Accident Analysis and Prevention 150 (2021) 105835

Contents lists available at ScienceDirect

Accident Analysis and Prevention

journal homepage: www.elsevier.com/locate/aap

Injury severity analysis of pedestrian and bicyclist trespassing crashes at non-crossings: A hybrid predictive text analytics and heterogeneity-based statistical modeling approach

Behram Wali [a,b,*], Asad J. Khattak [b], Numan Ahmad [b]

[a] Urban Design 4 Health, Inc., United States
[b] Department of Civil & Environmental Engineering, The University of Tennessee, United States

ARTICLE INFO

Keywords:
Non-motorist trespassing crashes
Non-crossings
Injury severity
Machine learning
Text analysis
Concept/Entity extraction
Dynamic factor analysis
Heterogeneity-based discrete outcome modeling

ABSTRACT

Non-motorists involved in rail-trespassing crashes are usually more vulnerable to receiving major or fatal injuries. Previous research has used traditional quantitative crash data for understanding factors contributing to injury outcomes of non-motorists in train involved collisions. However, usually overlooked crash narratives can provide useful and unique contextual crash-specific information regarding factors associated with injury outcomes. The main objective of this study is to harness the rapid advancements in more sophisticated qualitative analysis procedures for identifying thematic concepts in unstructured crash narrative data. A two-staged hybrid approach is proposed where text mining is applied first to extract valuable information from crash narratives followed by inclusion of the new variables derived from text mining in formulation of advanced statistical models for injury outcomes. By using ten-year (2006–2015) non-motorist non-crossing trespassing injury data obtained from the Federal Railroad Administration, statistical procedures and advanced machine learning text analytics are applied to extract unique information on contributory factors of trespassers' injury outcomes. The key concepts are systematically categorized into trespasser, injury, train, medical, and location related factors. A total of 13 unique variables are extracted from the thematic concepts that are not present in traditional tabular crash data. The analysis reveals a positive statistically significant association between presence of crash narrative and trespasser's injury outcome (coded as minor, major, and fatal injury). Compared to crashes with minor injuries, crashes involving major and fatal injuries are more likely to be reported with crash narratives. A crosstabulation of new variables derived from text mining with injury outcomes revealed that trespassers with confirmed suicide attempts, trespassers wearing headphones, or talking on cell phones are more likely to receive fatal injuries. Among other factors identified, trespassers under alcohol influence, trespasser hit by commuter train, and advance warnings by engineer are associated with more severe (major and fatal) trespasser injury outcomes. Accounting for unobserved heterogeneity and controlling for other factors, fixed and random parameter discrete outcome models are developed to understand the heterogeneous correlations between trespasser injury outcomes and the new crash specific explanatory variables derived from text mining – providing deeper insights. Practical implications and future research directions are discussed.

1. Introduction

According to the most recent available statistics, 95 % of all rail-related fatalities in the United States occurred at railroad crossings or included a trespasser - where about every three hours, a person or vehicle is hit by train (FRA, 2018). The Federal Railroad Administration defines trespasser as a "person who is on the part of railroad property used in railroad operations and whose presence is prohibited, forbidden, or unlawful" (FRA, 2015). Specifically, trespassing along highway-rail grade crossings and non-crossings is the leading cause of rail-related deaths in America, imposing billions of dollars in personal and societal costs (FRA, 2015). Non-motorist trespassers are usually more vulnerable to receiving serious or fatal injuries in such-like crashes (Zhang et al., 2018a). In the past ten years, overall crossing fatalities have declined but trespassing fatalities have substantially increased by 25 % (FRA, 2018). This highlights the vulnerability of non-motorized

* Corresponding author at: Urban Design 4 Health, Inc., United States.
E-mail addresses: bwali@ud4h.com, bwali@vols.utk.edu (B. Wali), akhattak@utk.edu (A.J. Khattak), nahmad1@vols.utk.edu (N. Ahmad).

https://doi.org/10.1016/j.aap.2020.105835
Received 10 January 2020; Received in revised form 13 August 2020; Accepted 3 October 2020
Available online 9 December 2020

trespassers (such as pedestrians and bicyclists) in receiving more severe injuries in rail-related crashes. Despite the unacceptably high fatalities to non-motorists in such-like collisions (FRA, 2018), remarkably little research has focused on injury outcomes sustained by non-motorists in train-involved crashes (Lobb, 2006; Khattak and Tung, 2015; Wang et al., 2016). Recognizing the huge economic and societal costs imposed by such collisions, a thorough understanding of factors that contribute to injury outcomes of non-motorists/pedestrians in train-involved collisions is warranted.

Reliable and adequate crash data has been fundamental to our understanding of factors contributing to unsafe outcomes (Elvik and Voll, 2014). The nature and quality of historical crash data (i.e., police crash reports) has not advanced significantly since inception as much as the methodological techniques have evolved. Traditionally, quantitative crash data has been extensively used for guiding safety policies pertaining to highway, highway rail-grade, and/or trespassing crashes (Abdel-Aty, 2003; Khattak and Luo, 2011; Savolainen et al., 2011; Khattak and Tung, 2015; Wang et al., 2016). However, in addition to the quantitative crash data, crash narratives may provide contextual crash-specific information that can enhance our understandings of the factors influencing trespasser injury severity outcomes. With the rapid advancements in more sophisticated qualitative analysis procedures for unstructured data (Feldman and Sanger, 2007), new knowledge critical to ongoing efforts of improving trespasser safety can be generated by systematically analyzing trespasser crash narratives. Injury severity is a critical aspect of railway safety-improvement programs (Khattak and Luo, 2011, Eluru et al., 2012; Zhao et al., 2016). An in-depth investigation of key correlates associated with pedestrian and bicyclist trespasser injury severity outcomes through advanced data and text mining techniques can facilitate development of goal-oriented trespassing injury prevention strategies. As such, the identified objectives of this study are to: (1) acquire non-crossing pedestrian and bicyclist trespassing injury severity data along with crash narratives, (2) apply advanced text data mining techniques to extract unique information on contributory factors embedded in context-specific crash narratives, and (3) develop appropriate heterogeneity-based statistical models to understand correlations between trespasser injury outcomes and key correlates derived from text mining of crash narratives.

2. Literature review

Trespassing is defined as presence of individual on railroad rights-of-way at any place other than a designated level crossing (Wang et al., 2016). According to Federal Railroad Administration, nationally more than 400 trespass fatalities and nearly as many injuries occur each year along railroad rights-of-way, making it the leading cause of rail-related deaths in America (FRA, 2015). Previous studies have focused on different themes pertaining to train-involved crashes including train pedestrian/bicyclist collisions and trespassing crashes.

2.1. Train pedestrian/bicyclist collisions

Among all types of rail-related crashes, collisions between rail-pedestrian/bicyclist crashes has received considerably more attention in previous research. Specifically, previous studies focused on analyzing frequency of fatalities in train-pedestrian collisions (Nichols et al., 1994; Silla and Luoma, 2012; Savage, 2016), severity of crashes sustained by pedestrians (Khattak and Tung, 2015; Zhao et al., 2016), and pedestrian/bicyclist behaviors and/or violations at highway-rail and pathway-rail grade crossings (Khattak and Luo, 2011, Metaxatos and Sriraj, 2013) – see Table 1 which provides detailed synthesis of key studies and their findings. The identification of manner of death in such-like collisions has gained extensive interest. But, due to lack of sufficient information required to make a definitive classification, it is challenging to determine manner of death (Nichols et al., 1994; Mishara, 2007). Recently, by using more accurate classification systems developed in Europe and United States for identification of intentional deaths, two important take-aways surfaced from the previous analyses (Silla and Luoma, 2012; Savage, 2016). First, 84 % of killed pedestrians in Finland were reported to have committed suicide (264 suicides out of 311 pedestrians killed) (Silla and Luoma, 2012). Second, young people (in their 20 s and 30 s) exhibited elevated risks of deaths at non-crossings in metropolitan Chicago (Savage, 2016). Other factors such as male pedestrians, afternoon to night time, 20−29 years old, and intoxicated pedestrians were associated with greater frequency of fatalities (Nichols et al., 1994; Silla and Luoma, 2012; Savage, 2016). From a behavioral perspective, higher intentional deaths were observed in high income and lower population density areas (Savage, 2016). From an injury severity perspective, relatively few studies have focused on systematic analysis of injury severities sustained by pedestrians in train-pedestrian collisions at highway rail-grade crossings (HRGC) (Khattak and Tung, 2015; Zhao et al., 2016). Several factors such as higher train speeds, rail equipment striking pedestrian, female pedestrians, freight trains, night time crashes, and commercial type areas were associated with higher injury outcomes (Khattak and Tung, 2015; Zhao et al., 2016). However, the influence of advanced warning signs at HGRC on injury severity outcomes is not clear, with standard flashing lights associated with lower injury outcomes (Khattak and Tung, 2015), and warning bells associated with higher injury outcomes (Khattak and Tung, 2015). In addition, presence of significant unobserved heterogeneity in train-pedestrian crash data was observed (Zhao et al., 2016), which is typically ignored in analysis of train-related crashes.

2.2. Trespassing crashes

The extant literature also contains some studies focusing on analysis of fatally injured railroad trespassers (Pelletier, 1997; Van Houwelingen and Beersma, 2001; Savage, 2007) and injury severity outcomes of trespassing crashes (Wang et al., 2016; Zhang et al., 2018a). Methodologically, descriptive statistics have been widely used for analyzing frequency of trespassing crashes, see for example (Pelletier, 1997; Van Houwelingen and Beersma, 2001). Generally, fatalities among trespassers are observed to exhibit geographic and temporal clustering (Pelletier, 1997; Savage, 2007). Interestingly, independent of time of year, suicide rates at night dropped by 10 % while the suicide rates were almost constant over daytime hours (Van Houwelingen and Beersma, 2001). Train-involved trespassing crashes are more likely to result in a fatality. Thus, a systematic analysis of injury severity outcomes in such-like collisions is of high interest. In addition, Zhang et al. (2018a) noted that several countermeasures can have different outcomes at HRGC and non-crossings, and thus the two should be analyzed separately (Zhang et al., 2018a).

2.3. Research gap & objective

While previous studies provided valuable insights into the factors associated with injury outcomes, much needs to be further examined. Diverse statistical tools are used in the relevant literature for extracting relationships embedded in traditional “quantitative” trespassing crash data. However, context-dependent crash narratives are generally overlooked in trespassing injury severity analysis. Crash narratives may provide contextual crash-specific information that may not be present in the traditional tabular data and can enhance our understanding of the factors correlated with injury severity outcomes. Given this potential gap in the literature, the present study proposes a two-staged hybrid approach where advanced text analytic procedures are applied first to extract valuable information from crash narratives. The new variables derived from text mining are then used as inputs (explanatory variables) in formulation of advanced statistical models for trespasser injury outcomes. In doing so, the study develops a fundamental framework for extracting valuable information in “unstructured crash narratives” by using advanced text analytics techniques. Key trespasser safety-relevant

Table 1
Synthesis of Current Body of Knowledge.

Key Focus	Authors/Year	Objective	Methodology	Key Findings		Location & N
Train-bicyclist/ pedestrian crashes: Frequency analysis	Anne Silla & Juha Luoma (2012)	Frequency of fatalities in train-pedestrian collisions	Descriptive statistics	84 % of killed pedestrians included suicide. Male (↑) 20−29 years old (↑) Intoxicated (↑) Afternoon to nighttime (↑)	Densely populated areas (↑)	2005−2009 Finnish data N = 311
	Ian Savage (2016)	Frequency of fatal train-pedestrian collisions (intentional, unintentional, and apparent intentional)	Descriptive statistics & Negative Binomial models	Pedestrian fatalities constitute 84 % of all railroad fatalities Females represent higher proportion of completed suicides. Higher intentional deaths in high income areas & lower population densities	Public crossings & stations per square mile (↑) Middle aged people in 40 s (↑) Males (↑)	2004−2012 Northeastern Illinois N = 338
Train-bicyclist/ pedestrian crashes: Injury severity analysis	Aemal Khattak & Li-Wei Tung (2015)	Severity of pedestrian crashes at HRGC	Descriptive statistics & ordered probability models	Train speed (↑) Rail equipment struck pedestrian (↑) Female pedestrians (↑) Commercial type areas (↑)	Temperature (↑) Clear weather (↓) Number of traffic lanes at crossings (↓) Standard flashing lights at crossings (↓)	FRA's 2007−2010 highway rail grade crossing crash and inventory data N = 399
	S Zhao, A Iranitalab, A Khattak (2016)	Severity of pedestrian crashes at HRGC	Descriptive statistics & latent class binary logit models	Highlighting unobserved heterogeneity in train-pedestrian crash data Higher train speeds (↑) Freight trains (↑) Warning bells at crossings (↑)	Warning bells at crossings (↑) Commercial areas (↑) Night-time crashes (↑) Driver age (↑) Adverse weather (↑)	FRA's 2009−2014 highway rail grade crossing crash & inventory data
	Aemal Khattak & Z Luo (2011)	Counts of four types (V1-V4)* of pedestrian & bicyclist violations at gated HRGC	Descriptive statistics and Poisson regression	Both for bicyclist & pedestrian separately, V2 & V3 constituted about 90 % of total violations. No meaningful differences in the occurrence of gate-related violations Gate-related violations increased with presence of more individuals	Young children violations were more than older pedestrians & bicyclists Young children involved in excessive gate-related violations in the absence of older crossing users.	Data collection at M Street grade crossing in Fremont, Nebraska. N = 633
Trespassing crashes: Frequency analysis	Andrew Palletier, 1997	Fatally injured railroad trespassers	Descriptive statistics	Out of all railroad-related deaths, 57 % involved trespassers 82 % of incidents occurred in trespasser county of residence 78 % of trespassers killed while intoxicated	Fatalities among trespassers exhibit geographic & temporal clustering. Unmarried male (↑) 20−29 years of age (↑) Less than high school education (↑)	North Carolina, 1990−1994 N = 224
	Cornelis A.J. van Houwelingena & Domien G.M. Beersma, 2001	Frequency of train suicides	Descriptive statistics & visualizations	Independent of time of year, suicide rates at night drop to 10 % of their daytime values Train suicide rates are almost constant over daytime hours	80 % increase in suicide rates starting at about 2 h after sunset	Netherlands, 1980−1994 Sample size = 2830
	Ian Savage, 2007	Frequency of trespassing casualties	Descriptive statistics & time-series count data models	Railroad road miles (↑) Average daily # of trains (↑) Proportion of population between 15−44 years old (↑)	Real gross domestic product per capita (↓) Walking & running (↑) Laying, sleeping (↑) Riding bicycle (↑)	2001−2004 FRA data Sample size = 3628
Trespassing crashes: Injury Severity analysis	X Wang, J Liu, A Khattak, D Clarke, 2016	Severity of non-crossing rail-trespassing crashes	Descriptive statistics & Geographically weighted logitistic regressions	Highlighting unobserved heterogeneity in trespassing crashes ≤ 16 years old (↓)55−64 years old (↑)Darkness (↓) Summer time (↓)	Railway yard (↓) Climbing, jumping (↓) Riding, operating bicycle (↓) Laying and sleeping (↑)	2004−2013 FRA's non-crossing trespassing data N = 8797

(*continued on next page*)

Table 1 (*continued*)

Key Focus	Authors/Year	Objective	Methodology	Key Findings		Location & N
	M Zhang, A Khattak, J Liu, & D Clarke, 2018	Injury severity analysis of rail-pedestrian & cyclist trespassing crashes at HRGC and non-crossings	Descriptive statistics & Mixed-effects ordered logit model	HRGC: Laying/sleeping (↑) Running/walking (↑) Crossing/crawling (↑) Sitting/standing (↑) Youth (< = 16 years) (↓) Darkness (↑)Winter (↓) Median household income (↑)	Non-Crossings: Laying/sleeping (↑) Running/walking (↑) Crossing/crawling (↑) Sitting/standing (↑) Youth (< = 16 years) (↓) Darkness (↓)Winter (↓) Median household income (↑)	2006−2015 FRA's rail-pedestrian trespassing crash data N = 7,157

Notes: ↑ = increases crash frequency or injury severity; ↓ = decreases crash frequency or injury severity; HRGC = Highway Rail Grade Crossing; * V1 = passing under descending gates, V2 = passing under lowered gates, V3 = passing under ascending gates, V4 = passing around lowered gates between successive trains or during train stoppage at the crossing; N = Sample size.

thematic concepts embedded in unstructured crash narratives are identified along with the creation of new variables. A crosstabulation analysis of injury outcomes with new variables derived from text mining is conducted. To better understand the correlations between text-analysis derived key variables and injury outcomes, ordered discrete outcome models are estimated in the second stage. Finally, it is evident that not all but only a subset of factors influencing trespasser injury outcomes could be derived from narratives. In presence of unobserved factors, it may happen that any correlations established between injury outcomes and variables derived from crash narratives are not real but an outgrowth of the unobserved factors. Thus, while the main focus of this study is to highlight the value of advanced text analytic techniques in extracting meaningful information from crash narratives, fixed and random parameter ordered probit models are also estimated to address methodological issue related to unobserved heterogeneity in trespasser injury severity analysis.

3. Methodology

3.1. Conceptual framework

The key idea is to extract valuable information embedded in trespassing crash narratives by applying advanced systematic data mining techniques. Fig. 1 conceptualizes the overall methodology used for achieving the study objectives. Through the application of machine learning natural language processing "NLP" techniques, valuable crash-specific information can be extracted from crash narratives, which in turn can be fused with traditional quantitative trespassing crash data (such as injury outcomes) for gaining deeper and more meaningful insights regarding factors associated with injury outcomes (Fig. 1). Along this line, the new information derived from text mining can be used as input to statistical models for estimating injury outcomes in trespassing crashes. Given the higher vulnerability of non-motorists, the study focuses on non-motorist trespassing crashes at non-grade crossings.

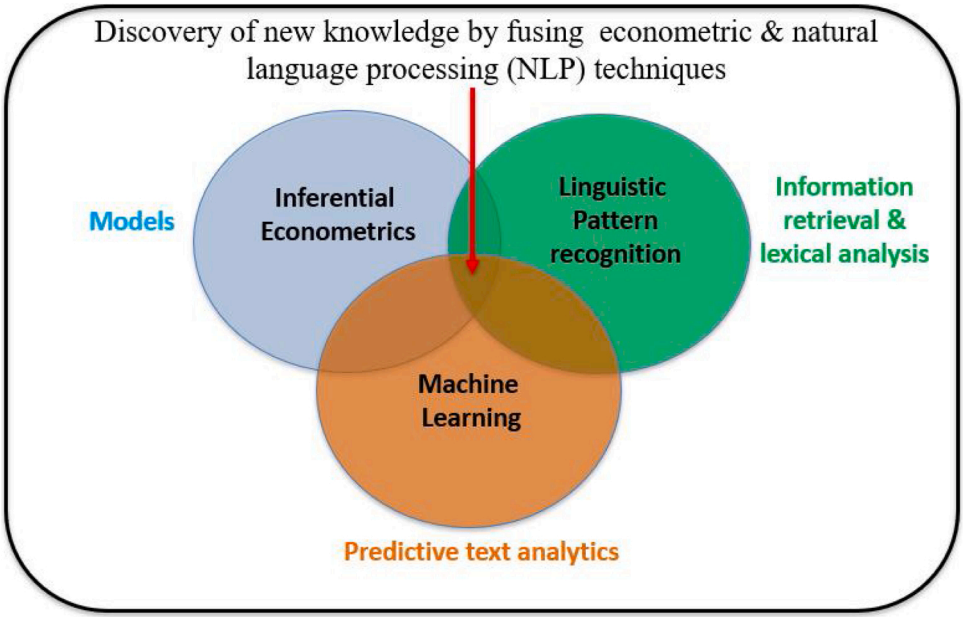


Fig. 1. Conceptual Framework.

3.2. Data

This study uses ten-year (2006−2015) trespassing injury data available on the Federal Railroad Administration's website (http://safetydata.fra.dot.gov/OfficeofSafety/default.aspx). The detail data description and descriptive statistics are presented later. In this section, we only discuss the crash narratives that are key input to the text data mining procedures. The FRA trespassing injury data (2006−2015) contains crash narratives generally reported by the crew member. A total of 12,242 crashes (including grade crossing and non-crossing) are considered, out of which 5,772 are non-crossing trespassing crashes, and 6,470 are pedestrian and bicyclist involved non-crossing trespassing crashes. This study focuses explicitly on pedestrian and bicyclist non-crossing trespassing crashes i.e., N = 6470. Out of 6470 such crashes, crash narratives are available for 41 % of crashes i.e. 2651 crashes. Injury severity data are also obtained for each of the 6470 crashes along with information on time of crash, location (region) of crash, age of trespasser, and year of crash. County level demographic and socioeconomic data are obtained from www.census.gov and linked with the trespassing crash data (details below).

3.3. Analysis method

The response outcome of specific interest is injury severity sustained by trespassers in collisions with trains (coded as minor, major, and fatal injury). Following standardized data analysis procedures, simple descriptive statistics and contingency tables are first developed to identify data anomalies (if any), conceptualize distributions, and spot patterns embedded in the data. Next, sophisticated text analytics techniques are used to extract meaningful information embedded in trespassing crash narratives. Specifically, advanced text analytics facilitate a systematic analysis of otherwise "random textual data" through application of statistical pattern learning techniques (Fig. 2a). A corpora of trespassing crash narratives is categorized and stemmed to spot unique concepts and entities for production of granular taxonomies (Fig. 2a). Fig. 2a summarizes the scheme of text mining framework employed in this study, whereas, Fig. 2b shows an example of actual trespasser crash narratives. Finally, through natural language processing of generated taxonomies and sentiment analysis of unstructured narratives, the overarching goal is to generate new meaningful knowledge regarding key risk factors associated with trespassers' injury outcomes. The methods used in this study pertaining to text analytics and econometric analysis (accounting for unobserved heterogeneity) are briefly described next.

3.3.1. Text analysis

3.3.1.1. Term frequency-inverse document frequency. The general way to encode text is to count the frequency of a term appearing in the documents, termed as term-frequency (TF) method. Note that terms with high frequency might not be necessarily important. Thus, the study weighs the terms with respect to local narrative, documents, or corpus. This study applies the term frequency-inverse document frequency (TF-IDF) weight method (Sebastiani, 2002), which is often used in information retrieval and text mining (Sebastiani, 2002). The importance of a term will increase proportionally to the number of times a term appears in the documents but offset by the frequency of the term in the corpus (Sebastiani, 2002). The TF-DIF weight consists of two parts: term frequency and inverse document frequency. Term frequency simply refers to the number of times a term appears in a document, while IDF measures how important the term is in the document. The formula for TF-IDF can be written as:

$$w_{ij} = tf_{i,j} + \log\left(\frac{N}{df_i}\right) \quad (1)$$

Where, $w_{i,j}$ is the weight of the term i in document j; $tf_{i,j}$ is the number of occurrences of term i in document j; N = total number of documents; and df_i = number of documents containing term i. To capture the context in which a word may be used, the study also analyzes "phrases" in addition to key words/topics. Before scanning for phrases, the maximum and minimum number of words in a phrase should be identified. This study extracts high-frequency phrases and add them to the currently active categorization dictionary. Note that the detailed trespassing crashes related categorization dictionaries are manually established by the users which are then used to calculate word frequencies (Provalis, 2014).

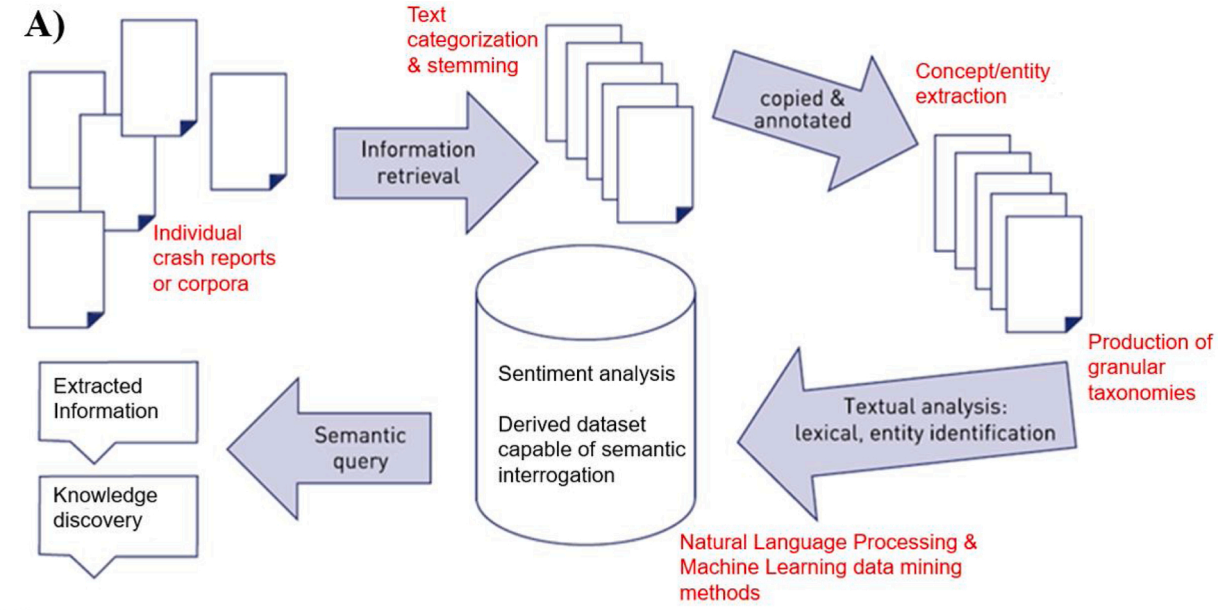
3.3.1.2. Factor analysis. The second element of text analysis is topic extraction, which attempts to uncover the hidden thematic structure of text collection. Based on the keywords obtained in frequency analysis, an inclusion dictionary is manually constructed to help software define more appropriate thematic concepts. The dictionary defines different categories and assign similar words to each category. This manual development of dictionary is used for concept/entity extraction; however, combination of natural language processing algorithms and statistical analysis is also used to complement and enhance the performance of manually developed dictionary (Provalis, 2014).

The main statistical procedure used for topic extraction is factor analysis. The key idea is to segment documents (in our case crash narratives) into smaller chunks and computing a word × segment frequency matrix. Once this matrix is obtained, a factor analysis with Varimax rotation is computed in order to extract small number of factors. In the factor analysis model, p represents the number of variables $(X_1, X_2, \ldots, X_p)$ and m represents the number of factors $(F_1, F_2, \ldots, F_m)$. This model assumes that each variable in the data is a linear function of m underlying factors with a residual variate. The mathematical formula can be written as follow (Jolliffe, 1986):

$$X_j = a_{j1}F_1 + a_{j2}F_2 + \ldots + a_{jm}F_m + e_j \quad (2)$$

Where $j = 1, 2, \ldots, p$; a_{jm} = factor loading of j^{th} variable on the m^{th} factor; e_j = specific or unique factor. This model shows the correlation of a variable wtih a particular factor. In the context under discussion, the variable refers to individual word obtained from a narrative whereas the factor refers to a key topic containing different words (variables). A high factor loading indicates high association between the factor and the variable. It is similar to the weight used in statistical regression, therefore, the factor loading determines the strength of correlation between the factor and variable (Jolliffe, 1986). As such, all words with a factor loading (or Eigen-values) higher than a specific criterion are then retrieved as part of an extracted topic. As opposed to hierarchical cluster analysis, topic modeling using factor analysis may result in a word being associated with more than one factor, a characteristic that may more realistically represent the polysemous nature of words, as well as the multiplicity of context of word usages (Jolliffe, 1986).

To control the topic modeling process (topic extraction), we use segmentation and loading procedures. Since the recorded crash narratives are open-ended fields, it may include several topics listed together as the narrative for a crash is recorded. Thus, we segment the topic extraction process by sentence to get a more precise extraction of the embedded topics in each crash narrative (Provalis, 2014). The maximum



B)

	A	B	C	D	E	F	G
1	X	Y	STATE	COUNTY	Countycode	InjSev	NARR4
2	-71.3918	42.4856	25	17	25C017	3	TRESPASSER DOVE INTO TRAIN MULTIPLE INJURIES DATE CORRECTED 9 21 2006
3	-80.4873	26.1523	12	11	12C011	3	TRAIN WAS APPROACHING HALLANDALE BEACH BLVD CROSSING WHEN EASTBOUND TRESPASSER, WEARING HEADPHONES, WALKED INTO PATH OF TRAIN. TRESPASSER WAS FATALLY INJURED. ALL CROSSING WARNING DEVICES WERE REPORTE D AS OPERATING PROPERLY
4	-86.1385	39.7817	18	97	18C097	3	CREW WAS MOVING WESTBOUND WHEN THEY CAME UPON SOMETHING IN MIDDLE OF TRACKS. THEY DETERMINED IT WAS A MALE SUBJECT LYING ON TRACKS. THEY IMMEDIATELY BEGAN BLOWING HORN REPEATEDLY AND PUT THE TRAIN I N EMERGENCY BUT TO NO AVAIL
5	-80.4873	26.1523	12	11	12C011	3	TRESPASSER STRUCK BY NORTHBOUND COMMUTER TRAIN. TRESPASSER STOPPED ON TRACKS AS TRAIN APPROACHED. TRESPASSER WAS FATALLY INJURED. CROSSING IS EQUIPPED WITH PEDESTRIAN GATES. GATES WERE REPORTED TO BE WORKING PROPERLY.
6	-81.8354	26.5774	12	71	12C071	2	A TRESPASSER WAS STRUCK AND INJURED BY TRAIN. THE TRESPASSER WAS SITTING BETWEEN THE RAILS, SHE STO OD UP AND FELL BACK DOWN.
7	-74.0744	40.9597	34	3	34C003	3	TRESPASSER SAT ON SOUTH RAIL AS TRAIN APPROACHED, MADE NO ATTEMPT TO MOVE.
8	-91.3414	42.8447	19	43	19C043	2	TRESPASSER WAS LAYING ON THE TRACK JUST EAST OF STREET CROSSING AND WAS STRUCK BY TRAIN. INDIVIDUAL EXPERIENCED SEVERE HEAD INJURIES AT THE SCENE OF THE ACCIDENT. THE TRAIN CREW WAS NOT DRUG/ALCOHOL TESTED AND THEY WERE NOT INJURED.

Fig. 2. A) Scheme of Text Mining Framework, B) Sample Crash Narratives.

number of topics (factors) to be extracted is set to 30 (Provalis, 2014). To retain a variable (a word in this case) in the final factor solution, we set the minimum factor loading to the recommended 0.4 value to balance the trade-off between representativeness and distinctiveness (Provalis, 2014). The text mining analysis is conducted in WordStat (version 8).

3.3.1.3. Term association. To conduct the text data-mining tasks related to clustering and classification, a notion of distance or similarity between documents is needed (Provalis, 2014). Many methods can be used to extract terms relevant to the same topics appearing in the same document or in the same paragraph or the same sentence (Provalis, 2014). The most commonly used heuristic rule is based on term co-occurrence (Salton, 1989). Given a set of n documents, an inverted term-document structure is created. This method assigns a vector form representing the weights of term in the document, i.e., $T_j = (d_{1j}, d_{2j}, \ldots, d_{nj})$, to each term. Therefore, this method can calculate the similarity between all useful term pairs. The similarity is measured by using a cosine function:

$$sim(T_j, T_k) = \frac{\sum_{i=1}^{n} d_{ij} d_{ik}}{\sqrt{\sum_{i=1}^{n} d_{ij}^2 \sum_{i=1}^{n} d_{ik}^2}} \quad (3)$$

Table 2 summarizes the key text analysis methods and associated outputs obtained from the analysis. Details pertaining to each method can be found in (Berry and Castellanos, 2008) for stemming & categorization process, (Breiger et al., 1975) and (Bridges, 1966) for concept and entity extraction, respectively.

Table 2
Text Analytic Methods and Outcomes.

Task	Procedure	Key outcome
Stemming & categorization process	Application of natural language processing routines to reduce inflected words to common stem or root firm. Categorizing specific words, word patterns, or phrases to a unique category, e.g. words such as "good", "excellent", or "satisfied" may be coded as instances of "positive evaluation".	Development of detailed specified inclusion dictionary related to trespassing crashes
Frequency/content analysis	Performing univariate frequency analysis on words or categories & updating the inclusion dictionary developed in previous task	Descriptive statistics, bar charts, pie charts
Concept/entity extraction	Application of machine learning algorithms & statistical procedures to uncover hidden thematic structure of text collection Mainly using factor analysis for concept/topic extraction	Most important topics, phrases, & Big Data concepts, categorization dictionary
Statistical analysis & visualizations	Conducting co-occurrence analysis for each specific concept/topic i.e. hierarchical cluster analysis & multidimensional scaling Correspondence & multivariate analysis to quantify high dimension correlations Automatically updating categorization dictionary	Correlations (within and between categories), Dendograms, 2D/3D concept maps, & proximity plots

3.3.2. Econometric modeling

Once the procedures described in the earlier section are applied, econometric models are estimated to link trespasser injury outcomes with variables derived from text mining and other controls. The key motivation is to identify statistically significant correlates (derived from text mining) of injury outcomes after accounting for multiple variables in a single specification. The response outcome is an ordinal variable with three levels: (1) minor injury, (2) major injury, and (3) fatal injury. Given the ordinal nature of the response outcome, an ordered probit modeling framework can be implemented (Duncan et al., 1998b; Quddus et al., 2002; Greene and Hensher, 2010; Jiang et al., 2017). Following the work presented in Greene and Hensher (2010), consider:

$$Y_i^* = \beta \boldsymbol{X}_i + \varepsilon_i, \tag{4}$$

Where: Y_i^* is the dependent variable coded as 0, 1, and 2 for trespasser crash i; β is the vector of parameter estimates; $\boldsymbol{X}_i$ is the vector of independent variables (mainly new variables derived from text mining of crash narratives); and ε_i is the normally distributed error term with density function $\varphi(.)$ and cumulative distribution $\varnothing(.)$. Given a specific injury severity outcome, an individual trespasser crash falls in category n if $\mu_{n-1} < y < \mu_n$. The observed injury outcome data, Y, are related to the underlying latent variable Y_i^* through thresholds μ_n, where $n = 1$ *and* 2 (Duncan et al., 1998a; Quddus et al., 2002). In this context, the ordered probability of each injury outcome for each crash i can be estimated as (Mannering et al., 2016; Khattak et al., 2020):

$$P(Y = n) = \varnothing(\mu_n - \beta \boldsymbol{X}) - \varnothing(\mu_{n-1} - \beta \boldsymbol{X}) \tag{5}$$

Where: $\mu_o = 0$, $\mu_1 < \mu_2$ are the two thresholds between which the ordered responses are estimated (Wali et al., 2017; Ahmad et al., 2019). To quantify the relationships of the independent variables derived from text mining on the probability of each injury-severity level of the trespasser, and especially on the intermediate levels, marginal effects can be computed as:

$$\frac{\partial Prob(Y = n)}{\partial \boldsymbol{X}} = -[\varphi(\mu_n - \beta \boldsymbol{X}) - \varphi(\mu_{n+1} - \beta \boldsymbol{X})]\beta, \; n = 0, 1, 2 \tag{6}$$

As the marginal effects can vary across the range of an explanatory variable, we report individual-level marginal effects averaged across all the observations (Train, 2009).

Table 3
Crosstab of trespasser injury severity with presence or absence of crash narrative.

			Injury Severity of Trespasser			Total
			1 (Minor)	2 (Major)	3 (Fatal)	
Narrative Dummy (1 if present)	0	Count	649_a	1000_b	2170_c	3819
		% within Narrative Dummy	16.9 %	26.2 %	56.8 %	100.0 %
		% within Injury Severity	68.5 %	59.3 %	56.5 %	59.0 %
	1	Count	298_a	686_b	1667_c	2651
		% within Narrative Dummy	11.2 %	25.8 %	62.8 %	100.0 %
		% within Injury Severity	31.4 %	40.6 %	43.45 %	41.0 %
Total		Count	947	1686	3837	6470
		% within Narrative Dummy	14.6 %	26.0 %	59.3 %	100.0 %
		% within Injury Severity	100.0 %	100.0 %	100.0 %	100.0 %

Pearson Chi-square (2 DOF) = 45.13; p value = 0.000
Likelihood-ratio Chi-square (2 DOF) = 46.18; p value = 0.012
Goodman and Kruskal's Gamma = 0.1352; ASE = 0.023
Kendall's τ_b = 0.0699; ASE = 0.012

Notes: Each subscript letter denotes a subset of injury severity of trespasser categories whose column proportions do not differ significantly from each other at a 95 % confidence level; Row and column percentages may not sum up to 100 due to rounding; ASE is asymptotic standard error.

This should be 3.3.2.13.4.1.1 Unobserved heterogeneity. The key focus of this study is to use emerging text mining procedures to extract new information from crash narratives that are often neglected in empirical analysis. As a next step, correlations between trespasser injury outcomes and text-mining derived key explanatory factors are examined. While a unique source of rich contextual information, there exists a real possibility that information on all the factors contributing to trespasser injury outcomes may not be captured in the crash narratives. These unobserved factors (potentially important) constitute what is referred to as "unobserved heterogeneity" in the safety literature (Mannering et al., 2016; Wali et al., 2018b, 2019b; Khattak and Fontaine, 2020), and which is reflective of the possibility of systematic variations in the correlations of explanatory factors across the sample population due to unobserved factors. Such unobserved factors can potentially introduce heterogeneity in the effects of observed explanatory factors (derived from crash narratives) on crash-injury severity. The statistical formulation shown in Eq. (4) does not account for unobserved heterogeneity, i.e., it assumes that the estimable correlations between explanatory factors and injury outcomes are fixed for all the trespassers. By allowing the exogenous explanatory factors to vary across individual crashes, more efficient, precise, and richer insights can be obtained (Wali et al., 2018b; Arvin et al., 2019; Wali et al., 2019a; Arvin and Khattak, 2020). To account for the unobserved heterogeneity in the ordered outcome probability process, random parameters can be introduced as (Sadri et al., 2013; Mannering and Bhat, 2014; Mannering et al., 2016; Wali et al., 2017; Saeed et al., 2019; Wali et al., 2020):

$$\beta_i = \beta + \mathbf{Y}\zeta_i \tag{4}$$

Where: β is now the mean of random parameter vectors, $\mathbf{Y}$ is the diagonal matrix with standard deviations for random parameters, and ζ_i is a randomly distributed random term that captures unobserved heterogeneity across crashes (Tay, 2015; Mannering et al., 2016). In particular, the distribution for ζ_i is specified by the analyst where different distributions can be tested (discussed later). The estimation proceeds with Maximum Simulated Likelihood procedures where Halton draws (compared to random draws) are used in the simulation process. In this study, 200 Halton draws are used for parameter estimation to produce accurate parameter estimates (Bhat, 2003). Regarding function form of the parameter density functions for random heterogeneity terms, we have tested normal, lognormal, triangular, uniform, and Weibull distributions – with normal distribution resulting in best fit.

4. Results – text mining

4.1. Descriptive statistics

Table 3 presents a crosstab of presence of crash narratives against observed injury outcomes - where the rows represent presence or absence of crash narratives and columns represent injury severity sustained by a trespasser. The crosstab in Table 3 shows the count (and percentage) of trespassers in each injury severity category for presence and absence of crash narrative. Several important insights can be obtained from Table 3. First, for crashes with narratives (Table 3), the percentages increase with increasing injury severity, suggesting that as injury severity sustained by a trespasser in a crash increases, the more likely that crash is to be reported with a crash narrative (denoted by different subscript letters in each row in Table 3). To determine if the differences in these percentages, and the observed upward trends are significant, we compare the column proportions for each row in Table 3. The results of statistical z-tests for comparing column proportions for each row reveals that the differences are significant, and the observed upward trend is real. In addition, the column proportion test assigns a subscript letter to each category of column variable. If a pair of values is significantly different, the values have different subscript letters assigned to them. Second, the distributions of injury severity against presence (and absence) of crash narratives are almost similar with minor differences. Third, chi-square test suggests that there is a positive (statistically significant) association between presence (and absence) of narrative and injury severity sustained by a trespasser[1] .

4.2. Frequency/Content analysis

First, we perform a simple univariate frequency analysis to obtain insights about the trends embedded in the crash narratives. The frequency analysis procedure reports descriptive statistics for all key words that appeared in at least 30 cases (where case is our unit of analysis, i.e. each row of trespassing crash). Table 4 provides descriptive statistics for 14 random keywords that have a case occurrence of greater than 30. While other columns in Table 4 are self-explanatory, the last column (TF.IDF) provides a more accurate measure of "keyword frequency" while balancing representativeness and discrimination (Berry and Castellanos, 2008). A keyword with high frequency "within" a narrative and lesser frequency "among" all narratives is more representative, but more discriminative as well (Berry and Castellanos, 2008). Specifically, keyword frequency is weighted by the inverse document frequency, i.e. the frequency of an item is adjusted to take into account the number of crash narratives containing this item, rather than the frequency of an item/word within one crash narrative (Berry and Castellanos, 2008). A keyword that appears more frequently within each narrative as well as in different crash narratives will have high TF.IDF scores (Table 3).

4.3. Concept/Entity extraction

Table 5 presents the results of factor analysis by summarizing the key topics embedded in crash narratives. To uncover the hidden thematic structure of text collection, factor analysis is used to conduct topic extraction (as explained in section 3.3.1). The topics are categorized into different categories such as:

- Trespasser–related.
- Injury-related.
- Train-related.
- Medical-related.
- Location and Advance warnings related topics.

For example, topic "Audible; Warning Signs" under warnings related topics (Table 5) is the second highest factor with an Eigenvalue of 3.47, and which explains 1 % of the variance in crash narratives. Likewise, topic "Wearing Headphones" under trespasser category has an Eigenvalue of 1.79 explaining around 0.6 % of the variance in narratives (Table 5). Finally, based on the results of content analysis and concept extraction, 13 variables (unique to crash narratives and not present in tabular form data) are constructed.[2] The descriptive statistics of new

[1] We could not find relevant details in FRA guidelines about recording processes of narratives. Given the limitations imposed by standardized reporting forms, the FRA Guide for Preparing Accident/Incident Reports strongly encourages collection of crash narratives. However, it does not mention if collection of narratives is mandatory (FRA, 2011). In its guidance, the FRA suggests the following items (to name a few) to be covered in narrative discussion when appropriate: drug/alcohol involvement, cause, diesel fuel tank (leakage, etc.), hazardous materials, train information, and other railroad equipment (Chapter 7, Page 16 in (FRA, 2011)).

[2] To understand the nature of a crash narrative more accurately, we extract idioms and phrases based on the topics extracted via factor analysis (shown in Table 5). In particular, the text corpus related to each crash narrative is scanned and the most frequent phrases or idioms are identified. To reduce redundancy, short idioms that are typically part of larger sentences are automatically removed from the subsequent categorization dictionary (Provalis, 2014). These newly extracted phrases serve as the basis for creation of new variables used in subsequent statistical modeling.

Table 4
Descriptive statistics for 14 random keywords.

	FREQUENCY*	% SHOWN	% PROCESSED	% TOTAL	NO. CASES	% CASES	TF ˙IDF
TRAIN	2482	14.88%	9.92 %	5.33 %	2124	32.83 %	1200.7
TRESPASSER	2390	14.33%	9.56 %	5.13 %	2063	31.89 %	1186.4
AGE	92	0.55 %	0.37 %	0.20 %	92	1.42 %	169.9
HOSPITAL	90	0.54 %	0.36 %	0.19 %	87	1.34 %	168.4
HORN	80	0.48 %	0.32 %	0.17 %	79	1.22 %	153.1
POLICE	66	0.40 %	0.26 %	0.14 %	60	0.93 %	134.2
FEMALE	53	0.32 %	0.21 %	0.11 %	47	0.73 %	113.4
IMPACT	52	0.31 %	0.21 %	0.11 %	52	0.80 %	108.9
JUMPED	52	0.31 %	0.21 %	0.11 %	52	0.80 %	108.9
TRANSPORTED	49	0.29 %	0.20 %	0.11 %	48	0.74%	104.4
INTOXICATED	44	0.26 %	0.18 %	0.09 %	44	0.68%	95.4
RAN	44	0.26 %	0.18 %	0.09 %	43	0.66%	95.8
LOCATION	43	0.26 %	0.17 %	0.09 %	43	0.66%	93.6
MIDDLE	43	0.26 %	0.17 %	0.09 %	42	0.65%	94.1

Note: *Frequency indicates the "case occurrences" of keywords. For example, word "train" is used in 2482 crash narratives.

Table 5
Results of Factor Analysis for Trespasser-Related, Injury-Related, Location, Warnings, Train-Related, and Medical-Related key Topics.

Category	Topic	Text Phrase	Eigenvalue	% VAR	Frequency	Cases	% Cases
Trespasser related	Old Male; Laying	Old; Male; Laying; Hit; Oncoming; Tracks	2.36	0.81	1001	648	10.02 %
	Body; Deceased	Body; Deceased; Discovered; Found; Reported; Police	1.83	0.6	291	191	2.95 %
	Wearing Headphones	Headphones; Wearing	1.79	0.6	31	22	0.34 %
	Attempting To Climb; Station Platform	Climb; Attempting; Platform; Station	1.68	0.61	194	151	2.33 %
	Child; Adult	Child; Adult; Minor	1.57	0.56	32	26	0.40 %
	Attempted To Cross	Cross; Attempted; Trying	1.53	0.6	112	91	1.41 %
	Age Unknown	Age; Unknown; Verified	1.7	0.6	244	158	2.44 %
	Riding; Railroad Property	Riding; Property; Atv; Railroad	1.95	0.66	108	81	1.25 %
	Standing	Standing; Tie	1.51	0.57	82	77	12.10 %
Injury related	Fatally Injured; Track At Milepost	Injured; Fatally; Milepost; Eastbound; Track; Gauge; Reportedly; Located; Westbound; Single	6.16	1	1468	783	4.16 %
	Fatal Injuries	Injuries; Multiple; Fatal; Expired; Sustained; Sustaining	3.03	0.84	497	269	3.15 %
	Trauma; Died	Trauma; Died; Due; Scene; Dead	1.55	0.58	94	76	1.17 %
	Head	Head; Treatment; Lacerations; Arms	1.54	0.57	28	26	0.40 %
	Left; Severed	Left; Severed; Arm; Leg; Hand; Foot; Sustained; Broken	2.24	0.69	211	144	2.23 %
Location	Main Line	Main; Line; Tons	1.85	0.67	238	166	2.57 %
	Grade Location	Grade; Location; Crossing	2	0.61	274	205	3.17 %
	Crossing	Crossing; Street; Avenue	1.63	0.62	92	75	1.16 %
Warnings	Audible; Warning Signals	Audible; Signals; Warning; Braking; Sounded; Immediately; Devices; Emergency	3.47	1	293	176	2.72 %
	Sounded Horn; Started Blowing	Horn; Blowing; Started; Bell; Engineer; Sounding; Whistle; Blew; Sounded; Emergency	2.17	0.78	315	179	2.77 %
	Unable; Prior To Impact	Unable; Prior; Stop; Impact; Emergency	2.5	0.94	220	116	1.79 %
	Lead Engine	Lead; Unit; Engine; Causing	2.03	0.66	101	75	1.16 %
	Tools/Machinery	Machinery; Tools; Physical; Activity; Event	2.92	0.96	150	65	1.00 %
Train related	Train Struck	Struck; Train; Trespasser	1.88	0.82	6581	2412	37.28 %
	Rail	Rail; West; East; Side; Head; Mainline; Observed; Feet	2.74	0.95	1181	701	10.83 %
	Operating; End	Heard; Operating; End; Southbound	2.08	0.69	165	139	2.15 %
Medical related	Hospital; Transported	Hospital; Transported; Removed; Treatment; Local; Emergency Medical Services; Medical; Brook	2.21	0.72	272	158	2.44 %

Notes: %VAR shows the percentage of variance explained by each topic. Note that the smaller the text phrases in each topic, the lower the percentage of variance explained, at least theoretically; Eigenvalue are calculated and used in deciding how many "text phrases" to extract for a specific topic. In factor analysis, "the topic" with largest Eigen value has the most variance, and topics with Eigen values greater than 1.00 are traditionally considered worth analyzing.

Table 6
Descriptive statistics of new variables derived from text mining of crash narratives.

Variables	N	Mean	Frequency	SD	Min	Max	Minor	Major	Fatal
Unable to stop prior to impact	2,651	0.044	116	0.205	0	1	12	16	88
Trespasser transported to hospital	2,651	0.044	117	0.205	0	1	20	72	25
Train put into emergency braking	2,651	0.051	136	0.221	0	1	21	24	91
Advance warnings, horns, whistles to trespassers	2,651	0.091	240	0.287	0	1	34	57	149
Male trespasser	2,651	0.072	192	0.259	0	1	24	52	116
Female trespasser	2,651	0.018	47	0.132	0	1	9	16	22
Trespasser on mainline	2,651	0.022	59	0.148	0	1	5	14	40
Trespasser jumped in front of train	2,651	0.081	215	0.273	0	1	26	54	135
Intoxicated/Under alcohol influence	2,651	0.020	53	0.140	0	1	11	33	9
Confirmed suicide attempt	2,651	0.010	26	0.099	0	1	4	3	19
Trespasser wearing headphones and talking on cellphone	2,651	0.008	22	0.091	0	1	3	7	12
Commuter train	2,651	0.007	18	0.082	0	1	1	1	16
Sleeping/sitting in rail gauge	2,651	0.018	47	0.132	0	1	8	13	26
Trespasser lying*	2,651	0.035	94	0.185	0	1	14	12	68
Trespasser walking*	2,651	0.052	139	0.223	0	1	12	38	89
Trespasser laying*	2,651	0.025	65	0.155	0	1	12	12	41

Notes: N is sample size; (*) variables likely to be present in traditional crash data as well.

variables derived from text mining of crash narratives are presented in Table 6. The last three columns in Table 6 provide distribution of key variables across injury severity levels, providing interesting insights. Finally, Fig. 3 presents the distribution of newly created variables across different injury severity levels.

5. Results – statistical models

5.1. Descriptive statistics

Before presenting the results of heterogeneity-based ordered discrete outcome models, we present the descriptive statistics of key control variables in Table 7. Note that these variables extracted from the traditional tabular data are used as controls besides the key text mining derived variables listed in Table 6. Referring to Table 7, crash narratives are available for around 41 % of the crashes. The mean trespasser age for this dataset is around 37 years. Note that age information is missing in 687 trespassing crashes. To avoid losing estimation sample and to estimate the effect of missing age, the missing values are replaced with 0 and subsequently a dummy indicator is created for missing age which is 1 if age was replaced (with 0) and 0 otherwise (see descriptive statistics in Table 7). Around 18 % of the crashes occurred during late night (12 AM – 4 AM) whereas another 39 % occurred during evening (4 P M – 8 PM) or nighttime (8 PM – 12 AM) (Table 7). The trespasser crashes are also widely distributed across the eight FRA regions – with most of the crashes (21.9 %) observed in Region 3 (Table 7). Finally, information on county level census attributes is also extracted and linked with the trespasser crash data. The average population per square mile is around 1253 with substantial variation across the counties. On average, around 16.04 % of the county residents live in poverty, whereas, the average annual income is USD 53,795. Other statistics are provided in Table 7.

5.2. Statistical modeling

Once unique information is extracted from the narratives through text mining procedures, statistical models are estimated for injury outcomes as a function of new variables derived from text mining and other controls (see Table 6 and 7). The discrete outcome statistical models are estimated using the entire data, i.e., N = 6470 which includes both cases with and without narratives. An indicator variable for presence of crash narratives is included in all the model specifications to quantify the correlation between presence of crash narratives and injury outcomes,[3] [4]. All the models are derived from a systematic process to include most important variables (available in the data) based on intuition, statistical significance, and specification parsimony. First, a fixed parameter ordered probit model was estimated where all the variables derived from text mining as well as control variables were included as explanatory factors (Table 8). A total of 39 parameters were estimated where several text-mining derived variables were found statistically significant. A total of 14 variables were found statistically significant at 90 % level of confidence (see Table 8). Next, a fixed parameter ordered probit model was estimated where the statistically insignificant variables were removed from the model specification. The results of the ordered probit model with statistically insignificant variables (based on 90 % confidence level) removed are shown in the second panel of Table 8. However, variables that were statistically insignificant in the fixed parameter model but exhibited significant heterogeneity in magnitude of effects in the random parameter model were retained. Doing so resulted in a better

[3] While the control data are available for all the 6470 crashes (see Table 7), the new variables derived from text mining are based on crashes with narratives available (i.e., N = 2651 crashes). As discussed earlier, the new explanatory variables derived from text mining are dummy coded as 0/1 for the crashes with narratives (Table 6). For crashes with no narratives (N = 3819 crashes), the corresponding fields for text mining-based variables are blank. Thus, to estimate the models using the entire data, we recode the text mining-based variables as '2' for cases where there was no narrative (N = 3819 crashes). Ultimately, the recoded categorical text mining based variables (0 – indicator variable not present in a crash narrative, 1 – indicator variable present in a crash narrative, and 2 – not applicable since no narrative is present for the crash) are tested in the discrete outcome model along with control variables. However, note that in estimation the recoded text mining-based variables (coded as 0, 1, and 2) automatically reduce to dummy variables (0/1) since the information captured in '2' is also perfectly captured in the indicator for presence of crash narratives (see Table 7).

[4] Note that text mining based new variables (e.g., "trespasser transported to hospital" or "Male trespasser" crash) cannot be set as "absence of crash narratives". Indicator variable for "absence of crash narratives" states if a crash has a narrative or otherwise (41% of crashes had narratives or N = 2,651 out of 6,470 crashes) (Table 7). On the other hand, "trespasser transported to hospital" (or any other text-mining based variable) dummy variable states if this variable was present in a particular crash narrative – among the crashes which had narratives (N = 2,651 crashes). For instance, for the crashes which had narratives (N = 2,651), 4.4% involved a trespasser transported to hospital (Table 7).

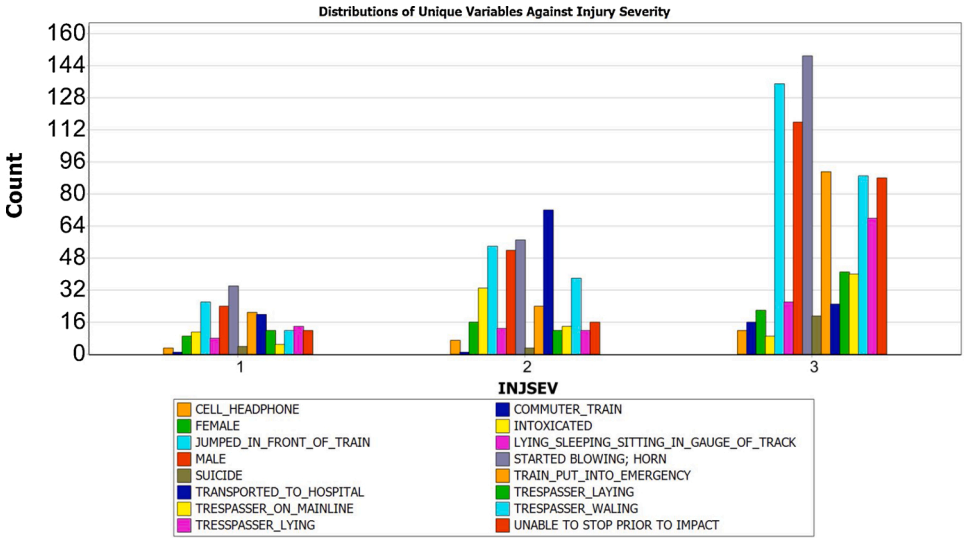


Fig. 3. Distribution of key variables across injury severity (1 – Minor injury, 2 – Major injury, 3 – Fatal injury).

Table 7
Descriptive Statistics of Control Variables Used in Subsequent Statistical Modeling.

Control Variables	N	Mean	SD	Min	Max
Presence of crash narratives (1/0)	6470	0.410	0.492	0	1
Trespasser age (in years)	5783	36.991	15.416	1	97
Age: Less than 19 years old	6470	0.120	0.176	0	1
Age: (19–35 years]	6470	0.321	0.467	0	1
Age: (35–65 years]	6470	0.421	0.493	0	1
Age: Greater than 65 years old	6470	0.031	0.174	0	1
Missing age	6470	0.106	0.308	0	1
Time of day					
Late night (12AM - 4 AM)	6470	0.179	0.384	0	1
Early morning (4 AM - 6 AM)	6470	0.055	0.229	0	1
Morning (6 AM - 9 AM)	6470	0.089	0.285	0	1
Midday (9AM - 2 PM)	6470	0.194	0.395	0	1
Afternoon (2 PM - 4 PM)	6470	0.089	0.284	0	1
Evening (4 PM - 8 PM)	6470	0.201	0.400	0	1
Night (8 PM - 12 AM)	6470	0.193	0.395	0	1
Region*					
Region 1	6470	0.101	0.302	0	1
Region 2	6470	0.143	0.350	0	1
Region 3	6470	0.219	0.414	0	1
Region 4	6470	0.119	0.324	0	1
Region 5	6470	0.134	0.341	0	1
Region 6	6470	0.048	0.213	0	1
Region 7	6470	0.181	0.385	0	1
Region 8	6470	0.055	0.228	0	1
County sociodemographic					
Percentage of residents with high school education	6470	85.276	6.090	52.90	97.50
Percentage of residents with college education	6470	27.327	10.344	5.10	71.70
Income (in thousands of USD)	6470	53.795	14.257	20.972	110.292
Percent in poverty	6470	16.041	5.558	3.90	40.70
Area (square miles) (in100 s)	6470	16.308	37.174	0.090	536.248
Population per square mile (in 100 s)	6470	12.530	32.356	0.003	694.675

Notes: (1) The control variables are obtained from the tabular crash data and/or linked from the U.S. Census. These variables are used as controls in subsequent statistical models in addition to the key variables derived from text mining shown in Table 6. (2) N is the entire sample size of pedestrian and bicyclist non-crossing trespassing crashes (including 2651 crashes with narratives present). (*) For definitions of the eight regions, see https://railroads.dot.gov/divisions/regional-offices/regional-offices.

data fit as indicated by a lower AIC (indicating better fit) of the model (with statistically insignificant variables removed) compared to the model with all explanatory factors included (see Table 8).

Finally, as discussed in section 3.3.2, unobserved heterogeneity is suspected due to the presence of several unobserved factors that could influence the injury outcomes but are not available in the data – and in presence of which the correlations obtained in the fixed parameter ordered probit models could be biased and inefficient. To this end, a random parameter ordered probit model was estimated which allowed the coefficients on explanatory variables to vary across individual trespassing crashes. In the random parameter model, a parameter estimate for a specific independent variable is treated as random if the estimable β's exhibited a statistically significant standard deviation or exhibited statistically significant mean as well as standard deviation (for details on this – see (Wali et al., 2018a; Boggs et al., 2020)). Referring to Table 8, a total of five text mining derived variables are found to be normally distributed random parameters suggesting that the associations of these variables with injury outcomes vary across individual crashes. Also, compared to fixed parameter counterparts, the random parameter ordered probit model resulted in substantial improvements in goodness of fit – the AIC for random parameter model is the lowest among all the alternative specifications (see Table 8). The likelihood-ratio test also provides conclusive evidence in favor of random parameter ordered probit model after accounting for the additional degrees of freedom (Table 8). Also, compared to the fixed parameter model with 14 statistically significant variables, a total of 19 variables are found to exhibit statistically significant fixed effects or statistically significant heterogeneity in magnitudes of correlations in the random parameter model. Altogether, these findings emphasize the importance of accounting for unobserved heterogeneity in injury outcome models where text-mining derived variables are included as key predictors. To better interpret the results, marginal effects are provided for the best fit random parameter model (Table 8).

6. Discussion

In this section, the key results obtained from text mining and statistical models for injury outcomes (with text mining derived variables as key predictors) are briefly discussed.

6.1. Text mining

In the first stage, text mining of non-crossing trespassing crash

Table 8
Estimation Results of Fixed and Random Parameter Injury Outcome Models Including New Variables Derived from Text-Mining.

Variables	Model 1 - Fixed Parameter Ordered Probit		Model 2 - Fixed Parameter Ordered Probit (statistically insignificant variables removed)		Model 3 - Random Parameter Ordered Probit Model				
	β	z-score	β	z-score	β	z-score	Minor	Major	Fatal
Constant	0.682	2.22	0.673	9.07	0.677	8.92	—	—	—
Threshold 1	0.834	46.17	0.833	46.16	0.855	45.99	—	—	—
Text-mining derived variables									
Unable to stop prior to impact	0.341	2.5	0.342	2.52	0.374	2.55	−0.0662	−0.0664	0.1326
Trespasser transported to hospital	−0.822	−7.66	−0.816	−7.85	−0.832	−5.95	0.2525	0.0597	−0.3123
Train put into emergency braking	−0.053	−0.41	−0.086	−0.73	0.031	0.24	−0.0067	−0.0051	0.0118
Scale parameter of normally distributed random parameter	**NA**	**NA**	**NA**	**NA**	***0.723***	***5.42***	—	—	—
Advance warnings, horns, whistles to trespassers	−0.072	−0.78	—	—	—	—	—	—	—
Male trespasser	0.008	0.09	—	—	—	—	—	—	—
Trespasser on mainline	0.221	1.3	0.217	1.29	0.251	1.33	−0.0476	−0.0435	0.0912
Trespasser jumped in front of train	0.025	0.28	—	—	—	—	—	—	—
Confirmed suicide attempt	0.059	0.23	0.062	0.24	0.446	1.2	−0.0753	−0.0797	0.15507
Scale parameter of normally distributed random parameter	**NA**	**NA**	**NA**	**NA**	***1.243***	***2.83***	—	—	—
Trespasser wearing headphones and talking on cellphone	−0.131	−0.51	—	—	—	—	—	—	—
Commuter train	0.824	2.14	0.852	2.22	0.874	2.56	−0.1136	−0.1537	0.2673
Sleeping/sitting in rail gauge	−0.298	−1.69	−0.282	−1.62	−0.252	−1.42	0.0621	0.035	−0.0971
Trespasser lying*	0.087	0.62	0.085	0.61	0.618	2.87	−0.0945	−0.1114	0.206
Scale parameter of normally distributed random parameter	**NA**	**NA**	**NA**	**NA**	***1.428***	***6***	—	—	—
Trespasser walking*	0.101	0.93	—	—	—	—	—	—	—
Trespasser laying*	−0.088	−0.58	−0.105	−0.69	0.139	0.73	−0.0281	−0.0234	0.0515
Scale parameter of normally distributed random parameter	**NA**	**NA**	**NA**	**NA**	***1.143***	***5.06***	—	—	—
Control Variables									
Presence of crash narratives (1/0)	0.188	5.18	0.192	5.69	0.193	8.92	−0.041	−0.031	0.072
Age: Less than 19 years old (base)	—	—	—	—	—	—	—	—	—
Age: (19–35 years]	0.095	1.91	0.094	1.91	0.109	2.16	−0.0232	−0.0176	0.0409
Scale parameter of normally distributed random parameter	**NA**	**NA**	**NA**	**NA**	***0.268***	***10.01***	—	—	—
Age: (35–65 years]	0.201	4.16	0.204	4.26	0.206	4.22	−0.0441	−0.0335	0.0776
Age: Greater than 65 years old	0.356	3.65	0.362	3.72	0.369	3.63	−0.0661	−0.0653	0.1314
Missing age	−0.041	−0.65	−0.036	−0.58	−0.043	−0.71	0.0095	0.0068	−0.0164
Time of day									
Early morning (4 AM - 6 AM)	0.168	2.28	0.145	2.17	0.159	2.36	−0.0321	−0.0268	0.0589
Morning (6 AM - 9 AM)	0.114	1.85	0.091	1.7	0.085	1.63	−0.0179	−0.0141	0.0321
Midday (9AM - 2 PM)	0.061	1.24	—	—	—	—	—	—	—
Afternoon (2 PM - 4 PM)	−0.007	−0.12	—	—	—	—	—	—	—
Evening (4 PM - 8 PM)	0.026	0.53	—	—	—	—	—	—	—
Night (8 PM - 12 AM)	0.021	0.41	—	—	—	—	—	—	—
Region									
Region 2	−0.017	−0.26	—	—	—	—	—	—	—
Region 3	0.033	0.48	—	—	—	—	—	—	—
Region 4	−0.008	−0.12	—	—	—	—	—	—	—
Region 5	−0.086	−1.19	−0.091	−2.01	−0.095	−2.15	0.0215	0.0148	−0.0364
Region 6	−0.129	−1.44	−0.137	−1.97	−0.142	−2.14	0.033	0.0214	−0.0545
Region 7	0.068	1	0.062	1.49	0.062	1.46	−0.0132	−0.0101	0.0234
Region 8	−0.005	−0.07	—	—	—	—	—	—	—
County sociodemographic									
Percentage of residents with high school education	−0.0006	−0.16	—	—	—	—	—	—	—
Percentage of residents with college education	−0.005	−2.27	−0.005	−2.66	−0.005	−2.79	0.0013	0.0009	−0.0022
Income (in thousands of USD)	0.007	3.87	0.006	4.06	0.006	4.21	−0.0015	−0.0011	0.0026
Area (square miles) (in100 s)	−0.0002	−0.52	—	—	—	—	—	—	—
Population per square mile (in 100 s)	−0.00008	−0.16	—	—	—	—	—	—	—
Summary Statistics									
Sample Size (N)	6470	6470	6470						
Number of Parameters (K)	39	23	28						
Loglikelihood at zero	−6091.92	−6091.92	−6091.92						
Loglikelihood at convergence	−5993.15612	−5995.8842	−5988.53						
AIC	12064.3	12037.8	12033.06						
Likelihood Ratio Test	*Model 2 Vs. Model 3*								
LR= -2[LL(β_{fixed}) – LL(β_{random})]	14.7084								
Degrees of freedom	5								
Critical $\chi^2_{0.050,5}$ (95% level of confidence)	11.07								

Notes: (*) These variables may be available in the traditional tabular data as well; NA indicates Not Applicable since the variable is considered a fixed parameter in the corresponding model; (—) indicates variables that are dropped from the subsequent models (Models 2 and 3) (hence, no estimable parameters available); All variables are dummy indicators unless otherwise noted.

narratives provided an overview of the key themes/concepts embedded in narrative data, as can be seen in Tables 4 and 5. For a clearer understanding of semantic themes, the key concepts embedded in narratives are structured into different categories such as trespasser, injury, train, medical, and location related factors. Next, unique variables (13 variables) are extracted from the thematic concepts that are not present in tabular crash data and can complement traditional empirical analysis. Table 6 provides some key insights regarding unique variables and their distributions across injury severity (last three columns in Table 6). Table 6 shows that trespassers with confirmed suicide attempts are more likely to be receiving fatal injuries; 73 % of trespassers who attempted suicide received fatal injuries compared to 15.3 % and 11.7 % of trespassers (who attempted suicides) receiving minor and major injuries, respectively (Table 6). Another important variable extracted is whether trespasser is wearing headphones or talking on cellphone. A total of 22 trespasser crashes are identified where trespasser is reported to be wearing headphones or talking on cellphone. Of such trespassers, 55 % received fatal injuries, 31.8 % receiving major injuries, and 13.2 % receiving minor injuries (Table 5). The above two findings show that trespassers attempting suicides or wearing headphones or talking on cellphone are more likely to sustain fatal injuries.

Trespassers that jumped directly in front of a train and trespassers hit by commuter train are also more likely to receive fatal injuries; 62.7 % of trespassers who jumped in front of train and 88.8 % of trespassers who are hit by train received fatal injuries. Likewise, if a train is unable to stop prior to impact and if a train is put into emergency braking, trespassers are more likely to receive higher-order injuries. Advance warnings by train engineer are more likely to be associated with fatal trespasser injuries. Regarding gender, more than 60 % and 45 % of male and female trespassers received fatal injuries. Based on distributions in Table 6 and Fig. 3, trespassers transported to hospitals have relatively more major injuries and lesser fatal injuries. This finding may indicate the effectiveness of timely medical care, i.e., providing timely medical care to trespassers who received major injuries can reduce the likelihood of major injuries turning into fatal ones. Finally, 62.2 % and 15.0 % percent of intoxicated or trespassers under alcohol influence received major and fatal injuries, respectively.

6.2. Heterogeneous correlations between text-mining derived variables and injury outcomes

In the second stage, heterogeneity-based statistical models are estimated to link injury outcomes with text mining derived key variables. The discussion of key findings is based on the random parameter ordered probit model given its relatively best fit. Several important insights can be obtained from the best-fit random parameter model. A trespasser having a confirmed suicide attempt had a higher likelihood of receiving fatal injury, i.e., the probability of fatal injuries increased by 0.155 units if the trespasser had a confirmed suicide attempt (see marginal effects in Table 8). However, with a mean of 0.446 and standard deviation of 1.243, this variable is found to be a normally distributed random parameter – suggesting a positive correlation between confirmed suicide attempt and injury outcomes for 64.01 % of the trespassers and negative for the rest (Table 8). Note that the suicide attempt related variable was statistically insignificant in the fixed parameter counterparts (see Table 8). Likewise, if a train was unable to stop prior to impact, the trespasser was more likely to receive fatal injuries – with the probability of fatal injury increasing by 13.2 percentage points in case a train did not stop prior to impact (Table 8). Contrarily, if a trespasser was transported to hospital, the probability of fatal injury decreased by 31.2 percentage points (Table 8). This finding confirms the observation obtained in the earlier section – suggesting that timely medical care could be effective in reducing fatal injuries. The best fit random parameter ordered probit model also sheds light on the correlation between emergency braking and trespasser injury outcomes. On average, the probability of fatal injury increased by 0.0118 units if emergency braking was activated. However, this finding does not imply causation since an engineer can activate emergency braking in cases where an unavoidable negative safety outcome is anticipated. In that sense, emergency braking is not causing the fatal injury outcomes. Notably, the emergency braking related variable is found to be a normally distributed random parameter with substantial heterogeneity observed not only in the magnitude but direction of correlations as well. In particular, a positive correlation is observed for 51.7 % of the cases whereas a negative correlation is observed for the rest (see the random parameter mean and standard deviation for this variable in Table 8). The fact that the correlations between emergency braking and injury outcomes are negative for around 48.2 % of the cases may indicate the usefulness of emergency braking in avoiding an otherwise fatal injury outcome. Again, such deeper insights cannot be obtained from traditional fixed parameter models. Likewise, in line with previous literature (Zhang et al., 2018a), more severe injury outcomes were observed for trespassers laying (on mainline) or if the trespasser was hit by a commuter train. However, the trespasser laying related variable was found to be normally distributed random parameter suggesting that the correlations vary across individual crashes (Table 8). Finally, variables related to trespasser wearing headphones/talking on cellphone, trespasser walking, and sleeping/sitting in rail gauge were found to be statistically insignificant.[5]

The best fit model also quantifies correlations between control variables and injury outcomes. Intuitively, older trespassers were more likely to receive fatal injuries. To account for potential non-linearities, the age variable was dummy coded into different categories (Zhang et al., 2018a). Compared to young trespassers ($\leq$ 19 years old), older trespassers - (19–35 years old], (35–65 years old], and greater than 65 years old – were more likely to receive fatal injuries. Among the older riders, the β coefficients increase with age indicating the greater vulnerability of older riders to receiving fatal injuries. Besides the nonlinear associations, significant unobserved heterogeneity is also observed in the associations of age with injury outcomes (Table 8). Note that the variable indicating missing values for age was statistically insignificant - suggesting that the missing values for age do not exhibit a statistically significant correlation with injury outcomes. Regarding time of day, trespassing crashes during early morning/morning were more likely to include a fatal injury. Finally, in line with the literature (Savage, 2016; Zhang et al., 2018a), counties with greater percentage of college graduated residents were less likely to have fatal injuries, whereas, counties with higher income were more likely to have fatal injuries.

7. Conclusions

Injury severity is a critical aspect of railway safety-improvement programs. Traditionally, quantitative crash data are used for guiding safety policies pertaining to highway rail-grade, and/or trespassing crashes. However, crash narratives, which are typically overlooked in trespassing injury severity analysis may provide contextual crash-specific information that can enhance our understandings of the factors influencing injury severity outcomes. Thus, the present study proposes a two-staged hybrid approach where advanced text analytic

[5] Note that several of the text mining derived variables seemed to have a correlation with injury outcomes based on analysis of cross-tabulation of key variables and injury outcomes (Table 6). For instance, of all the cases (N = 22) where a trespasser wore a headphone or talked on cellphone, 12 (54%) cases included a fatal injury (see descriptive analysis in Table 6). Likewise, major or fatal injuries were more likely outcomes for intoxicated or trespassers under the influence of alcohol (Table 6). Despite the somewhat clear trends in the descriptive analysis, these policy sensitive variables were found statistically insignificant in the empirical models. This finding could be an outgrowth of the limited sample sizes for these text mining derived variables and point out to the exigency of collecting larger amounts of crash narratives for trespasser collisions.

procedures are applied first to extract valuable information from trespasser crash narratives. The new variables derived from text mining are then used as inputs (explanatory variables) in formulation of advanced statistical models for trespasser injury outcomes. To achieve this, ten-year (2006−2015) trespassing injury data available at Federal Railroad Administration's website are used. The study focuses explicitly on pedestrian and bicyclist non-crossing trespassing crashes (N = 6470), out of which, crash narratives are available for 41 % of the crashes. The analysis reveals a positive statistically significant association between presence of crash narrative and trespasser's injury outcome. That is, more severe crashes were more likely to be reported with a narrative. To mine the unstructured textual data, content and univariate frequency analysis is first performed - spotting key words prevailing in the crash narratives. Next, machine learning algorithms and statistical procedures are applied to uncover hidden thematic structure of text collection. Specifically, factor analysis procedures are applied for concept/entity extractions. The key concepts emerging from the crash narratives are trespasser, injury, location and warnings related, train related, and medical related themes. A total of 13 unique variables are then extracted from the thematic concepts that are not present in tabular crash data and can complement traditional empirical analysis. A crosstabulation of injury outcomes and the text mining derived new variables revealed a higher chance of fatal injury for trespassers who attempted suicides and those wearing headphones or talking on cellphone. Likewise, trespassers jumping directly in front of the train and trespassers hit by a commuter train were also more likely to receive fatalities. Regarding train movement, a higher chance of higher-order injuries was observed if a train was unable to stop prior to impact or if a train was put into emergency braking. Finally, 62.2 % and 15 % of intoxicated or trespassers under alcohol influence received major and fatal injuries, respectively.

Once unique information is extracted from the narratives through text mining procedures, statistical models are estimated for modeling injury outcomes as a function of new variables derived from text mining and other controls. Owing to the potential presence of unobserved heterogeneity, fixed and random parameter ordered probit models are estimated in the second stage to quantify the heterogeneous correlations between text mining derived variables and injury outcomes. Compared to the fixed parameter counterparts, the random parameter ordered probit model resulted in substantial improvements in model goodness of fit. A total five (5) text mining derived variables are found to be normally distributed random parameters suggesting that the correlations of these variables with injury outcomes vary across individual crashes. Substantial heterogeneity is observed not only in the magnitude but direction of correlations as well. Altogether, these findings emphasize the importance of accounting for unobserved heterogeneity in injury outcome models where text-mining derived variables are included as key predictors. The best-fit heterogeneity-based model revealed that the probability of fatal injuries increased by 0.155 units if the trespasser had a confirmed suicide attempt albeit with substantial heterogeneity. Likewise, if a train was unable to stop prior to impact, the trespasser was more likely to receive fatal injuries – with the probability of fatal injury increasing by 13.2 percentage points. Contrarily, if a trespasser was transported to hospital, the probability of fatal injury decreased by 31.2 percentage points. This finding suggests the effectiveness of timely medical care in reducing fatal injuries. The best fit random parameter ordered probit model also quantifies the statistically significant correlations between emergency braking, trespasser pre-crash behaviors, control variables (age, time of day, and county-level sociodemographic), and injury outcomes.

Several of the findings relate to "preventable" risk factors. For example, the finding that confirmed trespasser suicide attempt is associated with higher injury outcome is concerning. Identifying trespassers with a suicide intent can be difficult especially at non-crossings and/or crowded platforms. Nonetheless, effective surveillance and monitoring systems based on video and emerging sensing technologies can potentially improve trespasser safety. Likewise, the finding related to intoxicated trespassers suggests that enforcement/awareness campaigns targeted at negative consequences of alcohol can be helpful in reducing deadly behaviors by trespassers. While it is not clear if other confounding factors may also be linked with alcohol intake (such as alcohol intake by a trespasser intending to suicide), there exists a possibility that trespassers under influence receive fatal injuries due to their unintentional (preventable) dangerous behaviors and not intentional actions per se. In a same way, the results suggest that trespassing wearing headphones or talking on cell are more likely to receive fatal injuries. Taking these results into a perspective, we think that preventive efforts may focus on separating trespassers from trains in motion. Trains at non-crossings are likely to be traveling at high speeds. Thus, physical barriers or diverting trains to relatively less accessible tracks may be helpful. Finally, the behaviors such as trespasser laying and/or walking on tracks are associated with higher injury outcomes and are "preventable" risk factors. With the rapid technological advancements, Connected and Automated Vehicles (Train) systems and data analytics seem to have the potential to distinguish these behaviors and generate timely warnings (Zhang et al., 2018b). For instance, with the long range Dedicated Short Range Communication (DSRC) systems, instrumented trains can sense trespassers laying or walking on tracks and proactive measures can be taken accordingly.

Overall, the results demonstrate the value of text mining procedures for extracting unique information specific to each trespassing crash. The newly constructed variables are not present in traditional tabular form data, and thus the new findings presented in this study are only possible with appropriate text analytics. While the injury severity distributions against presence (and absence) of crash narratives look similar, systematic bias in narrative reporting is possible which can preclude the generalizability of the findings.[6] Second, crash narratives were available for 41 % of the crashes so the contextual information for all the trespasser crashes could not be analyzed - but is indirectly controlled methodologically to the extent possible.[7] Third, the sample sizes for some of the new variables are small, providing conservative estimates;

[6] We also analyzed the presence of narratives across the eight FRA regions. The proportion of crashes with narratives was highest for region 1 (78.05% of crashes had narratives) and lowest for region 5 (19.40% of crashes had narratives). No strong correlation between presence/absence of narrative and trespasser age was found. Region 1 consists of Maine, Vermont, New Hampshire, New York, Massachusetts, Rhode Island, Connecticut, and New Jersey. Region 5 consists of New Mexico, Oklahoma, Arkansas, Louisiana and Texas. For definitions of the eight regions, see https://railroads.dot.gov/divisions/regional-offices/regional-offices.

[7] The text-mining based dummy variables (along with the random parameters' framework) would capture some level of unobserved heterogeneity due to missing data on narratives to the extent that such information would not be typically available in the traditional (quantitative data) variables used for modeling. The method created in this paper can also be viewed as creating new information, but with estimated coefficients that have missing data on the narratives and hence the text-mining variables. An ideal way to handling this omitted variable bias issue (due to missing crash narratives) require availability of narratives for all the crashes. Without complete data on crash narratives, one cannot test or fully correct for the omitted variable bias issue especially using traditional modeling methods (such as fixed parameter models). Given these constraints, a rigorous methodological remedy is to use heterogeneity-based modeling methods (as is done in the present study) to account for the issue of omitted variable bias. In the random parameter models presented, the omitted information (due to missing crash narratives) is tracked as unobserved heterogeneity – this has the potential to mitigate the adverse impacts of missing narratives. However, as explained in (Mannering et al., 2016; Wali et al., 2018b), the β coefficients in the heterogeneity-based models presented herein may not track the omitted variable bias attributable unobserved heterogeneity as well as if the text-mining based variables were based on complete (no missing) data for crash narratives. Nonetheless, the heterogeneity-based methodology presented in this study is a rigorous method to account for omitted variable issue as much as possible keeping in view the fact that crash narratives are not available for all the crashes (Wali et al., 2018b).

occurrences of the key risk factors in reality are likely to be larger. While fatal crashes are more likely to be reported with narrative compared to minor crashes, the fact that ~56 % of fatal crashes did not have narrative is concerning. Based on this, safety agencies can encourage strategies that facilitate collection of larger amounts of crash narratives for such-like collisions, and this in future, may provide richer insights into the context specific mechanisms leading to specific crash outcomes. Notably, the idea is to demonstrate the potential of text and data mining techniques to extract important information in unstructured crash narratives. With more data coming in future, these limitations can be easily addressed. The present study uses a 10-year crash data to analyze as much crash narratives as possible. However, using such a large timespan also introduces issues related to temporal instability of the explanatory variables (Mannering, 2018) and potential changes in the definitions of variables across years. The associations between text-mining derived variables and injury outcomes may show temporal heterogeneity across different years. One likely reason behind the temporal instability (or heterogeneity) could be the continuously evolving technology over a 10-year period. A step further, the associations may exhibit time-of-day variations as well (Alnawmasi and Mannering, 2019; Behnood and Mannering, 2019). Thus, in future, with the availability of more crash narrative data, the methodology in this study can be expanded to account for potential temporal heterogeneity – allowing even better interpretation of crash narrative data-analysis findings.

CRediT authorship contribution statement

Behram Wali: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Data curation, Writing - original draft, Writing - review & editing, Visualization, Project administration. **Asad J. Khattak:** Resources, Writing - review & editing, Project administration. **Numan Ahmad:** Validation, Visualization, Methodology, Writing - review & editing.

Declaration of Competing Interest

The authors declare that they have no competing interests.

Acknowledgements

This paper is based on work supported by the US Department of Transportation through the Collaborative Sciences Center for Road Safety (CSCRS), a consortium led by The University of North Carolina at Chapel Hill (UNC) in partnership withThe University of Tennessee. The authors sincerely acknowledge the contribution of Dr. Meng Zhang in helping with portions of Section 3.3.1 of the manuscript on text analysis methods. The opinions and findings presented herein are those of the authors and does not represent the official views of U.S. Federal Railroad Administration and/or U.S. Department of Transportation.

Appendix A. Supplementary data

Supplementary material related to this article can be found, in the online version, at doi:https://doi.org/10.1016/j.aap.2020.105835.

References

Khattak, A., Luo, Z., 2011. Pedestrian and bicyclist violations at highway-rail grade crossings. Transp. Res. Rec.: J. Transp. Res. Board 2250, 76–82.
Abdel-Aty, M., 2003. Analysis of driver injury severity levels at multiple locations using ordered probit models. J. Saf. Res. 34 (5), 597–603.
Ahmad, N., Ahmed, A., Wali, B., Saeed, T.U., 2019. Exploring factors associated with crash severity on motorways in Pakistan. In: Proceedings of the Institution of Civil Engineers-Transport. Thomas Telford Ltd..
Alnawmasi, N., Mannering, F., 2019. A statistical assessment of temporal instability in the factors determining motorcyclist injury severities. Anal. Methods Accid. Res. 22, 100090.
Arvin, R., Khattak, A.J., 2020. Driving impairments and duration of distractions: assessing crash risk by harnessing microscopic naturalistic driving data. Accid. Anal. Prev. 146, 105733.
Arvin, R., Kamrani, M., Khattak, A.J., 2019. The role of pre-crash driving instability in contributing to crash intensity using naturalistic driving data. Accid. Anal. Prev. 132, 105226.
Behnood, A., Mannering, F., 2019. Time-of-day variations and temporal instability of factors affecting injury severities in large-truck crashes. Anal. Methods Accid. Res. 23, 100102.
Berry, M.W., Castellanos, M., 2008. Survey of Text Mining II. Springer.
Bhat, C.R., 2003. Simulation estimation of mixed discrete choice models using randomized and scrambled Halton sequences. Transp. Res. Part B Methodol. 37 (9), 837–855.
Boggs, A.M., Wali, B., Khattak, A.J., 2020. Exploratory analysis of automated vehicle crashes in California: a text analytics & hierarchical Bayesian heterogeneity-based approach. Accid. Anal. Prev. 135, 105354.
Breiger, R.L., Boorman, S.A., Arabie, P., 1975. An algorithm for clustering relational data with applications to social network analysis and comparison with multidimensional scaling. J. Math. Psychol. 12 (3), 328–383.
Bridges Jr., C.C., 1966. Hierarchical cluster analysis. Psychol. Rep. 18 (3), 851–854.
Duncan, C., Khattak, A., Council, F., 1998a. Applying the ordered probit model to injury severity in truck-passenger car rear-end collisions. Transp. Res. Rec. 1635, 63–71.
Duncan, C.S., Khattak, A.J., Council, F.M., 1998b. Applying the ordered probit model to injury severity in truck-passenger car rear-end collisions. Transp. Res. Rec. 1635 (1), 63–71.
Eluru, N., Bagheri, M., Miranda-Moreno, L.F., Fu, L., 2012. A latent class modeling approach for identifying vehicle driver injury severity factors at highway-railway crossings. Accid. Anal. Prev. 47, 119–127.
Elvik, R., Voll, N.G., 2014. Challenges of improving safety in very safe transport systems. Saf. Sci. 63, 115–123.
Feldman, R., Sanger, J., 2007. The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data. Cambridge University Press.
FRA, 2011. FRA Guide for Preparing Accidents/incidents Reports, Federal Railroad Administration, United States. URL: https://safetydata.fra.dot.gov/OfficeofSafety/ProcessFile.aspx?doc=FRAGuideforPreparingAccIncReportspubMay2011.pdf.
FRA, 2015. Rail Safety. U. D. o. T. F. R. Administration.
FRA, 2018. Railroad Crossing Safety & Trespass Prevention. URL: https://cms8.fra.dot.gov/divisions/highway-rail-crossing-and-trespasser-programs/railroad-crossing-safety-trespass-0.
Greene, W.H., Hensher, D.A., 2010. Modeling Ordered Choices: a Primer. Cambridge University Press.
Jiang, X., Zhang, G., Zhou, Y., Xia, L., He, Z., 2017. Safety assessment of signalized intersections with through-movement waiting area in China. Saf. Sci. 95, 28–37.
Jolliffe, I.T., 1986. Principal component analysis and factor analysis. Principal Component Analysis. Springer, pp. 115–128.
Khattak, Z.H., Fontaine, M.D., 2020. A Bayesian modeling framework for crash severity effects of active traffic management systems. Accid. Anal. Prev. 145, 105544.
Khattak, A., Tung, L.-W., 2015. Severity of pedestrian crashes at highway-rail grade crossings. J. Transp. Res. Forum.
Khattak, Z.H., Fontaine, M.D., Smith, B.L., 2020. Exploratory investigation of disengagements and crashes in autonomous vehicles under mixed traffic: An endogenous switching regime framework. IEEE Trans. Intell. Transp. Syst.
Lobb, B., 2006. Trespassing on the tracks: a review of railway pedestrian safety research. J. Saf. Res. 37 (4), 359–365.
Mannering, F., 2018. Temporal instability and the analysis of highway accident data. Anal. Methods Accid. Res. 17, 1–13.
Mannering, F.L., Bhat, C.R., 2014. Analytic methods in accident research: methodological frontier and future directions. Anal. Methods Accid. Res. 1, 1–22.
Mannering, F.L., Shankar, V., Bhat, C.R., 2016. Unobserved heterogeneity and the statistical analysis of highway accident data. Anal. Methods Accid. Res. 11, 1–16.
Metaxatos, P., Sriraj, P., 2013. Pedestrian/bicyclist Warning Devices and Signs at Highway-rail and Pathway-rail Grade Crossings. FHWA-ICT-13-013.
Mishara, B.L., 2007. Railway and metro suicides: understanding the problem and prevention potential. Crisis 28 (S1), 36–43.
Nichols, C.A., Koelpin, J., Conradi, S.E., Cina, S.J., 1994. A decade of train-pedestrian fatalities: the Charleston experience. Journal of Forensic Science 39 (3), 668–673.
Pelletier, A., 1997. Deaths among railroad trespassers: the role of alcohol in fatal injuries. JAMA 277 (13), 1064–1066.
Provalis, 2014. WORDSTAT 7 User's Guide. https://provalisresearch.com/Documents/WordStat7.pdf.
Quddus, M.A., Noland, R.B., Chin, H.C., 2002. An analysis of motorcycle injury and vehicle damage severity using ordered probit models. J. Saf. Res. 33 (4), 445–462.
Sadri, A.M., Ukkusuri, S.V., Murray-Tuite, P., 2013. A random parameter ordered probit model to understand the mobilization time during hurricane evacuation. Transp. Res. Part C Emerg. Technol. 32, 21–30.
Saeed, T.U., Hall, T., Baroud, H., Volovski, M.J., 2019. Analyzing road crash frequencies with uncorrelated and correlated random-parameters count models: An empirical assessment of multilane highways. Anal. Methods Accid. Res. 23, 100101.
Salton, G., 1989. Automatic Text Processing: the Transformation, Analysis, and Retrieval of Information by Computer. Addison-Wesley, Reading.
Savage, I., 2007. Trespassing on the railroad. Res. Transp. Econ. 20, 199–224.
Savage, I., 2016. Analysis of Fatal Train-pedestrian Collisions in Metropolitan Chicago 2004–2012. Accid. Anal. Prev. 86, 217–228.
Savolainen, P.T., Mannering, F.L., Lord, D., Quddus, M.A., 2011. The statistical analysis of highway crash-injury severities: a review and assessment of methodological alternatives. Accid. Anal. Prev. 43 (5), 1666–1676.

Sebastiani, F., 2002. Machine learning in automated text categorization. ACM Comput. Surv. (CSUR) 34 (1), 1–47.
Silla, A., Luoma, J., 2012. Main characteristics of train–pedestrian fatalities on Finnish railroads. Accid. Anal. Prev. 45, 61–66.
Tay, R., 2015. A random parameters probit model of urban and rural intersection crashes. Accid. Anal. Prev. 84, 38–40.
Train, K.E., 2009. Discrete Choice Methods with Simulation. Cambridge University Press.
Van Houwelingen, C.A., Beersma, D.G., 2001. Seasonal changes in 24-h patterns of suicide rates: a study on train suicides in the Netherlands. J. Affect. Disord. 66 (2), 215–223.
Wali, B., Ahmed, A., Ahmad, N., 2017. An ordered-probit analysis of enforcement of road speed limits. Proceedings of the Institution of Civil Engineers-Transport.
Wali, B., Khattak, A.J., Bozdogan, H., Kamrani, M., 2018a. How is driving volatility related to intersection safety? A Bayesian heterogeneity-based analysis of instrumented vehicles data. Transp. Res. Part C Emerg. Technol. 92, 504–524.
Wali, B., Khattak, A.J., Khattak, A.J., 2018b. A heterogeneity based case-control analysis of motorcyclist's injury crashes: Evidence from motorcycle crash causation study. Accid. Anal. Prev. 119, 202–214.
Wali, B., Khattak, A.J., Ahmad, N., 2019a. Examining correlations between motorcyclist's conspicuity, apparel related factors and injury severity score: evidence from new motorcycle crash causation study. Accid. Anal. Prev. 131, 45–62.
Wali, B., Khattak, A.J., Karnowski, T., 2019b. Exploring microscopic driving volatility in naturalistic driving environment prior to involvement in safety critical events—Concept of event-based driving volatility. Accid. Anal. Prev. 132, 105277.
Wali, B., Khattak, A.J., Karnowski, T., 2020. The relationship between driving volatility in time to collision and crash-injury severity in a naturalistic driving environment. Anal. Methods Accid. Res. 28, 100136.
Wang, X., Liu, J., Khattak, A.J., Clarke, D., 2016. Non-crossing rail-trespassing crashes in the past decade: A spatial approach to analyzing injury severity. Saf. Sci. 82, 44–55.
Zhang, M., Khattak, A.J., Liu, J., Clarke, D., 2018a. A comparative study of rail-pedestrian trespassing crash injury severity between highway-rail grade crossings and non-crossings. Accid. Anal. Prev. 117, 427–438.
Zhang, Z., Trivedi, C., Liu, X., 2018b. Automated detection of grade-crossing-trespassing near misses based on computer vision analysis of surveillance video data. Saf. Sci. 110, 276–285.
Zhao, S., Iranitalab, A., Khattak, A., 2016. Investigation of pedestrian injury severity in crashes at highway-rail grade crossings using latent class analysis. Transportation Research Board 95th Annual Meeting.