of the location data. If the location information is determined to be accurate, the information is accepted as reported. Otherwise, the crash location data is reported as an error and employees at LA DOTD and HSRG manually review the error crash reports, specifically looking over the crash narrative and diagram, and correct the location information.

At the end of this process, crashes which occur on state routes (interstates, highways, and state roads) are assigned valid latitudes (lat), longitudes (long), control sections, and milepost information. The lat/long information is used to electronically locate and map crashes and the control section and log mile information is used to integrate the crash and roadway databases.

## **Roadway Data**

The roadway data used in this study comes from the LA DOTD highway section and curve databases. Within Louisiana, the LA DOTD collects and maintains all the state's roadway data. All information concerning state routes resides in these databases and are identified by control section, beginning log mile, and ending log mile. Each state route is divided into sections based on similar road characteristics and the control section log mile information uniquely identifies each road segment. Using control section log mile data for crashes occurring on state routes, crashes can be assigned to the state route where the crash occurred. This data integration allows crashes occurring on state routes to be linked with roadway data.

Information within the highway section database includes roadway features such as average daily traffic (ADT), control section, log mile begin, log mile end, lane width, medium type, number of lanes, road type, and shoulder width. This information is updated yearly by the LA DOTD.

The curve database contains information of all curves on state routes and was made available for Louisiana in 2013. The curve database contains curve attribute information such as control section, log mile begin, log mile end, curve percentage grade, and curve radius. The curve data represents the characteristics of curves in Louisiana as of 2012. For this study, the 2012 curve data was used to represent curve information for each year between 2005 and 2012. Since curves are more permanent in nature and are not normally modified over time, using the 2012 curve data to represent the curve's characteristics for all years was considered to be adequate for this study.

LA DOTD manually reviewed each record in the curve database against Google Earth imagery and marked each record identifying it as a true curve or a missed categorized curve. Only true curves were used in this study.

The crash data is integrated with the curve database using the control section and log mile information, similar to the way the crash and roadway database were integrated. Since the crash database is integrated with both the roadway and curve databases, analysis of crashes can be performed based on roadway and curve data elements. For this study, crashes between 2005 and 2012 occurring within a curve on rural two-lane roads were analyzed.

### **Crash Trend**

An overview of single vehicle crashes involving a young driver occurring within a curve on a two-lane rural highway is presented in Table 3. The total number of crashes, as well as the crashes per severity level, have been decreasing since 2007. While this may first seem like the problem is improving, this conclusion may not be accurate. Table 4 shows the overall number of crashes in Louisiana has also been declining since 2007, with a small increase in 2012.

**Table 3 Overview of Single Vehicle Young Driver Curve Crashes on Two-Lane Rural Highways**

Year	Fatal	Serious Injury	Moderate Injury	Possible Injury	No Injury	Total
2005	14	10	144	306	469	943
2006	20	11	120	333	477	961
2007	21	5	134	376	514	1054
2008	13	9	111	294	476	905
2009	13	4	109	278	469	873
2010	9	9	81	238	486	823
2011	13	5	85	264	446	814
2012	9	4	70	249	440	772
Total	112	57	854	2,338	3,777	7,145

Table 4 Summary Data for all Crashes in Louisiana

Year	Fatal	Serious Injury	Moderate Injury	Possible Injury	No Injury	Total
2005	875	1,530	10,804	37,154	108,063	158,493
2006	890	1,505	10,143	37,116	112,237	162,125
2007	900	1,567	10,434	36,165	105,107	159,717
2008	820	1,499	10,244	34,789	104,825	157,485
2009	729	1,434	9,972	33,945	104,854	155,930
2010	643	1,223	9,082	32,178	104,545	147,678
2011	630	1,223	9,100	33,023	101,138	149,737
2012	654	1,172	9,260	34,144	103,265	153,215
Total	6,141	11,153	79,039	278,514	844,034	1,244,380

## CHAPTER 5. MODEL ESTIMATION

There were 18,796 single vehicle crashes within curves on two-lane rural highways involving drivers between the ages of 15 – 54 from 2005 to 2012 in Louisiana. These crashes resulted in 2,913 (15.5%) injuries, where injury is defined as a fatality, serious, or moderate injury to the driver. Young drivers accounted for 38% of drivers and 35% of injuries.

### **Binary Logistic Regression**

With the creation of dummy variables, as explained in Section 3.3.1, there are thirty-four potential independent variables within the logistic regression model; thirteen driver variables, eight environmental variables, ten roadway variables, and four vehicle variables. The following variables are controlled for; highway type, number of vehicles, and segment type. Highway type is limited to only rural two-lane roadways. Only single vehicle crashes are included in the study. Segment types of tangent (straight) are excluded, focusing only on crashes occurring within curves. Summary descriptions and characteristics of the factors and variables used in the logistic regression model are displayed in Table 1.

### Correlation

Before performing the binary logistic regression model, a correlation matrix of the dependent and potential independent variables, minus the dummy variables, was generated using JMP statistical software from SAS. An examination of the partial correlations indicates relatively weak correlations, except for a few variables. Highly correlated variables along with their correlation values are shown in Table 5.



Table 5 Correlated Variables

Variable 1	Variable 2	Correlation Value
Substance Suspected	Predicted Alcohol	0.8918
Inattentive	Distracted	0.8383
Dark	6:00 – 19:00	0.7192
No Protection System	Ejected	0.4326
Injured	Ejected	0.3951
Injured	No Protection System	0.3789
Substance Suspected	Inattentive	0.3359
Substance Suspected	Distracted	0.3182
Lane Width LT 12	ADT LT 3000	0.2944
Predicted Alcohol	Inattentive	0.2918
Predicted Alcohol	Distracted	0.2795
6:00 – 19:00	Predicted Alcohol	0.2712
6:00 – 19:00	Substance Suspected	0.2490
Predicted Alcohol	No Protection System	0.2469
Substance Suspected	No Protection System	0.2363
Dark	Predicted Alcohol	0.2136
Lane Width LT 12	Shoulder Width LT 4	0.2049
Dark	Substance Suspected	0.2020

A high correlation exists among the variables substance suspected/predicted alcohol, inattentive/distracted, and dark/6:00 – 19:00. Since part of the HSRG's definition of predicted alcohol (see Appendix) is substance suspected, a high correlation between these two variables can be expected. The correlation between inattentive and distracted is explained by officers selecting inattentive on most crash reports where the driver is also identified as being distracted. Likewise, dark (daylight) and 6:00 – 19:00 (time of day) is expected to be highly correlated since it is mostly daylight in Louisiana between the hours of 6:00 am and 7:00 pm.

A strong correlation also exists between the variables no protection system/ejected. This can be expected since driver ejection is dependent on the seat belt usage of the driver. When drivers use their seatbelts, their chances of being ejected is greatly reduced. These two variables are also the only potential independent variables shown to be highly correlated with the dependent variable injured.

Strong correlations are produced among the variables inattentive, distracted, and no protection system usage with predicted alcohol and substance suspected. This implies drivers who tend to be under the influence of alcohol and/or drugs tend to also be inattentive/distracted and not wearing a seatbelt when involved in crashes.

Predicted alcohol and substance suspected are also found to be correlated with dark and t6:00 – 19:00. Part of the HSRG’s definition of predicted alcohol (Appendix 1) is dependent on the time of the crash, contributing to the high correlation among these two variables.

The correlation between lane width LT 12/ADT LT 3000 and lane width LT 12/shoulder width LT 4 is expected since only crashes occurring on rural two-lane curve roadways are examined. This implies roadways’ ADT and shoulder widths are dependent on lane width for rural two-lane curve roadways in Louisiana.

#### Binary Logistic Regression Modeling

The dependent variable, driver injury level, was coded as 1 for injured (fatal, serious injury, and moderate injury) and 0 for no injury (possible injury and no injury). A list of the independent variables and their codes is displayed in Table 1. Of the possible thirty-four potential independent variables, thirty-one are selected for the logistic regression model. The three variables; predicted alcohol, inattentive, and 6:00 – 19:00 are removed due to high correlations with substance suspected, distracted, and lighting respectively. The results of the logistic regression model using the remaining thirty-one variables are presented in Table 6.

Table 6 Logistic Regression Coefficient Table for Driver Injury as a Function of 31 Predictors

Term	Estimate	Std Error	Chi Square	Prob > ChiSq
Intercept	0.82477	0.42555	3.76	0.0526
Airbag Non-Deployed	-0.53268	0.03847	191.75	<.0001
Distracted	-0.10937	0.02783	15.44	<.0001
Ejected	0.96831	0.08116	142.33	<.0001
Male	-0.16600	0.02837	34.25	<.0001
No Protection System	0.60477	0.03911	239.16	<.0001
African American	0.05522	0.05922	0.87	0.3511
Substance Suspected	0.48813	0.04245	132.20	<.0001
Youth Driver	-0.05650	0.02563	4.86	0.0275
Violation Careless Operations	0.11829	0.02848	17.25	<.0001
Violation Speeding	0.27870	0.06563	18.03	<.0001
Weekend	0.00017	0.02408	0.00	0.9943
Harm Event Culvert Ditch	-0.17117	0.04023	18.10	<.0001
Harm Event Other Fixed Object	-0.11714	0.04428	7.00	0.0082
Harm Event Pole or Tree	0.22900	0.03287	48.53	<.0001
Harm Event Roll Over	0.33864	0.03787	79.97	<.0001
Dark	-0.23756	0.13330	3.18	0.0747
Non-Clear Weather	-0.09275	0.02635	12.39	0.0004
ADT GT 3000	-0.08975	0.02686	11.17	0.0008
Curve CMF LT .5	0.06568	0.03122	4.43	0.0354
Curve Length Small	-0.11535	0.04924	5.49	0.0192
Curve Length Medium	-0.09578	0.04715	4.13	0.0422
Curve Length Large	0.00000	0.00000	.	.
Curve Radius Small	-0.05913	0.05830	1.03	0.3105
Curve Radius Medium	0.03628	0.02947	1.52	0.2182
Curve Radius Large	0.00000	0.00000	.	.
Lane Width LT 12	0.05418	0.02592	4.37	0.0366
Shoulder Width LT 4	0.04444	0.02542	3.05	0.0805
Vehicle Type Passenger Car	-0.09127	0.04868	3.51	0.0608
Vehicle Type Light Truck	0.01375	0.04731	0.08	0.7713
Vehicle Type SUV	0.03060	0.05603	0.30	0.5849
Vehicle Year LT 2000	-0.55397	0.38001	2.13	0.1449

The model has 18,716 observations, was found to be significant with a p-value of 0.0001, and has a misclassification rate of 0.1264. A review of the independent variables finds nineteen predictors to be significant with a p-value less than or equal to .05. A list of the nineteen predictors is shown in Table 7.

Table 7 Significant Variables within the 31 Predictors Model

Term	Estimate	Std Error	Chi Square	Prob > ChiSq
Airbag Non-Deployed	-0.53268	0.03847	191.75	<.0001
Distracted	-0.10937	0.02783	15.44	<.0001
Ejected	0.96831	0.08116	142.33	<.0001
Male	-0.16600	0.02837	34.25	<.0001
No Protection System	0.60477	0.03911	239.16	<.0001
Substance Suspected	0.48813	0.04245	132.20	<.0001
Violation Careless Operations	0.11829	0.02848	17.25	<.0001
Violation Speeding	0.27870	0.06563	18.03	<.0001
Harm Event Culvert Ditch	-0.17117	0.04023	18.10	<.0001
Harm Event Pole Tree	0.22900	0.03287	48.53	<.0001
Harm Event Roll Over	0.33864	0.03787	79.97	<.0001
Non-Clear Weather	-0.09275	0.02635	12.39	0.0004
ADT GT 3000	-0.08975	0.02686	11.17	0.0008
Harm Event Other Fixed Object	-0.11714	0.04428	7.00	0.0082
Curve Length Small	-0.11535	0.04924	5.49	0.0192
Youth Driver	-0.05650	0.02563	4.86	0.0275
Curve CMF LT .5	0.06568	0.03122	4.43	0.0354
Lane Width LT 12	0.05418	0.02592	4.37	0.0366
Curve Length Medium	-0.09578	0.04715	4.13	0.0422

A new model is formed using the nineteen significant variables identified above. The new model has 18,716 observations, was found to be significant with a p-value of 0.0001 and has a misclassification rate of 0.1265. The results of the new logistic regression model using only nineteen variables are presented in Table 8. All variables remained significant in the new model. The receiver operating characteristic (ROC) curve for the model is displayed in Figure 2 showing the model has good predictive ability.

Table 8 Logistic Regression Coefficient Information for Driver Injury

Term	Estimate	Std Error	Chi Square	Prob > ChiSq
Intercept	0.09032	0.12040	0.56	0.4531
Airbag Non-Deployed	-0.49354	0.03612	186.72	<.0001
Distracted	-0.10953	0.02774	15.59	<.0001
Ejected	0.95750	0.08041	141.78	<.0001
Male	-0.14331	0.02652	29.20	<.0001
No Protection System	0.60699	0.03885	244.11	<.0001
Substance Suspected	0.46119	0.04121	125.22	<.0001
Youth Driver	-0.07223	0.02511	8.27	0.004
Violation Careless Operations	0.12403	0.02839	19.08	<.0001
Violation Speeding	0.27894	0.06551	18.13	<.0001
Harm Event Culvert Ditch	-0.17790	0.03997	19.81	<.0001
Harm Event Other Fixed Object	-0.11186	0.04403	6.45	0.0111
Harm Event Pole Tree	0.23709	0.03265	52.72	<.0001
Harm Event Roll Over	0.35071	0.03779	86.14	<.0001
Non-Clear Weather	-0.08674	0.02624	10.93	0.0009
ADT GT 3000	-0.08996	0.02678	11.29	0.0008
Curve CMF LT .5	0.08442	0.02828	8.91	0.0028
Curve Length Small	-0.10563	0.04276	6.10	0.0135
Curve Length Medium	-0.09136	0.04479	4.16	0.0414
Lane Width LT 12	0.06153	0.02555	5.80	0.016

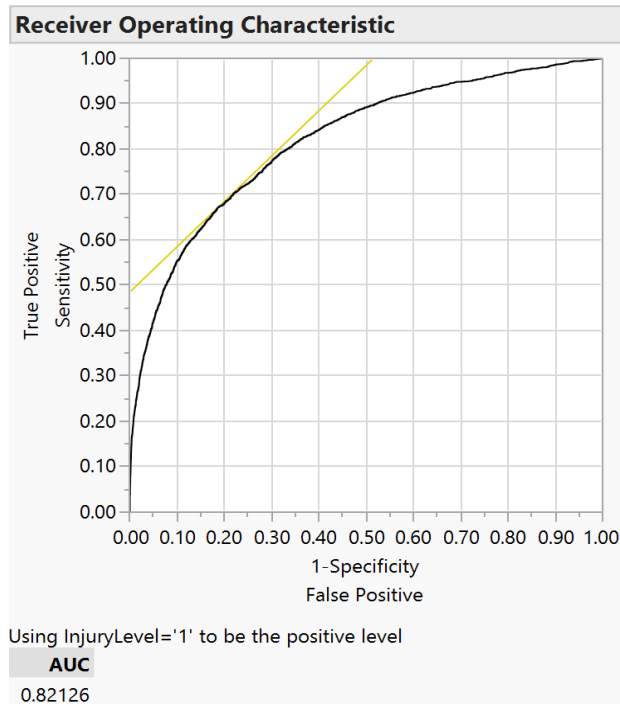


Figure 2 ROC Curve Information

### Logistic Regression Results

This binary logistic regression model produces the following equation:

$$\begin{aligned}\text{logit} = & 0.090 - 0.988(\text{Air Bag Non-Deployed}) - 0.218(\text{Distracted}) \\ & + 1.916(\text{Partially or Totally Ejected}) - 0.286(\text{Male}) \\ & + 1.214(\text{No or Improper Seatbelt Usage}) + 0.922(\text{Alcohol and/or Drugs Suspected}) \\ & - 0.144 (\text{Youth Driver}) + 0.248(\text{Careless Operations}) + 0.560(\text{Speeding}) \\ & - 0.356(\text{Hitting a Culvert or Ditch}) \\ & - 0.222 (\text{Hitting a Fixed Object Other Than a Culvert/Ditch/Pole/Tree}) \\ & + 0.474(\text{Hitting a Pole or Tree}) + 0.700(\text{Rollover}) - 0.174(\text{Non-Clear Weather}) \\ & - 0.180(\text{ADT GT 3000}) + 1.708 \text{ Curve CMF LT .05} - 0.212(\text{Small Curve Length}) \\ & - 0.182(\text{Medium Curve Length}) + 0.124(\text{Lane Width LT 12})\end{aligned}$$

Note: The parameter estimates above were multiplied by 2 since JMP codes two-level nominal variables as 1 and -1, as opposed to the typical 0 and 1.

Positive coefficients on the dummy variables (careless operations, speeding, harm event pole or tree, harm event roll over), while holding everything else constant, are associated with higher probabilities of drivers having an injury. Likewise, negative coefficients on the dummy variables (harm event culvert or ditch, harm event other fixed object, curve length small, curve length medium), while holding everything else constant, are associated with lower probabilities of drivers having an injury. For the dichotomous variables, positive (negative) coefficients indicate a higher value on that predictor is associated with a higher (lower) probability value of drivers obtaining an injury.

The nine predictors ejected, no or improper protection system, substance suspected, careless operation, speeding, harm event pole or tree, harm event roll over, curve CMF LT .05, and lane width LT 12, all have a positive impact on drivers sustaining an injury. The remaining

ten predictors air bag non-deployed, distracted, male, youth driver, harm event culvert or ditch, harm event other fixed object, non-clear weather, ADT GT 3000, curve length small, and curve length medium have a negative impact on drivers sustaining an injury.

With linear regression the coefficients represent the change in the response variable for a unit change in the predictor variable, when all else remains the same. For logistic regression models, the regression coefficient represents the change in the logit for a unit change in the predictor variable. Using the odds ratio formula, identified with formula 3.1, to calculate the response variables is more intuitive.

$$\begin{aligned} \text{Odds(driver injury)} = & e^{0.090} \times 0.347^{(\text{Air Bag Non-Deployed})} \times 0.803^{(\text{Distracted})} \\ & \times 9.581^{(\text{Partially or Totally Ejected})} \times 0.751^{(\text{Male})} \times 4.033^{(\text{No or Improper Seatbelt Usage})} \\ & \times 1.963^{(\text{Alcohol and/or Drugs Suspected})} \times 0.865^{(\text{Youth Driver})} \times 1.281^{(\text{Careless Operations})} \\ & \times 1.747^{(\text{Speeding})} \times 0.701^{(\text{Harm Event Culvert or Ditch})} \times 0.799^{(\text{Harm Event Other Fixed Object})} \\ & \times 1.607^{(\text{Harm Event Pole or Tree})} \times 2.017^{(\text{Harm Event Roll Over})} \times 0.841^{(\text{Non-Clear Weather})} \\ & \times 0.835^{(\text{ADT GT 3000})} \times 1.184^{(\text{Curve CMF LT .5})} \times 0.809^{(\text{Curve Length Small})} \times 0.833^{(\text{Curve Length Medium})} \\ & \times 1.131^{(\text{Lane Width LT 12})} \end{aligned}$$

The greater the predictor's odds ratio is from 1, the greater the effect the predictor has on driver injury levels. Predictors with an odd ratios greater (less) than 1 indicates the predictor is more (less) likely to contribute to drivers becoming injured. For instance, drivers are 4.033 times more likely to have an injury, compared to no injury, when not wearing or improperly wearing their seatbelt. Evaluating the odds ratios in the above equation shows that being partially or totally ejected, not or improperly wearing a seatbelt, driving under the suspicion of alcohol and/or drugs, speeding, and vehicle rolling over greatly increase the odds of drivers being injured. Likewise, air bags not deploying have the least odds on injuring the driver. When the air bag does not deploy, drivers' are 0.347 times less likely to have an injury compared to no

injury. While this might seem counter intuitive at first, some research has shown this to be true. Bosch Automotive Handbook (2011) states airbag systems are designed in such a way that their deployment threshold is adjusted when occupants are not wearing their seat belts. Stated differently, not wearing a seat belt causes the airbag to trigger differently, which may influence injury risk. Donaldson III (2008) states occupants in motor vehicle crashes resulting in airbag deployment who are not wearing seatbelts are at higher risk of cervical spine fractures and other spinal cord injuries.

The remaining variables distracted, male, young drivers, careless operating, harm event culvert or ditch, harm event other fixed object, non0clear weather, ADT GT 3000, curve CMF LT .5, curve length small, curve length medium, and lane width LT 12all have minimal effect of driver injury levels.

Of the nineteen variables in the final model, nine represent driver characteristics, five are concerned with environment factors and five signify roadway elements.

Evaluating driver characteristics shows being partially or totally ejected, not wearing or improperly wearing a seatbelt, and driving under the suspicion of alcohol and/or drugs are strongly associated with higher injury severity levels. This study also identifies males as being 0.751 times less likely to be injured. These findings are similar to previous research studies (Dissanayake & Lu 2002, Clarke et al. 2006, Shinar 2007, Schneider IV et al. 2009, Barua 2010, de Oña et.al 2010, Hummer et al. 2010, Peek-Asa et al. 2010, and Zhang 2010). Previous research also shows females are more susceptible to injuries than males (Mercier et al. 1997, Dissanayake & Lu 2002, Clarke et al. 2006, Shinar 2007, Schneider IV et al. 2009, Barua 2010, de Oña et.al 2010, and Zhang 2010) and that higher crash speeds lead to more severe driver injuries (Simoncic 2004 and , Zhang 2010).



When evaluating environmental factors, rollover crashes and hitting a pole or tree were associated with an increased likelihood of contributing to a driver injury. Research has shown more severe injuries occur when the vehicle overturns (Hummer et al. 2010 and de Oña et al. 2010) or when the vehicle leaves the roadway and strikes a tree (Schneider IV et al. 2009, Chen 2010, and Hummer et al. 2010). Driver injuries also tend to be more severe in crashes occurring in clear weather (Hummer et al. 2010 and Schneider IV et al. 2009). This research is similar to the findings of this study which shows non-clear weather conditions are 0.841 times less likely to contribute to driver injuries.

While logistic regression is sensitive to high correlations among predictor variables, Bayesian Networks (BN) are not influenced by multicollinearity. The knowledge discovery algorithms utilized in BayesiaLab software, use information-theoretic measures to search for probabilistic relation between variables (Conrady, S. & Jouffe, L. 2013b). The nature of learning used in BNs automatically considers multiple relationship types among all variables, including collinear relationships, and can handle processing each without any issues (Conrady, S. & Jouffe, L. 2013b).

### Bayesian Network Modeling

The twenty-four potential independent variables shown in Table 2 plus injury level, the primary variable of interest, are used in the BN model. Each variable is discrete, ranging from two to ten possible outcomes. Unlike the logistic regression model where the variables are dichotomous, BNs allow for variables to have multiple outcome levels.

BayesiaLab software is used to construct the BN models. Using the crash, location, and roadway data, 18,796 records are used in modeling the networks. An initial unconnected network of all variables is displayed in Figure 3, where each variable is represented by a node.



Figure 3 Initial Bayesian Network as Unconnected Nodes

Within BayesiaLab, a machine learning algorithm is used to learn the probabilistic relationships between the variables in the network. This knowledge based discovery method relies on the computer to process the data and build a network structure without any assumptions.

The first BN was built using all twenty-four potential independent variables (factors) identified in Table 2. Injury level was excluded in this network, as the purpose of this first network is to develop an understanding of how the independent variables directly relate to one another. Figure 4 shows the BN for the twenty-four independent factors.

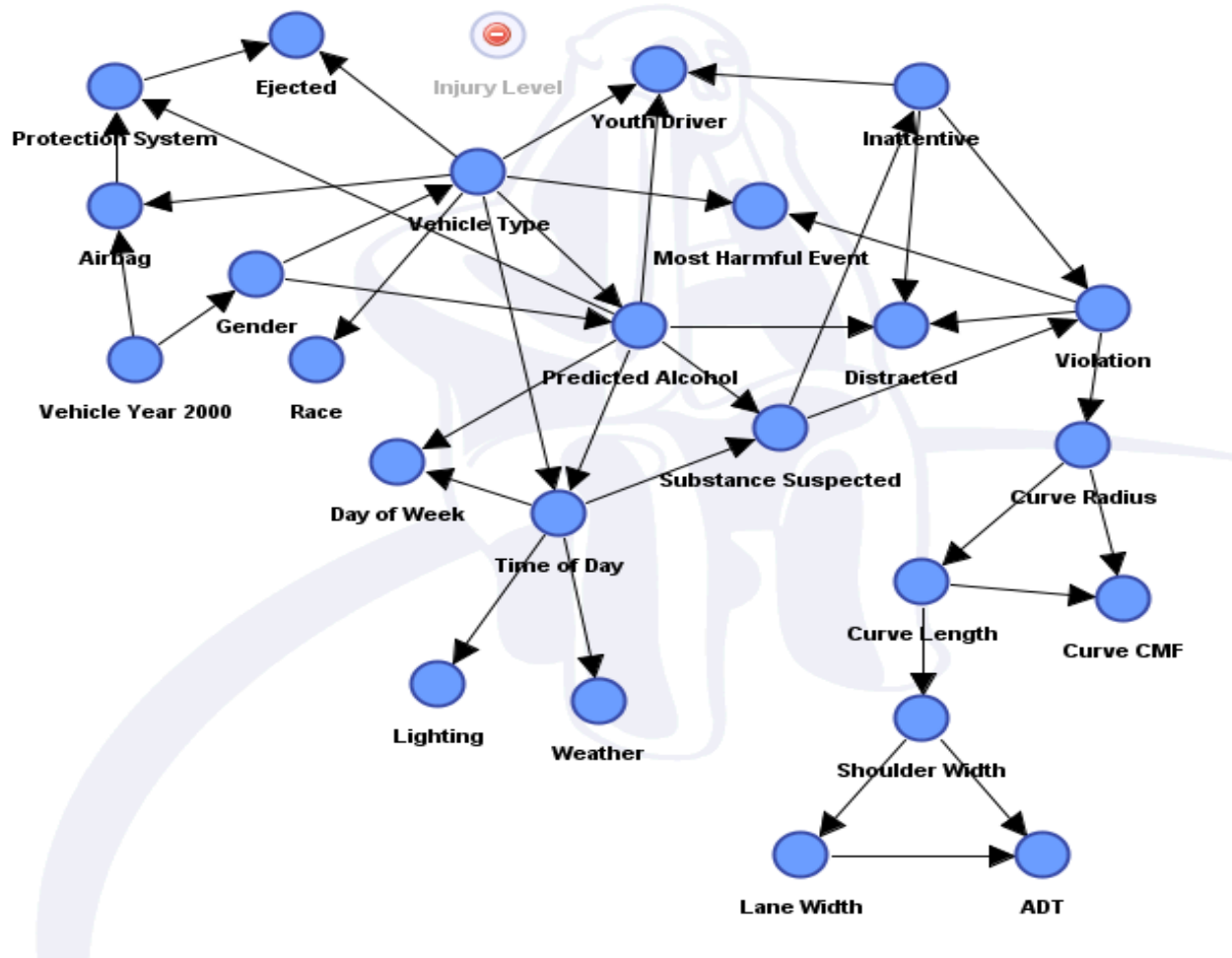


Figure 4 Bayesian Network for Potential Factors

Within this network, there are thirty-four arcs between twenty-four nodes showing a large amount of interaction between the variables. To identify highly correlated variables, the amount of mutual information shared between connected nodes was analyzed. Mutual information  $(X,Y)$ , measured as  $P(X|Y)/P(X)$ , shows how much knowing of variable  $Y$  reduces the uncertainty about variable  $X$ . Figure 5 displays the variables that share a high amount of mutual information within the network.

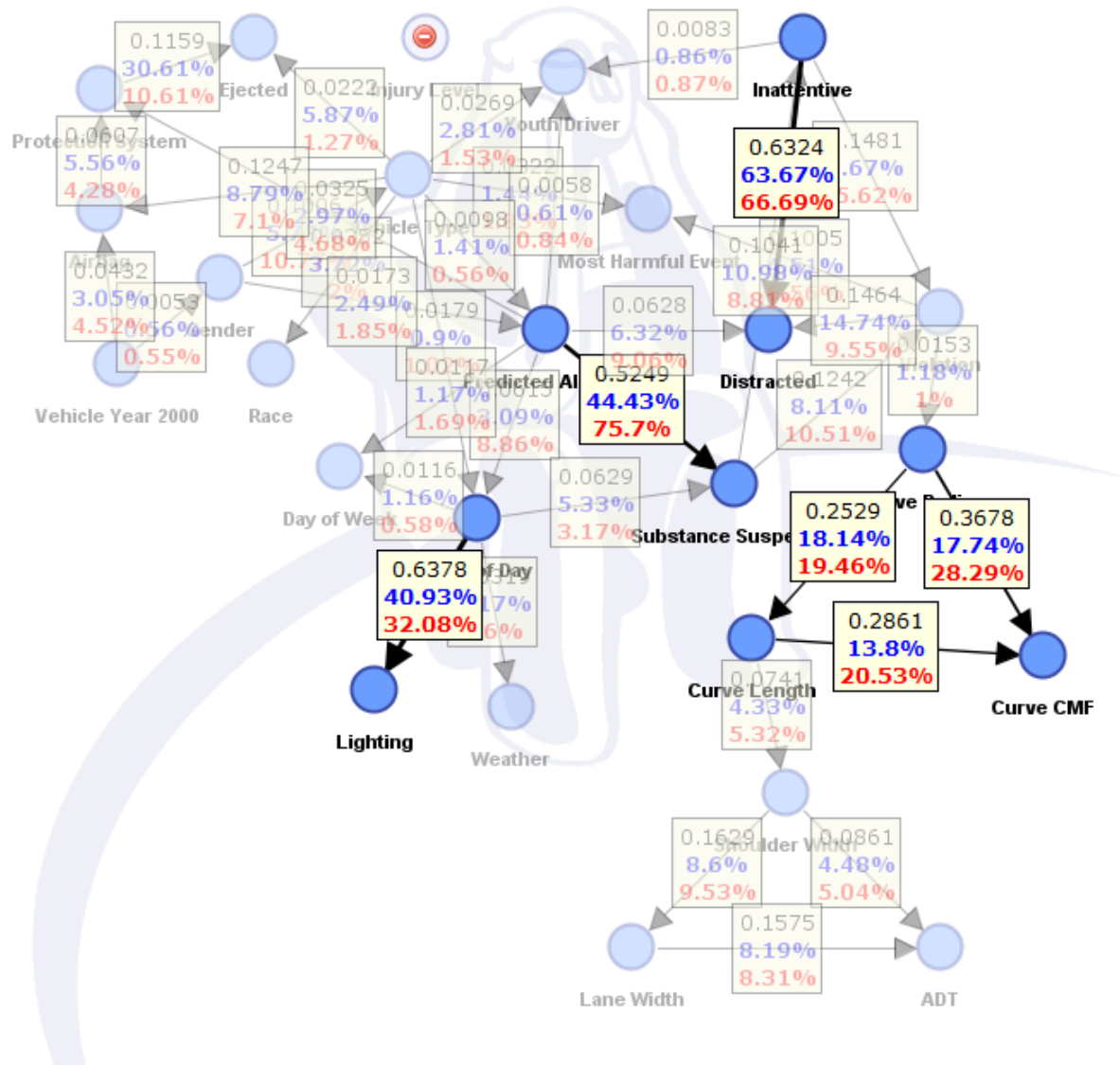


Figure 5 Mutual Information Shared Between Nodes

As seen in Figure 5 a large amount of mutual information is shared between the variables substance suspected/predicted alcohol, time of day/lighting, inattentive/distracted, curve radius/curve length, curve radius/curve CMF, and curve CMF/curve length. This is very similar to the correlation between variables found using the logistic regression model.

The top number represents the mutual information shared between the two variables. The middle number is the relative mutual information in the direction of the arc, whereas the bottom number shows the relative mutual information in the opposite direction of the arc. If the two

variables are totally independent, then knowing about X would not provide any information about Y and the mutual information amount would be 0. Likewise, if the two variables were totally correlated, then knowing about X would provide all information about Y and the mutual information amount would be 1. In the BN model, knowing the value of substance suspected on average reduces the uncertainty of predicted alcohol by 75.7%. Conversely, knowing the value of predicted alcohol reduces the uncertainty of substance suspected by 44.43%

Based on this information, a new BN was built excluding the variables predicted alcohol, inattentive, and lighting. Since these variables share a large amount of mutual information with substance suspected, distracted, and time day, little additional information is gained from keeping these variables in the model. The new BN can be seen in Figure 6.

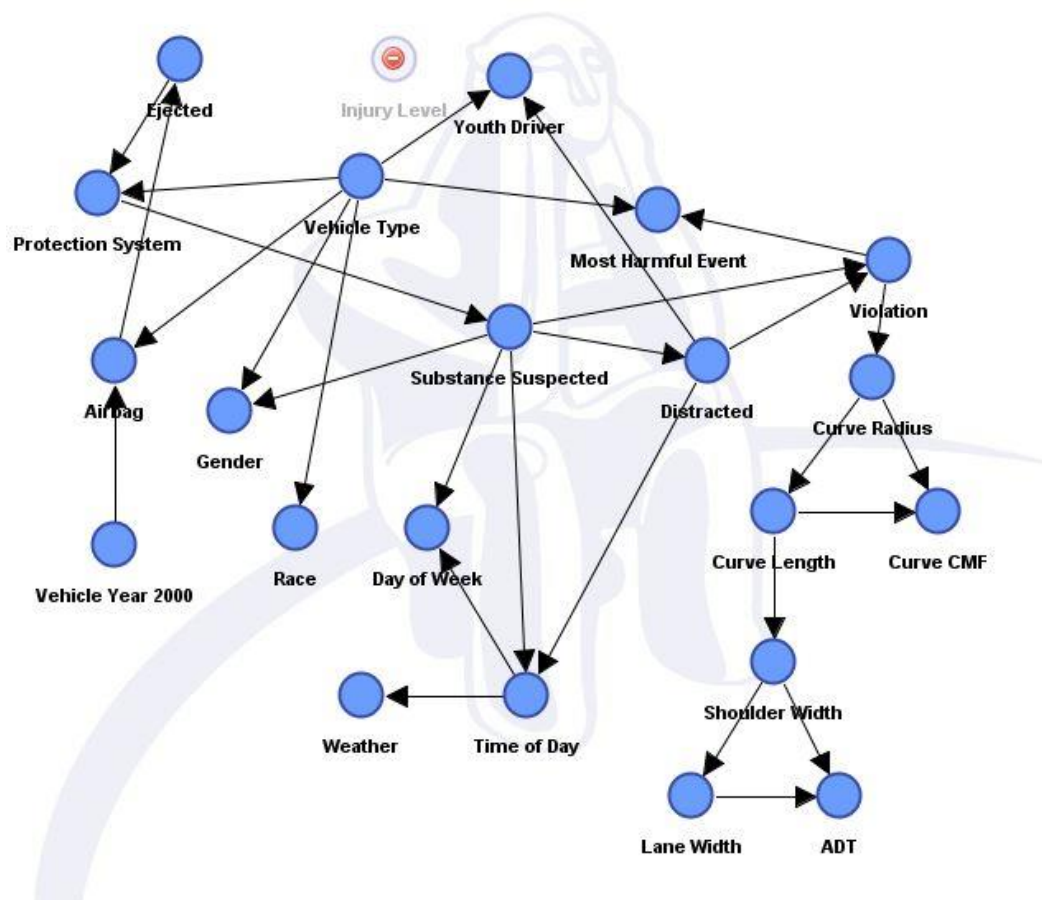


Figure 6 Bayesian Network without Highly Correlated Variables

Clustering among related variables are easily identified when viewing the network. For instance, roadway variables are gathered together towards the bottom right corner of the network. Likewise, environmental variables capturing information about the weather, time of day, and day of week are grouped together at the lower middle of the network. In all, five sets of clusters are identified with BayesiaLab. To better identify clustering among the factors, Figure 7 shows the BN as related clusters.

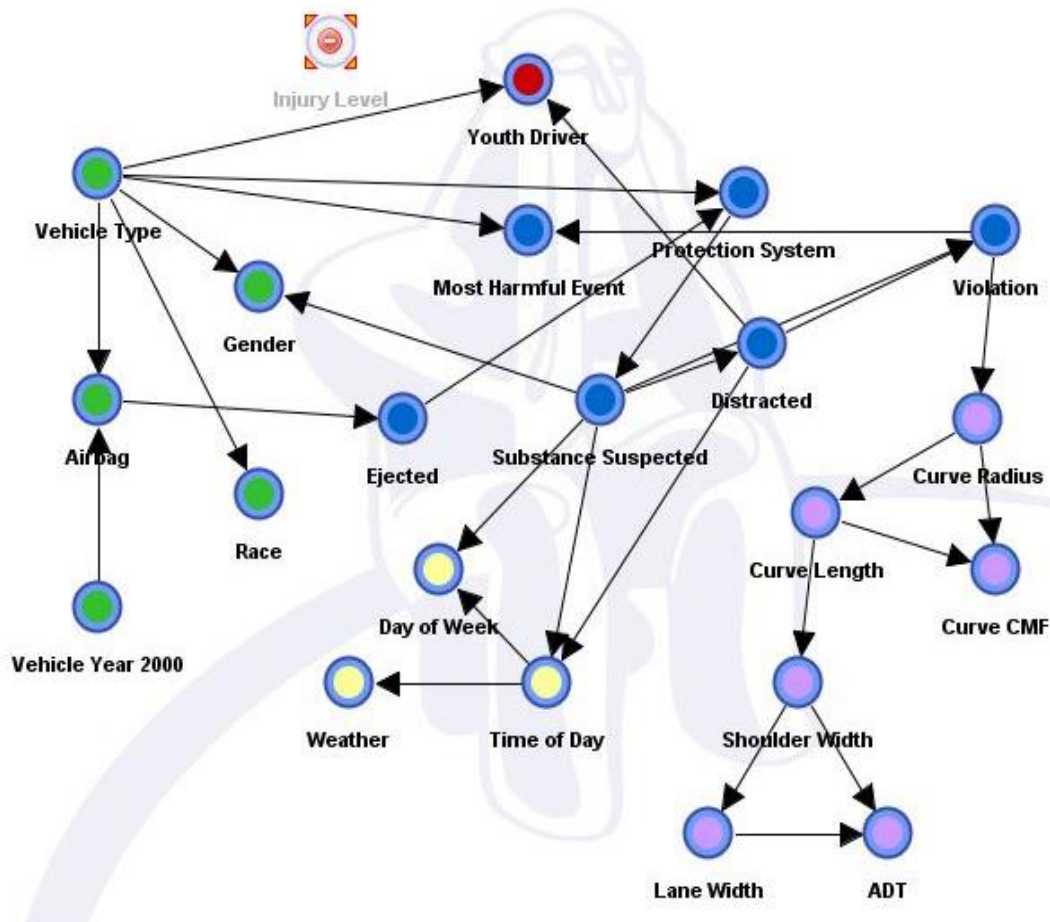


Figure 7 Bayesian Network with Clustering of Factors

Using the data structure produced with clustering on the nodes and identifying injury level as the target variable of interest, the final BN is shown in Figure 8.

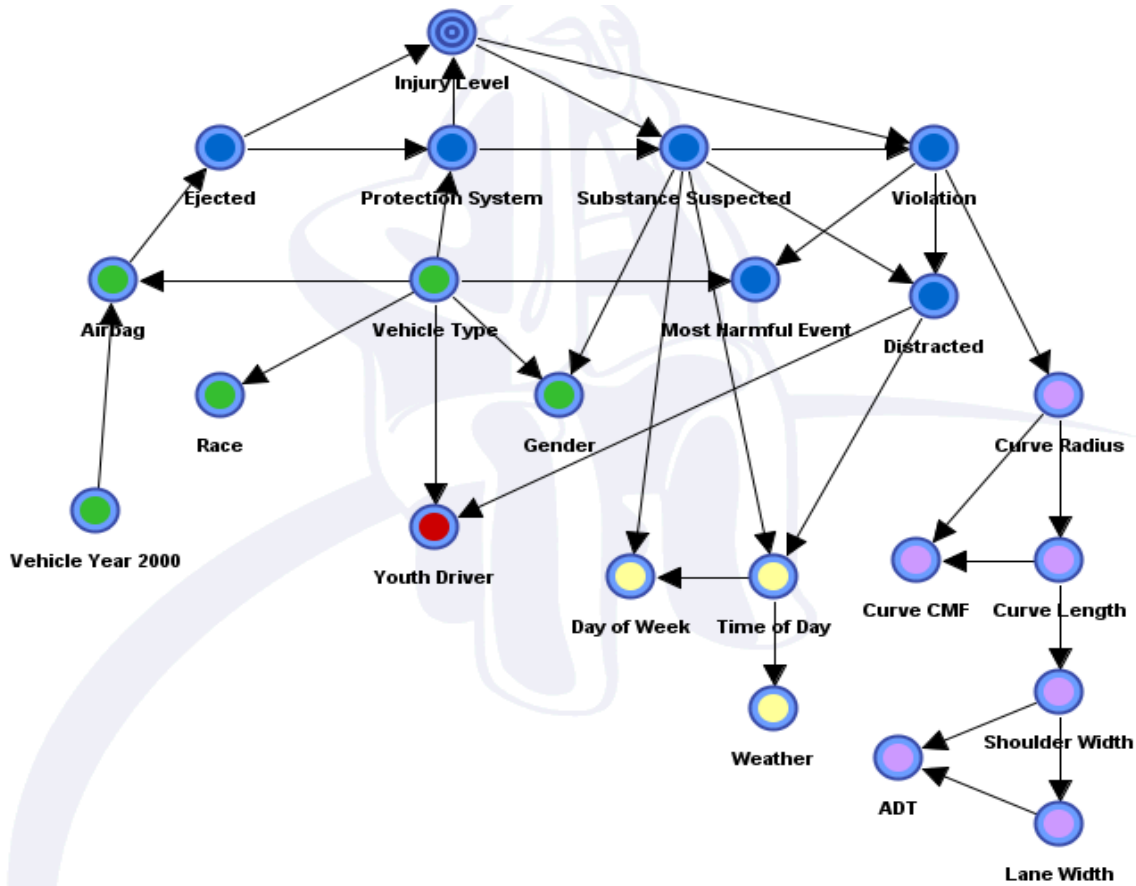


Figure 8 Final Bayesian Network Model

### Bayesian Network Results

Based on the final BN, Table 9 displays the node significance and p-value with respect to the information gain brought by the node to the knowledge of injury level. Driver related variables are at the top of the list and have the most significance. Environmental, vehicle, and roadway variables complete the list in respective order and while some have statistical significance, all have little to no relative significance.

Besides determining the nodes relative significance on injury level, the BN can be used to help measure the node's direct effect on injury level. In order to transition from association to exploratory, a more in depth knowledge of the BN is required.

Table 9 Node Significance with Injury Level

Node	Mutual Information	Relative Significance	p-value
Protection System	0.0903	1.0000	0.00%
Ejected	0.0879	0.9728	0.00%
Substance Suspected	0.0336	0.3720	0.00%
Violation	0.0200	0.2211	0.00%
Airbag	0.0083	0.0920	0.00%
Time of Day	0.0022	0.0246	0.00%
Most Harmful Event	0.0018	0.0197	0.00%
Gender	0.0015	0.0165	0.00%
Vehicle Type	0.0012	0.0130	0.00%
Distracted	0.0009	0.0095	0.00%
Vehicle Year 2000	0.0007	0.0083	0.02%
Day of Week	0.0005	0.0060	0.02%
Curve Radius	0.0002	0.0024	12.88%
Curve CMF	0.0001	0.0010	81.53%
Youth Driver	0.0001	0.0008	17.09%
Curve Length	0.0001	0.0007	63.05%
Weather	0.0000	0.0004	99.98%
Race	0.0000	0.0002	97.97%
Shoulder Width	0.0000	0.0001	99.83%
Lane Width	0.0000	0.0001	99.96%
ADT	0.0000	0.0001	99.97%

Arcs within BNs correspond to direct probabilistic relations between connected variables (nodes). For instance, viewing the arc direction between protection system and substance suspected in network shows the arc pointing from protection system to substance suspected.

The arc direction in Figure 9 is derived based from machine learning on the crash data and may not always represent causation. The crash data over the years has shown that drivers who are suspected of being under the influence of a substance are less likely to wear their seatbelts. This means substance suspected has more of a causal effect on protection system, not vice versa.

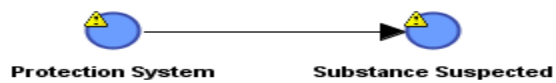


Figure 9 Relationship between Protection System and Substance Suspected in the BN



Unfortunately, software package can only provide causality direction based on their interpretation of the data provided. It is the responsibility of the researcher to review the results and ensure the output is consistent based with their domain knowledge.

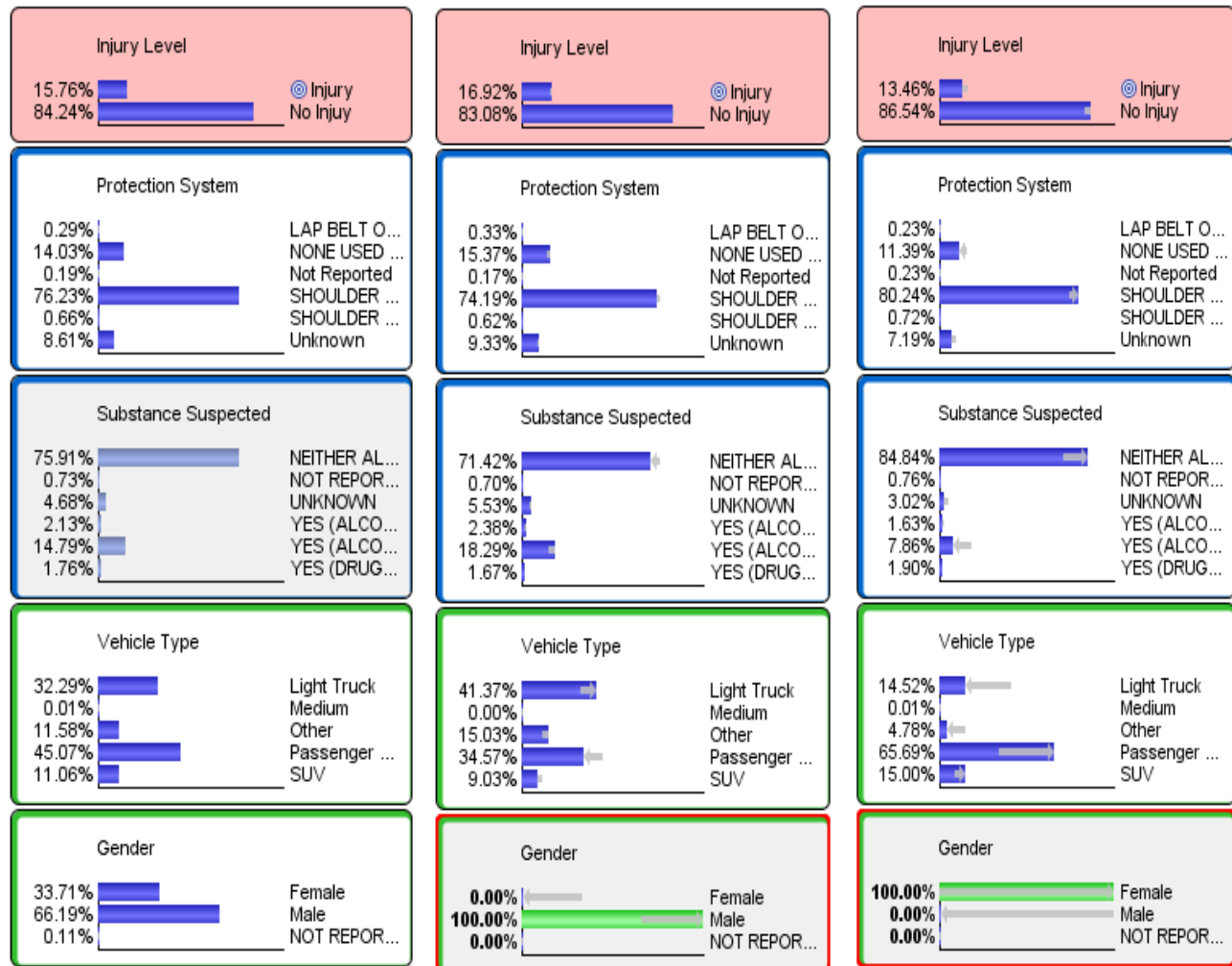
Arc directions within BNs may be reversed as long as doing so does not introduce loops between the nodes and the inversion does not modify the joint probability distribution. The direction of the arc is very important when transitioning from a general BN to causal network. That is a network where the parents of each node are its direct cause (Conrady & Jouffe, 2013c). Having a causal network, is the only way to truly evaluate causation.

The network developed in this study, using machine learning techniques, is not a causal network since the parents of each node are not always its direct cause. This means the arc directions in Figure 8 cannot be interpreted as causal direction. However, the network can be used to explore the data and make causal inferences. Using Jouffe's Likelihood Matching (LM) algorithm within BayesiaLab, casual inference may be measured by manipulating the probability distribution of any variable, while holding the probability distribution of all ascending nodes constant, and evaluating the effect the change has on the probability distribution of the target variable (Conrady & Jouffe, 2013c).

#### Jouffe's Likelihood Matching

Figure 10 shows that overall 15.76% of drivers were injured and 33.71% of all drivers were female. If the evidence of gender is set to 100% male, the injury rate increases to 16.92%. However, the injury rate decreases to 13.46% when gender is set to 100% female.

However, this does not mean that being female reduces the risk of driver injury by 20.5%. There are numerous other relevant factors that must be controlled before causal inference can be implied.



Observational Inference for youth drivers:

$P(\text{InjuryLevel}=\text{Injury}|\text{Gender}=\text{Male}) = 16.92\%$

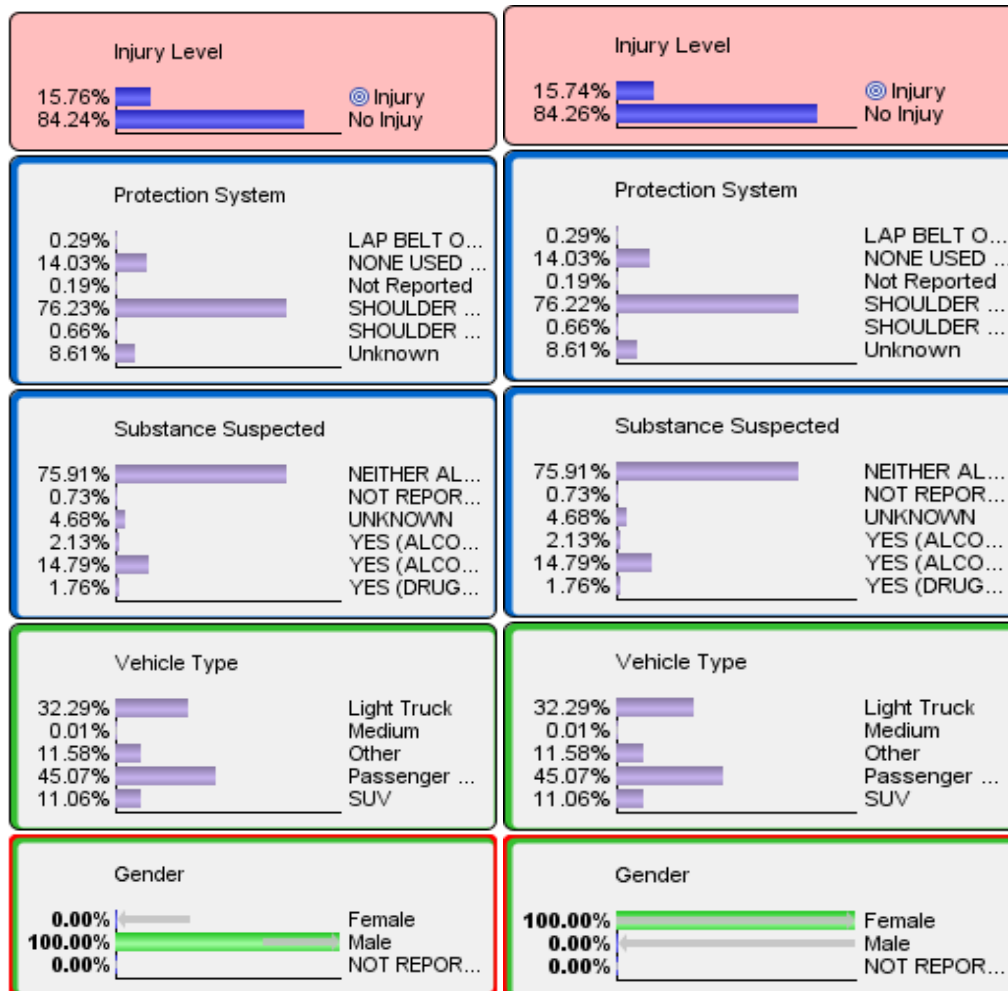
$P(\text{InjuryLevel}=\text{Injury}|\text{Gender}=\text{Female}) = 13.46\%$

Figure 10 Evaluating Driver Injury Based on Gender

For instance, male drivers are more likely to drive light trucks, drive under the suspicion of alcohol and/or drugs, and not wear their seat belts when compared to females. These differences are demonstrated in Figure 10 which highlights that males and females are quite different in driver characteristics and thus are not directly comparable on injury levels. This is a problem associated with observational studies.

To overcome this challenge, Jouffe's LM algorithm may be utilized. Jouffe's LM within Bayesia Lab allows the probability distributions of covariates to remain fixed, thus measuring the

direct effect a node has on a target node. By setting the probability distributions for vehicle type, substance suspected, and protection system to remain unchanged, the direct effect of gender on injury level can be measured as shown in Figure 11.



Causal Inference:

$$P(\text{InjuryLevel}=\text{Injury}|\text{do}(\text{Gender}=\text{Male})) = 15.76\%$$

$$P(\text{InjuryLevel}=\text{Injury}|\text{do}(\text{Gender}=\text{Female})) = 15.74\%$$

The casual effect can then be calculated as:

$$P(\text{InjuryLevel}=\text{Injury}|\text{do}(\text{Gender}=\text{Male})) - P(\text{InjuryLevel}=\text{Injury}|\text{do}(\text{Gender}=\text{Female})) = 0.02\%$$

Figure 11 Direct Effect of Gender on Driver Injury

The difference 0.02% is the “gender effect” with regard to the probability of a male driver, compared to a female driver, sustaining an injury as a result of a single vehicle crash within a curve on a rural two-lane road in Louisiana. This means that given the same crash

factors, a male driver has a 0.02% increased chance of being injured in the crash compared to a female driver. It is not gender that has a direct effect on injury, but the behavior of the gender (seat belt use, substance suspected, vehicle type, etc.) that directly effects the level of injury.

#### Driver Factors

The driver variables ejected, protection system, substance suspected, and violation were the only factors to have a direct effect on driver injury, as displayed in Table 10. When a driver is partially or totally ejected from the vehicle, their chances of sustaining an injury are greatly enhanced. Likewise not wearing a safety belt or being suspected of alcohol and/or drugs increases a driver's chance of injury.

Table 12 Direct Effect of Driver Factors on Driver Injury

Factor	Variable	Injury	No Injury
Ejected	Not Ejected	12.93%	87.07%
	Not Reported	0.49%	99.51%
	Partially Ejected	80.99%	19.01%
	Totally Ejected	59.35%	40.65%
	Unknown	42.67%	57.33%
Protection System	Lap Belt Only	14.31%	85.69%
	None Used	33.29%	66.71%
	Not Reported	2.74%	97.26%
	Shoulder and Lap Belt	12.28%	87.72%
	Shoulder Belt Only	17.50%	82.50%
Substance Suspected	Unknown	17.11%	82.89%
	Alcohol	29.52%	70.48%
	Alcohol and Drugs	20.15%	79.85%
	Drugs	22.73%	77.27%
	Neither Alcohol nor Drugs	13.56%	86.44%
Violation	Not Reported	14.97%	85.03%
	Unknown	21.74%	78.26%
	Careless Operations	16.89%	83.11%
	No Violation	8.16%	91.84%
	Other	15.22%	84.78%
	Speeding	20.47%	79.53%
	Unknown	30.29%	69.71%

### Environmental Factors

None of the environmental factors have a direct effect on injury level.

### Roadway Factors

Likewise, no roadway factors have a direct effect on injury level.

### Vehicle Factors

Vehicle types other than light truck, passenger car, and SUV have a very slight decrease in injury levels as shown in Table 11. Also, vehicles manufactured after the year 2000 slightly decrease driver injuries.

Table 13 Direct Effect of Vehicle Factors on Driver Injury

Factor	Variable	Injury	No Injury
Vehicle Type	Light Truck	15.84%	84.16%
	Other	15.61%	84.39%
	Passenger Car	15.87%	84.13%
	SUV	15.84%	84.16%
Vehicle Manufacture Year	After 2000	15.71%	84.29%
	Other	15.76%	84.24%
	Before 2000	15.78%	84.22%

### **Identify Factors Affecting Driver Injury Level**

Variables that have a significant impact on driver injury levels identified using either logistic regression or BN models are listed in Table 12. This table shows both models recognize eight of the same contributing factors. The BN model found time of day to be statistically significant, however it was excluded from the logistic regression model due to high correlation with lighting, which was not found significant. The logistic regression model found youth, weather, ADT, curve CMF, and curve length to be significant whereas the BN did not. A more detailed analysis on how these factors influence driver injury level is discussed in Chapter 6.

Table 14 Driver Injury Contributing Factors

	Logistic Regression	Bayesian Network
Driver		
Airbag	Y	Y
Distracted	Y	Y
Ejected	Y	Y
Gender	Y	Y
Protection System	Y	Y
Substance Suspected	Y	Y
Youth	Y	
Violation	Y	Y
Careless Operation		
Speeding		
Environmental		
Most Harmful Event	Y	Y
Culvert or Ditch		
Other Fixed Object		
Pole or Tree		
Rollover		
Time of Day	N/A	Y
Weather	Y	
Roadway		
ADT	Y	
Curve CMF	Y	
Curve Length Size	Y	
Small		
Medium		
Lane Width	Y	
Vehicle		
Vehicle Type		Y
Vehicle Year		Y

## CHAPTER 6: ANALYSIS AND DISCUSSION

### Impact of Identified Contributing Factors

There were eight factors identified by both models as contributing to driver injury levels: air bag, distracted, ejected, gender, protection system, substance suspected, violation, and most harmful event. Among these factors: distracted, protection system, substance suspected, and violation are driver factors which can be altered by educational countermeasures. Overall injury level for youth drivers is displayed in Figure 12.

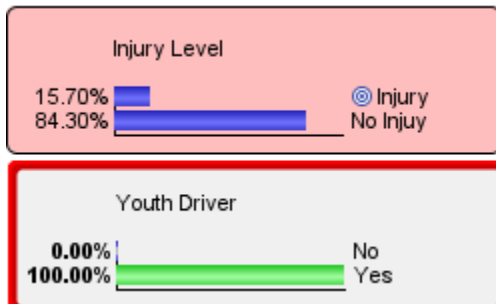


Figure 15 Youth Driver Injury Levels

### Protection System

Not wearing a seatbelt is identified in both models as being a major significant factor contributing to driver injuries. Figure 13 show that if everything remains constant the direct effect of seatbelt use on youth driver injuries is -21.13%.

### Protection System and Driver Ejection

Driver ejection is also found in both models as being a major significant factor contributing to driver injuries. Seatbelts are a driver's best defense to prevent ejection in a car crash, a necessary factor in reducing driver injuries. If all youth drivers wear their seatbelt, the probability distribution of being totally ejected from the vehicle would decrease from 5.09% to 0.33% and injury distribution would decrease from 15.70% to 9.26% as shown in Figure 14.

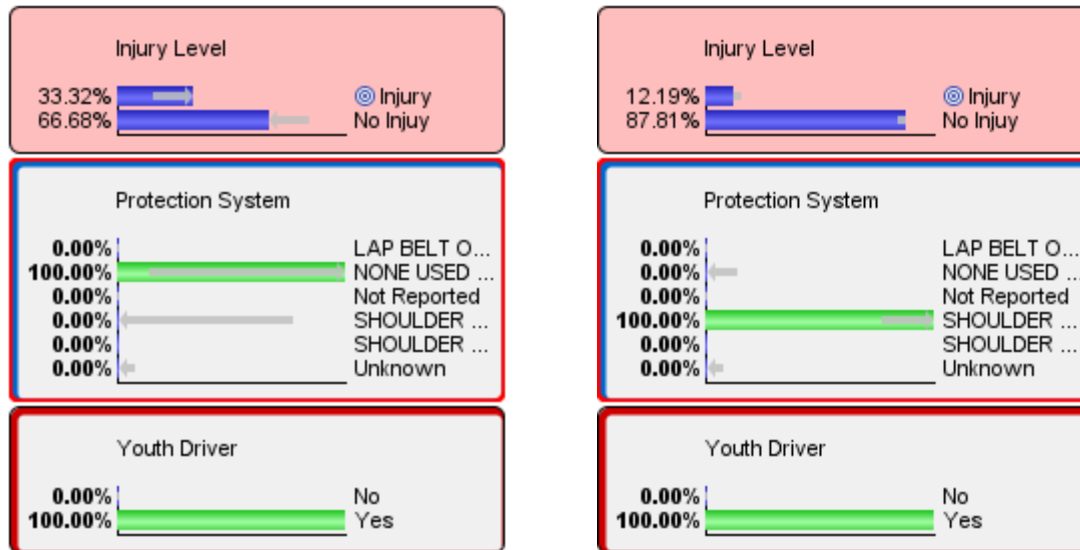


Figure 16 Direct Effect of Seatbelt Use on Youth Driver Injury

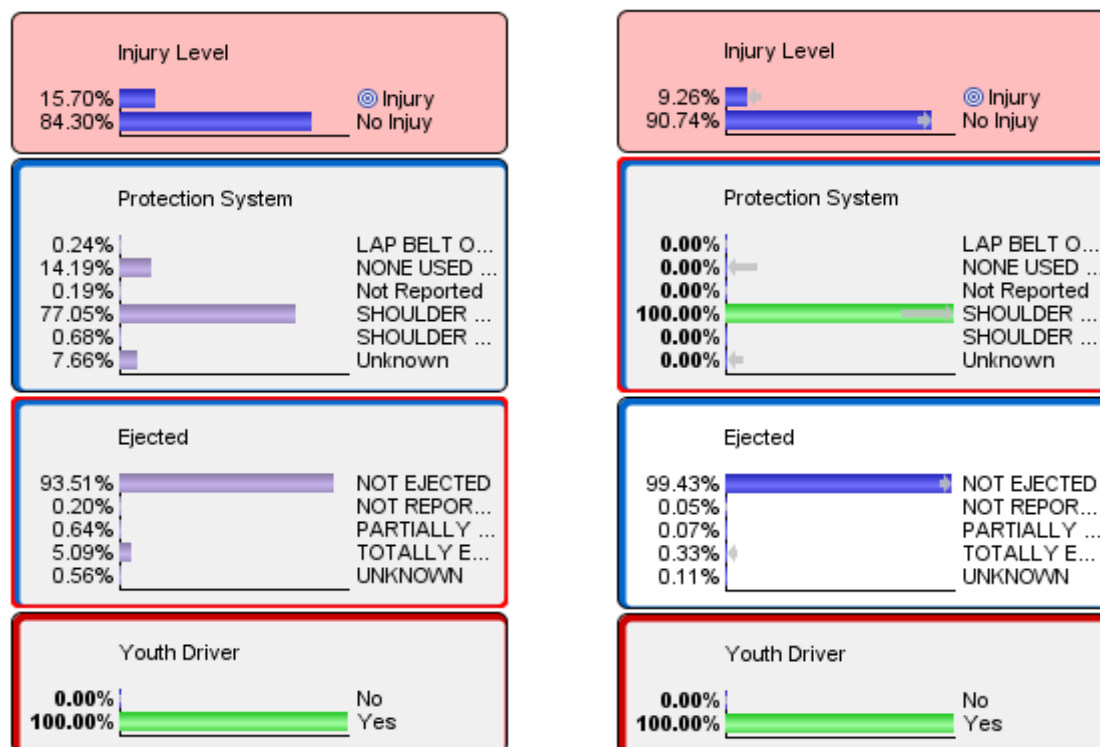


Figure 17 Youth Driver and Ejection/Seatbelt Information

### Protection System and Airbags

The effect of airbags as a safety device has little effect on young driver injuries when utilized in conjunction with seatbelts. If all young drivers wear their seatbelts, Figure 15



demonstrates that airbag deployment would very slightly increase from 24.09% to 24.20%.

Likewise, the injury distribution of 12.09% for seatbelt only (Figure 13) would also very slightly increase to 12.28%.

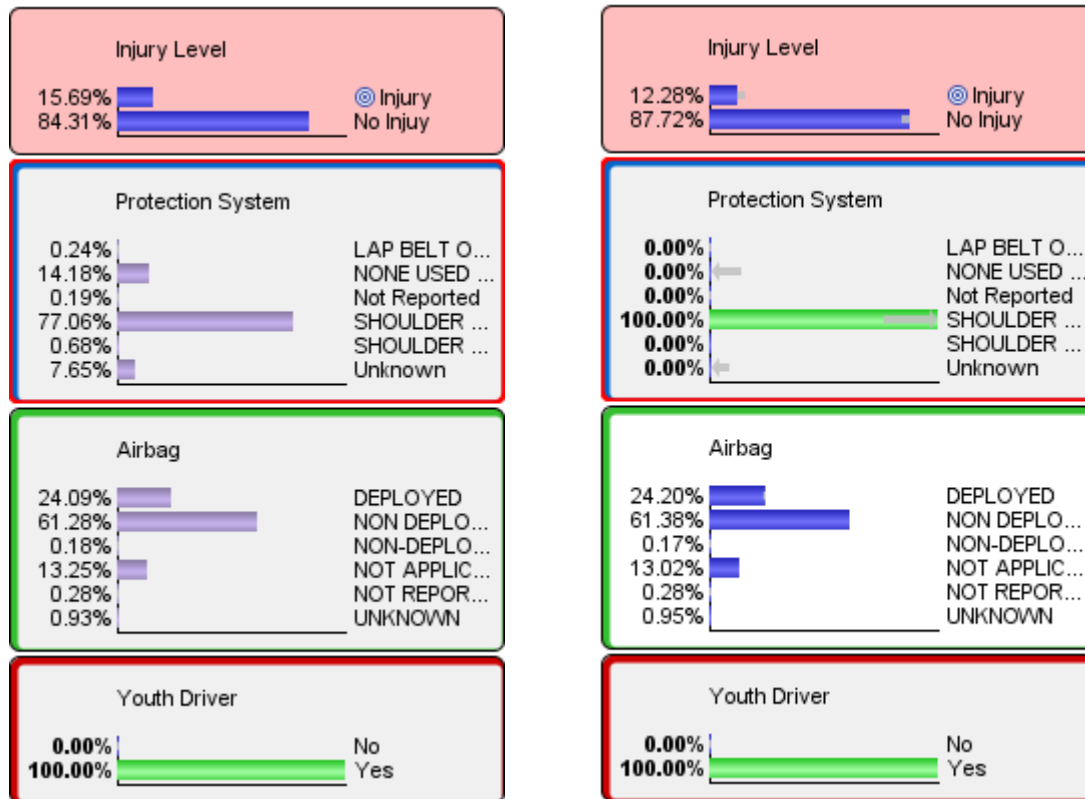


Figure 15 Youth Driver and SeatBelt/Airbag Information

### Substance Suspected

Alcohol use is also identified as having a strong effect on drivers' injury levels in both models. Figure 16 demonstrates suspicion of alcohol has a 16.19% direct effect on youth driver injury distribution.

### Substance Suspected and Protection System Usage

When young drivers drive under the influence of alcohol, they tend to not wear their seatbelts. This can be seen in Figure 17 where young drivers suspected of alcohol only use their seatbelt 63.59% compared to 80.61% for young drivers not suspected of alcohol.

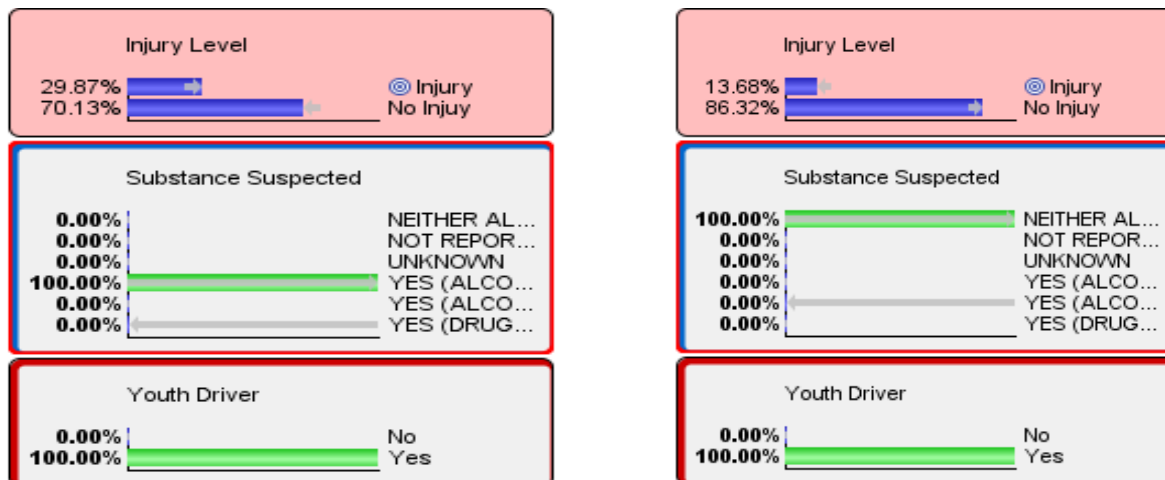


Figure 18 Direct Effect of Substance Suspected On Youth Driver Injury

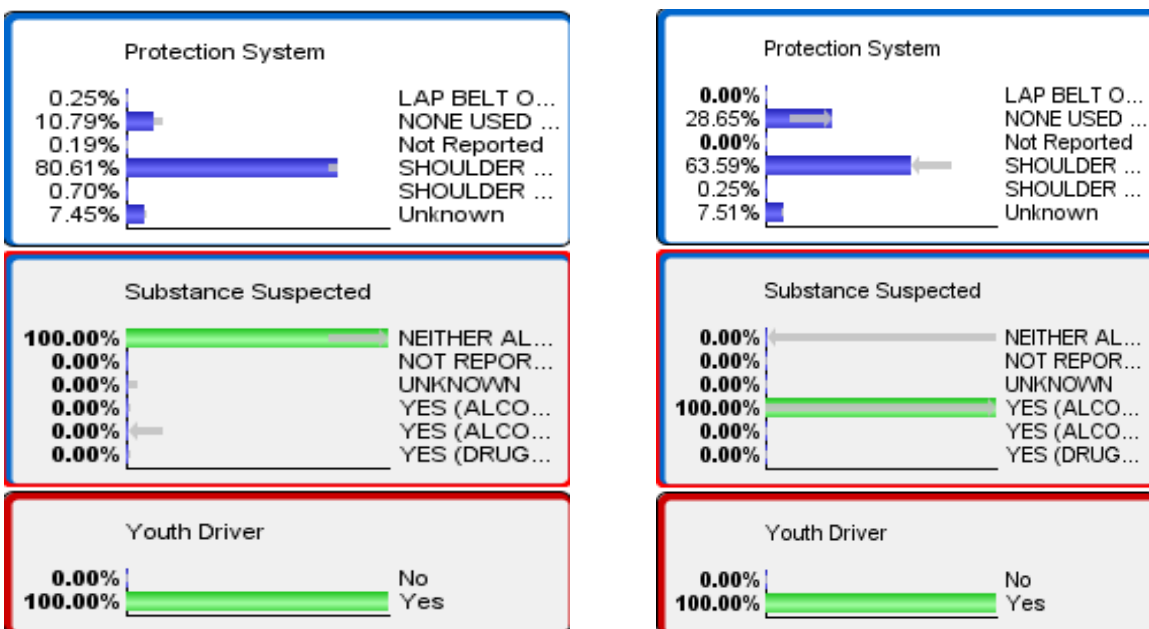


Figure 17 Youth Drivers Protection System Usage and Substance Suspected Information

Violations

Violations, particularly careless operation and speeding, are found to be significant in both models. Within the current model, the injury rate of youth drivers is 15.70% (Figure 12).

This rate increases to 16.76% and 20.17% for careless operation and speeding respectively, as shown in Figure 18.

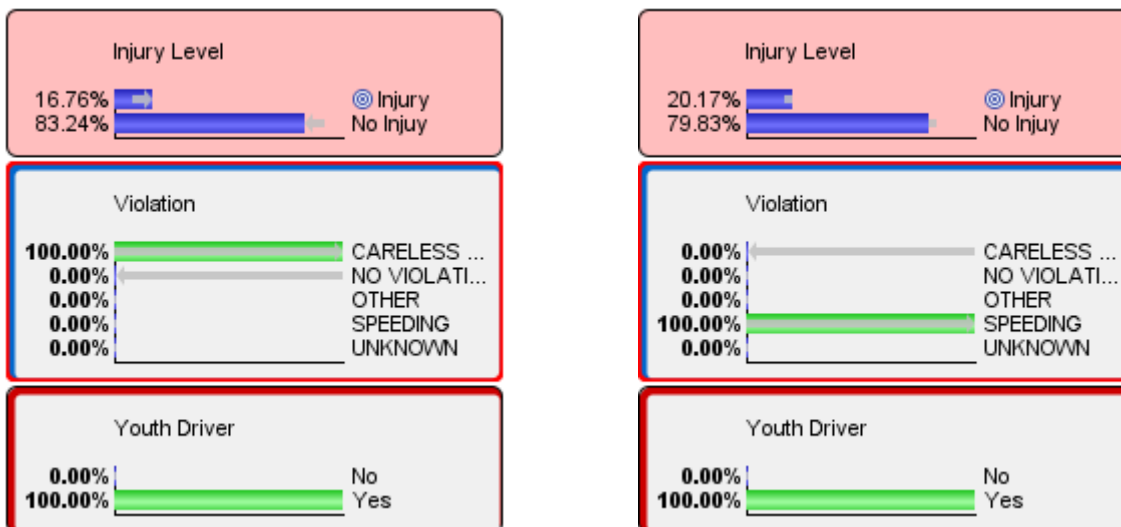


Figure 18 Youth Driver Injury Percentages for Violation Information

#### Violations and Protection System Usage

Seatbelt use is 78.56% for youth drivers when no violation is committed in a crash, as seen in Figure 19. This number decreases to 76.83% and 76.07% for crashes involving careless operation and speeding respectively, also shown in Figure 20.

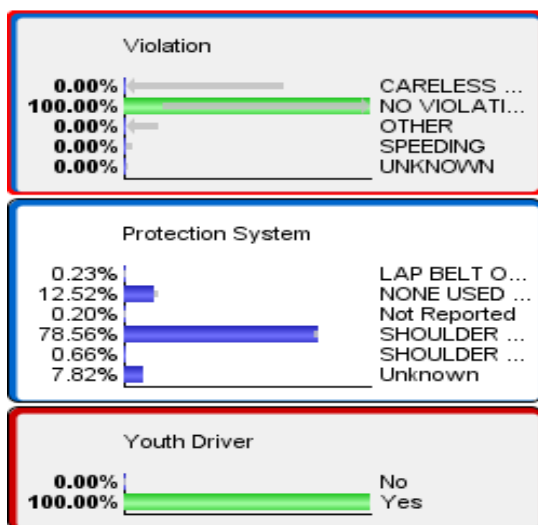


Figure 19 Protection System Usage When No Violation

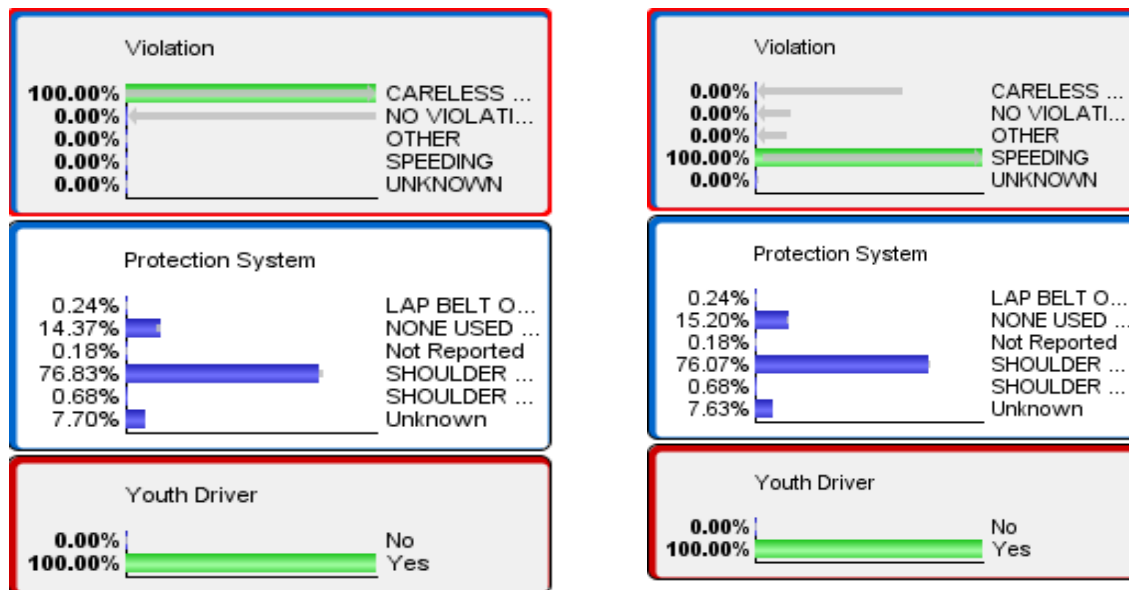


Figure 20 Protection System Usage With Violation

### Violations and Substance Suspected

Violations are also attributed to alcohol use. In crashes with no violation, 97.37% of youth drivers are not suspected of alcohol (Figure 21). However, when a youth driver is in violation of speeding, only 86.38% are not suspected of being under the influence of alcohol. This number further drops to 74.03% for crashes involving careless operation (Figure 21).

### Distraction

While distraction for youth drivers has a minimal effect on injury levels (Figure 22), youth drivers tend to be more distracted than adult drivers (Figure 23). Distraction for youth drivers is also highly attributed to alcohol consumption. When youth drivers are not suspected of alcohol, distraction is only 38.22%, compared to 84.03% when alcohol is suspected (Figure 24).

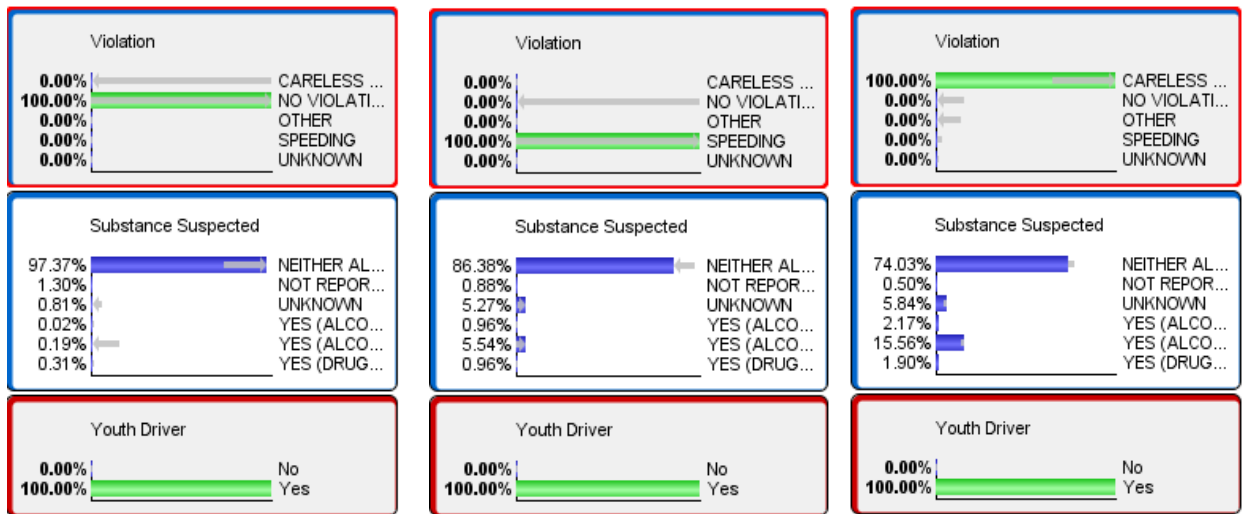


Figure 21 Substance Suspected and Violation Information

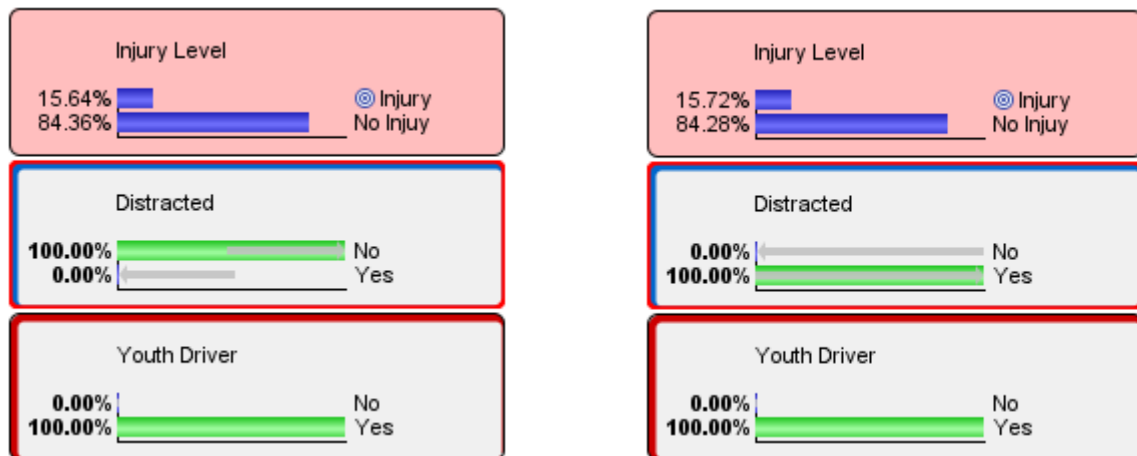


Figure 22 Youth Driver Injury and Distraction Information

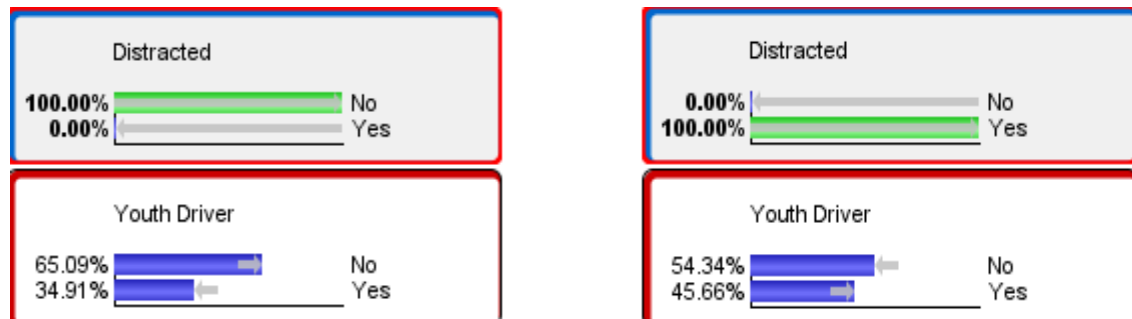


Figure 23 Youth Driver and Distraction Information

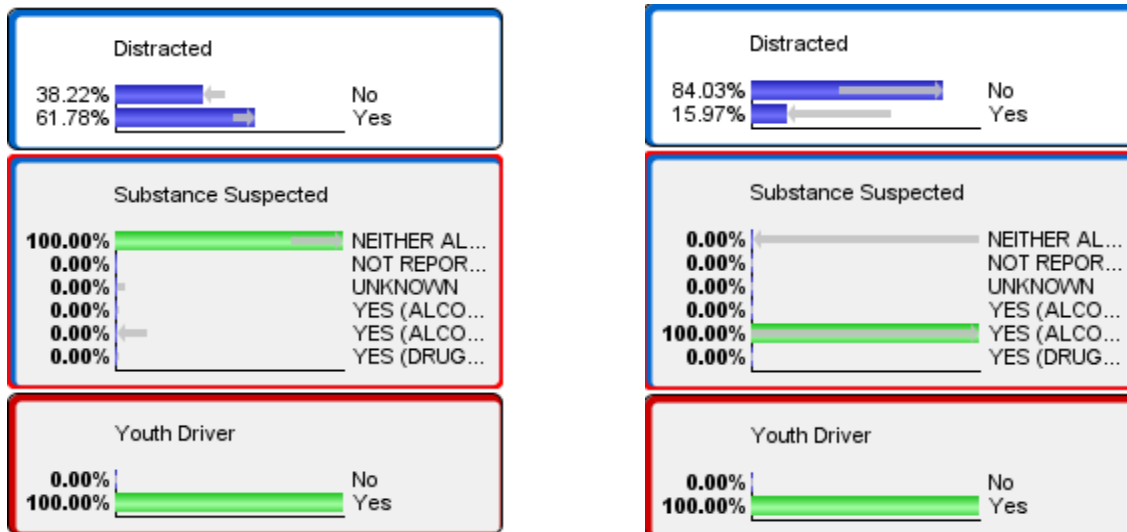


Figure 24 Youth Driver, Distraction, and Substance Suspected Information

## Gender

Male youth drivers are more likely to be distracted and be under the influence of alcohol, while female youth drivers are more likely to drive carelessly and speed as shown in Figure 25.

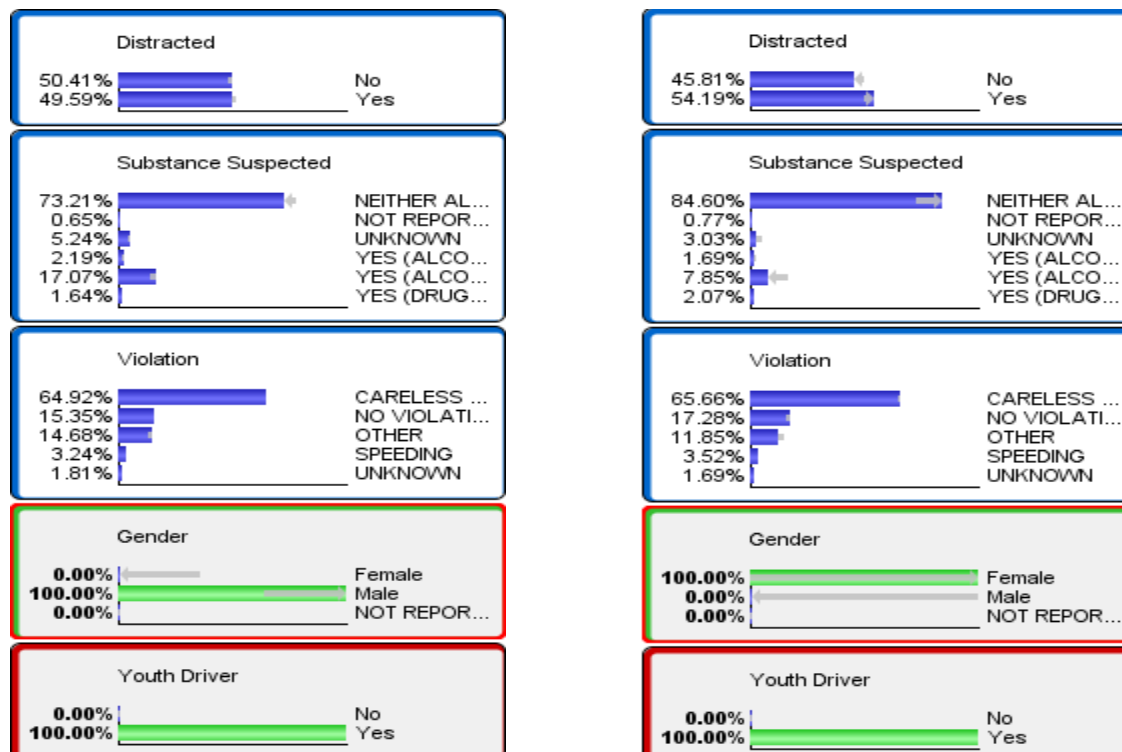


Figure 25 Youth Driver Gender Information

Data from the HSRG in 2012 show that the alcohol fatal crash rate increases as the age group of young drivers increases (Table 13) and males have higher rates than their female counterparts for each age category.

Table 13 LA Alcohol Related Crash Information for Young Drivers

	LICENSED DRIVERS		ALCOHOL RELATED FATAL CRASHES		ALCOHOL FATAL CRASH RATE	
AGE	FEMALE	MALE	FEMALE	MALE	FEMALE	MALE
15-17	33930	34394	3	4	8.84	11.63
18-20	69132	71194	6	13	8.68	18.26
21-24	107593	101477	10	31	9.29	30.55

### Most Harmful Event

Most harmful event is found significant in both models. Within the logistic regression model, hitting a culvert/ditch and hitting a fixed object other than a pole/tree both decrease a driver's chance of injury compared against hitting a pole/tree or the vehicle rolling over, which increase a driver's chance of injury. Evaluating youth drivers' injury levels within the BN, most harmful event is associated with driver violation.

Figure 26 shows driver injuries are lowest when there is no violation. When a youth driver does not have a violation, most harmful event is something other than rolling over or hitting a fixed object.

When a youth is driving carelessly, driver injuries increase along with roll overs and hitting a fixed object (Figure 27). Likewise, when a youth driver is speeding, driver injuries and hitting a pole/tree are at their highest levels (Figure 27).

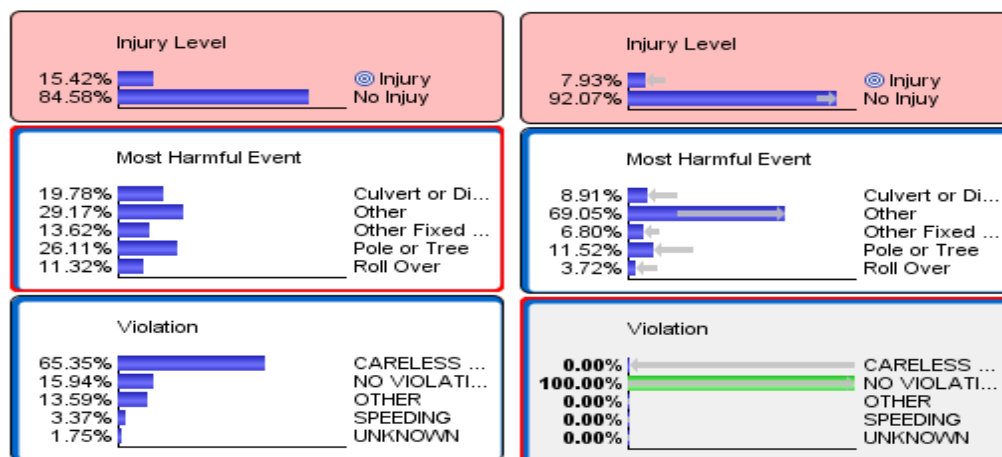


Figure 26 Youth Driver and Most Harmful Event with No Violation Information

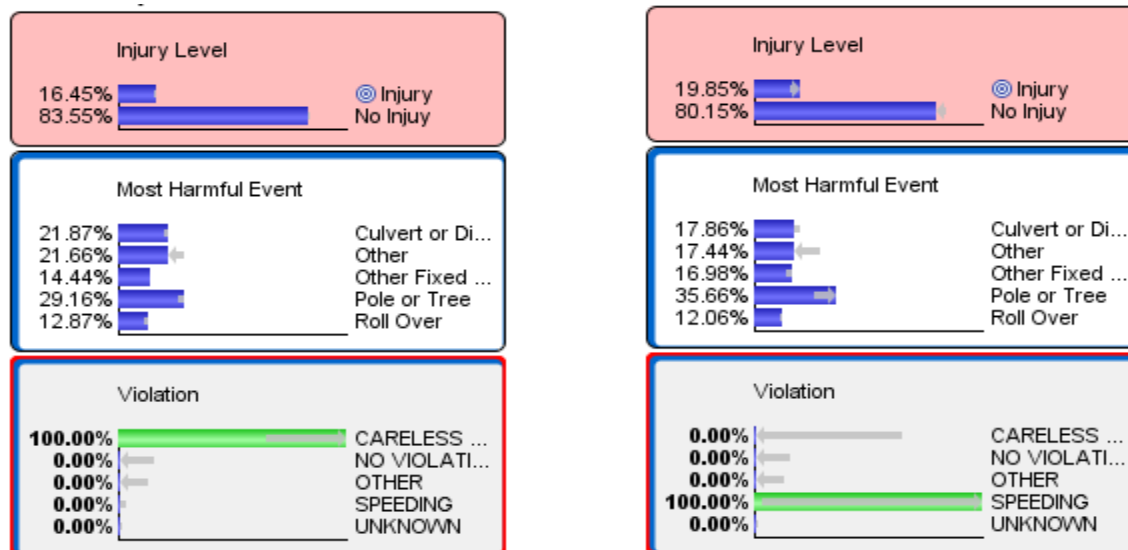


Figure 27 Youth Driver and Most Harmful Event with Violation Information

### Time of Day

Time of day was significant within the BN model, but was excluded from the logistic regression model due to high correlation with lighting. Within the BN, time of day is associated with substance suspected. The number of youth drivers suspected of alcohol increases during the hours of 12:00 am - 6:00 am and 6:00 pm – 12:00 pm, as seen in Figure 28.



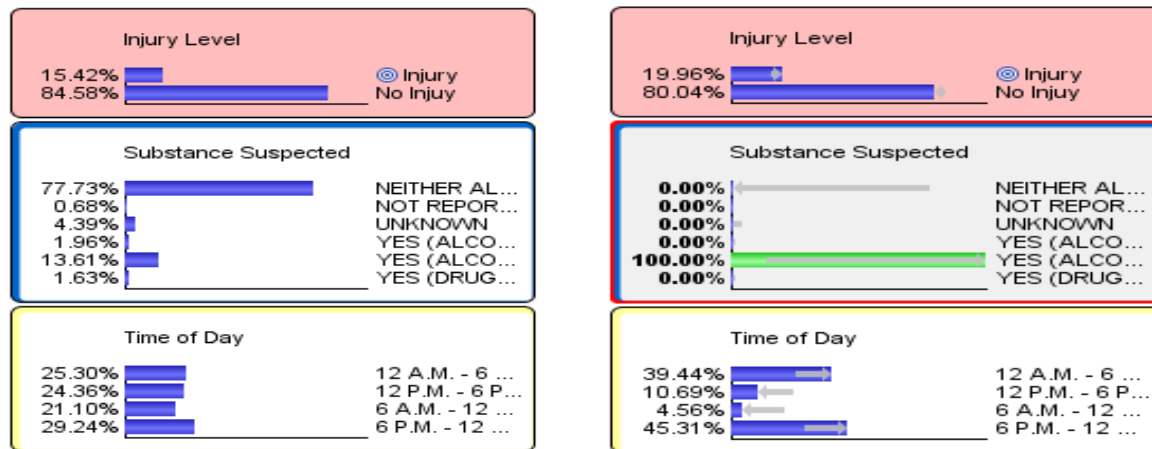


Figure 28 Youth Driver Injury and Substance Suspected/Time of Day Information

### Vehicle Type

The BN model finds vehicle type statistically significant, but with little relative significance with driver injury. Within the BN, vehicle type is associated with driver injury through protection system, showing seatbelt use has more of an effect on injury level than the vehicle type.

The youth driver variable is associated with vehicle type within the BN and Figure 29 shows youth drivers tend to drive more passenger cars and fewer SUVs compared to adult drivers.

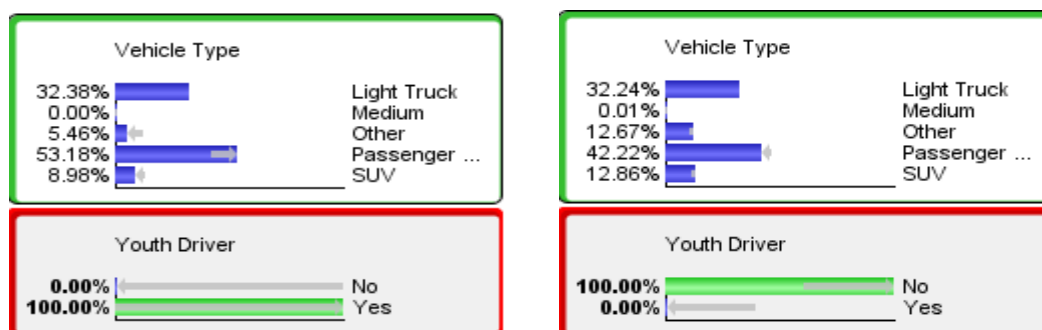


Figure 29 Driver and Vehicle Type Information

### Vehicle Year

Vehicle year was also found statistically significant, but offers little relative significance within the BN. In 1998, the National Highway Safety Transportation Administration required vehicles to have dual front airbags. As safety technology advances, cars are manufactured with more safety features. The safety features in cars manufactured before 2000 is limited compared to vehicles manufactured since 2000.

Within the BN model, vehicle type is associated with driver injury through ejection and air bag. As more drivers remain in the vehicle and utilize safety devices, their chances of serious injuries decrease. The BN places more association of driver injury with not being ejected and using safety devices than the manufacturing year of the vehicle the driver is driving.

## CHAPTER 7. CONCLUSIONS AND FUTURE RESEARCH DIRECTIONS

This study set out to address three objectives. First, identify and quantify the main contributing factors of driver injury levels for single vehicle curve crashes on rural two-lane roadways in Louisiana including driver, environmental, roadway, and vehicle factors using the traditional binary logistic regression modeling technique. Second, using BN modeling, also identify and quantify the main contributing factors of driver injury levels for single vehicle curve crashes on rural two-lane roadways in Louisiana and compare the results against the findings produced from the binary logistic regression model. Third, identify benefits of the BN model over the traditional binary logistical regression model.

There were nineteen significant factors identified in the binary logistic regression model. Of these seven factors have an odds ratio greater than 1.25 with concern to driver injuries; ejected (partially or totally ejected), protection system (none or improper seatbelt usage), substance suspected (alcohol or drugs suspected), violations (speeding & careless operation), and harmful events (rollover & hitting a pole or tree).

Within the BN, only twelve variables were found to be significant. Four of which had relative significance greater than or equal to 0.2 towards driver injury levels; protection system (none used), ejected (totally ejected), substance suspected (alcohol), and violation (careless operations).

### **Driver Factors**

Overall, driver factors are shown by both models as being significant and important as related to driver injury levels. Comparing the results from the two models shows they each identified the following four driver factors as the primary and most dominate factors; ejected (totally ejected), protection system (none used,) substance suspected (alcohol), and violation

(careless operation). The binary logistic regression model also identified violation (speeding) as another significant driver factor.

Driver factors are the only factors that can potentially be altered through educational programs. Data from the HSRG from 2009 – 2013 shows that lack of seatbelt use were attributed to nearly 60% of driver fatalities and nearly 30% of all fatal crashes involving a youth driver suspected of alcohol (HSRG, 2014). Protection system and alcohol are two of the biggest problems in LA when analyzing fatal and serious/moderate injuries.

The LA Strategic Highway Safety Plan created emphasis areas to address these factors, along with youth drivers. This study helps confirm and measure the direct effect these factors have on youth drivers, and drivers overall. Through effective countermeasures, educational programs should be created to modify driver behaviors to help reduce driver fatalities and serious/moderate injuries.

#### Ejection and Protection System

Being ejected from the vehicle and lack of seatbelt usage are the top two factors effecting driver injury identified in within the binary logistic regression and BN models. The raw data shows that 1,103 drivers were partially or totally ejected in curve crashes. Of these 797 drivers (72.26%) were injured. While being ejected from a vehicle significantly increases a driver's odds of being injured, the BN shows that when young drivers wear their seatbelts, the distribution of ejections are drastically reduced from 5.09% to 0.33%. The BN also measures the direct effect of seatbelt usage for young drivers as -21.13%. If no young drivers wear their seatbelt, the injury distribution is 33.32%, however this number falls to only 12.19% if all young drivers were to utilize their seatbelts.

### Substance Suspected

Drivers being suspected of alcohol is the third ranked factor in each model. The BN model estimates the direct effect of alcohol as 16.19% for young drivers. Alcohol use is shown to be associated with lack of seatbelt usage and violations. The BN shows only 63.59% of young drivers use their seatbelt when suspected of alcohol compared to 80.61% of young drivers not suspected of alcohol. Also, the BN demonstrates that 0.19% of young drivers are suspected of alcohol when no violation is given. However, when a young driver is charged with speeding or careless operation, 5.54% and 15.56%, respectively, are suspected of alcohol. While this does not mean alcohol causes the violations, it does show there is strong association.

### Violation (Careless Operation and Speeding)

Careless operation is the seventh ranked factor within the binary logistic regression model and the fourth ranked factor in the BN model. A direct effect of 8.9% is found in the BN for careless operations. While this is a significant and important factor, a clear definition of careless operation is required. In LA, careless operation is defined as “Whereas, careless operation of a vehicle means driving so as to endanger the life, limb, or property of any person” (NHTSA 2014). In order to educate drivers of the effect of careless operation, a better understanding of this factor is needed. If law enforcement officers use this violation to cover a wide range of incidents (speeding, run off road, driving recklessly, improper lane change, etc.) it will be difficult to pinpoint the exact problem and alter driver behavior through education. The raw data shows that careless operation violations account for over 63% of all violations.

Speeding is the fourth ranked factor in the binary logistic regression and is not shown to be significant within the BN. However, the BN does show injuries increase from the current rate of 15.70% to 20.17% if all violations were attributed to speeding.

## Youth Drivers

The binary logistic regression models finds youth drivers as a significant factor with a slight decrease in odds for being injured, while the BN does not show age to be significant. Within the BN, youth drivers are associated with driver injury through vehicle type and protection system usage. This implies vehicle type and protection system usage can serve as confounding factors when evaluating driver age and injury.

Protection system usage is shown to be one of the most significant and important factors in reducing driver injury. The use of a seatbelt in decreasing driver injury is independent of the driver's age, meaning seatbelts decrease the odds of driver injuries for youth drivers and adult drivers alike. The main factor is seatbelt use, not the driver's age.

Vehicle type also plays a role in youth driver injury levels within the BN and will be addressed when evaluating vehicle factors.

## **Environmental Factors**

The binary logistic regression model identifies harmful events (roll over and hitting a pole or tree) as significant factors. While the BN shows harmful event (hitting a pole or tree) as significant, it has little relative significance in regard to driver injury.

## Harmful Events (Roll Over and Hitting a Pole or Tree)

Harmful events, such as roll overs and hitting a pole or tree, were identified as fifth and sixth in the binary logistic model and were not found important within the BN. While these harmful events are significant when evaluating driver injury levels, the underlying cause of the harmful event needs to be considered. While hitting a tree or pole can cause serious injury to a driver, a more important question to ask is "What caused the driver to leave the roadway and hit the pole or tree?" Factors such as driver distraction, speeding, careless operation, rain, and/or alcohol should be considered as confounding factors.

The BN shows whenever a youth is in violation of careless operations driver injuries increase along with rollovers and hitting a fixed object. Likewise, when a youth driver is speeding, driver injuries and hitting a pole/tree are at their highest levels. As the BN demonstrates, the driver behavior is the cause of the harmful event and should be studied more than the harmful event itself. By altering driver behaviors such as alcohol use and violations, drivers are more likely to not leave the roadway thus decreasing the chance of hitting a pole or tree.

#### Other Environmental Factors

The BN model finds time of day (6:00 pm – 12:0 am) and day of week (Friday – Sunday) to be significant, but places little relative significance on these factors. Likewise, the binary logistic regression model shows non-clear weather to decrease the odds of driver injury. Environmental factors are things that cannot be altered or controlled by researchers.

While these variables may be significant, they lack importance on driver injury level. However, while they may not show importance, they should be investigated along with confounding factors. The HSRG's data shows alcohol related crashes occur more often on Fridays through Sundays and between the hours of 6:00 pm and 12:00 am (HSRG 2014). This relationship can also be seen in the BN as day of week and time of day are related to driver injury through substance suspected.

More investigation into alcohol as a confounding factor should be investigated in future research. The same can be said for weather and violations, do drivers speed more in clear weather? Is there something about clear weather that makes drivers feel they can drive more aggressively?

### **Roadway Factors**

The binary logistic regression model found ADT (greater than or equal to 3,000), curve CMF (less than .5), curve length (small and medium), and lane width (less than 12) to be significant, but not as important as driver behavior factors. Only the factors curve CMF and lane width had odds ratios greater than 1, meaning an increased odds of driver injury.

Roadway characteristics themselves would not have a direct effect on driver injury levels, but could contribute to certain types of crashes. For example, it would be expected that there are more rollover and roadway departure crashes within curves than non-curves. In these cases, it would then be the characteristics of rollover and roadway departure crashes that would have a direct effect on driver injury. Based on this reasoning, it was thought harmful events would be a confounding factor for curve crashes and roadway data would have been associated with driver injury through harmful events with the BN. However, the BN developed through machine learning in this study found violations as the confounding variable, not harmful events. Based on the raw data, careless operations account for 63% of all violations. This may be explained in law enforcement officers code most violations within curves as careless operation. Further research should be conducted to determine the relationship between violation and curves.

### **Vehicle Factors**

Only the BN model found the vehicle manufacturing date (less than 2000) and vehicle type as significant, but finds very little relative significance of these factors with driver injury. Within the BN, vehicle type is associated with driver injury through protection system, showing seatbelt use had more of an effect on injury level than the vehicle type. Likewise, the vehicle year is associated with driver injury through airbag and ejection. This implies driver injury are more associated with not being ejected and using safety devices more than the manufacturing year of the vehicle.



## **Benefits of Bayesian Networks**

The BN model produced in this study was developed using machine learning techniques within BayesiaLab software. While software can produce a BN model, domain knowledge is required to understand and interpret the network. For instance, when reviewing the general BN (Figure 8) users cannot interpret the arc directions as causation. This can be seen when evaluating the youth driver and gender nodes. These nodes should have arcs flowing out, not into them. Vehicle type and substance suspected do not influence gender, rather gender influences vehicle type and substance suspected. Likewise, vehicle type and distracted do not influence youth driver, rather youth driver influences vehicle type and distracted. This demonstrates that before concluding causation, correcting the network arcs and establishing a causal network is required.

However, even without having a causal network, a general BN still offers many advantages. First and foremost, for exploratory purposes and making general causal inferences, a general BN can utilize Jouffe's Likelihood Matching technique.

### Causal Inference

Within observational studies, the focus is on what we observe. Binary logistical regression modeling techniques allow the researcher to make observational inferences from the data. However, this is not the same as causal inference, where the focus is on what we do. BNs allow a researcher to explore a domain, with the help of human knowledge, to move from statistical correlation to causal inference (Conrady & Jouffe, 2013b).

Randomized experiments are the gold standard in research studies for concluding causal inference. However, in many areas, it is not feasible, ethical, or practical to perform randomized experiments. One such area is the study of driver injury levels. Research would never be conducted using humans in a randomized experiment to study injury levels due to different crash

factors. Utilizing BayesiaLab, a software used to produce and explore BNs from observational data, a researcher can perform causal inference computations to measure the impact of intervening on a variable (causal inference), rather than simply observing the variable's state (observational inference).

This can be demonstrated when looking into the gender factor. Gender was found to be significant in both models, but with different results. The binary logistic regression model finds males to be .751 times less likely to be injured in a crash than their female counterparts, while the BN model shows gender only has a .016 relative significance with driver injury. From an observational perspective, the binary logistic regression model concludes that there is a difference in driver injury levels based on gender, with males being less likely to be injured. However, this does not mean that gender has a causal influence on driver injury levels.

Using the BN model, the direct effect gender has on driver injury can be measured. While holding the probability distributions fixed for all variables except gender and injury levels, BayesiaLab finds males have a 0.02% increased chance of injury over female drivers. Further investigation of the BN reveals males and females have different driver characteristics. Referring back to Table 10, males are more likely not to wear their seatbelts and drive under the influence of alcohol. They also tend to drive more light trucks, where females drive more passenger cars.

While this does not show a direct causation between gender and injury, it does offer exploratory evaluation of the relationship. The BN can be used to identify differences within the driver characteristics of males and females which may have a casual effect on injury levels.

### Directed Acyclic Graphs

BNs have the benefit of displaying the variables and their relationship through a directed acyclic graphs (DAG), as shown in Figure 8. A DAG represents the structure of a domain

displaying the variables as nodes and their relations with arcs. The graph does not visually display the data, rather it visualizes the structure. The BN is meant to generalize the underlying data, not be a perfect replica of the raw data (Conrady & Jouffe, 2013b). DAGs enable a researcher to visualize the domain and acquire a deeper understanding of the variables and how they relate to one another. However, having domain knowledge is an important requirement before interpreting the DAG as previously discussed.

#### Investigation of Multiple Variable Interactions

Within BN, the researcher can manipulate the probability distribution on any variable to evaluate the behavior or causal effect the intervention has on other related variables in the network, not just the target variable. Within binary logistic regression, the intervention can only measure the observational effect on the target variable.

#### Variables can Support Multiple Outcome Values

Within binary logistic regression models, the variables must be dichotomous. When any variable has more than two possible outcomes, dummy variables must be created to analyze the different possible outcomes. This can be seen in the creation of four dummy variables for most harmful event in this study. However, with BN, the variables are not restricted to be dichotomous and can have multiple values.

### **Direction of Future Research**

#### Data

Complete and accurate data is crucial for quality research projects. The crash data collected and analyzed in this study was taken from crash reports completed by law enforcement officers and as such may be prone to errors. Of particular concern is the accuracy of driver injury level. In LA, driver injury level on the crash form ranges from fatal to no injury, with

three levels of injuries in between; serious, moderate, and possible complaint. Most officers receive little medical training and may have difficulty in properly diagnosing accurate injury levels. Since this study classifies no injury as possible complaint and no injury, the correct classification of possible complaint versus moderate injury is a possible concern and can influence the results of the study.

Completeness of the data is another area of concern. Blood Alcohol Content (BAC) information is missing in most crash reports since officers do not test all drivers involved in a crash. Also, when tests are given, the results are not always updated within the crash report. Without having adequate BAC results to prove drivers were under the influence of alcohol, this study had to use predicted alcohol and substance suspected variables to determine alcohol use. While these variables do indicate alcohol involvement, they do not indicate true impairment. Having more accurate BAC data will help improve the strength of causal relations with driver injury levels in future research.

#### Creating a Causal Network

This study produced a general BN which is utilized to explore the relationship between the different factors and generally infer causation based on the established relationships. However, having a causal network would allow a deeper understanding of the causation between the variables of interest. Further research should be conducted to transition the current general BN into a causation network.

#### Establishing a Quantitative Relationship between Driver Behavior and Crashes

Human factors are identified as the main contributing factors to driver injury levels in this study. This was also concluded in previous studies (Shinar 2007, GAO 2003, Hendericks et al. 1999, Treat et al., 1979). Future research using driving simulators and/or videotaping driver

behavior could help quantify human behavior characteristics and crashes. Looking into seatbelt usage, alcohol, distraction, inattentive, and other driver behaviors and crash occurrences could lead to quantifying the relation between driver behavior and crashes. For example, further research could evaluate the causal relationship of alcohol consumption and seatbelt usage and/or violations. Do drivers who normally wear their seatbelt and drive safely, not buckle up and/or drive carelessly after drinking? Does alcohol cause the driver not to use their safety belt?

## REFERENCES

- AASHTO (2005). *Strategic Highway Safety Plan: A Comprehensive Plan to Substantially Reduce Vehicle-Related Fatalities and Injuries on the Nation's Highways*. AASHTO, Washington D.C.
- AASHTO (2010a). *Highway Safety Manual*. 1st Edition, Volume 1. AASHTO Washington D.C.
- AASHTO (2010b). *Highway Safety Manual*. 1st Edition, Volume 2. AASHTO Washington D.C.
- Al-Ghamdi, A. S. (2002). Using logistic regression to estimate the influence of accident factors on accident severity. *Accident Analysis & Prevention*, 34(6), 729-741.
- Barua, U., Azad, A. K., and Tay, R. (2010). Fatality Risk of Intersection Crashes on Rural Undivided Highways in Alberta, Canada. *Transportation Research Record: Journal of the Transportation Research Board*, 2148(1), 107-115.
- Becker, L. R., Zaloshnja, E., Levick, N., Li, G., and Miller, T. R. (2003). Relative risk of injury and death in ambulances and other emergency vehicles. *Accident Analysis & Prevention*, 35(6), 941-948.
- Beirness, D. J., Mayhem, D. R., Simpson, H. M., and Desmond, K. (2004). The road safety monitor 2004: young drivers. *Traffic Injury Prevention*. 5(3), 237-240.
- Bosch, R. (2011). Bosch Automotive Handbook Eighth Edition. Cambridge, MA: Bentley Publishers.
- Boyce, T. E. and Geller, E. S. (2002). An instrumented vehicle assessment of problem behavior and driving style do younger drivers actually take more risk? *Accident Analysis & Prevention*. 34 (1), 51-64.
- CDC. Centers for Disease Control and Prevention, 2010.  
[http://www.cdc.gov/MotorVehicleSafety/Teen\\_Drivers/teendrivers\\_factsheet.html](http://www.cdc.gov/MotorVehicleSafety/Teen_Drivers/teendrivers_factsheet.html), Accessed December 2013.
- Chang, L. Y., and Wang, H. W. (2006). Analysis of traffic injury severity: An application of non-parametric classification tree techniques. *Accident Analysis & Prevention*, 38(5), 1019-1027.
- Charniak, E. (1991). Bayesian networks without tears. *AI Magazine*, 12(4), 50.
- Chen, H., Cao, L., and Logan, D. B. (2012). Analysis of risk factors affecting the severity of intersection crashes by logistic regression. *Traffic injury prevention*, 13(3), 300-307.
- Chen, S. (2010). Mining Patterns and Factors Contributing to Crash Severity on Road Curves. Doctoral Dissertation, Brisbane, Australia.

- Clarke, D., Patrick, W., Bartle, C., and Truman, W. (2006). Young driver accidents in the UK: The influence in age, experience, and time of day. *Accident Analysis and Prevention*, 38, 871-878.
- Conrady, S. and Jouffe, L. (2013a). Introduction to Bayesian Networks & BayesiaLab.
- Conrady, S. and Jouffe, L. (2013b). Vehicle Size, Weight, and Injury Risk.
- Conrady, S., and Jouffe, L. (2013c). Causal Inference with Bayesian Networks.
- Conrady, S. and Jouffe, L. (2014). Where is my Bag?
- Darwiche, A. (2009). *Modeling and reasoning with Bayesian networks*. Cambridge University Press.
- de Oña, J., López, G., Mujalli, R., and Calvo, F. J. (2013). Analysis of traffic accidents on rural highways using Latent Class Clustering and Bayesian Networks. *Accident Analysis & Prevention*, 51, 1-10.
- de Oña, J., Mujalli, R. O., and Calvo, F. J. (2011). Analysis of traffic accident injury severity on Spanish rural highways using Bayesian networks. *Accident Analysis & Prevention*, 43(1), 402-411.
- Deery, H. (1999). Hazardous and Risk Perception among Young Novice Drivers. *Journal of Safety Research*, 30(4), 225-236.
- Dissanayake, S. (2003). *Young Drivers and Run-Off-the-Road Crashes*. Proceedings of the 2003 Mid-Continent Transportation Research Symposium, Ames, Iowa.
- Dissanayake, S., and Lu, J. J. (2002). Factors influential in making an injury severity difference to older drivers involved in fixed object-passenger car crashes. *Accident Analysis & Prevention*, 34(5), 609-618.
- Donaldson III, W. F., Hanks, S. E., Nassr, A., Vogt, M. T., and Lee, J. Y. (2008). Cervical spine injuries associated with the incorrect use of airbags in motor vehicle collisions. *Spine*, 33(6), 631-634.
- Donnell, E. T., and Mason, J. M. (2004). Predicting the severity of median-related crashes in Pennsylvania by using logistic regression. *Transportation Research Record: Journal of the Transportation Research Board*, 1897(1), 55-63.
- Eluru, N., Bhat, C. R., and Hensher, D. A. (2008). A mixed generalized ordered response model for examining pedestrian and bicyclist injury severity level in traffic crashes. *Accident Analysis & Prevention*, 40(3), 1033-1054.
- Gabauer, D. J., and Gabler, H. C. (2008). Comparison of roadside crash injury metrics using event data recorders. *Accident Analysis & Prevention*, 40(2), 548-558.

- General Accounting Office (2003). Research Continues on a Variety of Factors That Contribute to Motor Vehicle Crashes. GAO-03-436, General Accounting Office, Washington, D.C.
- Hauer, E. (2006). The frequency–severity indeterminacy. *Accident Analysis & Prevention*, 38(1), 78-83.
- Hendricks (2010). Traffic Accidents Are Top Cause of Teen Deaths. WebMD Health News. (see: <http://www.webmd.com/parenting/news/20100505/traffic-accidents-are-top-cause-of-teen-deaths>) , Accessed December 2013.
- Hendericks, D. L, Fell, J. C., and Freedman, M. (1999). *Identifying Unsafe Driver Actions that Lead to Fatal Car-Truck Crashes*. Summary Technical Report, DTNH22-94-C-05020, U.S. Department of Transportation.
- HSRG (2014). Highway Safety Research Group Crash Data Warehouse. Louisiana State University, Baton Rouge, LA.
- Huang, H. F., Schneider, R. J., Zegeer, C. V., Amerlyck, A. J., and Lacy, J. K. (2002). Identification of Severe *Crash Factors and Countermeasures in North Carolina-Final Report*. FHWA/NC/2001-03, University of North Carolina.
- Hummer, J. E., Rasdorf, W., Findley, D. J., Zegeer, C. V., and Sundstrom, C. A. (2010). Curve Collisions: Road and Collision Characteristics and Countermeasures. *Journal of Transportation Safety & Security*, 2, 2003-220.
- Jones, A. P., and Jørgensen, S. H. (2003). The use of multilevel models for the prediction of road accident outcomes. *Accident Analysis & Prevention*, 35(1), 59-69.
- Lenguerrand, E., Martin, J. L., and Laumon, B. (2006). Modelling the hierarchical structure of road crash data—Application to severity analysis. *Accident Analysis & Prevention*, 38(1), 43-53.
- Maycock, G. (2002). Novice driver accidents and driving test. TRL Research Report 527. Transport Research Laboratory, Crowthorne, Berkshire.
- Mayhew, D. Simpson, H. and Singhal, D. (2005). Best practices for Graduated Driver Licensing: in Canada. Traffic Injury Research Foundation, Ottawa, CA.
- Mercier, C. R., Shelley, M. C., Rimkus, J. B., and Mercier, J. M. (1997). Age and gender as predictors of injury severity in head-on highway vehicular collisions. *Transportation Research Record: Journal of the Transportation Research Board*, 1581(1), 37-46.
- Mujalli, R. O. and de Oña, J., (2011). A method for simplifying the analysis of traffic accidents injury severity on two-lane highways using Bayesian networks. *Journal of Safety Research*, 42, 317-326.



- Mujalli, R. O. and de Oña, J. (2011b). Injury severity models for motor vehicle accidents: a review.
- NHTSA (2010). Traffic Safety Facts 2010 Data. National Highway Transportation Safety Administration Report DOT-HS-811-622. U.S. Department of Transportation, Washington DC.
- NHTSA (2011). Traffic Safety Facts 2011 Data. National Highway Transportation Safety Administration Report DOT-HS-811-754. U.S. Department of Transportation, Washington DC.
- NHTSA (2014). Louisiana State Statues. (see: <http://www.nhtsa.gov/people/injury/enforce/stspdlaw/laspeed.htm>), Accessed October 2014.
- OECD (2006). Young Drivers: the road to safety. Organizations of Economic Cooperation and Development and the European Conference of Ministries of Transport Report ITRD. OECD Publishing, Paris, France.
- Pai, C. W. (2009). Motorcyclist injury severity in angle crashes at T-junctions: Identifying significant factors and analysing what made motorists fail to yield to motorcycles. *Safety Science*, 47(8), 1097-1106.
- Peek-Asa, C., Britton, C., Young, T., Pawlovich, M., and Falb, S. (2010). Teenage driver crash incidence and factors influencing crash injury by rurality. *Journal of Safety Research*, 41(6), 487-492.
- Pollatsek, A., Narayanaan, V., Pradhan, A., and Fisher, D. L. (2006). Using Eye Movements to Evaluate a PC-Based Risk Awareness and Perception Training Program on a Driving Simulator. *Human Factors: The Journal of Human Factors and Ergonomics Society*. 48: 447.
- Porter M. and Whitton, M. J.. (2002). Assessment of Driving With Global Positioning Systems and Video Technology in Young, Middle-Aged, and Older Drivers. *Journal of Gerontology: Medical Science*, 57A(9), M578-M582.
- Robertson, R., and Vanlaar, W. (2008). Elderly drivers: Future challenges?. *Accident Analysis & Prevention*, 40(6), 1982-1986.
- Treat, J. R., Tumbas, N. S., McDonald, S. T., Shinar, D., Hume, R. D., Mayer, R. E., ... and Castellan, N. J. (1979). Tri-level study of the causes of traffic accidents: final report. Executive summary.
- Queensland Transport, Q. (2006). Webcrash 2.3: Queensland Government.

- Qin, X., Wang, K., and Cutler, C. E. (2013). Logistic Regression Models of the Safety of Large Trucks. *Transportation Research Record: Journal of the Transportation Research Board*, 2392(1), 1-10.
- Savolainen, P. T., Mannering, F. L., Lord, D., and Quddus, M. A. (2011). The statistical analysis of highway crash-injury severities: A review and assessment of methodological alternatives. *Accident Analysis & Prevention*, 43(5), 1666-1676.
- Schneider IV, H., Savolainen, P., and Zimmerman, K. (2009). Driver Injury Severity Resulting from Single-Vehicle Crashes Along Horizontal Curves on Rural Two-Lane Highways. In *Transportation Research Record: Journal of the Transportation Research Board*, NO. 2102, Transportation Research Board of the National Academics, Washington, D.C, pp.85-92.
- Shinar, D. (2007). *Traffic Safety and Human Behavior*. Howard House, Bingley BD16 1WA, UK.
- Simoncic, M. (2004). A Bayesian network model of two-car accidents. *Journal of transportation and Statistics*, 7(2/3), 13-25.
- Tabachnick, B. J. and Fidell, L. S. (2006). *Using Multivariate Statistics Fifth Edition*. Needham Heights, MA: Allyn & Bacon, Inc.
- Tay, R., Rifaat, S. M., and Chin, H. C. (2008). A logistic model of the effects of roadway, environmental, vehicle, crash and driver characteristics on hit-and-run crashes. *Accident Analysis & Prevention*, 40(4), 1330-1336.
- Treat, J. R., Tumbas, N. S., McDonald, S. T., Shinar, D., Hume, R. D., Mayer, R. E., ... and Castellan, N. J. (1979). Tri-level study of the causes of traffic accidents: final report. Executive summary.
- Torbic, D. J., Harwood, D. W., Gilmore, D. K., Pfefer, R., Neuman, T. R., Slack, K. L., and Hardy, K. K. (2004). *Guidance for implementation of the AASHTO strategic highway safety plan. Volume 7: A guide for reducing collisions on horizontal curves* (No. Project G17-18 (3) FY'00).
- WISQARS. Centers for Disease Control and Prevention's Web-based Injury Statistics Query and Reporting System, 2010. <http://www.cdc.gov/injury/wisqars/>, Accessed December 2013.
- Xie, Y., Zhang, Y., and Liang, F. (2009). Crash injury severity analysis using Bayesian ordered probit models. *Journal of Transportation Engineering*, 135(1), 18-25.
- Yagil, D. (1998). Gender and age-related differences in attitudes towards traffic laws and traffic violations. *Transportation Research Part F*, 1, 123-135.

- Yan, X., Radwan, E., and Abdel-Aty, M. (2005). Characteristics of rear-end accidents at signalized intersections using multiple logistic regression model. *Accident Analysis & Prevention*, 37(6), 983-995.
- Zhang, H. (2010). Identifying and Quantifying Factors Affecting Traffic Crash Severity in Louisiana. Doctoral Dissertation, Louisiana State University. Baton Rouge, LA.
- Zhu, H., Dixon, K. K., Washington, S., and Jared, D. M. (2010). Predicting single-vehicle fatal crashes for two-lane rural highways in Southeastern United States. *Transportation Research Record: Journal of the Transportation Research Board*, 2147(1), 88-96.

## APPENDIX: HSRG PREDICTED ALCOHOL FORMULA

A driver is predicted to have alcohol using the following logistic regression model:

$$P(\text{Alcohol} | \mathbf{x}) = \frac{1}{1 + e^{-\beta(\mathbf{x})}}$$

with

$$\begin{aligned}\beta(\mathbf{x}) = & -0.761 - 0.9246x_1 - 0.2647x_2 + 0.804x_3 - 0.1514x_4 - 2.5984x_5 + 2.889x_6 + 1.662x_7 \\ & + 1.662x_8 + 0.7132x_9 - 0.3123x_{10} - 0.5066x_{11} + 0.476x_{12}\end{aligned}$$

where  $\mathbf{x} = (x_1, x_2, \dots, x_{12})$  is a 12-tuple of binary variables whose values correspond to the truth values of

- $x_1$  = The crash happened between 5:00am and 5:00pm
- $x_2$  = The crash happened between 5:00pm and 8:00pm
- $x_3$  = The crash happened between 12:00am and 5:00am
- $x_4$  = The crash happened Monday - Thursday
- $x_5$  = The officer suspects neither drugs nor alcohol
- $x_6$  = The officer suspects alcohol
- $x_7$  = The officer suspects drugs
- $x_8$  = The officer suspects alcohol and drugs
- $x_9$  = The crash was a 'Non-Collision with Another Vehicle' crash
- $x_{10}$  = The crash was a 'Collision with Another Vehicle' crash
- $x_{11}$  = The crash type was 'Other'
- $x_{12}$  = No driver restraint was used

## VITA

Cory Hutchinson was born in Houma, Louisiana. He graduated with a Bachelor of Science in Quantitative Business Analysis in December 1991 from Louisiana State University (LSU). Cory then entered the graduate program at LSU and received a master's degree in Quantitative Business Analysis in December 1993. From March 1994 to October 2007, Cory worked for the University Information Systems Department at LSU working his way from an Applications Analyst I to Manager. During this time, he also received a master's in Business Administration from LSU in May 1998. In November 2007, Cory joined the Highway Safety Research Group, a division of Information Systems and Decision Sciences (ISDS), at LSU as a Manager and currently serves as Associate Director. From 2008 to the present, he taught multiple ISDS courses within the area of business intelligence. In January 2008, he started the doctoral program at LSU. In August 2014, Cory received a master's degree in Human Resource Education and Workforce Development and expects to receive the degree of Doctor of Philosophy in Human Resource Education and Workforce Development in December 2014 from LSU.