Using Adjuster Notes & Text Mining for Auto Insurance Predictive Analytics

presented by:
Philip S. Borba, Ph.D.
Milliman, Inc.
New York, NY

March 31, 2014

Casualty Actuarial Society, Ratemaking & Product Management Seminar, Washington, DC

Milliman

Casualty Actuarial Society -- Antitrust Notice

The Casualty Actuarial Society is committed to adhering strictly to the letter and spirit of the antitrust laws. Seminars conducted under the auspices of the CAS are designed solely to provide a forum for the expression of various points of view on topics described in the programs or agendas for such meetings.

Under no circumstances shall CAS seminars be used as a means for competing companies or firms to reach any understanding – expressed or implied – that restricts competition or in any way impairs the ability of members to exercise independent business judgment regarding matters affecting competition.

It is the responsibility of all seminar participants to be aware of antitrust regulations, to prevent any written or verbal discussions that appear to violate these laws, and to adhere in every respect to the CAS antitrust compliance policy.

Overview

- Starting Considerations and Definitions

- Reasons to be Interested in Text Data

- National Motor Vehicle Crash Causation Survey

- NMVCCS Definition of “Distracted Driving”

- Accident Descriptions:
  - 3 examples where cell phone use mentioned
  - NMVCCS Accident Descriptions compared to Claim Adjuster Notes
  - Breaking Text Data into Manageable Units – Creating NGrams

- Flags for Cell Phone Use Created from Text Data

- Cell Phone Use: Structured Data v. Text Data

- Multivariate (Logit) Analyses

Starting Considerations

- Key Facts and Statistics (NHTSA and CTIA)
  - In 2012, 3,328 people killed in motor vehicle accidents involving a distracted driver (2011: 3,360 people killed).
  - In 2012, 421,000 persons injured in distracted-driver accidents (2011: 387,000 persons injured).
  - At any given daylight moment in the US, approximately 660,000 drivers are using cell phones or manipulating electronic devices while driving (a number that has held steady since 2010).
  - 4.6 seconds = the length of a football field: The amount of time sending or receiving a text takes a driver’s eyes from the road and the distance covered at 55 miles per hour.
  - 20% of teens and 10% of parents admit they have extended, multi-message text conversations while driving.
  - Use of hand-held phones and other portable devices increases the risk of getting into an accident by three times.

Definitions

- NHTSA – National Highway Traffic Safety Administration
  - Federal agency established in 1970 to carry out safety programs.

- NMVCCS – National Motor Vehicle Crash Causation Survey
  - Research-designed survey by NHTSA collecting information on accidents between July 3, 2005 and December 31, 2007.
  - On-scene and post-accident data collection.

- Structured data
  - Data reported in numeric or categorical form.
  - Numeric data includes dollar amounts, age, number of vehicles in an accident.
  - Categorical data includes assignment of other types of information to a specific character or number (such as a “rear-end crash” assigned to “22” or “weather-snow” to “2”, in fields for accident type or weather condition).

- Text data
  - Data provided in text form, such as a claim adjuster note, accident description, deposition, or other reports. Books, magazine articles, and research reports are other examples of text data.

Reasons to be Interested in Text Data

- General considerations

- Reasons specific to cell phone use

- State laws on cell phone use and texting while driving

Reasons to be Interested in Text Data

- Capture information not in structured data
  - New concepts not on claim-reporting forms
  - Many structured data-reporting forms do not capture cell phone use
  - Drivers / occupants may be averse to reporting cell phone use at time of the accident

- Capture information that has not reached structured data (such as payments for medical treatments or claimant attorneys) (information available in claim adjuster notes or case manager notes)
  - Medical treatments: “needs surgery”, “surgery scheduled”, “referred for MRI”
  - Attorney involvement: “call from claimant’s attorney”, “notice from attorney”

- Claim stratification
  - Can identify claims with “dialing on cell phone,” “talking on cell phone”, etc.
  - Can identify claims with “on medication,” “taking prescription”, etc. (other demonstrations)

- Predictive Analytics
  - Does information from text data improve the predictability of target outcomes?

Reasons to be Interested in Text Data re Cell Phone Use

- Newly developed area for factors that may be associated with accidents. Claim data-capture forms do not have a standardized coding scheme for cell phone use.

- Difficult to accurately capture at the time of the accident (drivers averse to reporting cell phone use – often obtained from post-accident investigations).

- Subtle distinctions may be important.
  - hand-held v. hands-free
  - If hands-free, position of controls (built-in or after market)
  - use of speaker phone
  - driver or occupant using phone

- State laws are different re cell phone use and texting while driving.

State Laws on Cell Phone Use and Texting While Driving

- 12 states and D.C. prohibit all drivers from using handheld cell phones.
- 42 states and D.C. prohibit all drivers from text messaging.
- Table presents laws for DC and the mid-Atlantic states.

State	Hand-Held Ban	Cell Phone Ban for School Bus and Novice Drivers (Primary)	Texting Ban
Delaware	Yes (primary)	School bus drivers, Learner or intermediate license	Yes (primary)
D.C.	Yes (primary)	School bus drivers, Learners permit	Yes (primary)
Maryland	Yes (primary)	<18 w/ Learner or provisional license (secondary)	Yes (primary)
New Jersey	Yes (primary)	School bus drivers, Permit or provisional license	Yes (primary)
New York	Yes (primary)	----------	Yes (primary)
North Carolina	----------	School bus drivers, <18 (primary)	Yes (primary)
Pennsylvania	----------	----------	Yes (primary)
Virginia	----------	School bus drivers, <18 (secondary)	Yes (primary)

Limitations

- Results in this presentation are for demonstration purposes only.
- Data are from public sources and have been reviewed for consistency but have not been audited.
- The analyses and statistical results are intended to demonstrate the principles of text-mining and predictive analytics. Presented methodologies and results may not be appropriate for all applications in the property-casualty insurance industry. Users are strongly advised to review the underlying methodology and data sources when performing a text-mining extraction or predictive analytics.

National Motor Vehicle Crash Causation Survey

- Data collection process
- Data files
- Case weights
- Files of special interest to this presentation
- NMVCCS summary statistics

National Motor Vehicle Crash Causation Survey

- National Motor Vehicle Crash Causation Survey (NMVCCS)
  - Conducted by the National Highway Traffic Safety Administration (NHTSA).
  - Sample of accidents investigated between July 3, 2005 and December 31, 2007.
  - Primary focus of Survey: Determine the critical pre-accidents events and reasons underlying the critical factors at the accident.
  - Looked into factors related to drivers, vehicles, roadways, and the environment.
  - Considerable attention to behavioral considerations and factors.

- Data collection process
  - On-site data collection by NMVCCS researchers.
  - Accidents occurring between 6am and midnight.
  - Accident must have resulted in a harmful event.
  - EMS must have been dispatched.
  - Police present when NMVCCS researcher arrived.
  - At least one of the first 3 vehicles involved must be present at the accident scene.
  - Completed police report.

National Motor Vehicle Crash Causation Survey

- Data files
  - 22 files
  - Occupant: demographics on driver and occupants
  - Case Vehicle: police reported alcohol or drug presence
  - Pre-Crash Assessment: identified "internal distraction", read-end collision, weather conditions, driver on medication, driver fatigue
  - Crash Description: narrative description of accident
  - Contents are static (not updated)

- Case weights
  - To make the sample representative of all similar types of accidents in the US.
  - Case weights not used in present analyses.
  - Present analyses are from the prospective of an insurer's book of business, rather than a research or policy analysis.

National Motor Vehicle Crash Causation Survey

- Structured data
  - Date and time of accident
  - Type of accident (e.g., rear end)
  - Police report indicated whether there were injuries
  - Vehicle equipment: presence of a cell phone
  - PCA: whether the driver was engaged in a conversion, weather conditions
  - Drivers: use of medications, drugs, driver fatigue

- Text data
  - Crash Description file
  - One record per crash
  - 8,000 bytes
  - Vehicles are identified in various references: V1, Vehicle 1, Vehicle #1, Vehicle One
  - References not always consistent within the same crash description

NMVCCS Sample -- Summary Characteristics

- 6,949 accidents
  - 74% involved multiple vehicles
  - 73% of the police reports reported an injury or possibility of an injury
  - 18% were rear-end accidents
  - 24% occurred where weather may be been a contributing factor
  - 22% occurred on a weekend
  - 47% involved at least one driver on meds
  - 13% involved at least one driver reported to be fatigued
  - 2% involved at least one driver reported to be using drugs
  - 6% involved at least one driver possibly under the influence of alcohol
  - 3% involved at least one driver talking on a cell phone

NMVCCS Definition for “Distracted Driving”

- Present definition limited to internal sources of distraction and non-driving cognitive activities

- Internal sources (examples)
  - Dialing/hanging up phone
  - Adjusting radio/CD player
  - Conversing with passenger
  - Driver talking on phone
  - Text messaging

- Non-driving cognitive activities
  - Inattentive, though focus unknown
  - Financial problems
  - Family or personal problems

- Distractions captured in categorical fields (structured data)

NMVCCS Accident Descriptions

- Characteristics of NMVCCS accident descriptions

- Three examples that include “cell phone in use”

- Breaking text data into manageable units

NMVCCS Accident Descriptions

- One record for each accident. Maximum length = 7,800 bytes.

- Three examples in the following slides.
  - Examples are typical of the NMVCCS accident descriptions.
  - Selected to demonstrate different ways the same concept may be expressed.

- In claim adjuster notes, much greater variations in expressions (less consistency among adjusters for same insurer, differences in style across insurers)

Accident Description #1

Accident #1: This crash took place during the early afternoon of a holiday on a four lane divided roadway. There were two eastbound lanes and two westbound lanes divided by a median. Conditions were daylight and dry and the roadway had a posted speed limit of 30mph (48kmph).

V1, a 1992 Honda Accord, was traveling west in lane one negotiating a curve right. Just after passing the apex of the curve this vehicle lost control and departed the roadway to the right. V1 struck the curb, then struck an overhead light pole before re entering the roadway and coming to rest in its original travel lane.

V1 was driven by a 17 year-old male who stated that his mother had left the house and left her keys to the car at home. He took the car without her permission and was going to his friends house. The driver stated that as well as being fun, he was driving too fast to get back home before his mother. **Just prior to the crash the driver was on his hand held cell** phone telling his friend that he was almost there. This driver was operating the vehicle with a drivers permit which had a restriction demanding proper supervision.

(236 words, 1,281 bytes)

Accident Description #2

Accident #2: The crash occurred on an east / west urban interstate in the eastbound lanes. …. The roadway was straight and level with paved shoulders on either side. The crash occurred at mid-afternoon on a weekend under daylight and dry conditions. The posted speed limit was 55 MPH.

Vehicle 1, a 1997 Honda Civic, was traveling in the second eastbound lane when it crossed the dashed line to its right and impacted the left rear side of Vehicle 2, a 2003 Ford Mustang. After impact, Vehicle 1 crossed the right fog line and paved shoulder and went off the right side of the roadway …..

Vehicle 2 went into a counter-clockwise spin and crossed the left two lanes of traffic, onto the left shoulder and impacted a guardrail with the its right rear corner, coming to rest about 120 meters east of POI facing southwest. Both vehicles were towed due to damage.

Vehicle 1 was driven by a 35-year old male who was the beneficiary of deployed frontal air bags while wearing his lap and shoulder belt. He was uninjured in the crash. **The driver of Vehicle 1 was charged by police with DUI.** The driver had 2 different narcotics in his system at the time of the crash and also admitted to using marijuana that day.

Fatigue was coded since the driver had slept only 2 ½ hours the morning of the crash and that was 10 hours pre-crash. The driver stated he was in a hurry to get home and **had been on the phone just before the crash**. He then dropped his phone on the floor, went to look for it and that was when his car departed his lane to the right.

Vehicle 2 was driven by a 20-year old female who was belted and uninjured in the crash. Her airbag was not deployed.   (471 words, 2,603 bytes)

Accident Description #3

Accident #3: The crash occurred in the intersection of two roadways. …. Both roadways were five-lane, two-way, with a posted speed 35 mph. It was early afternoon on a weekday and the road was dry and the sky was clear. Traffic was flowing.

V1, a 2004 Chevrolet Trailblazer four door with one occupant was traveling eastbound in lane two. V2 a 1994 Chevrolet G-series van with two occupants was traveling southbound in lane one. The **driver of V1** stated that he looked at the light and it was green. He started **dialing his cell phone** and when he looked back up the light had turned red. He stated that he did not have time to stop. The **driver of V2 stated that he was talking on the phone** when V1 entered the intersection. He stated that he did not see V1 until impact. The front of V2 contacted the left of V1 both vehicles then rotated and the right of V2 contacted the left of V1 before they both came to final rest in the roadway.

The driver of V1 …. was **getting ready to call his wife on his cell phone**. The light was green so he looked for her number on his phone. He was going to go straight through the intersection. He looked back up at the light as he was going through and he saw the light was red. It was too late, he was already in the intersection. There was nothing he could do. He stated that he was traveling between 31-40 mph when he struck V2.

The Critical Reason for the Critical Pre-crash Event was a driver related factor: “internal distraction”, because he did not see the light turn red because **he was dialing his cell phone**. Associated factors for the driver of V1 was that the driver of V1 was fatigued, he had only had four hours of sleep, and he had taken medication prior to the crash.

The driver of V2 was a 25-year old male who reported injuries and was transported to a local trauma facility. He advised that he had just left his home and was on his way to the hospital. He was **talking on his cell phone as he was driving** down the street. He advised that he had been traveling between 31-40 mph prior to being struck by V1. He stated that he did not see V1 prior to impact and therefore had no time to attempt any avoidance actions.

…… Associated factors for the driver of V2 was that he failed to look far enough ahead and that **he was talking on his cell phone** at the time of the crash. Another factor is that the driver rarely drove that roadway. (585 words, 3,060 bytes)

NMVCCS Accident Descriptions

- From the three examples, there are notable differences.

- References to “vehicle”:
  - V1, V2 (#1, #3)
  - Vehicle 1, Vehicle 2 (#2)
  - Other accident descriptions: insert “#” before the number (e.g., V#1), spell numeric (e.g., Vehicle One)
  - Reference not always consistent within the same accident description. (Significant problem with claim adjuster notes.)

- References to cell phone with common “cell phone use” implication:
  - driver was on his cell phone (#1)
  - had been on the phone (#2)
  - dialing his cell phone (#3)
  - talking on this cell phone (#3)
  - With claim adjuster notes, would need to be careful about “cell phone” and “on the phone” referring to adjuster trying to contact claimant or other party (eg, attorney, medical provider)

Summary Characteristics of Accident Descriptions

- 6,949 cases (accidents)
  - 438 : average number of words in accident descriptions
  - 330 / 514: first and third quartiles for words in accident descriptions
  - 2,436: average number of bytes in accident descriptions
- Similar numbers for cases with weights

	All Cases	With Case Weights
Number of accidents	6,949	5,470
Number of words in accident descriptions		
Average number of words	438	444
Median number of words	411	416
Q1 / Q3 number of words	330 / 514	336 / 520
Maximum number of words	1,294	1,294
Number of bytes in accident descriptions		
Average number of bytes	2,436	2,471
Median number of bytes	2,300	2,324
Q1 / Q3 number of bytes	1,843 / 2,869	1,874 / 2,911
Maximum number of bytes	7,800	7,800

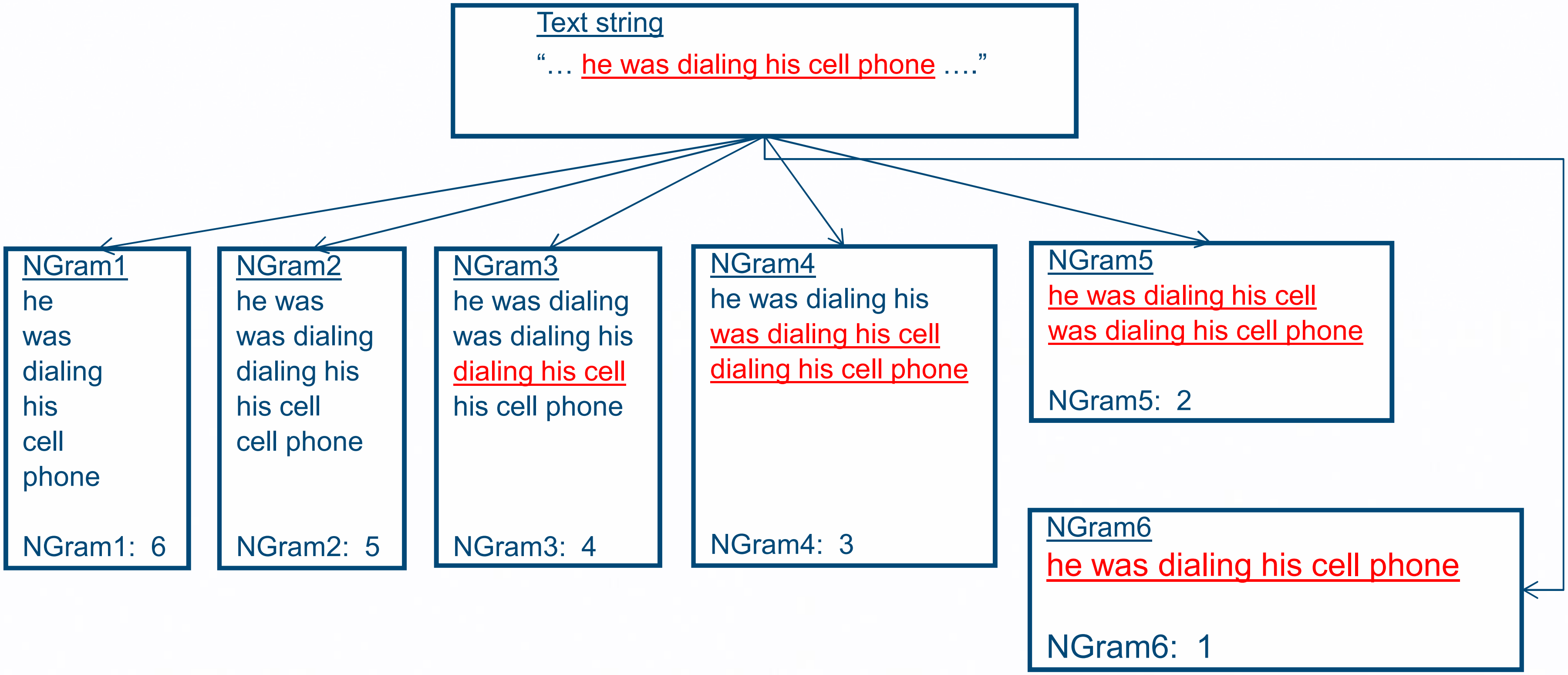
NMVCCS Accident Descriptions compared to Claim Adjuster Notes

- NMVCCS accident descriptions are “cleaner” than the typical claim adjuster notes.

- Claim adjuster notes provide a longer and wider scope of information:
  - Typically span more than one record.
  - Include considerable amount of ancillary information (eg, phone numbers, addresses).
  - Provide claim activity, often with dates (open, closed).
  - Provide insurer-liability information (e.g., subrogation).
  - Provide post-accident information.
  - Provide medical care (e.g., “scheduled surgery”).
  - Provide attorney involvement (e.g., “received call from claimant’s attorney”).
  - May identify co-morbidities.

- Insurer text data can go beyond claim adjuster notes (e.g., medical case manager notes, underwriting notes, depositions, statements, e-mails, memorandums).

Breaking Text Data into Manageable Units – Creating "NGrams"



Text string

"… he was dialing his cell phone …."

NGram1

he
was
dialing
his
cell
phone

NGram1: 6

NGram2

he was
was dialing
dialing his
his cell
cell phone

NGram2: 5

NGram3

he was dialing
was dialing his
dialing his cell
his cell phone

NGram3: 4

NGram4

he was dialing his
was dialing his cell
dialing his cell phone

NGram4: 3

NGram5

he was dialing his cell
was dialing his cell phone

NGram5: 2

NGram6

he was dialing his cell phone

NGram6: 1

- 1 six-word phrase produced 21 NGrams.

NGrams of Interest from NMVCCS Accident Descriptions

- Accident Description #1:
  - driver was on his hand held
  - on his hand held cell
- Accident Description #2:
  - had been on the phone
- Accident Description #3:
  - he was dialing his cell phone
  - dialing his cell phone
  - talking on the phone
  - on this cell phone
  - talking on his cell phone

NGrams Created from NMVCCS Accident Descriptions

- Each accident description was parsed into NGram1-NGram6.

- Process removes certain NGram1-NGram3 not expected to be needed in any claim segmentation or analytics.

- For each accident description, unique NGrams are retained. (Repeats can produce misleading emphasis on a particular NGram. Same concept can be expressed with different words.)

	All Cases
Number of accidents	6,949
Size of NGram	
NGram1	607,260
NGram2	1,998,412
NGram3	2,578,495
NGram4	2,689,556
NGram5	2,725,082
NGram6	2,737,144
Total	**13,335,949**

Flags for Cell Phone Use Created from Text Data

- Flags for the theme “cell phone in use”
- Structured data v. text data
- Flags for theme “adjusting radio/cd”

Flags for Cell Phone Use Created from Text Data

- “Conversing With” captures text that includes:
  - conversing with
  - conversing on
  - conversation with
  - conversation on
  - All of the above replacing “conversing” with “talking”

- “Cell Phone Conversing” captures text that includes:
  - cell, cellular, hand-held, handsfree, hands free, mobile, phone
  - on his cell (or phone, hand held, handheld, etc.)
  - on a cell, use of a, holding a, ending a, using a, …..

- “Cell Phone Other” captures text that includes:
  - cell, but excludes if there are references to anemia, disease, sickle, or “cell” is part of excellent, cancellation, et. al.

Use of Cell Phone: Structured Data v. Text Data

- Table at right presents information for "cell phone in use".

- Structured data: NMVCCS
  - 196 claims with cell phone in use (2.8%)

- Text data: accident descriptions
  - 264 crashes with cell phone in use (4.0%)

- Overlap between structured data and text data: 171 crashes

Number of Accidents		Text Data		
		Not in Use	In Use	Total
Structured Data	Not in Use	6,660	93	6,753
	In Use	25	171	196
	Total	6,685	264	6,949

Row Percents	Text Data	
Structured Data	Not in Use	In Use
Not in Use	98.6%	1.4%
In Use	12.8%	87.2%

Column Percents	Text Data	
Structured Data	Not in Use	In Use
Not in Use	99.6%	35.2%
In Use	0.4%	64.8%

Flags for Adjusting Radio/CD

- 0/1 flags for “adjusting radio/CD” were created using the same process

- “Adjusting Radio/CD” captures text that includes
  - adjusted / adjusting the radio
  - reached / reaching for the radio
  - turn down / turning down / turndown the radio
  - All of the above replacing “the” with “his” and “her”
  - All of the above replacing “radio” with “CD”

Incidence of Cell Phone Use and Adjusting Radio/CD

- 6,949 accidents
  - 74% were multi-veh
  - 18% were rear-end
- Cell phone
  - 83% were multi-veh
  - 24% were rear-end
- Adjusting radio/CD
  - 65% were multi-veh
  - 32% were rear-end

Incidence of Cell Phone Use in Accidents	All Accidents	Multi-Vehicle Accidents	Rear-End Accidents
All Accidents	100.0%	**73.8%**	**18.0%**
Conversing on Cell Phone - No	96.2%	73.4%	17.8%
Conversing on Cell Phone - Yes	3.8%	**83.3%**	**23.9%**

Incidence of Adjusting Radio/CD in Accidents	All Accidents	Multi-Vehicle Accidents	Rear-End Accidents
All Accidents	100.0%	**73.8%**	**18.0%**
Adjusting Radio/CD – No	99.1%	73.9%	17.9%
Adjusting Radio/CD - Yes	0.9%	**64.5%**	**32.3%**

Multivariate (Logit) Analyses

- Proof of Concept
  - Does the inclusion of text data improve the results from predictive analytics?

- Modeling considerations
  - Two outcome measures
  - Explanatory variables
    - Environmental controls
    - Driver conditions
    - Adjusting radio/CD
    - Cell phone in use
  - Logit regressions
  - Estimated probabilities using results from logit regressions

Multivariate (Logit) Analyses

- Two outcome measures
  - Multiple vehicles in accident (0/1)
    - Are accidents where a cell phone was in use more likely to be multiple-vehicle accidents? (Distraction associated with cell phone use may be more difficult to manage with other moving objects in the vicinity.)
    - Are accidents where a driver was adjusting radio/CD more likely to be multiple-vehicle accidents?
  - Rear-end collision (0/1)
    - Does a cell phone in use influence the type of accident (e.g., a rear-end accident)?
    - Does adjusting radio/CD influence the type of accident (e.g., a rear-end accident)?

Multivariate (Logit) Analyses

- Explanatory variables
  - Environmental Controls
    - Night: accident occurred before 7am or after 6pm.
    - Weekend: accident occurred on a Saturday or Sunday
    - Weather: on or more adverse conditions (eg., snow, rain, ice)
  - Driver Conditions
    - Driver fatigue: at least one driver in accident was reported to be fatigued
    - Medications: one or more drivers reported taking drugs/medications within 24 hours preceding the accident
    - Drugs: police report recorded illegal drug(s) in driver’s system
    - Alcohol: police report recorded presence of alcohol with the driver

Multivariate (Logit) Analyses

- Explanatory variables (continued)
  - Adjusting radio/CD (developed from text data)
  - Three 0/1 indicators for cell phone in use
    - Text data: conversing on cell phone (0/1 developed from NGrams)
    - Structured data: conversing on cell phone (reported in NMVCCS Pre-Crash Assessment file)
    - Structured data: any cell phone use (reported in NMVCCS Pre-Crash Assessment file)

Probability the Accident Involved Multiple Vehicles

- Outcome Measure: multiple vehicles in accident
  - Are accidents where a cell phone was in use more likely to involve multiple vehicles?
  - Are accidents where a driver was adjusting radio/CD more likely to involve multiple vehicles?

- Principal Findings:
  - Use of cell phone is associated with an increased likelihood of being in a multi-vehicle accident.
  - Coefficients statistically significant and consistent across the different cell-phone-use variables.
  - The distraction caused by cell phone use may impair a driver's ability to avoid an accident.
  - Adjusting radio/CD not statistically significant for association with a multiple-vehicle accident.

	Accident Descriptions (text)	Structured Field	Structured Field
	On Cell Phone	Conversing on Cell Phone	Cell Phone in Use
Intercept	1.236*	1.239*	1.238*
NIGHT	-0.440*	-0.440*	-0.439*
WEEKEND	-0.419*	-0.416*	-0.415*
WEATHER	-0.520*	-0.520*	-0.519*
DRIVER FATIGUE	-0.577*	-0.575*	-0.575*
MEDICATIONS	0.590*	0.590*	0.589*
DRUGS	-0.563*	-0.560*	-0.559*
ALCOHOL	-0.539*	-0.536*	-0.541*
ADJUSTING RADIO/CD	**-0.423**	**-0.418**	**-0.413**
CELL PHONE	**0.615***	**0.649***	**0.567***
-2 log Likelihood	7,598	7,603	7,602

Probability the Accident Involved Multiple Vehicles

- Table presents probabilities for cell-phone-use being associated with a multiple-vehicle accident.
  - Probabilities are for “reference group” (daytime, weekday, good weather, etc.)
  - Probabilities are for reference group + cell phone use
  - Probabilities for cell phone use from text data (accident descriptions) and structured data

- After controlling for other factors, probabilities for cell phone use are approximately 18 percentage points higher. (Statistically significant at 5% level.)

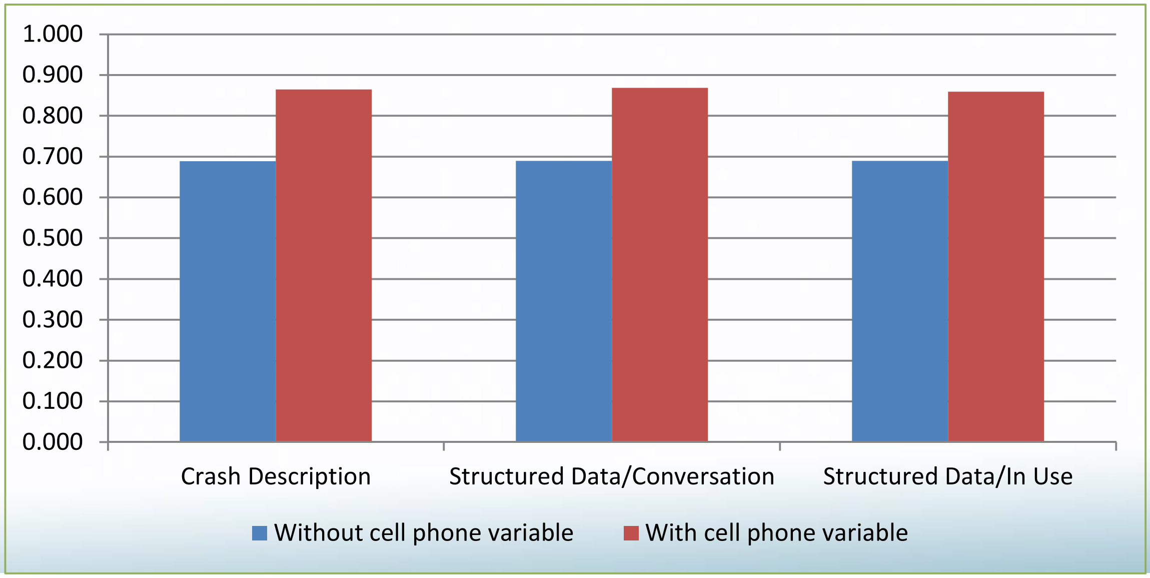
Probability Accident was a Multi-Vehicle Accident	Accident Desc -- On Cell Phone	Structured --- Conv on Cell Phone	Structured -- Cell Phone in Use
Reference group	0.689	0.690	0.690
Daytime			
Weekday			
Good weather			
Driver not fatigued			
No medications			
No drugs			
No alcohol			
Not adjusting radio			
No cell phone use			
Cell Phone Use	0.864	0.869	0.859

Probability the Accident Involved Multiple Vehicles

- Graph presents the probability the accident involved multiple vehicles.
  - Left-hand (blue) bars: no cell phone variable in the model.
  - Right-hand (red) bars: cell phone variable in the model.

- Three sets of probabilities:
  - Cell phone variable from text data (accident descriptions)
  - Cell phone variable from structure data (conversation)
  - Cell phone variable from structure data (in use)

- Implications:
  - **Including cell phone variable increased the probability of predicting of a multiple-vehicle accident**.
  - **Cell phone variable from text data produced results similar to variables from structured data.**



Probability the Accident was a Rear-End Collision

- Outcome Measure: Rear-end collision (0/1)
  - Does a cell phone in use influence the type of accident (e.g., a rear-end accident)?

- Principal Findings
  - Use of cell phone is associated with an increased likelihood of being in a multi-vehicle accident.
  - Coefficients statistically significant and consistent across the different cell-phone-use variables.
  - The distraction caused by cell phone use may impair a driver's ability to avoid an accident.

	Accident Descriptions (text)	**Structured Field**	**Structured Field**
Variable	**On Cell Phone**	**Conversing on Cell Phone**	**Cell Phone in Use**
Intercept	-1.391 *	-1.389 *	-1.391 *
NIGHT	-0.409 *	-0.409 *	-0.408 *
WEEKEND	-0.375 *	-0.374 *	-0.373 *
WEATHER	-0.329 *	-0.330 *	-0.329 *
DRIVER FATIGUE	0.008	0.010	0.010
MEDICATIONS	0.186 *	0.187 *	0.185 *
DRUGS	-0.688 *	-0.685 *	-0.685 *
ALCOHOL	-0.114	-0.112	-0.117
ADJUSTING RADIO/CD	**0.767 ***	**0.769 ***	**0.771 ***
CELL PHONE	**0.341 ***	**0.358 ***	**0.361 ***
-2 log Likelihood	6,447	6,448	6,447

Probability the Accident was a Rear-End Collision

- Table presents probabilities for cell-phone-use being associated with a rear-end collision.
  - Probabilities are for “reference group” (daytime, weekday, good weather, etc.)
  - Probabilities are for reference group + cell phone use
  - Probabilities for cell phone use from text data (accident descriptions) and structured data

- After controlling for other factors, probabilities for cell phone use are approximately 12 percentage points higher (almost double the probabilities for the reference group). (Statistically significant at 5% level.)

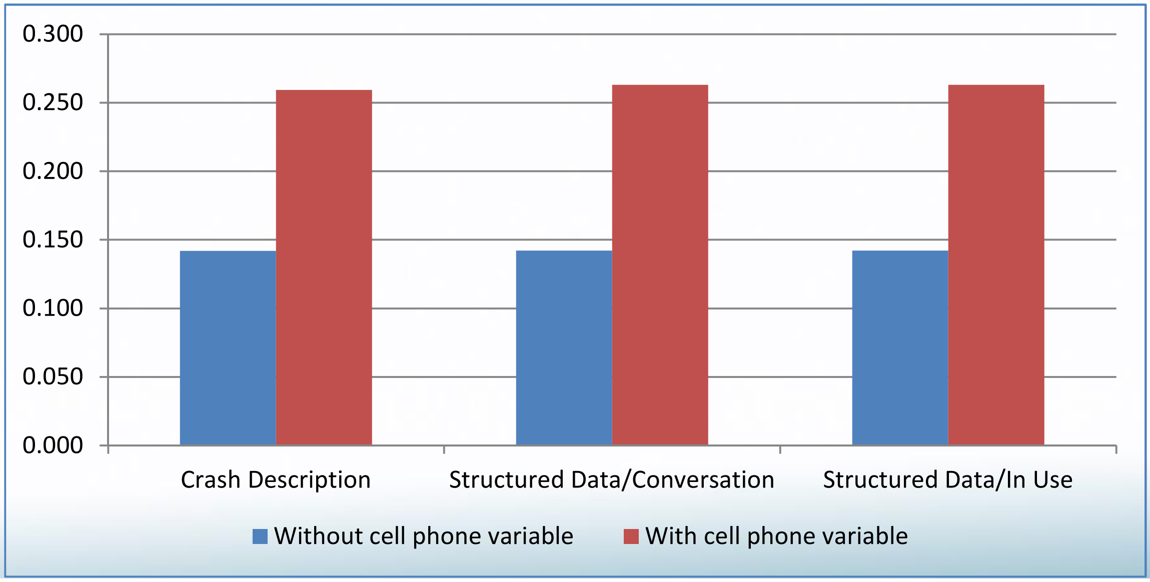
Probability Accident was a Rear-End	Accident Desc (text) - On Cell Phone	Structured - Conversation on Cell Phone	Structured - Cell Phone in Use
Reference group	0.142	0.142	0.142
Daytime			
Weekday			
Good weather			
Driver not fatigued			
No medications			
No drugs			
No alcohol			
Not adjusting radio/CD			
No cell phone use			
Cell Phone Use	0.259	0.263	0.263

Probability the Accident was a Rear-End Collision

- Graph presents the probability the accident involved multiple vehicles.
  - Left-hand (blue) bars: no cell phone variable in the model.
  - Right-hand (red) bars: cell phone variable in the model.

- Three sets of probabilities:
  - Cell phone variable from text data (accident descriptions)
  - Cell phone variable from structure data (conversation)
  - Cell phone variable from structure data (in use)

- Implications:
  - **Including cell phone variable increased the probability of predicting of a rear-end collision**.
  - **Cell phone variable from text data produced results similar to variables from structured data.**

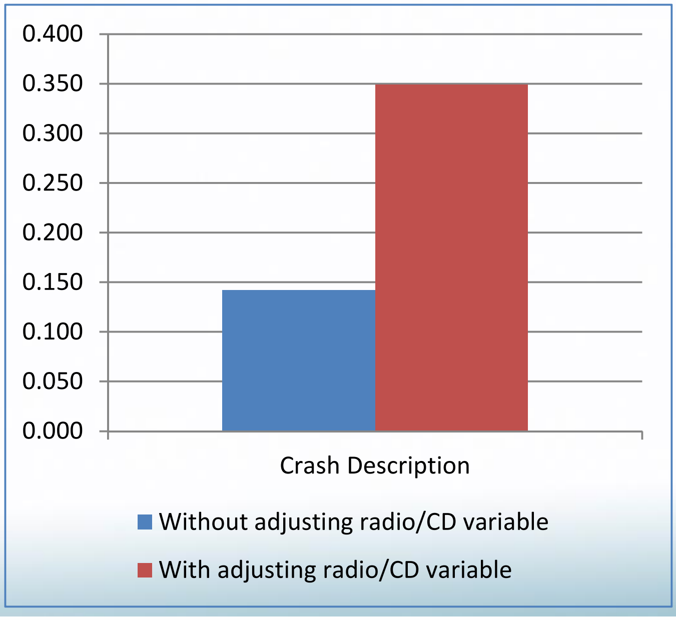


Probability the Accident was a Rear-End Collision

- Table presents probabilities for adjusting radio/CD being associated with a rear-end collision.
  - Probability for “reference group” (daytime, weekday, good weather, etc.)
  - Probability for reference group + adjusting radio/CD

- After controlling for other factors, probability for adjusting radio/CD is approximately 20 percentage points higher (almost 2.5x the probability for the reference group). (Statistically significant at 5% level.)

Probability of a Multi-Vehicle Accident	Accident Desc (text) - Adjusting Radio/CD
Reference group	**0.142**
Daytime	
Weekday	
Good weather	
Driver not fatigued	
No medications	
No drugs	
No alcohol	
Not adjusting radio/CD	
No cell phone	
Adjusting Radio/CD	**0.349**



Predictive Analytics -- Summary

- Two outcome measures
  - Rear-end collision
  - Multiple vehicles
- Control variables: structured data
  - Environmental
  - Driver
  - Adjusting radio/CD
- Additional information from text data
  - Cell phone in use
- **Proof of Concept**
  - **Text data improved results for predictive analytics**

Summary

- Reasons to be Interested in Text Data
- National Motor Vehicle Crash Causation Survey
- Accident Descriptions: 3 examples where cell phone use mentioned
- NMVCCS Definition of "Distracted Driving"
- Flags for Cell Phone Use Created from Text Data
- Cell Phone Use: Structured Data v. Text Data
- Results from Multivariate (Logit) Analyses
- **Text Data Improved Results for Predictive Analytics**