Distracted Driving, Text Data, and Predictive Analytics

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

March 20, 2012

Casualty Actuarial Society, Ratemaking & Product Management Seminar, Philadelphia, PA

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
- Crash Descriptions:
  - 3 examples where cell phone use mentioned
  - NMVCCS Crash Descriptions compared to Claim Adjuster Notes
  - Breaking Text Data into Manageable Units – Creating NGrams
- NMVCCS Definition of “Distracted Driving”
- Flags for Cell Phone Use Created from Text Data
- Cell Phone Use: Structured Data v. Text Data
- Multivariate (Logit) Analyses

Milliman

Starting Considerations

- NHTSA has issued a policy statement:
  - Advising drivers to resist using any activity that distracts from the operation of a motor vehicle, specifically mentioning cell phones, and
  - Recommending that states prohibit “novice” drivers from using electronic devices during the learners and intermediate stages of a driver license program.
- March 5: NHTSA began a national telephone survey on driving habits and attitudes related to distracted driving.
- NHTSA has proclaimed April to be “National Distracted Driving Awareness Month.”

Milliman

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 crashes 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 a crash.
  - 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 adjustor note, crash description, deposition, or other reports. Books, magazine articles, and research reports or other examples of text data.

Milliman

Reasons to be Interested in Text Data

- Able to capture concepts in text data not captured in structured data
  - Many structured data-reporting forms do not capture cell phone use
  - Drivers / occupants may be averse to reporting cell phone use at time of crash
- Claim stratification
  - Able to identify claims with “dialing on cell phone,” “talking on cell phone”, etc.
- Univariate and bi-variate analyses
  - How often does cell phone use occur while driving?
  - What types of accidents do cell phones appear to be an associated (possibly, contributing) factor?
  - Is there a difference by age of driver?
- Multivariate analyses (“predictive analytics”)
  - Does the inclusion of information from text data improve the predictability for target outcomes?

Milliman

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.

- 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.

Milliman

State Laws on Cell Phone Use and Texting While Driving

- Table below presents laws for selected states.
- Considerable differences across states.

State	Hand-Held Ban	All Cell Phone Ban	Texting Ban
California	All drivers	School and transit bus drivers, Drivers under 18	All drivers
Connecticut	All drivers	Learner's permit holders Drivers under 18 School bus drivers	All drivers
Florida	No	No	No
Illinois	Drivers in construction and school speed zones	Learner's permit holders under 19 Drivers under 19 School bus drivers	All drivers
Massachusetts	Local option	School bus drivers Passenger bus drivers Drivers under 18	All drivers
Texas	Drivers in school cross zones	Bus drivers Drivers under 18	Bus drivers with passengers under 18. Intermediate license holders for first 12 months. Drivers in school crossing zones.

Milliman

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.

Milliman

National Motor Vehicle Crash Causation Survey

- National Motor Vehicle Crash Causation Survey (NMVCCS)
  - Conducted by the National Highway Traffic Safety Administration (NHTSA)
  - Sample of crashes investigated between July 3, 2005 and December 31, 2007.
  - Primary focus of Survey: Determine the critical pre-crash events and reasons underlying the critical factors.
  - 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.
  - Crashes occurring between 6am and midnight.
  - Crash 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 crash scene.
  - Completed police report.

Milliman

National Motor Vehicle Crash Causation Survey

- Data files
  - 22 files
  - Crash Description, Pre-Crash Assessment (PCA), Occupant
  - Contents are static (not updated)

- Case weights
  - To make the sample representative of all similar types of crashes 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

- Files of special interest to this presentation
  - Structured data
  - Date and time of accident
  - Type of accident (eg, 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
    - > One record per crash
    - > 8,000 bytes
    - > Vehicles are identified in various references: V1, Vehicle 1, Vehicle #1, Vehicle One
    - > References not always consistent with the same crash description

NMVCCS Sample -- Summary Characteristics

- 6,949 crashes
  - 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

Milliman

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

Milliman

NMVCCS Crash Descriptions

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

- Three examples in the following slides.
  - Examples are typical of the NMVCCS crash 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)

Crash Description #1

Crash #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)

Crash Description #2

Crash #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)

Milliman

Crash Description #3

Crash #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)

Milliman

NMVCCS Crash Descriptions

- From the three examples, differences are notable.

- References to “vehicle”:
  - V1, V2 (#1, #3)
  - Vehicle 1, Vehicle 2 (#2)
  - Other crash descriptions: insert “#” before the number (eg., V#1), spell numeric (eg., Vehicle One)
  - Reference not always consistent within the same crash 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 Crash Descriptions

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

	All Cases	With Case Weights
Number of crashes	6,949	5,470
Number of words in crash 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 crash 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

Milliman

NMVCCS Crash Descriptions compared to Claim Adjuster Notes

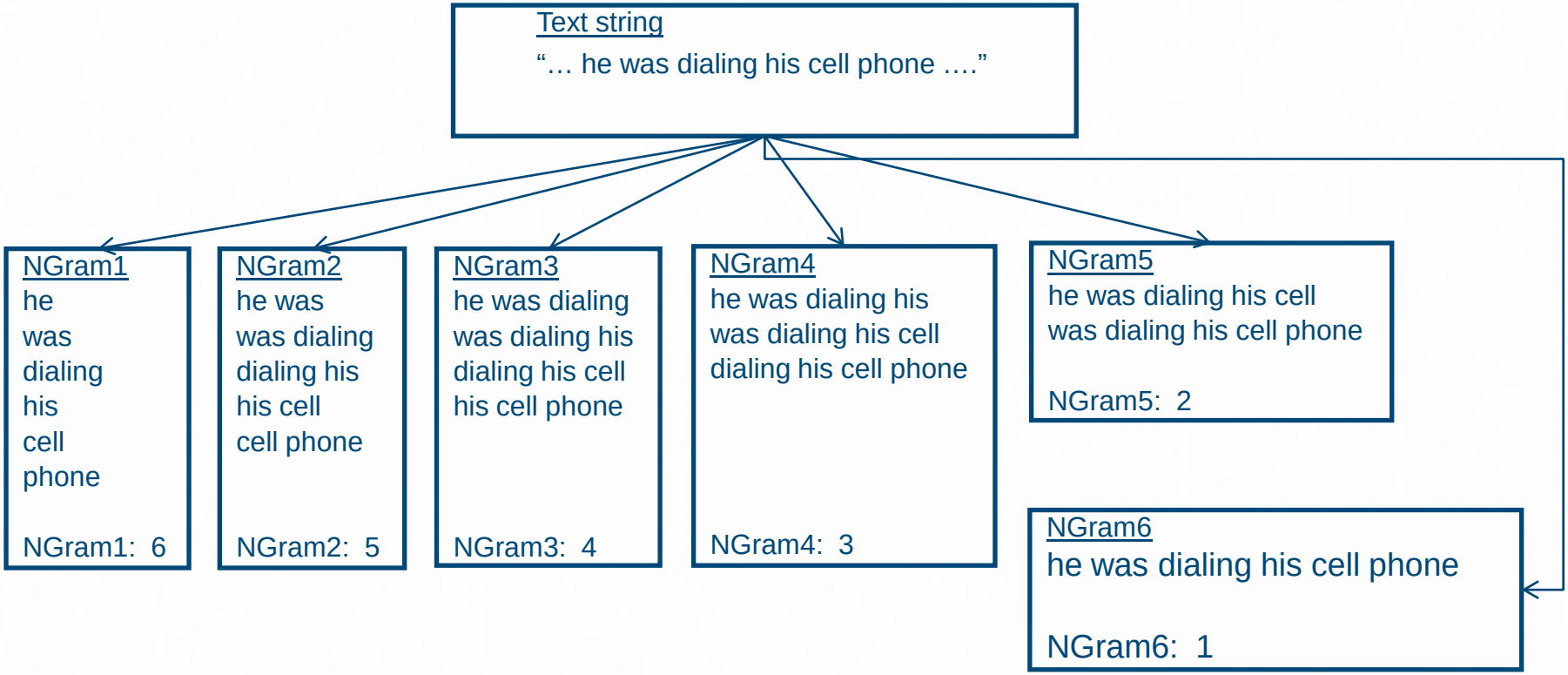
- NMVCCS crash descriptions are “cleaner” than the typical claim adjuster notes.

- Distinctions with Claim Adjuster notes :
  - 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 (eg., subrogation).

- Compared to the NMVCCS data, many of these points provide for a much wider scope of information.

- Insurer text data can also include text data beyond claim adjuster notes (eg, medical case manager notes, underwriting notes, depositions, statements).

Breaking Text Data into Manageable Units – Creating "NGrams"



- 1 six-word phrase produced 21 NGrams.

Milliman

NGrams Created from NMVCCS Crash Descriptions

- Each crash description was parsed into NGram1-NGram6.
- Process removes certain NGram1-NGram3 not expected to be needed in any claim segmentation or analytics.
- For each crash 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 crashes	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

Milliman

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.

Milliman

Use of Cell Phone: Structured Data v. Text Data

- Table below presents information for “cell phone in use”.
- Structured data: NMVCCS
  - 196 claims with cell phone in use (2.8%)
- Text data: crash descriptions
  - 264 crashes with cell phone in use (4.0%)
- Overlap between structured data and text data: 171 crashes

Number of Claims	Text Data	
Structured Data	Not in Use	In Use
Not in Use	6,660	93
In Use	25	171

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%

Milliman

Multivariate (Logit) Analyses

- Three outcome measures
  - Injury may have occurred (police report)
    - Are crashes where a cell phone was in use more likely to result in an injury?
  - Multiple vehicles in crash
    - Are crashes where a cell phone was in use more likely to involve multiple vehicles? (Distraction associated with cell phone use may be more difficult to manage with other moving objects in the vicinity.)
  - Rear-end collision
    - Does a cell phone in use influence the type of accident (eg, a rear-end accident)?

Multivariate (Logit) Analyses

- Explanatory variables
  - Environmental Controls
    - Night: crash occurred before 7am or after 6pm.
    - Weekend: crash 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 the crash was reported to be fatigued
    - Medications: one or more drivers reported taking drugs/medications within 24 hours preceding the crash
    - Drugs: police report recorded illegal drug(s) in driver's system
    - Alcohol: police report recorded presence of alcohol with the driver

Milliman

Multivariate (Logit) Analyses

- Explanatory variables
  - Three 0/1 indicators for cell phone 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)

Logit Regressions: Injury May Have Occurred

- Outcome measure: Injury may have occurred (police report)
  - Are crashes where a cell phone was in use more likely to result in an injury?
- Principal finding: use of cell phone does not significantly change the likelihood of an injury.
  - Signs on the coefficients for the three cell phone measures were mixed and none close to be statistically significant at the 5% level.
  - Finding may be because drivers using cell phones typically are not using excessive speed or placing the vehicle in a seriously dangerous position.

	Crash Descriptions (text)	Structured Field	Structured Field
INJURY POSSIBLE	On Cell Phone	Conversing on Cell Phone	Cell Phone in Use
Intercept	0.699*	0.704*	0.704*
NIGHT	-0.241*	-0.242*	-0.242*
WEEKEND	0.035	0.034	0.034
WEATHER	-0.138*	-0.139*	-0.139*
DRIVER FATIGUE	0.101	0.103	0.103
MEDICATIONS	0.720*	0.720*	0.720*
DRUGS	0.065	0.064	0.064
ALCOHOL	0.644*	0.645*	0.645*
CELL PHONE	0.148	-0.003	0.008
-2 log Likelihood	7,937	7,938	7,938

Milliman

Logit Regressions: Injury May Have Occurred

- Table below presents starting frequencies for cell-phone-use derived from the text data and probability after adjusting for other factors captured in the logit analyses ("estimated difference" in bottom of table on the right).

- After controlling for other factors, estimated difference associated with cell phone use is an increase of 2.9 percentage points. (Not statistically significant at 5% level.)

Number of Claims	Injury May Have Occurred		
Cell Phone Conversation	No	Yes	Total
No	1,839	4,846	6,685
Yes	63	201	264
Total	1,902	5,047	6,949

Row Percents	Injury May Have Occurred		
Cell Phone Conversation	No	Yes	Total
No	27.5	72.5	100.0
Yes	23.9	76.1	100.0
Estimated Difference		2.9	

Milliman

Logit Regressions: Multiple Vehicles in Crash

- Outcome Measure: multiple vehicles in crash
  - Are crashes where a cell phone was in use more likely to involve multiple vehicles?

- Principal Findings:
  - Use of cell phone is associated with an increased likelihood of being in a multi-vehicle crash.
  - Coefficients are 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 a crash.

	Crash Descriptions (text)	Structured Field	Structured Field
	On Cell Phone	Conversing on Cell Phone	Cell Phone in Use
Intercept	1.232*	1.235*	1.234*
NIGHT	-0.440*	-0.440*	-0.439*
WEEKEND	-0.416*	-0.414*	-0.413*
WEATHER	-0.519*	-0.520*	-0.519*
DRIVER FATIGUE	-0.582*	-0.580*	-0.579*
MEDICATIONS	0.591*	0.591*	0.589*
DRUGS	-0.561*	-0.559*	-0.558*
ALCOHOL	-0.540*	-0.537*	-0.542*
CELL PHONE	0.612*	0.646*	0.566*
-2 log Likelihood	7,601	7,603	7,604

Milliman

Logit Regressions: Multiple Vehicles in Crash

- Table below presents starting frequencies for cell-phone-use derived from the text data and probability after adjusting for other factors captured in the logit analyses ("estimated difference" in bottom of table on the right).

- After controlling for other factors, estimated difference associated with cell phone use is an increase of 11.6 percentage points. (Statistically significant at 5% level.)

Number of Claims	Multiple Vehicles in Crash		
Cell Phone Conversation	No	Yes	Total
No	1,778	4,907	6,685
Yes	44	220	264
Total	1,822	5,127	6,949

Row Percents	Multiple Vehicles in Crash		
Cell Phone Conversation	No	Yes	Total
No	26.6	73.4	100.0
Yes	16.7	83.3	100.0
Estimated Difference		11.6	

Milliman

Logit Regressions: Rear-End Collision

- Outcome Measure: Rear-end collision
  - Does a cell phone in use influence the type of accident (eg, a rear-end accident)?

- Principal Findings
  - Use of cell phone is associated with an increased likelihood of being in a rear-end collision.
  - Coefficients are 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 a crash.

	Crash Descriptions (text)	Structured Field	Structured Field
	On Cell Phone	Conversing on Cell Phone	Cell Phone in Use
Intercept	-1.383*	-1.380*	-1.383*
NIGHT	-0.406*	-0.406*	-0.405*
WEEKEND	-0.379*	-0.378*	-0.377*
WEATHER	-0.330*	-0.331*	-0.330*
DRIVER FATIGUE	0.020	0.022	0.021
MEDICATIONS	0.185*	0.185*	0.184*
DRUGS	-0.692*	-0.688*	-0.689*
ALCOHOL	-0.111	-0.110	-0.114
CELL PHONE	0.346*	0.363*	0.363*
-2 log Likelihood	6,454	6,455	6,454

Milliman

Logit Regressions: Rear-End Collision

- Table below presents starting frequencies for cell-phone-use derived from the text data and probability after adjusting for other factors captured in the logit analyses ("estimated difference" in bottom of table on the right).

- After controlling for other factors, estimated difference associated with cell phone use is an increase of 5.0 percentage points. (Statistically significant at 5% level.)

Number of Claims	Rear-End Collision		
Cell Phone Conversation	No	Yes	Total
No	5,498	1,187	6,685
Yes	201	63	264
Total	5,699	1,250	6,949

Row Percents	Rear-End Collision		
Cell Phone Conversation	No	Yes	Total
No	82.2	17.8	100.0
Yes	76.1	23.9	100.0
Estimated Difference		5.0	

Milliman

Logit Regressions -- Summary

- Three outcome measures
  - Injury may have occurred
  - Rear-end collision
  - Multiple vehicles
- Control variables
  - Environmental
  - Driver
- Preliminary Findings
  - Presence of cell phone use influences the type of accident

Summary

- Reasons to be Interested in Text Data
- National Motor Vehicle Crash Causation Survey
- Crash 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
- Multivariate (Logit) Analyses