TRR
JOURNAL OF THE TRANSPORTATION RESEARCH BOARD

Article

Transportation Research Record
1–10
© National Academy of Sciences:
Transportation Research Board 2018
Reprints and permissions:
DOI: 10.1177/0361198118772703
journals.sagepub.com/home/trr
SAGE

Crash and Near-Crash Risk Assessment of Distracted Driving and Engagement in Secondary Tasks: A Naturalistic Driving Study

Peter R. Bakhit[1], **BeiBei Guo**[2], **and Sherif Ishak**[3]

Abstract

Distracted driving behavior is a perennial safety concern that affects not only the vehicle's occupants but other road users as well. Distraction is typically caused by engagement in secondary tasks and activities such as manipulating objects and passenger interaction, among many others. This study provides an in-depth analysis of the increased crash/near-crash risk associated with different secondary tasks using the largest real-world naturalistic driving dataset (SHRP2 Naturalistic Driving Study). Several statistical and data-mining techniques were developed to analyze the distracted driving and crash risk. First, a bivariate probit model was constructed to investigate the relationship between engagement in a secondary task and the safety-critical events likelihood. Subsequently, two different techniques were implemented to quantify the increased crash/near-crash risk because of involvement in a particular secondary task. The first technique used the baseline-category logits model to estimate the increased crash risk in terms of conditional odds ratios. The second technique used the a priori association rule mining algorithm to reveal the risk associated with each secondary task in terms of support, confidence, and lift indexes. The results indicate that reaching for objects, manipulating objects, reading, and cell phone texting are the highest crash risk factors among various secondary tasks. Recognizing the effect of different secondary tasks on traffic safety in a real-world environment helps legislators enact laws that reduce crashes resulting from distracted driving, as well as enabling government officials to make informed decisions about the allocation of available resources to reduce roadway crashes and improve traffic safety.

Driving is a daily complex task that requires a driver's full attention. Despite the complexity associated with this task, it is not uncommon to observe drivers perform other secondary tasks while operating a vehicle. These secondary tasks might include reading a newspaper in slow-moving traffic, shaving to be ready for work, and discussing important topics with a passenger, among many others. Although these tasks might seem trivial, they degrade driving performance and increase the likelihood of a crash or near-crash event. Moreover, the technological features embedded in vehicles nowadays, in addition to the advanced wireless communication devices, have brought a new level of distraction to the driving environment (*1*). Thus, it is important to estimate the relative crash/near-crash risk for better understanding of the effect of different types of secondary tasks on traffic safety.

Recognizing the effect of different secondary tasks on traffic safety is imperative. It can help legislators initiate laws that reduce crashes resulting from distracted driving. It can also help government officials make informed decisions about the allocation of available resources to reduce roadway crashes. Most of the previous researches attempted to gather information about secondary tasks through driving simulators and interview studies. Limited research, however, has tapped into comprehensive naturalistic driving datasets. In this study, a large naturalistic driving dataset (SHRP2 Naturalistic Driving Study) is employed to investigate the relationship between engagement in a secondary task

[1] Department of Civil and Environmental Engineering, Louisiana State University, Baton Rouge, LA
[2] Department of Experimental Statistics, Louisiana State University, Baton Rouge, LA
[3] Civil and Environmental Engineering, University of Alabama in Huntsville, Huntsville, AL

Corresponding Author:
Address correspondence to Peter R. Bakhit: pbakhi1@lsu.edu

and the crash/near-crash likelihood using different statistical and data-mining techniques. This dataset includes information collected from more than 3,000 drivers recruited in six different states. By far, this is the largest naturalistic driving study conducted in the United States to date. Thus, this dataset is considered to be one of the most representative datasets of the driving population. Furthermore, in this study, the increased crash risk resulting from distracted driving is quantified so that secondary tasks with a higher crash risk are recognized.

Background

To assess the impact that a particular secondary task has on the crash risk, three different approaches are frequently used, namely, experimental studies, interview studies, and observational studies. Experimental studies are usually conducted in either a driving simulator or in a controlled traffic environment (*2, 3*). During the experiments, participants are asked to perform a specific secondary task at a given time according to the scenarios designed by the experimenter. A comprehensive overview of distracted driving experimental studies can be found in Regan et al. (*3*). Although the experimental studies were successful in recognizing the degradation in driving performance because of the engagement in a secondary task, they were not helpful in making a valid estimate of the actual crash risk for two main reasons. First, the participants did not decide where, when, and how to engage in the secondary task, which is not an accurate representation of real-world secondary task involvement. Second, the transferability of the outcomes from the driving simulators to real life remains questionable. Thus, experimental studies are not considered the best approach to determine the increased crash risk resulting from engagement in different secondary tasks.

Interview studies represent another approach in collecting secondary tasks information (*4–8*). In these studies, information is gathered using telephone surveys, as in McEvoy et al. and Royal (*5, 6*), or online surveys as in Lansdown, and Young and Lenné (*4, 8*). In Sullman and Baas (*7*), a sample of 287 New Zealand drivers were asked about their cell phone use while driving and the perceived risk. The results showed that the percentage of drivers who never used a cell phone or used a cell phone occasionally while driving were 43% and 43%, respectively. The percentage of drivers who used the cell phone frequently was 14%. Although Sullman and Baas's study (*7*) was interested in cell phone use as the sole secondary task, the remaining four studies (*4, 5, 6, 8*) were concerned with all secondary tasks. In McEvoy's study (*5*), the participants were asked to list all secondary tasks that lasted for 5 min or more over the last trip. Young et al. (*8*) conducted another survey asking participants to report all kinds of secondary tasks and how often they engaged in them. Eating/drinking, smoking, clothing/body care, integrated devices, passenger-related, and outside distractors are the most common distractors that contribute to crashes. Despite the valuable information reported in these studies, self-reporting bias (*9*) is a major concern.

Observational studies or naturalistic driving studies (NDS) are the most realistic approaches for gathering secondary tasks data (*10–13*). In these studies, vehicles are equipped with advanced data-collection devices to record normal driving behavior. For example, Klauer (*11*) studied the distracted driving and its relationship to safety-critical events (SCEs). For this 1-year study project, 82 crash events and 761 near-crash events were observed. This data was then analyzed as a part of the 100-Car Naturalistic Driving Study. The results indicated that passenger interaction, embedded devices in a vehicle, and manipulating objects are the most common secondary tasks. Moreover, in three other studies, secondary tasks such as eating/drinking, smoking-related, clothing/body care, integrated devices, passenger-related, outside distractors, and other in-vehicle devices were listed as the most frequent secondary tasks (*10–13*). Another study conducted by Olson et al. investigated the impact of distracted driving on traffic safety (*14*). In Olson et al.'s study, the increased crash risk was reported in terms of odds ratio (OR) estimates. Key findings were that commercial drivers were engaged in secondary tasks in 71% of crashes, 46% of near-crashes, and 60% of all SCEs. The most risky behavior identified was "Cell phone texting" with an OR of 23.2, followed by "Interacting with a device" (OR = 9.9), and "Cell phone dialing" (OR 5.9). Although research approaches derived from NDS data are considered more realistic, a high number of equipped vehicles in addition to long observation periods are required. As a result, in this study, the SHRP2 NDS data (the largest NDS conducted so far) is employed. This real-world driving study project was successful in collecting approximately two petabytes of video and sensor data from more than 3,000 drivers over a 3-year period (2010–2013).

The literature review shows evidence of a relationship between engagement in a secondary task and crash likelihood. However, there are some limitations that need to be addressed. First, most of the statistical models did not take into account the correlation between interrelated variables, such as engagement in a secondary task and the crash likelihood. More specifically, in previous statistical models, the multiple dependent variables were modeled separately, each as a function of a set of independent variables, and therefore, the correlations among the dependent variables were ignored. Second, recent data-mining techniques have captured researchers' attention as they outperform traditional modeling techniques. Thus, this study is the first of its kind to introduce the market basket analysis (association modeling) into the distracted driving area. In this context, the objectives of this study are to (a) construct a statistical model that considers the correlation between engagement in a secondary task and the crash/near-crash occurrence (bivariate probit model); (b) estimate valid crash risk measures for different types of secondary tasks using a large naturalistic driving dataset without sampling bias (baseline-category logits model); and (c) offer a new methodology for investigating the relationship between the different secondary tasks and the crash/near-crash risk using a new data-mining technique.

The rest of the paper is organized as follows. "Data Description" describes the data used in this study. "Model Development" presents model development, detailing the bivariate probit model and results, and the two alternative models used to quantify the increased crash-risk estimate using the baseline-category logits model, and the association rule mining model, respectively. "Discussion" discusses the models' outcomes and their implications. Finally, conclusions are provided in "Conclusion."

Data Description

In this study, the SHRP2 NDS dataset was employed to achieve the research objectives. SHRP2 NDS data is considered the largest study of its kind that has been conducted in the United States to date. The data include observations from more than 3,000 drivers recruited in six different states, namely New York, Washington, Indiana, Florida, Pennsylvania, and North Carolina. The dataset consists of 21,292 driving events (observations) categorized in three groups: crash (1,415 events), near-crash (2,616), and baseline events (18,873) (*15*). For each driving event, different data types are collected, such as event summary data, and driver sociodemographic data. Event summary data include information related to the vehicle and environmental conditions during the event incidence time, such as event type, event severity, surface condition, weather conditions, and whether a secondary task existed or not, among many other factors. Driver sociodemographic data include information related to a driver's socioeconomic and driving characteristics, such as gender, age, education, working status, marital status, and years of driving, among others. Table 1 lists all the different variables used in this research. These variables are selected based on previous distraction-safety-related studies (*11, 14, 16*).

Prior to the models' development, the SHRP2 NDS dataset was reduced to remove any biases that might have affected the crash risk estimates. First, a crash event that did not involve injuries or property damage was excluded from the final dataset. Second, driving events were filtered out to remove any events associated with observable driver impairment. Finally, driving records with missing driver information were excluded; leaving 905 crashes, 2,558 near-crashes, and 18,544 baseline events as a final dataset. It should be noted that the "Secondary Task" variable is the key variable in the rest of the study. This variable shows the type of the secondary task in which the driver was engaged prior to the crash/near-crash time or during the selected normal driving event. The type of the secondary task was manually coded by reviewing video by Virginia Tech Transportation Institute (VTTI) according to the SHRP

Table 1. List of Input Variables

Type	Variable	# of Categories	Categories
Driver characteristics	Age	11	(16–19), (20–24), … (85–89)
	Gender	2	Male/Female
	Working status	3	Full-time, Part-time, Not working
	Marital status	5	Single, Married, Divorced, Widowed
	Education	5	High School, College degree, Advanced degree
	Driver training	6	Through a private company, Through school, Informal training, No informal training by parents, No training, Other
	Years of driving		Quantitative measure
Event characteristics (summary data)	Event type	3	Crash/Near Crash/Baseline
	Event duration		Quantitative measure
	Secondary task	14	Cell phone, Texting, Passenger interaction, and so forth
	Presence of passenger	2	Yes/No
Roadway characteristics (summary data)	Relation to junction	10	Intersection, Parking entrance, Non-junction, Rail grade crossing, and so forth
	Intersection influence	2	Yes/No
	Alignment	3	Curve left, Curve right, Straight
	Grade	5	Level, Grade up, Grade down, Dip, Hill crest
	Traffic lighting	6	Daylight, Dusk, Darkness lighted, darkness unlighted, Dawn, Other
	Locality	11	Business, School, Interstate, Residential, and so forth

Table 2. Secondary Tasks Classification

Secondary task type	Description
Eating/drinking	Eating with utensils, Eating without utensils, Drinking with lid and straw, Drinking from an open container … and so forth
Smoking	Smoking cigar/cigarette, Lighting cigar/cigarette, Extinguishing cigar/cigarette
Passenger interaction	Passenger in adjacent seat – interaction, Passenger in rear seat – interaction
Manipulating objects	Object dropped by driver, Object in vehicle, Other, and so forth
Reaching for objects	Reaching for food-related or drink-related item, Reaching for cigar/cigarette, Reaching for personal body-related item, and so forth
Vehicle integral devices	Adjusting/monitoring climate control, Adjusting/monitoring radio, Inserting/retrieving CD (or similar), and so forth
Personal hygiene	Combing/brushing/fixing hair, Applying make-up, Shaving, Brushing/flossing teeth, and so forth
Outside distractors	Looking at previous crash or incident, Distracted by construction, Looking at pedestrian, and so forth
Other secondary tasks	Other non-specific internal eye glance, Other known secondary task, Unknown type (secondary task present)
Dancing	
Reading	
Writing	
Pet in vehicle	
Cell phone, talking/listening handheld	
Cell phone, talking/listening hands-free	
Cell phone, texting	
Cell phone, dialing handheld	
Cell phone, locating/reaching/answering	
Cell phone, other	
No secondary tasks	

2 data dictionary. If no secondary task existed, the variable showed the "No Secondary Task" outcome. In this study, the secondary task activities were classified according to Stutts et al.'s study (*13*) (Table 2).

According to the NHTSA (*17*), distracted driving is responsible for 30% of all crashes. In general, distraction has many sources. Engagement in a secondary task while driving is one of the major sources of distraction. Thus, in this study, distracted driving will be defined as driver engagement in a secondary task. In most of the previous studies, the responsibility of distracted driving as a main cause of accidents was measured by descriptive statistics (for example, mean, standard deviation, chi squared test, etc.). Despite the significant correlation between distracted driving and crash likelihood, these descriptive statistical analyses cannot clearly identify the association among multiple factors in complex relationships. Therefore, this study proposes a new methodology to identify the correlation among responses that are made simultaneously using a discrete choice model.

Model Development

Bivariate Probit Model

One of our objectives was to predict the crash/near-crash likelihood given that the driver was distracted, that is, engaged in a secondary task. Since the occurrence of distraction and crash may both depend on various explanatory variables including the driver's demographic characteristics, vehicle characteristics, and roadway characteristics, the multivariate approach was chosen to link these two variables to the explanatory variables. In particular, a bivariate probit model was constructed to identify the factors affecting these two responses and also capture the correlation between them. Let y_1 be the distraction index with $y_1 = 1$ if the driver is distracted and $y_1 = 0$ otherwise, y_2 be the safety-critical index with $y_2 = 1$ if a crash/near-crash occurs and $y_2 = 0$ otherwise. The bivariate probit model with a latent variable formulation takes the following form:

$$\begin{aligned} z_1 &= \beta_1 X_1 + \epsilon_1, & y_1 &= 1 \textit{ if } z_1 \geq 0, & y_1 &= 0 \textit{ otherwise} \\ z_2 &= \alpha z_1 + \beta_2 X_2 + \epsilon_2, & y_2 &= 1 \textit{ if } z_2 \geq 0, & y_2 &= 0 \textit{ otherwise} \end{aligned} \tag{1}$$

where

z_1 is the latent variable that indicates whether the driver is distracted ($y_1 = 1$ if $z_1 \geq 0$) or not ($y_1 = 0$ if $z_1 < 0$)

X_1 is the vector of explanatory variables for the first response

z_2 is the latent variable that indicates whether the driver was involved in a SCE ($y_2 = 1$ if $z_2 \geq 0$) or not ($y_2 = 0$ if $z_2 < 0$)

X_2 is the vector of explanatory variables for the second response

β_1, α, β_2 are the parameters to be estimated
ϵ_1, ϵ_2 are two random errors that follow a normal distribution with mean 0 and variance 1.

If the two responses are interrelated, the coefficient α should be significantly different from 0. By implementing this model, the interrelationship between the distracted driving and the crash likelihood could be investigated but from a different perspective.

Distracted Driving—Crash/Near-Crash Involvement. In this model, SAS® software was employed to investigate the association between distracted driving and the SCE involvement (crash/near-crash). The two response variables are modeled as a function of a set of independent variables, as described in Equation 1. In the first equation to model distraction, the set of independent variables included driver's age, gender, marital status, working status, driver training, education, years of driving, relation to junction, and locality, as shown in Table 1. These independent variables were selected in accordance with previous studies, which examined the driver's willingness to be engaged in a secondary tasks based on different personal and traffic flow factors (*18–21*). Table 3 displays the outcomes of the constructed bivariate probit model (only significant independent variables are shown). In the first equation, it was found that drivers between the ages of 16 and 34 years are more likely to be engaged in a secondary task while driving. The results also showed that the tendency of drivers of either full-time or part-time working status to be distracted while driving is higher than that of non-working drivers. This result is logical, as full-time and part-time drivers are more involved in the driving task. Moreover, Table 3 indicates that drivers are willing to engage in a secondary task when they have passengers on board. The table also depicts that drivers prefer to be engaged in a secondary task while they approach intersections. This might suggest that there is a potential relationship between traffic density and secondary task engagement.

Unlike the first equation, the second equation to model crash/near-crash uses all the event and roadway characteristics shown in Table 1, in addition to driver age and gender as independent variables. The model showed that engagement in a secondary task is significantly correlated to the crash/near-crash likelihood. The positive coefficient implies that secondary task engagement increases the probability of crash/near-crash occurrence. Hence, there is strong statistical evidence of the impact of distracted driving on travel safety. The results also showed that parking and intersection locations are most prone to crash/near-crash occurrences. To summarize, the bivariate probit model found that distracted driving, driver's age, and intersection influence are the most significant predictors of crash/near-crash likelihood. In the next section, further analysis will quantify the increased crash/near-crash risk that results from different types of secondary tasks.

Table 3. Bivariate Probit Model Results

	Coefficient	*t*-statistics	*p*-Value (0.05*)
First Model: Engaged in a secondary task (distracted/not distracted)			
AgeGroup 16–19	0.546	6.11	<0.0001
AgeGroup 20–24	0.653	7.39	<0.0001
AgeGroup 25–29	0.563	4.83	<0.0001
AgeGroup 30–34	0.317	2.22	0.0265
FullTime	0.239	3.54	0.0063
PartTime	0.129	1.96	0.0492
Intersection_Related	0.295	2.19	0.0287
Presence of Passengers	0.419	6.17	<0.0001
Second Model: Involved in crash/near-crash event or not			
Engaged	**1.335**	**14.78**	<**0.0001**
AgeGroup 16–19	1.622	9.258	<0.0001
AgeGroup 20–24	1.17	10.529	<0.0001
AgeGroup 25–29	0.293	2.49	0.009
Parking_Related	0.673	4.23	<0.0001
Intersection	0.26	1.96	0.0492
Intersection Influence	0.629	6.35	<0.0001
Grade Up	–0.318	–2.02	0.037

*Significance level.

Secondary Tasks Risk Assessment

In this section, the increased crash/near-crash risk that results from the different secondary tasks is investigated. For this purpose, two different models were developed: a multinomial logit model, and an association analysis model. The two models attempted to quantify the increased crash/near-crash risk from two different perspectives. The multinomial logit model is a traditional statistical technique that is based on probability theory, whereas association analysis is a new powerful data-mining technique that reveals patterns in big data such as SHRP2 NDS data. The model results will be presented and the merits of each model detailed.

Multinomial Logit Model (Baseline-Category Logits Model). The multinomial logit model is a statistical technique that is employed when the response variable has more than two categories. In this model, the response variable is the event severity (normal, near-crash, or crash event), whereas the explanatory variables are the secondary tasks listed in Figure 1. The baseline-category logits model pairs each response category with a reference response category. As the SHRP2 NDS dataset provides the distribution of secondary tasks in SCEs as well as in

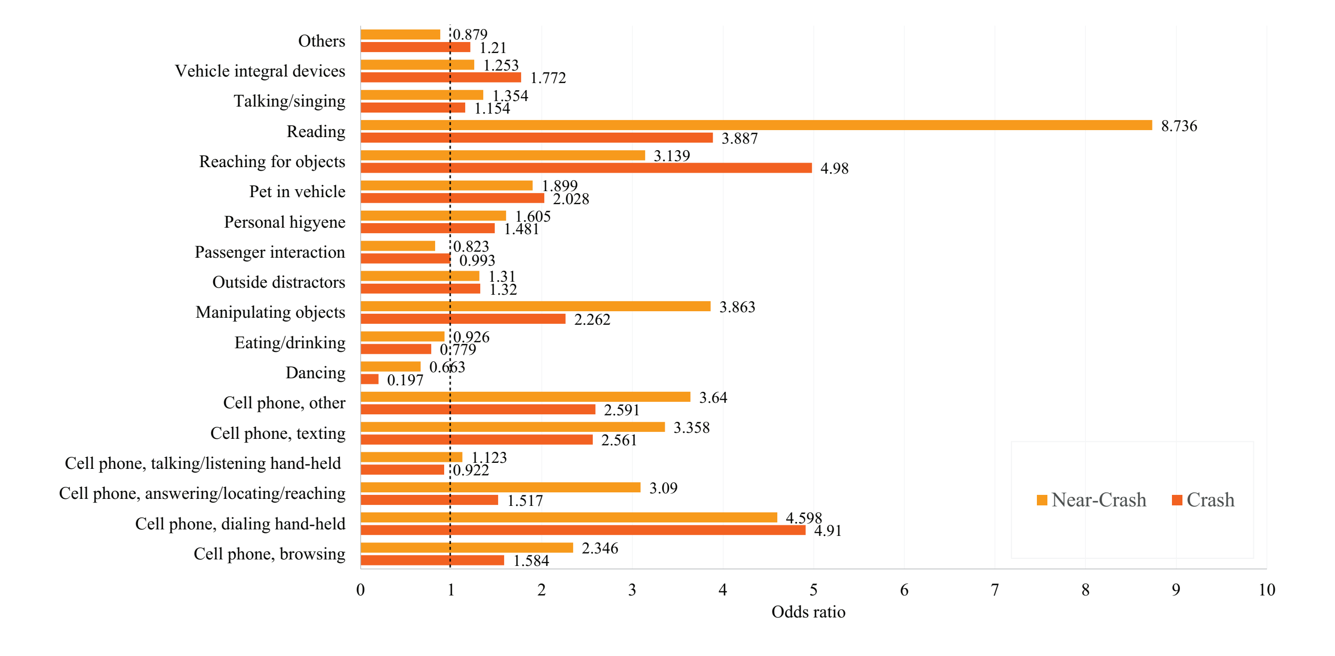

Figure 1. Odds ratios of different secondary tasks.

non-SCEs, the increased crash/near-crash risk could be recognized and quantified. When the "normal" category is the baseline, the baseline-category logits are,

$$\log\left(\frac{\pi_j}{\pi_J}\right), \text{where } j = \text{Near-crash, Crash, and } J = \text{Normal} \tag{2}$$

where π_j is the probability of the jth category. The baseline-category logits model with a set of predictor variables X (secondary tasks in our case) is defined as,

$$\log\left(\frac{\pi_j}{\pi_J}\right) = \beta_j X, \quad j = 1, \ldots, J-1 \tag{3}$$

This model has J-1 equations with separate parameters for each. The effects vary with the category paired with the reference category. If $J = 2$, this model simplifies to an ordinary logistic regression. The main advantage of the baseline-category logits model is the simultaneous fit of all the equations together. This advantage produces parameter estimates with smaller standard errors compared with fitting each equation separately using an ordinary logistic regression. In this regard, a PROC QLIM statement was recalled in SAS platform to achieve the modeling requirements.

According to previous studies (*3, 11, 12*), the OR is frequently used to estimate the relative risk of the secondary tasks while driving. ORs in a baseline-category logits model are defined as in a binary logistic model, except that they describe conditional odds. For instance, Figure 1 shows that the OR for a driver engaged in cell phone texting is 3.358 for near-crash. This means that the odds that drivers who are engaged in cell phone texting will be involved in a near-crash event rather than a normal driving event are about 3.35 times the odds for drivers who are not engaged in cell phone texting, adjusting for all the other secondary tasks. Similarly, the OR for manipulating objects is 2.262 for a crash event. Hence, we may say that the drivers who manipulate objects while driving have the odds of being involved in a crash event vs. normal event that is about 2.262 times the odds for those who are not engaged in manipulating an object, adjusting for all the other secondary tasks. Figure 1 displays the ORs of all secondary tasks in crash and near-crash events.

It should be noted that if the OR for a particular secondary task is less than 1.00 (dashed line), then the secondary task has no harmful effect on traffic safety. Accordingly, passenger interaction, eating/drinking, and dancing show a protective effect rather than a risk effect. However, Figure 1 indicates that the remaining secondary tasks are all within the risk range (OR $>$ 1). For near-crash events, reading while driving showed the highest risk with an OR of 8.739, followed by cell phone dialing handheld, and manipulating objects with ORs of 4.508 and 3.863, respectively. Cell phone texting, cell phone other, and cell phone answering/locating/reaching follow, with ORs of 3.358, 3.64, and 3.00 respectively.

Association Analysis Model (A Priori Algorithm). Data-mining techniques have been receiving increased attention from transportation researchers. These techniques have shown successful implementation in addressing safety problems compared with traditional statistical analyses (*22–27*). Detection of association rules is one of the powerful tools in data-mining techniques. It is considered the most frequent tool employed in web mining within the retail industry (also known as market basket analysis). However, this method has a wide variety of useful applications such as transportation safety. The goal of association analysis is to find rules in the form of conditions (antecedents) and results (consequents). The rules are developed based on the a priori algorithm. More details about the a priori algorithm can be found in Agrawal et al.'s study (*28*). Each developed rule is then evaluated using three performance measures: support, confidence, and lift. For instance, if an extracted rule states that "if ($Var1 = x$), → ($Var2 = y$), 30%, 80%", it means that if variable 1 is equal to x, then the probability (Prob) that variable 2 will be equal to y is 80%, and the joint event ($Var1 = x, Var2 = y$) occurs in 30% of the observations. Accordingly, Support (S), the first percentage in the rule, is defined as the probability of antecedent and consequent,

$$\text{Support(S)} = \text{Prob(antecedent and consequent)} \tag{4}$$

whereas Confidence (C), the second percentage in the rule, is the conditional probability of the consequent given the antecedent, and is denoted by,

$$\begin{aligned}\text{Confidence(C)} &= \text{Prob(consequent}|\text{antecedent)}\\ &= \frac{\text{Prob(antecedent and consequent)}}{\text{Prob(antecedent)}}\end{aligned} \tag{5}$$

Lift (L) is a performance measure that was presented in Brin et al.'s study (*29*). Lift displays the ratio of confidence for the rule to the marginal probability of having the consequent. To illustrate, suppose that 10% of the entire population buys a product X, then a rule that predicts whether people will buy product X with 20% confidence will have a lift of 20/10 = 2.00. If another rule tells you that people will buy X with 11% confidence, then the rule has a lift close to 1.00, meaning that having antecedent(s) makes little difference to the probability of having a consequent. Therefore, lift is a measure of how helpful the rule is. Rules with a lift index different from 1.00 are more interesting. Equation 6 shows the mathematical expression for the lift measure,

$$\begin{aligned}\text{Lift(L)} &= \frac{\text{Prob(consequent}|\text{antecedent)}}{\text{Prob(consequent)}}\\ &= \frac{\text{Prob(antecedent and consequent)}}{\text{Prob(antecedent)*Prob(consequent)}}\end{aligned} \tag{6}$$

To conclude, support (S) is a measure of frequency, confidence (C) is the measure of belief, and lift (L) is the measure of the improvement brought by the rule. In the marketing industry, sellers are more interested in finding rules with high support levels, high confidence indexes and lifts greater than 1.00. In transportation safety, SCEs (crashes/near-crashes) have much lower frequencies than non-SCEs. As the main objective is to find the association between the SCEs and the associated secondary tasks, support of the rules could be quite low. Therefore, a lift performance measure is used in rules evaluation.

In this research, the authors were more interested in finding rules connecting the secondary task activities (cell phone texting, eating, writing, manipulating objects, etc.) with the event severity. In essence, the study dataset was transformed into a tabular format, in which the columns represented indicator variables for the secondary task activities and the target variable was an SCE or non-SCE. Before interpreting the results, it is important to mention that the minimum support level specified for the proposed model was set at 0.1%. Regardless of the lift value, this means that no rule would have been extracted if it had a support level lower than 0.1%. This low value was selected because of the interest in extracting information related to rare events (crashes/near-crashes). Table 4 shows the rules extracted for secondary tasks and the SCE outcome. The rules are ranked based on the lift index. For better understanding the risk associated with different secondary tasks, rules should be compared with each other. For instance, Table 4 includes the following two rules:

(*Cell phone Texting = 1*), → (*Event = SCE*), 2.56%, 27.64%",
(*Vehicle-embedded devices = 1*), → (*Event = SCE*), 3.46%, 13.44%"

This means that the risk associated with cell phone texting is higher than that of operating vehicle-embedded devices. In other words, the probability of observing an SCE given that a driver is engaged in cell phone texting, (27.64%), is higher than that of vehicle-embedded devices (13.44%). Following the same criteria, all secondary tasks could be ranked based on how risky they are. The results indicated that reaching for objects, manipulating objects, reading, and other cell phone interaction activities are the riskiest secondary task activities. However, passenger interaction, eating/drinking, and dancing do not indicate a risk factor for SCE occurrence.

Discussion

This study provides quantitative insight into the risk associated with crash/near-crash events when drivers are engaged in secondary task activities. One of the most striking results in Figure 1 is the magnitude of the ORs (risk estimate). The figure shows that some activities can increase the crash or near-crash risk by four- to

Table 4. Association Analysis Results

Consequent	Antecedent	Support (S) %	Confidence (C) %	Lift (L)
EVENTSEVERITY = SCE	Reaching for objects	1.41	40.79	3.35
EVENTSEVERITY = SCE	Manipulating objects	5.78	39.77	3.27
EVENTSEVERITY = SCE	Reading	0.11	39.13	3.21
EVENTSEVERITY = SCE	Cell phone other	0.31	38.81	3.19
EVENTSEVERITY = SCE	Cell phone locating/reaching/answering	0.81	30.06	2.47
EVENTSEVERITY = SCE	Cell phone dialing handheld	0.19	30.00	2.46
EVENTSEVERITY = SCE	Cell phone texting	2.56	27.64	2.27
EVENTSEVERITY = SCE	Pet in vehicle	0.19	26.83	2.20
EVENTSEVERITY = SCE	Cell phone browsing	0.93	23.12	1.90
EVENTSEVERITY = SCE	Personal hygiene	4.01	15.45	1.27
EVENTSEVERITY = SCE	Talking/singing	7.73	13.72	1.13
EVENTSEVERITY = SCE	Vehicle-embedded devices	3.46	13.44	1.10
EVENTSEVERITY = SCE	External distractor	10.94	12.72	1.04
EVENTSEVERITY = SCE	Cell phone talking/listening handheld	3.25	12.02	0.99
EVENTSEVERITY = SCE	Eating/drinking/smoking	4.28	10.34	0.85
EVENTSEVERITY = SCE	Passenger interaction	14.94	9.22	0.76
EVENTSEVERITY = SCE	Other known secondary task	3.72	9.00	0.74
EVENTSEVERITY = SCE	Dancing	1.11	7.11	0.58

eight-fold (such as reaching for objects, reading, and cell phone dialing handheld). These activities are considered high-risk distractors as they not only require multiple steps to be completed but also longer eyes-off-road time (such as, reaching for objects and reading). Surprisingly, other secondary tasks such as passenger interaction showed unexpected impacts. Cooper et al. found that passenger interaction increases the crash risk (*30*). However, this study found that the presence of a passenger has a protective effect rather than a risk effect. This could be explained as the presence of a passenger on board being equivalent to having more eyes on the road, which could reduce the crash or near-crash probability. This result is consistent with the findings in Geyer and Regland's study (*31*).

As shown in Figure 1 and Table 4, some secondary tasks have a similar risk impact in relation to ORs or lift values. Thus, it is preferable to group these secondary tasks together. As a result, the *k*-means clustering algorithm was employed to group secondary tasks with similar risk effects together. *k*-means is a common unsupervised-learning clustering technique, which partitions *n* observations of unlabeled data into *k* clusters in which each observation belongs to the cluster with the nearest mean. Three clustering models were developed using SPSS Modeler with a predetermined number of clusters ($k = 4$). The clustering models used either the OR, obtained from the baseline-category logits model, or the lift index (L), obtained from the association model, as a clustering-based variable. The results of the clustering

Table 5. Secondary Tasks Ranking

Multinomial Logits Model		A Priori Association Model
Crash (*k* = 4)	Near-Crash (*k* = 4)	SCE (Crash/Near-Crash) (*k* = 4)
Reaching for objects^{a}	Reading^{a}	Reaching for objects^{a}
Cell phone, dialing handheld^{a}	Cell phone, dialing handheld^{b}	Manipulating objects^{a}
Reading^{a}	Manipulating objects^{b}	Reading^{a}
Cell phone, texting^{b}	Cell phone, texting^{b}	Cell phone, other^{a}
Cell phone, other^{b}	Cell phone, other^{b}	Cell phone, answering/reaching^{b}
Manipulating objects^{b}	Reaching for objects^{b}	Cell phone, dialing handheld^{b}
Pet in vehicle^{b}	Cell phone, answering/reaching^{b}	Cell phone, texting^{b}
Vehicle-embedded devices^{c}	Cell phone, browsing^{c}	Pet in vehicle^{b}
Cell phone, browsing^{c}	Pet in vehicle	Cell phone, browsing^{c}
Personal hygiene^{c}	Personal hygiene	Personal hygiene
Cell phone, answering/reaching^{c}	Talking/singing	Talking/singing
Outside distractors	Outside distractors	Vehicle-embedded devices
Talking/singing	Vehicle-embedded devices	External distractors

^{a,b,c}Indicates the *k*-clusters.

analysis are shown in Table 5. Although the highest impact secondary tasks are similar in both models, each model has its own advantages and disadvantages. To have considered more variables in baseline-category logits models would have exposed the developed model to multicollinearity. This problem can produce incorrect parameter estimates and hence, lead to incorrect results and conclusions. However, multicollinearity does not represent a problem for the association analysis model. In the association analysis model, no particular variable is defined as a response variable. Consequently, all rules that describe the association between the SCE/non-SCE event attributes can be extracted. In this study, only the one-product association rules were requested. Rules were then filtered to present only the rules connecting the secondary task activities with the event severity. To the authors' knowledge, this is the first study to employ and adjust the a priori algorithm settings in a distracted driving analysis.

It is essential to understand the impact of distracted driving in the larger context of naturalistic driving to provide useful suggestions for countermeasures. The outcomes of this study can help different sectors (automobile industry, decision makers, safety campaigns, etc.) to address distracted driving behavior. For instance, the automobile industry needs to reduce the in-vehicle features that require visual and physical interaction. This, in turn, will increase driver focus and decrease the eyes-off-road time. One of the possible recommendations is to lock out all the complex in-vehicle features while the vehicle is in motion. Moreover, as cell phone interaction activities are also one of the main internal distraction sources, it would be preferable to develop a new cell phone mode that prohibits all complex features while the vehicle is in motion (similar to airplane mode). Additionally, drivers should be aware of all the relative risks that are associated with the various secondary task activities so that they can adjust their behavior or consider alternatives. Safety campaigns that carry the message "all distractions are bad" are unrealistic and ineffective. Identifying the most serious secondary tasks can help safety campaigns to achieve their goals effectively. Finally, policymakers and legislative institutions should devise their acts (texting bans, handheld cell phone bans, etc.) based on NDS information, not on unrealistic experiments.

Conclusion

This study analyzed the increased crash and near-crash risk associated with multiple secondary tasks using a variety of statistical and data-mining models. First, a bivariate model was constructed using the SHRP2 NDS data to examine the relationship between distracted driving and SCE likelihood from different perspectives. The model indicated that distracted driving is a major contributor to an SCE occurrence. Subsequently, two different models were employed to quantify the increased risk associated with each secondary task: a baseline-category logits model, and a rule mining association model. The baseline-category logits model identified the increased risk in terms of ORs, while the a priori association algorithm detected the associated risks in terms of rules. Each rule was then evaluated based on the lift index (L). The two models succeeded in ranking all the secondary task activities according to the associated increased crash/near-crash risk efficiently. Both models revealed that reading while driving and reaching for objects are the highest crash risk among all secondary tasks. Furthermore, the *k*-means algorithm was implemented to cluster secondary tasks with similar risk impacts. Based on the results, a table was constructed to identify the *k*-means groups and the riskiest secondary tasks within each group. This study's outcomes could help drivers understand the relative risks associated with the various secondary task activities so that they can adjust their behavior or consider alternatives. The study could also help legislators initiate laws that reduce the crashes resulting specifically from distracted driving. Finally, it could help government officials make informed decisions about the allocation of available resources to reduce roadway crashes and improve traffic safety.

Author Contributions

The authors confirm contribution to the paper as follows: study conception and design: P.B., S.I.; analysis and interpretation of results: P.B., B.G; draft manuscript preparation: P.B., S.I., and B.G. All authors reviewed the results and approved the final version of the manuscript.

References

1. Young, K. L., and P. M. Salmon. Examining the Relationship between Driver Distraction and Driving Errors: A Discussion of Theory, Studies and Methods. *Safety Science*, Vol. 50, No. 2, 2012, pp. 165–174.
2. Caird, J. K., C. R. Willness, P. Steel, and C. Scialfa. A Meta-Analysis of the Effects of Cell Phones on Driver Performance. *Accident Analysis & Prevention*, Vol. 40, No. 4, 2008, pp. 1282–1293.
3. Regan, M. A., J. D. Lee, and K. Young. *Driver Distraction: Theory, Effects, and Mitigation*. Boca Raton, Fla., CRC Press, 2008.
4. Lansdown, T. C. Frequency and Severity of In-Vehicle Distractions: A Self-Report Survey. *Proc., 1st International Conference on Driver Distraction and Inattention (DDI 2009)*, 2009.
5. McEvoy, S. P., M. R. Stevenson, and M. Woodward. The Impact of Driver Distraction on Road Safety: Results

from a Representative Survey in Two Australian States. *Injury Prevention*, Vol. 12, No. 4, 2006, pp. 242–247.

6. Royal, D. *National Survey of Speeding and Unsafe Driving Attitudes and Behaviors: 2002*. Volume II-Findings Report No: HS-809 688. National Highway Traffic Safety Administration, Washington, D.C., 2003.
7. Sullman, M. J., and P. H. Baas. Mobile Phone Use amongst New Zealand Drivers. *Transportation Research Part F: Traffic Psychology and Behaviour*, Vol. 7, No. 2, 2004, pp. 95–105.
8. Young, K. L., and M. G. Lenné. Driver Engagement in Distracting Activities and the Strategies Used to Minimise Risk. *Safety Science*, Vol. 48, No. 3, 2010, pp. 326–332.
9. af Wåhlberg, A., L. Dorn, and T. Kline. The Manchester Driver Behaviour Questionnaire as a Predictor of Road Traffic Accidents. *Theoretical Issues in Ergonomics Science*, Vol. 12, No. 1, 2011, pp. 66–86.
10. Johnson, M. B., R. B. Voas, J. H. Lacey, A. S. McKnight, and J. E. Lange. Living Dangerously: Driver Distraction at High Speed. *Traffic Injury Prevention*, Vol. 5, No. 1, 2004, pp. 1–7.
11. Klauer, S. G., T. A. Dingus, V. L. Neale, J. D. Sudweeks, and D. J. Ramsey. *The Impact of Driver Inattention on Near-Crash/Crash Risk: An Analysis Using the 100-Car Naturalistic Driving Study Data*. Report No. DOT HS 810 594. National Highway Traffic Safety Administration, Washington, D.C., 2006.
12. Sayer, J. R., J. M. Devonshire, and C. A. Flannagan. *The Effects of Secondary Tasks on Naturalistic Driving Performance*. Report No. UMTRI-2005-29. The University of Michigan Transportation Research Institute, Ann Arbor, 2005.
13. Stutts, J., J. Feaganes, D. Reinfurt, E. Rodgman, C. Hamlett, K. Gish, and L. Staplin. Driver's Exposure to Distractions in their Natural Driving Environment. *Accident Analysis & Prevention*, Vol. 37, No. 6, 2005, pp. 1093–1101.
14. Olson, R., R. Hanowski, J. Hickman, and J. Bocanegra. Driver Distraction in Commercial Operations Vehicle. *Federal Motor Carrier Safety Administration, DC, FMCSA-RRR-09*, Vol. 42, 2009.
15. Transportation Research Board of the National Academies; Virginia Tech Transportation Institute. The 2nd Strategic Highway Research Program Naturalistic Driving Study InSight Dataset [Data set]. VTTI Dataverse, 2016. https://doi.org/10.15787/VTT1/3YVSY4.
16. Dingus, T. A., F. Guo, S. Lee, J. F. Antin, M. Perez, M. Buchanan-King, and J. Hankey. Driver Crash Risk Factors and Prevalence Evaluation Using Naturalistic Driving Data. *Proceedings of the National Academy of Sciences*, Vol. 113, No. 10, 2016, pp. 2636–2641.
17. National Highway Traffic Safety Adminstration. *Distracted Driving 2014*. Report No. DOT HS 812 260. National Highway Traffic Safety Adminstration, Washington, D.C., 2016.
18. Goldberg, L. R., J. A. Johnson, H. W. Eber, R. Hogan, M. C. Ashton, C. R. Cloninger, and H. G. Gough. The International Personality Item Pool and the Future of Public-Domain Personality Measures. *Journal of Research in Personality*, Vol. 40, No. 1, 2006, pp. 84–96.
19. Lane, W., and C. Manner. The Impact of Personality Traits on Smartphone Ownership and Use. *International Journal of Business and Social Science*, Vol. 2, No. 17, 2011 pp. 22–28.
20. Lansdown, T. C. Individual Differences and Propensity to Engage with In-Vehicle Distractions–A Self-Report Survey. *Transportation Research Part F: Traffic Psychology and Behaviour*, Vol. 15, No. 1, 2012, pp. 1–8.
21. Titchener, K., and I. Y. Wong. Driver Distractions: Characteristics Underlying Drivers' Risk Perceptions. *Journal of Risk Research*, Vol. 13, No. 6, 2010, pp. 771–780.
22. Abdel-Aty, M., and J. Keller. Exploring the Overall and Specific Crash Severity Levels at Signalized Intersections. *Accident Analysis & Prevention*, Vol. 37, No. 3, 2005, pp. 417–425.
23. Abdelwahab, H., and M. Abdel-Aty. Development of Artificial Neural Network Models to Predict Driver Injury Severity in Traffic Accidents at Signalized Intersections. *Transportation Research Record: Journal of the Transportation Research Board*, 2001. 1746: 6–13.
24. Bayam, E., J. Liebowitz, and W. Agresti. Older Drivers and Accidents: A Meta Analysis and Data Mining Application on Traffic Accident Data. *Expert Systems with Applications*, Vol. 29, No. 3, 2005, pp. 598–629.
25. Chang, L.-Y. Analysis of Freeway Accident Frequencies: Negative Binomial Regression Versus Artificial Neural Network. *Safety Science*, Vol. 43, No. 8, 2005, pp. 541–557.
26. Chang, L.-Y., and W.-C. Chen. Data Mining of Tree-Based Models to Analyze Freeway Accident Frequency. *Journal of Safety Research*, Vol. 36, No. 4, 2005, pp. 365–375.
27. Golob, T. F., and W. W. Recker. A Method for Relating Type of Crash to Traffic Flow Characteristics on Urban Freeways. *Transportation Research Part A: Policy and Practice*, Vol. 38, No. 1, 2004, pp. 53–80.
28. Agrawal, R., T. Imieliński, and A. Swami. Mining Association Rules between Sets of Items in Large Databases. *ACM SIGMOD Record*, No. 22, 1993, pp. 207–216.
29. Brin, S., R. Motwani, J. D. Ullman, and S. Tsur. Dynamic Itemset Counting and Implication Rules for Market Basket Data. *ACM SIGMOD Record*, Vol. 26, 1997. pp. 255–264.
30. Cooper, D., F. Atkins, and D. Gillen. Measuring the Impact of Passenger Restrictions on New Teenage Drivers. *Accident Analysis & Prevention*, Vol. 37, No. 1, 2005, pp. 19–23.
31. Geyer, J., and D. Ragland. Vehicle Occupancy and Crash Risk. *Transportation Research Record: Journal of the Transportation Research Board*, 2005. 1908: 187–194.

The Standing Committee on Safety Data, Analysis and Evaluation (ANB20) peer-reviewed this paper (18-01562).