Demonstrating the Value of Text Data for Predictive Modeling

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

May 28, 2015

Southern California Casualty Actuarial Club, Spring 2015 Meeting, Pasadena, CA

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.

Objective of the Presentation

Demonstrating the Value of Text Data for Predictive Modeling

Overview

- Business Issue and Value Proposition
- From Text Data to Modeling File
- Breaking Text Data into Manageable Units – Creating NGrams
- NGram-Flag Dictionary
- Uses for the Expanded Modeling File
- Proof of Concept Results

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.

Starting Considerations

- **Business Issue**
  - Adjusters' notes and other text data contain information useful for predictive modeling outcomes that are not readily available in structured data.

- **Value Proposition**
  - **Early identification of claim characteristics** for improved predictive modeling results, claim triage opportunities, subrogation opportunities, fraud detection, and other property-casualty claim initiatives and analytics.
  - Identification of claim **characteristics not typically captured in structured data**, including comorbidities.
  - Identification of **newly developing claim characteristics** not part of incumbent structured coding systems (e.g., new work-related diseases, distracted driving, driving under the influence of medications).

Proof of Concept Demonstrations

Multivariate Analyses Using the Modeling File

- **Auto - Distracted Driving: Use of Cell Phone**
  - Evaluation: probability the accident was a rear-end collision
  - Results compared to structured data created by NHTSA

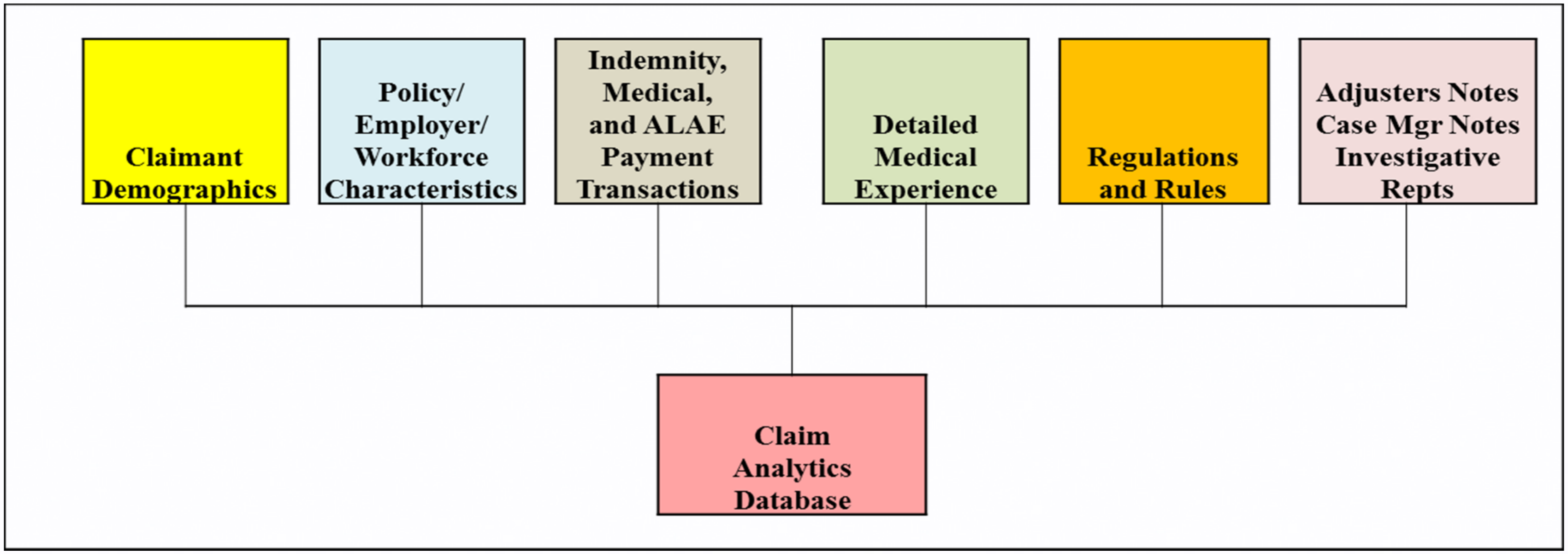
- **Auto - Driving under the Influence of Medications, Rx, or Narcotics**
  - Accident causes not known at time of accident and not easily coded in structured data
  - Analyses for different levels of inferred severity per text data

- **Workers Compensation - Claim Severity at 30 Days**
  - Limited information on payment history and medical experience
  - Information from accident description (e.g., FNOL) and adjusters' notes

Available Data Feeds

One Approach to Claim Analytics and Predictive Modeling

- Gather all data at the transaction level, including text data (e.g., Adjusters Notes)



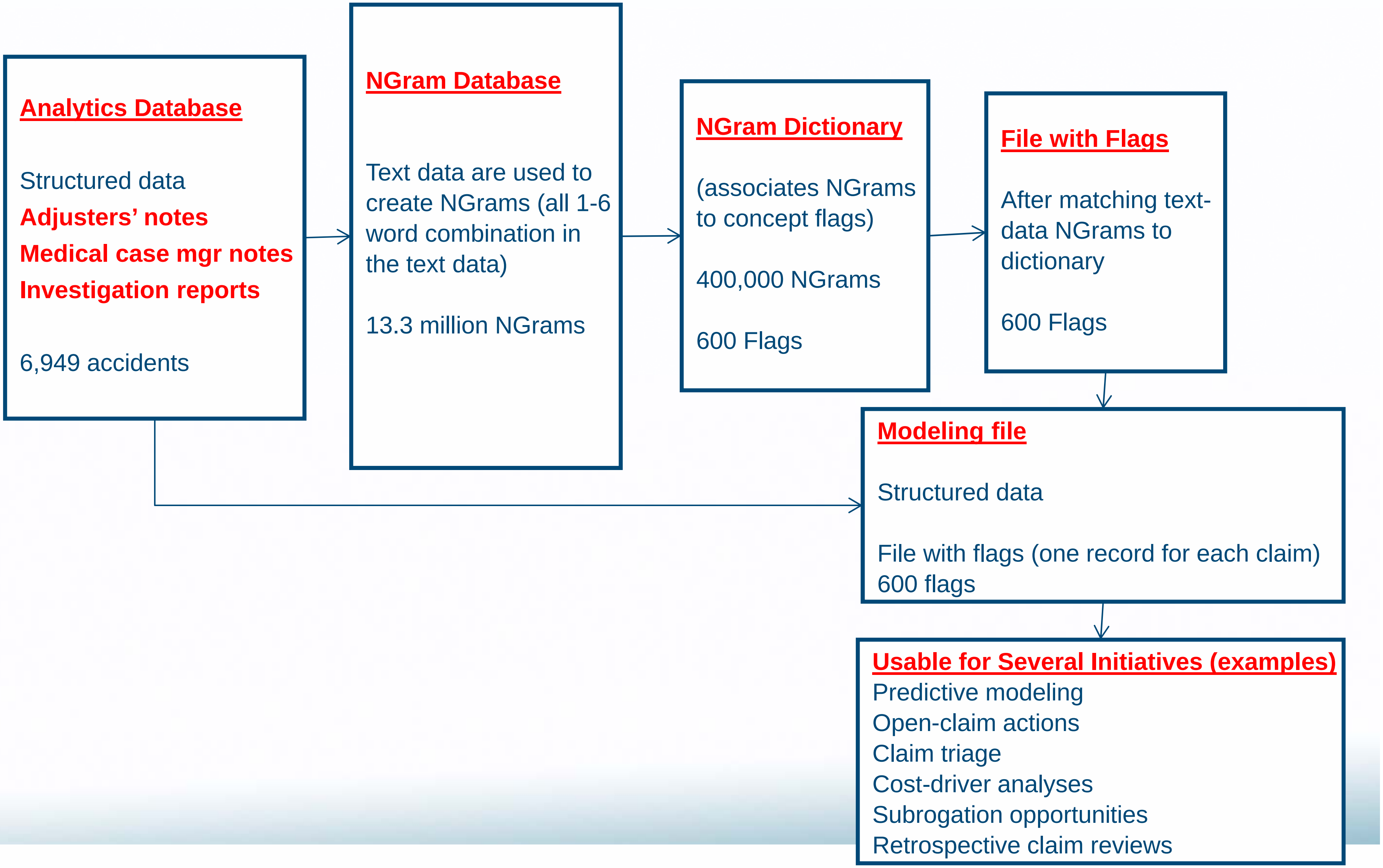
Claimant Demographics
Policy/ Employer/ Workforce Characteristics
Indemnity, Medical, and ALAE Payment Transactions
Detailed Medical Experience
Regulations and Rules
Adjusters Notes Case Mgr Notes Investigative Repts
Claim Analytics Database

–Considerations

- Little payment transaction or detailed medical available at 30 days from FNOI
- Considerable info available in Accident Descriptions, Adjusters Notes, Case Manager Notes, etc.

Process Flow

From Text Data to Modeling File



Analytics Database

Structured data

Adjusters' notes

Medical case mgr notes

Investigation reports

6,949 accidents

NGram Database

Text data are used to create NGrams (all 1-6 word combination in the text data)

13.3 million NGrams

NGram Dictionary

(associates NGrams to concept flags)

400,000 NGrams

600 Flags

File with Flags

After matching text-data NGrams to dictionary

600 Flags

Modeling file

Structured data

File with flags (one record for each claim)
600 flags

Usable for Several Initiatives (examples)

Predictive modeling
Open-claim actions
Claim triage
Cost-driver analyses
Subrogation opportunities
Retrospective claim reviews

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-accident 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.
  - 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
  - Accident Description, Pre-Crash Assessment (PCA), Occupant
  - 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

- 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: driver fatigue, presence of alcohol

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

Accident Description #1 (distracted driving)

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)

Accident Description #2 ("…taking several meds")

Accident #1: V1, a 2002 Dodge Stratus, was traveling westbound on a four-lane, two-way, dry, asphalt roadway with a level grade in daylight conditions. V1 was intending to go straight. V2, a 2004 Honda Accord, was traveling eastbound in the second lane of travel on the same roadway in similar conditions, also intending to go straight. The posted speed limit was 56 kmph (35 mph). The driver of V1 was experiencing low blood sugar and passed out at the wheel, relinquishing control of the car. V1 crossed the double yellow lines and the front of V1 contacted the front of V2. V2 came to final rest on the roadway facing west. V1 came to final rest off the south side of the roadway facing north.

The driver of V1 was a 43-year old diabetic male who reported that he had blacked out due to low blood sugar. Medical records indicated that immediately after the crash, his blood sugar was 32, a dangerously low level. The driver of V1 sustained serious injuries during the crash and was transported to a local trauma facility. The driver of V1 told doctors that he had skipped a meal earlier in the day but had still taken his insulin.

The Critical Pre-crash Event for the driver of V1 was when he traveled over the lane line on the left side of the travel lane. The Critical Reason for the Critical Pre-crash Event was a critical non-performance error due to the diabetic blackout. The driver of V1 was taking several medications for various health problems, including heart problems, high cholesterol, thyroid problems, and diabetes.

The driver of V2 was a 44-year old female who had reported that she had been traveling between 50-64 kmph (31-40 mph) prior to the crash. She had no health related problems and was rested and traveling back to work. She was wearing her prescribed lenses that corrected a myopic (near-sighted) condition. She sustained minor injuries during the crash and was transported to a local trauma facility.

The Critical Pre-crash Event for the driver of V2 was other motor vehicle encroachment, from opposite direction-over left lane line. The Critical Reason for the Critical Pre-crash Event was not coded to the driver of V2 and she was not thought to have contributed to the crash. (380 words, 2,224 bytes)

NMVCCS Accident Descriptions

- Notable differences across accident descriptions.

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

- References to medications, Rx, and drugs with common “under the influence” implications:
  - was taking several medications (#1)
  - use of prescription medications (#2)
  - diuretic side effects (#2)
  - health medication with possible side effects (#2)
  - takes prescription anti-inflammatory drug (#2)
  - tested positive for amphetamines (#3)
  - mention of “red flags” (#3)
  - With claim adjuster notes, some meds/Rx may not be contributing factors to the accident.

NMVCCS Accident Descriptions compared to Claim Adjuster Notes

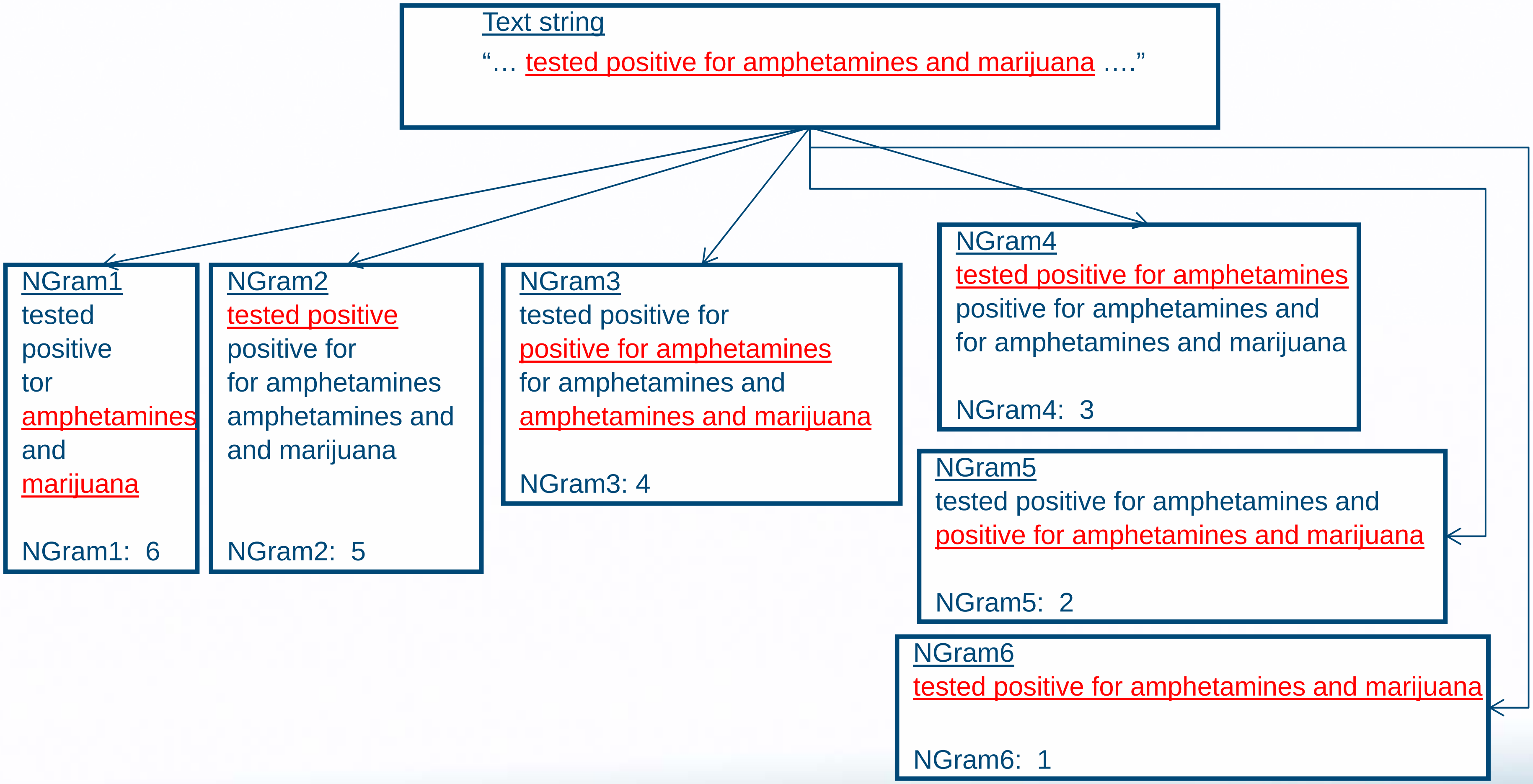
- NMVCCS accident descriptions are “cleaner” than the typical claim adjuster notes.

- 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”



Text string

“… tested positive for amphetamines and marijuana ….”

NGram1

tested
positive
tor
amphetamines
and
marijuana

NGram1: 6

NGram2

tested positive
positive for
for amphetamines
amphetamines and
and marijuana

NGram2: 5

NGram3

tested positive for
positive for amphetamines
for amphetamines and
amphetamines and marijuana

NGram3: 4

NGram4

tested positive for amphetamines
positive for amphetamines and
for amphetamines and marijuana

NGram4: 3

NGram5

tested positive for amphetamines and
positive for amphetamines and marijuana

NGram5: 2

NGram6

tested positive for amphetamines and marijuana

NGram6: 1

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

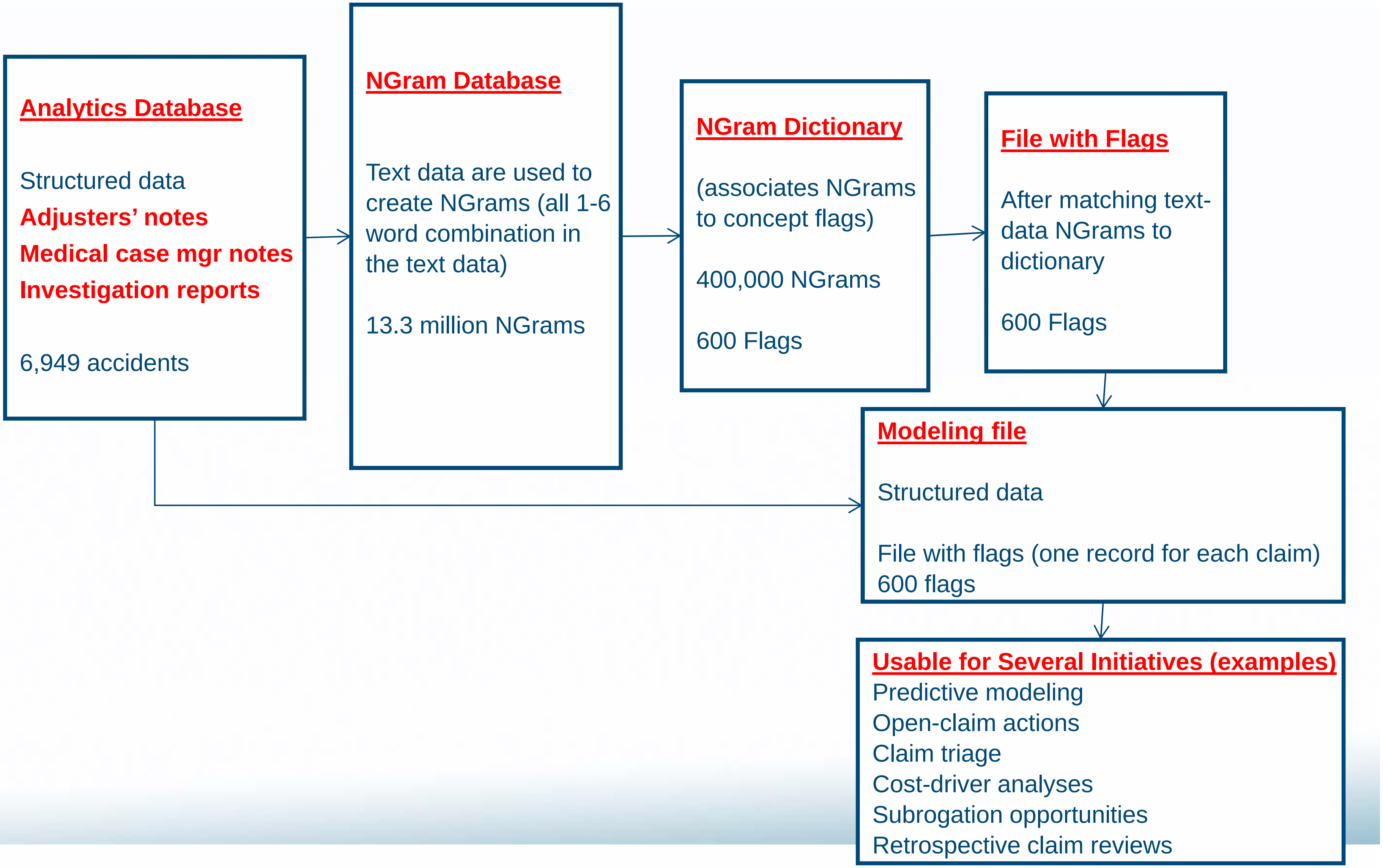
Strategies for Identifying Target Ngram-Flag Combinations

- **Most general: reference to a general term**
  - Mention of “medication” or “prescription”
    - “was taking ….”
    - “had taken ….”
  - “Medication” or “prescription” can refer to broad set of OTC, Rx, or other meds
  - Present analysis: approximately 1,100 phrases

- **Action associated with a term: action + noun**
  - Action associated with a drug name
    - “had taken his [drug name]”
    - “was on [drug name]”
  - With subgrouping, able to control combinations of action+drug
  - Present analysis: 3,590 phrases (10 actions x 395 drug names)

- **Most specific: target list of words**
  - List of drugs (esp. narcotics) that are red flags
    - Cocaine, heroin, marijuana
  - Present analysis: 52 narcotics

Expanded Modeling File Usable for Several Initiatives



Analytics Database

Structured data

Adjusters' notes

Medical case mgr notes

Investigation reports

6,949 accidents

NGram Database

Text data are used to create NGrams (all 1-6 word combination in the text data)

13.3 million NGrams

NGram Dictionary

(associates NGrams to concept flags)

400,000 NGrams

600 Flags

File with Flags

After matching text-data NGrams to dictionary

600 Flags

Modeling file

Structured data

File with flags (one record for each claim)
600 flags

Usable for Several Initiatives (examples)

Predictive modeling
Open-claim actions
Claim triage
Cost-driver analyses
Subrogation opportunities
Retrospective claim reviews

Multivariate Analyses Using the Modeling File

- **Auto - Distracted Driving: Use of Cell Phone**
  - Evaluation: probability the accident was a rear-end collision
  - Results compared to structured data created by NHTSA

- **Auto - Driving under the Influence of Medications, Rx, or Narcotics**
  - Accident causes not known at time of accident and not easily coded in structured data
  - Analyses for different levels of inferred severity per text data

- **Workers Compensation - Claim Severity at 30 Days**
  - Limited information on payment history and medical experience
  - Information from accident description (e.g., FNOL) and adjusters’ notes

Auto – Distracted Driving: Use of Cell Phone

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

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

Auto – Distracted Driving: Use of Cell Phone

- Multivariate (logit) analyses: Explanatory variables
  - Time if day/week
    - Night: accident occurred before 7am or after 6pm.
    - Weekend: accident occurred on a Saturday or Sunday
  - Environment
    - Weather: on or more adverse conditions (eg., snow, rain, ice)
    - Wet roads
  - Nature of the accident
    - Multiple vehicles
    - Rear end
    - Head on
    - Turned into path
  - Driver Conditions
    - Driver fatigue: at least one driver in the accident was reported to be fatigued
    - Alcohol: police report recorded presence of alcohol with the driver

Multivariate (Logit) Analyses

- Explanatory variables (continued)
  - 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)

Proof of Concept Demonstration - 1

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

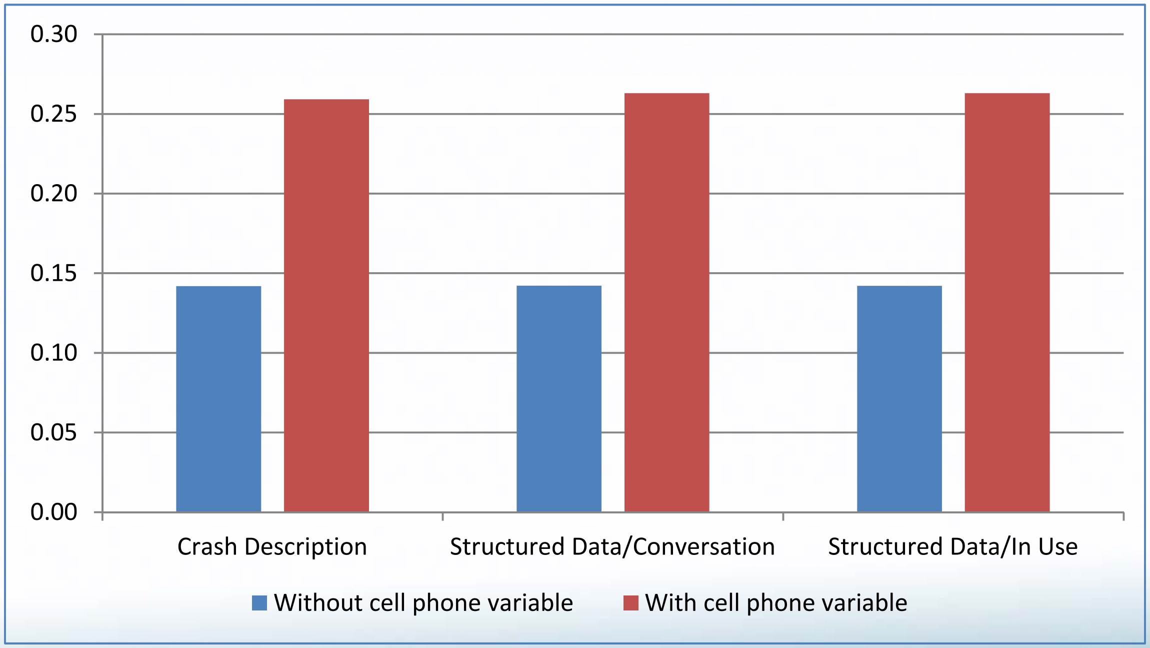
Proof of Concept Demonstration - 1

Auto - Distracted Driving: Use of Cell Phone

- Three sets of probabilities for cell-phone variable: from text data (accident descriptions), from structured data (conversation), from structured data (in use)

- Graph presents the **probability the accident was a rear-end collision.**
  - Left-hand (blue) bars: no cell phone variable in the model.
  - Right-hand (red) bars: cell phone variable in the model.

- Findings:
  - **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.**



Auto – DUI of Medications, Rx, or Narcotics

- Multivariate (logit) analyses: Explanatory variables
  - Four 0/1 indicators:
    - Medications: mention of driver taking or on “medication”
    - Prescription: mention of driver taking or on “prescription”
    - Drugs: action + drug name (“taking [drug name]”)
    - Narcotics: single-word “red flag” (or per se) references
- Outcome measure
  - Injury may have occurred (police report)
  - Are accidents where one of the drivers has been taking meds, Rx, a drug, or a narcotic more likely to result in an injury?

Logit Regressions: Injury May Have Occurred

- Outcome measure: Injury may have occurred (police report)
  - Are accidents where a driver was taking or on a med, Rx, drug, or narcotic more likely to result in an injury?

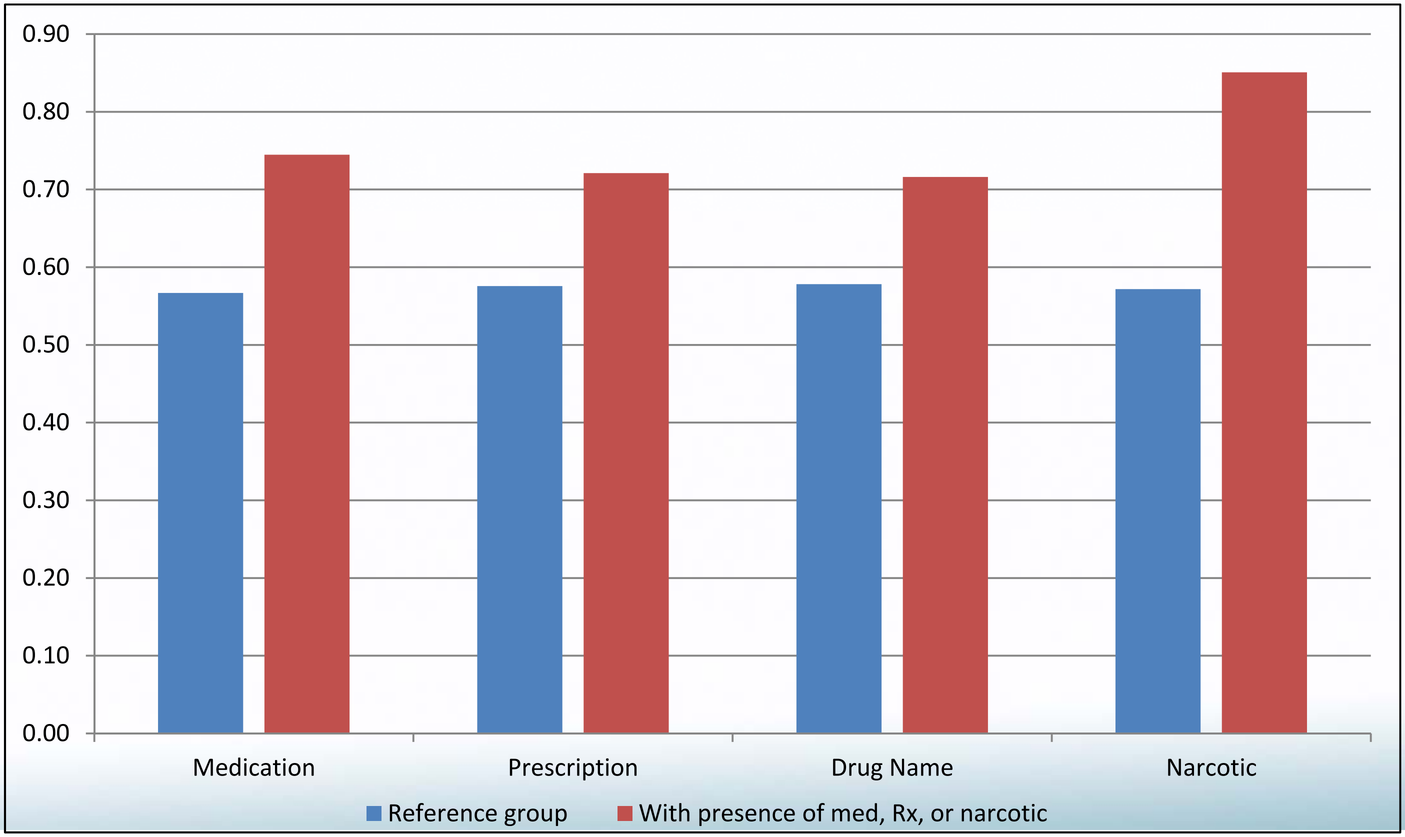
- Principal finding:
  - taking or on a med, Rx, drug, or narcotic increases the likelihood of an injury
  - coefficient for each of the four measures statistically significant at the 5% level.

Variable	Medication		Prescription		Drug Name		Narcotic	
Intercept	0.5220	*	0.5726	*	0.5811	*	0.5679	*
Night	-0.2527	*	-0.2672	*	-0.2657	*	-0.2789	*
Weekend	0.0584		0.0509		0.0490		0.0459	
Weather	0.0394		0.0615		0.0569		0.0599	
Wet road surface	-0.2179	*	-0.2341	*	-0.2336	*	-0.2322	*
Multiple vehicles	0.5403	*	0.5395	*	0.5359	*	0.5569	*
Rear end	-0.3009	*	-0.2978	*	-0.3059	*	-0.2942	*
Head on	0.6660	*	0.6675	*	0.6663	*	0.6352	*
Turned into path	0.2957	*	0.3011	*	0.2989	*	0.3156	*
Driver fatigue	0.2262	*	0.2643	*	0.2588	*	0.2452	*
Alcohol	0.7155	*	0.7192	*	0.7076	*	0.6655	*
Medications	0.5488	*	----		----		----	
Prescription	----		0.3771	*	----		----	
Drugs	----		----		0.3439	*	----	
Narcotics	----		----		----		1.1729	*

Proof of Concept Demonstration – 2

Auto - Driving under the Influence Meds, Rx, Narcs

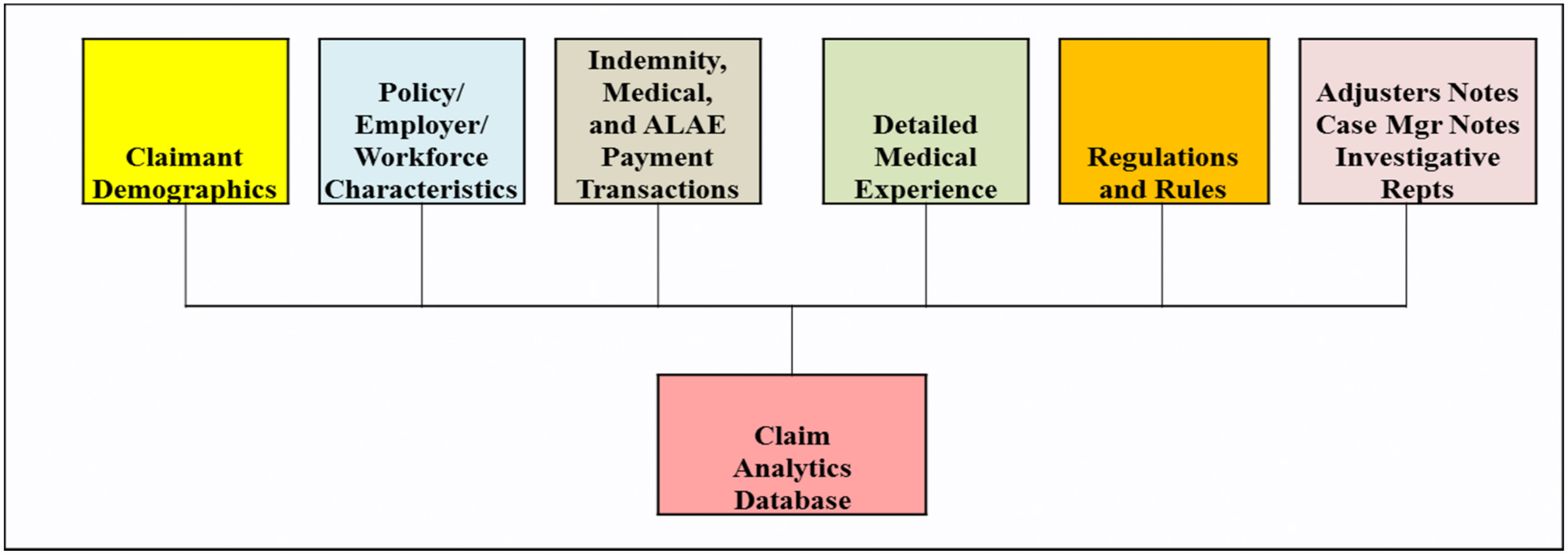
- **Probability the accident involved an injury** (NHTSA data does not have financial losses)
- Blue bars: “Reference group”: 0 values for control variables in the logistic regression
- Red bars: Probabilities for presence of meds, Rx, a drug, or narcotic included in regression
- **Finding**: Variables from text data are associated with higher probabilities the accident involved an injury



Proof of Concept Demonstration – 3 (Workers Comp)

One Approach to Claim Analytics and Predictive Modeling

- Gather all data at the transaction level, including text data (e.g., Adjusters Notes)


Claimant Demographics
Policy/ Employer/ Workforce Characteristics
Indemnity, Medical, and ALAE Payment Transactions
Detailed Medical Experience
Regulations and Rules
Adjusters Notes Case Mgr Notes Investigative Repts
Claim Analytics Database

–**Considerations**

- Little payment transaction or detailed medical available at 30 days from FNOI
- Considerable info available in Accident Descriptions, Adjusters Notes, Case Manager Notes, etc.

Proof of Concept Demonstration - 3

Workers Compensation - Adjuster Notes

- **Claim #1 (adjuster note 14 days after injury):**
  - Example for early surgery.
  - "The treating doctor reports the employee has decreased range of motion and increased swelling and pain in the left knee. The employee is **scheduled for an arthroscopic surgery** of the left knee to repair the torn medial meniscus ….."

- **Claim #2 (adjuster note 14 days after injury):**
  - Example for claimant represented by an attorney.
  - "Claimant noticed pain in the upper back that radiates to both shoulders..…has been experiencing intermittent migraines….symptoms have been getting progressively worse….an increase in stress at work....**is represented by an attorney**."

- **Claim #3 (adjuster note 9 days after injury):**
  - Example for early medical, injury severity, prior injury, co-morbidity.
  - "He slipped on the floor and went down on his back…. An **ambulance was called**…They **performed X-rays** on his back and right arm…He describes having **sustained pain**… He describes having a **prior back injury**… This was a workers comp injury… He **has hypertension as well as diabetes**…."

Sample NGram-Flag Combinations in Dictionary

NGram Dictionary

(associates NGrams to concept flags)

400,000 NGrams

600 Flags

Flag	Ngram
Surgery (>19,000 NGrams)	clmt had surgery needs surgery neurosurgery surgery will be needed scheduled for surgery ……….
Attorney (>2,000 NGrams)	attorney called received call from atty report from attorney …..
Hospital (>26,000 NGrams)	adjuster has requested hospital admitted ee to hospital requested iw contact hospital sent her to hospital ….
Overweight (>3,500 NGrams)	moderately overweight morbidly overweight …..

Value Proposition – Summary Statistics

Increase in Medical Payments After 30 Days From FROI		
	Mention of Co-Morbidity in Adjuster Notes before 30 Days	
Co-Morbidity	**No**	**Yes**
Arthritis	$5,500	**$16,300**
Cancer	5,700	**17,100**
Diabetes	5,300	**19,300**
Overweight	5,700	**16,000**
Prior injury	5,300	**13,200**
Smoker	5,500	**17,900**
Medical Service		
Surgery	$3,700	**$25,700**
Radiology	4,100	**20,300**

Average Incurred (after more than 3 years from FROI)		
Characteristic	**Book 1**	**Book 2**
All Claims	$28,900	$19,300
Mentioned in Adj Notes 0-3 days from FROI		
Attorney	**$191,100**	**114,100**
Ambulance	**172,600**	**61,500**
Hospital	**134,100**	**110,400**
Prior injury	**37,300**	**30,200**
Surgery	**145,400**	**155,500**
Ambulance, Hosp, Surg	**124,900**	**98,200**

Payments and Incurred Losses

for Early Identifications from Adjusters’ Notes

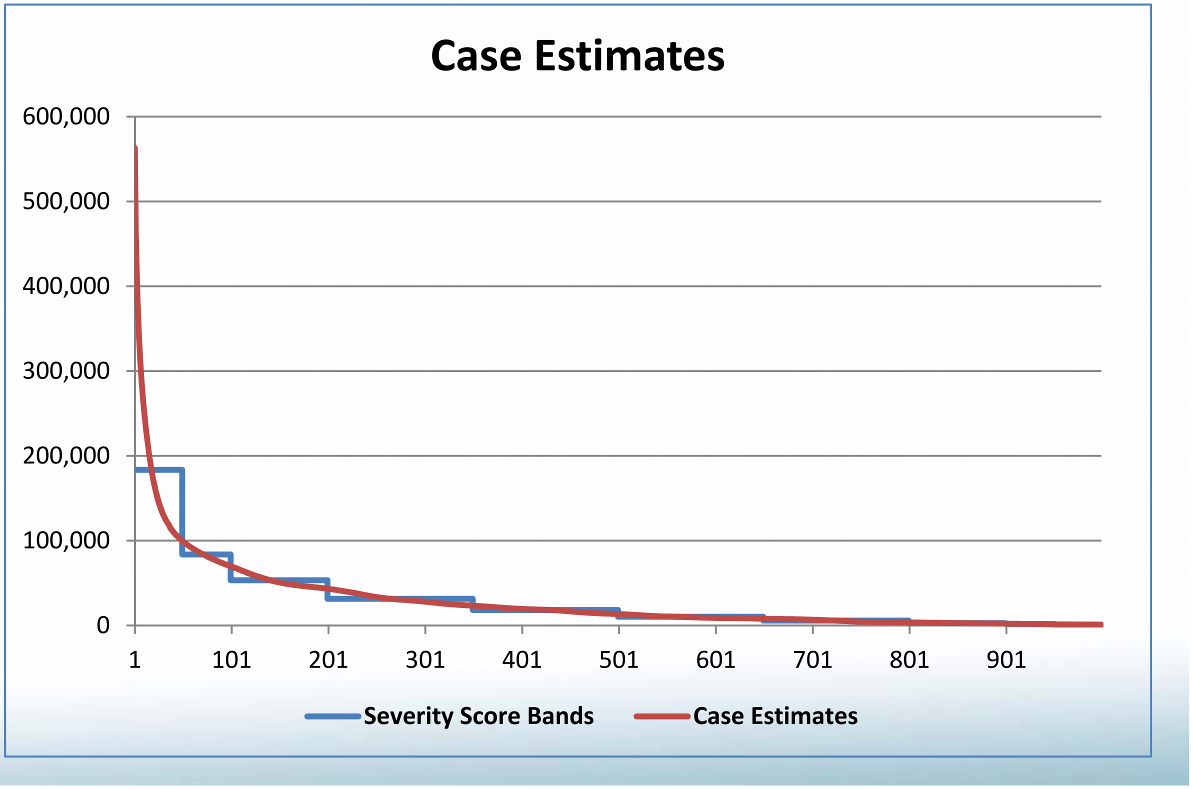
- Top table:
  - Co-morbidities and medical services
  - Adjusters’ Notes at 28 days from FROI
  - Significant increases in post -28 days medical payments for claims with identified characteristics in adjusters’ notes

- Bottom table:
  - Attorney involvement, medical services, and prior injury
  - Adjusters’ Notes at 3 days from FROI
  - Incurred losses after more than 3 years are significantly higher for claims with characteristics identified at 3 days from FROI

Workers Compensation Claim Severity – 30-Day Information

File Group	Claim Characteristic	Influence on 30-Day Severity
Master	AGE Class Group Gender Tenure Wage	31%
First Report of Loss	Body Part Nature of Accident Nature of Injury Reporting Lag	27%
Policy	Payroll Premium Premium / Payroll	10%
Payment Trans	Indemnity at 30 days Medical at 30 days	7%
Accident Desc	**Laceration** **Low Back** **Multiple Natures of Injury** **Number of Injd Body Parts** **Number of Natures of Injury** **Sprain** **Strain**	**7%**
Adjusters Notes	**Ambulance** **Hospital** **Multiple Body Parts** **No Losttime** **MRI** **Number of Body Parts** **Number of Natures of Injury** **Surgery**	**18%**

- **Objective**: Severity model using information as of 30 days from FROL
- **Methodology**: machine-learning approach that used a set of decision-tree results in an ensemble model.
- **Results**: Severity estimates ranged from $360 to over $550,000.
- **Text data**: Acc Descriptions and Adjusters Notes contributed 25% to model results



Summary

- Starting Considerations
- Available Data Feeds
- Breaking Text Data into Manageable Units – Creating NGrams
- NGram-Flag Dictionary
- Modeling File Creation
- Proof of Concept Results