Accident Analysis and Prevention 62 (2014) 119–129

Contents lists available at ScienceDirect

Accident Analysis and Prevention

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

Near-miss narratives from the fire service: A Bayesian analysis☆

Jennifer A. Taylor[a,*], Alicia V. Lacovara[a,3], Gordon S. Smith[b,1], Ravi Pandian[a,3], Mark Lehto[c,2]

[a] Department of Environmental & Occupational Health, Drexel University School of Public Health, 1505 Race Street, MS 1034, Philadelphia, PA 19102, United States
[b] University of Maryland School of Medicine, Department of Epidemiology & Public Health, 110 S. Paca Street, 4th floor, Rm 4-S-125, Baltimore, MD 21201, United States
[c] Purdue University, School of Industrial Engineering, 315 N. Grant Street, West Lafayette, IN 47907-2023, United States

ARTICLE INFO

Article history:
Received 11 June 2013
Received in revised form 4 September 2013
Accepted 17 September 2013
Available online 1 October 2013

Keywords:
Text-mining
Near-miss narratives
Fire fighter injury
Bayesian models

ABSTRACT

Background: In occupational safety research, narrative text analysis has been combined with coded surveillance, data to improve identification and understanding of injuries and their circumstances. Injury data give, information about incidence and the direct cause of an injury, while near-miss data enable the, identification of various hazards within an organization or industry. Further, near-miss data provide an, opportunity for surveillance and risk reduction. The National Firefighter Near-Miss Reporting System, (NFFNMRS) is a voluntary reporting system that collects narrative text data on near-miss and injurious, events within the fire and emergency services industry. In recent research, autocoding techniques, using Bayesian models have been used to categorize/code injury narratives with up to 90% accuracy, thereby reducing the amount of human effort required to manually code large datasets. Autocoding, techniques have not yet been applied to near-miss narrative data.
Methods: We manually assigned mechanism of injury codes to previously un-coded narratives from the, NFFNMRS and used this as a training set to develop two Bayesian autocoding models, Fuzzy and Naïve. We calculated sensitivity, specificity and positive predictive value for both models. We also evaluated, the effect of training set size on prediction sensitivity and compared the models' predictive ability as, related to injury outcome. We cross-validated a subset of the prediction set for accuracy of the model, predictions.
Results: Overall, the Fuzzy model performed better than Naïve, with a sensitivity of 0.74 compared to 0.678., Where Fuzzy and Naïve shared the same prediction, the cross-validation showed a sensitivity of 0.602., As the number of records in the training set increased, the models performed at a higher sensitivity, suggesting that both the Fuzzy and Naïve models were essentially "learning". Injury records were, predicted with greater sensitivity than near-miss records.
Conclusion: We conclude that the application of Bayesian autocoding methods can successfully code both near misses, and injuries in longer-than-average narratives with non-specific prompts regarding injury. Such, coding allowed for the creation of two new quantitative data elements for injury outcome and injury, mechanism.

1. Introduction

1.1. Collection and analysis of narrative text

In occupational safety research, narrative text analysis has been combined with coded surveillance data to improve identification and understanding of injuries and their circumstances. Narrative text analysis identifies more target events than can be found using injury codes alone, thus reducing the problem of undercounting—a critical concern in injury surveillance. Further, narrative text analysis provides a means to check coding accuracy, and provides important information on circumstances surrounding injuries and unknown risk factors (Lipscomb et al., 2004; Bondy et al., 2005;

* Corresponding author. Tel.: +1 215 762 2590.
E-mail addresses: Jat65@drexel.edu (J.A. Taylor), Avl24@drexel.edu (A.V. Lacovara), gssmith@som.umaryland.edu (G.S. Smith), Rsp46@drexel.edu (R. Pandian), lehto@purdue.edu (M. Lehto).
[1] Tel.: +1 410 328 3847; fax: +1 410 328 2841.
[2] Tel.: +1 765 49 45428; fax: +1 765 494 1299.
[3] Tel.: +1 215 762 2590.

0001-4575/$ – see front matter © 2013 The Authors. Published by Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.aap.2013.09.012

Smith et al., 2006; Bunn et al., 2008). New risk factors identified through narrative text analysis are an important source of variables to be added to administrative coding systems (Bunn et al., 2008). Narrative data analysis can also be a basis for comparing data among systems and countries that use different coding schemes, or to study historical data that include narrative text (Stout, 1998).

The large-scale study of narrative text has only recently been made possible by advances in computerized information retrieval techniques. This is particularly important for large, growing datasets which adds to increased time, cost and labor, in order to code these narratives. Computerized coding algorithms have enabled large-scale analysis of narrative text, presenting an efficient and plausible way for individuals to code large narrative datasets. Although computer coding is a cost-efficient alternative to manual coding with an accuracy of up to 90%, it does not eliminate the need for human review entirely (Lehto and Sorock, 1996; Wellman et al., 2004; Lehto et al., 2009; Bertke et al., 2012; Patel et al., 2012).

The most critical bottle-neck is that computer coding methods require a learning set of previously coded cases. The accuracy of computer coding also tends to improve when larger training sets are used to develop the algorithms. The latter issue is especially important when the coded categories differ greatly in frequency, as it may become difficult to obtain enough training cases for the small, rarely occurring codes. For this and other reasons, computer coding algorithms tend to predict some codes much more accurately than others. One solution strategy is for the coding algorithm to assign the "easy" cases and flag the remaining potentially ambiguous cases for human review (Lehto et al., 2009). This approach allows computer coding errors to be efficiently identified and corrected during use. The results of the human review can also be fed back into the system, allowing the model to learn over time after implementation.

1.2. The importance of near-miss data

A near-miss is an incident that had the capacity to cause injury but did not, due to either intervention or chance (Aspden et al., 2004). Both injury and near-miss data are important to collect in surveillance systems. While injury data give information about incidence and the direct cause of an injury, near-miss data enable the identification of various hazards within an organization or industry while providing an opportunity for surveillance and risk reduction. Near-miss narratives in particular provide insight to the upstream causes of injury (Rivard et al., 2006). Near-miss reporting can capture the successful recovery from potentially harmful incidents. In the field of healthcare, research has found that even a few reports can be sufficient to detect and communicate a hazard that is actionable for prevention (Leape, 2002) and prompt an organizational response. Importantly, near-misses occur more frequently than adverse events (Barach and Small, 2000), and can be combined with injuries to increase statistical power for analysis as supported by the common cause hypothesis (Alamgir et al., 2009).

1.3. Purpose of this study

Injury narratives are frequently coded for mechanism of injury (using ICD-9-CM or ICECI codes), but there is an absence of literature that addresses application of mechanism-of-injury coding to near-miss narratives. In theory, assigning a mechanism-of-injury code to a near-miss narrative should be straight forward—the reporter explains briefly the circumstances, what led to the event, and why it was a near-miss. Coding of near-misses will help to construct hazard scenarios, and inform development of appropriate interventions to prevent future injury and harm (Lincoln et al., 2004).

Our objective was to manually code narratives from the National Firefighter Near Miss Reporting System (NFFNMRS) and use this coded set to train a computer algorithm to assign mechanism of injury codes to un-coded narratives. Since no variable currently exists on the NFFNMRS reporting form to capture the presence or absence of an injury, the study also sought to create a quantitative variable to identify injury and near-miss events.

2. Method

2.1. Data source

In order to improve understanding of the circumstances leading to firefighter injuries, the International Association of Fire Chiefs (IAFC) (with funding from the Assistance to Firefighters Grant Program of the U.S. Department of Homeland Security) launched the NFFNMRS in 2005. Reporting to the system is voluntary and non-punitive. The NFFNMRS defines a near-miss as "an unintentional, unsafe occurrence that could have resulted in an injury, fatality, or property damage" (www.firefighternearmiss.com). Despite this definition, the NFFNMRS captures a number of actual injuries, including fractures, back injuries, hypothermia, burns, and cyanide poisoning, as well as melted equipment and destroyed engines.

The reporting form consists of 22 fields. Two of these fields are narrative sections, asking the reporter to "Describe the event", and to share "Lessons Learned". Within these fields, reporters can submit as much text as they wish.

2.2. Selection of narratives for manual coding

The quantitative component of the near-miss forms contains a field called "Event Type" in which the reporter selects whether the incident occurred during a fire emergency event, a vehicle event, a training activity, etc. (the form can be viewed at http://www.firefighternearmiss.com/Resources/NMRS-Mail.pdf). In order to reduce cognitive shifts required for coding of different event types (hazards described in vehicle event narratives are different than those in fire event narratives), we limited our analysis to only include those indicated as fire emergency events, as identified by the reporter. This data set contained 2285 narratives. Of these "Fire Emergency Events", we manually coded 1000 narratives, which resulted in 764 fire-related events considered suitable as training narratives for the algorithm. The 236 narratives discarded from the training set were not "Fire" related cases (e.g., neither the precipitating nor proximal cause was a fire event), or they were fire-related but lacked specific information for sub-categorization (e.g., fire-burn, fire-struck-by/against), or they fell into a category that ended up having fewer than five narratives (e.g., motor vehicle-rollover, hot substance or object, caustic or corrosive material, and steam). Fig. 1 shows the case inclusion criteria for our analysis.

2.3. Manual coding rubric

The initial rubric was a set of mechanism of injury codes from the International Classification of Disease 9 Clinical Modification Manual (ICD-9-CM), selected by the Principal Investigator (JAT) as codes that were possible within the fire-fighting/EMS occupational field. The rubric was modified over time in an iterative, consensus-driven process. Whenever a change was made the Project Manager (AVL) went back over the previously coded narratives and amended the code in accordance with the revised rule when necessary. A precipitating mechanism (what set the injury chain of events in motion) and a proximal mechanism (what caused the injury or near-miss) were assigned to each narrative.

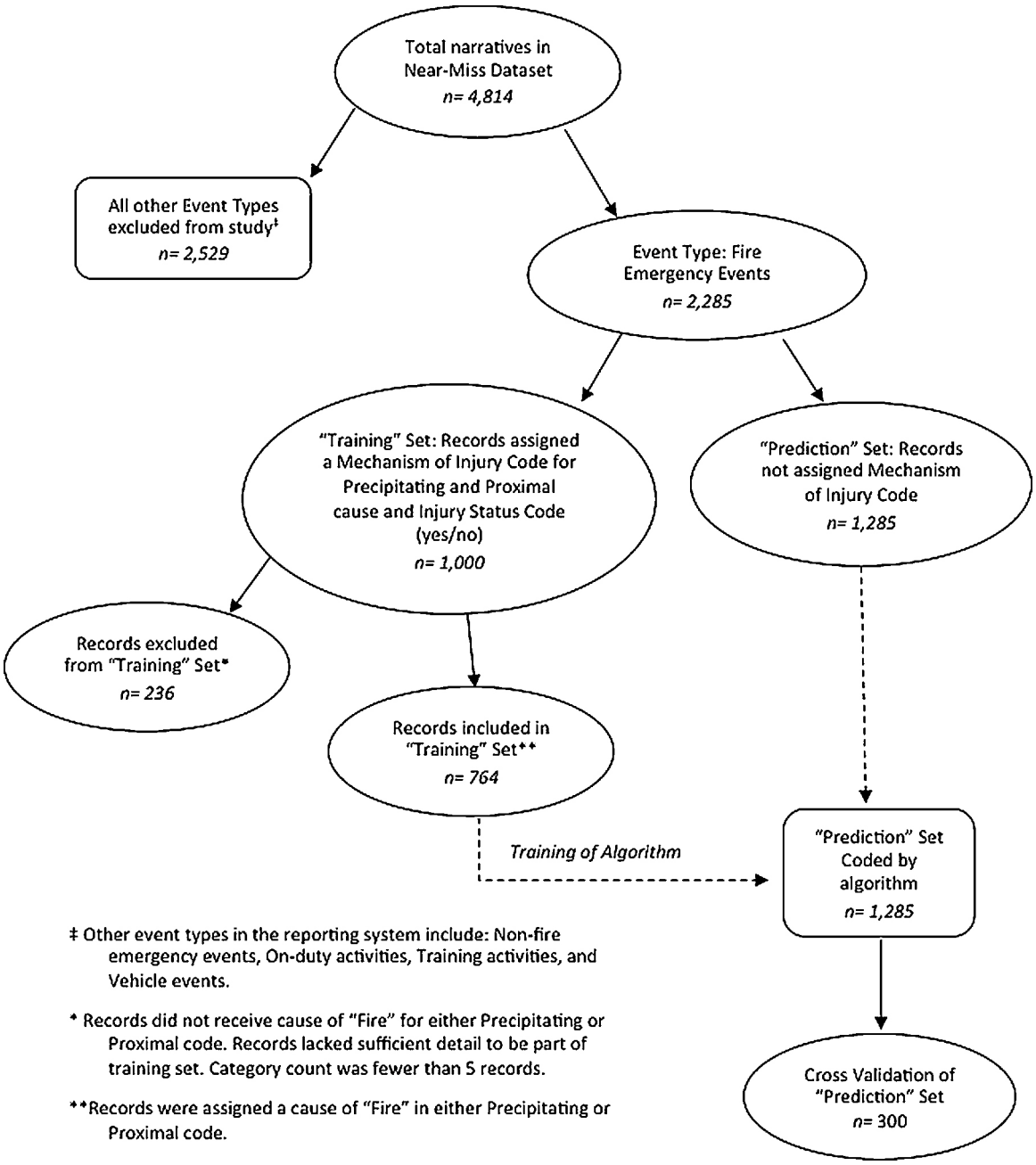


Fig. 1. Case inclusion flow chart.

In creating our coding rubric, it became evident that the ICD-9-CM is not granular enough for firefighting. For example, since fire fighters encounter fire frequently, coding the majority of cases to Conflagration (i.e., E890–E899) would mask hazards that occur during fires such as electrocutions, falls, smoke inhalation, struck-by motor vehicles, etc. Therefore, we created subcategories within conflagration (fire) to further capture specific firefighting hazards (Table 1). The resultant coding scheme extracted more detail from each narrative while honoring the ICD-9-CM hierarchy by retaining the overall cause category as conflagration (fire). Because this process was iterative, we re-coded previous cases as necessary updating them to the newer rubric.

2.4. Manual coding of narratives

In the field of autocoding, there has not been an established minimum size of the training set with regard to the total dataset. Therefore, we decided to code a minimum of 20% of our dataset to act as the training set for the algorithm, similar to Bertke et al. (2012). Based on these recommendations, we calculated that we needed to manually code a minimum of 456 narratives for our training set and aimed to complete more than this.

Three of the authors (JAT, AVL, GS) coded each narrative for a (1) whether an injury occurred (yes/no), (2) the cause of the injury/near-miss (proximal cause), and (3) what lead to the injury/near-miss (precipitating cause). By asking the above three questions in this order, we were able to consistently evaluate each narrative for injury outcome, proximal cause, and precipitating cause. The order was important because in near-miss narratives, the proximal cause is often difficult to discern since no actual injury occurred. It took each coder approximately 25 h to assign mechanism of injury codes to 1000 narratives. The narratives were coded in seven batches. After each batch, the three coders reconciled their

Table 1
List of mechanism of injury categories used to classify narratives.

Original list of cause codes (pre-coding)	Final list of cause codes
Accidents caused by machinery	Accidents caused by explosive material (gas leak, dynamite, etc)
Air and space transport accidents	Accidents caused by machinery
Caught accidentally in or between objects	Air and space transport accidents
Cutting and piercing instruments or objects	Caught accidentally in or between objects
Drowning/submersion	Cutting and piercing instruments or objects
Electric current	Drowning/submersion
Exposure to radiation	Electric current
Explosive material	Exposure to radiation
Explosion of pressure vessel	Fall
Fall	Fire
Fire	Fire-Burn
Firearm	Fire-caught-in/between
Hot substance or object, caustic or corrosive material, and steam	Fire-CO, smoke, fumes from PVC, etc
Motor vehicle non-traffic accident	Fire-collapse
Motor vehicle traffic (MVT)	Fire-electric current
Natural/environmental	Fire-equipment/machinery
Other	Fire-explosion caused by fire
Other road vehicle accidents	Fire-fall (through floor, from ladder, jump)
Overexertion	Fire-medical condition (MI, Asthma, etc)
Poisoning	Fire-struck-by
Railway accidents	Fire-vehicle
Struck by, against	Fire-wildland, etc
Suffocation	Firearm/ammunition
Water transport accidents	Hot substance or object, caustic or corrosive material, steam
	Motor vehicle non-traffic accident
	Motor vehicle traffic (MVT)
	MV-collision
	MV-FF struck by vehicle
	MV-other
	MV-rollover
	N/A
	Natural/environmental
	Other
	Other road vehicle accidents
	Overexertion
	Poisoning
	Railway accidents
	Struck-by, against
	Suffocation
	Water transport accidents

individual scores for each narrative, assigning a final Mechanism of Injury code. Reconciliation of the seven batches took approximately 25 h. The entire coding, reconciliation, and rubric revision process occurred over a one year interval. Overall coder agreement statistics were calculated and kappa values were obtained.

The final set used as the training set consisted of 764 narratives. A total of 236 narratives were not included in the training set because they were not assigned a code of "fire" for either the precipitating or proximal code ($n = 214$), or they were assigned a mechanism of injury code that existed in fewer than five total narratives ($n = 22$). For example, many of the narratives were categorized by the reporter as "Fire emergency events", but the narrative actually describes a motor vehicle accident on the way to a structure fire. Other narratives lacked enough detail or information to either classify them as a fire event, or assign a mechanism of injury code. The categories with fewer than five narratives were not included in the analysis because after dropping rare words, which is standard practice to reduce model noise, these small categories would no longer have strong predictor words.

2.5. Model development

Two different Bayesian models, referred to as Naïve Bayes and Fuzzy Bayes, were developed and evaluated using the TextMiner program (developed by author ML). The models and software have been described elsewhere (Lehto et al., 2009). Both models used the statistical relationship between the words present in the injury narratives of the training set ($n = 764$) and the manually assigned mechanism of injury code to predict a particular code for a new narrative. This prediction is essentially the probability of a particular code given the words within the new narrative. The two models differ in that the Naïve Bayes prediction is a weighted function of all the words present, while the Fuzzy Bayes prediction is based on the single strongest predictive word for each category. Specifically, the Naïve Bayes model calculates the probability of an injury code using the following expression:

$$P(E_i|n) = P(E_i)\Pi(P(n_i|E_i))/P(n_j) \quad (1)$$

where $P(E_i|n)$ is the probability of event code category E_i given the set of n words in the narrative. $P(n_i|E_i)$ is the probability of word n_j given category E_i. $P(E_i)$ is the probability of category E_i, and $P(n_j)$ is the probability of word n_j in the entire list of keywords.

The Fuzzy model is similar, except that it estimates $P(E_i|n)$ using the 'index term' most strongly predictive of the category, instead of multiplying the conditional probabilities as in the Naïve model:

$$P(E_i|n) = \mathrm{MAX}_j(P(n_i|E_i)P(E_i))/P(n_j) \quad (2)$$

The two models were both tested using the TextMiner Software which runs on a Microsoft Access platform. After all the Fire-Events narratives were manually coded, the database was prepared for analysis in TextMiner. Narratives that were non-fire related (as coded by the researchers, see Fig. 1) were removed from the dataset. For the remaining narratives, all non-alphanumeric symbols were removed (e.g., Fire-Eqpt/Mach became FireEqptMach). A training flag was used to denote all manually coded narratives that were part of the training set.

Once the training set ($n = 764$) and prediction set ($n = 1285$) were divided, the words from the narratives within the training set were used to generate a wordlist. The wordlist was contained in a table listing every word in the entire dataset, starting with the first word in the first narrative and ending with the final word of the last narrative. The dataset was cleaned by removal of words occurring fewer than three times. Each narrative was edited for spelling mistakes during the initial report submission process. No additional modifications were made such as assigning synonyms to words or removing common stop words such as "A,THE, . . ." The purpose of this was to see how well the algorithm could perform on a raw dataset with little to no human input.

2.6. Model evaluation

2.6.1. Training set

The two models generated predictions for every narrative in our Fire Emergency Events dataset, including the training set. The results of the predictions were compared to the manually assigned "gold standard" codes by expert coders, and model sensitivity, specificity, and positive predictive value (PPV) was calculated for each category Sensitivity is simply the proportion of correctly identified codes for a particular category. For example, if 100 cases should have been coded as x, and model correctly assigned 50 of these cases, the sensitivity would be 50% for this category [i.e., 50 correct identifications/100 cases where x is correct]. Specificity measures how often a code is correctly not assigned, when some other code should have been assigned. For example, if we assume code x should not be assigned to 100 cases, and found it was correctly not assigned to 99 of these cases, the specificity would be 99% [i.e., 99 correct rejections/100 cases where x is not correct]. Positive predictive value measures prediction accuracy, and corresponds to the proportion of correct responses given a particular prediction of the model. For example, if the model predicted category y 50 times, and each of these predictions was correct, the PPV for category y would be 100% [i.e., 50 correct/50 times predicted]. Note that these measures are complementary to each other. Ideally, the predictive model will score high on all three of these measures, demonstrating that it is likely to assign the correct code for each of the categories.

2.6.2. Prediction set and cross validation

The Fuzzy and Naïve Bayes models were also both run on a prediction set of 1285 previously unclassified narratives (Fig. 1). In order to test the accuracy of the algorithm's predictions for these new (not originally manually classified cases), we performed a cross validation study in which 300 narratives from the prediction set were manually coded by the reviewers. The cases in the cross validation set were equally divided into three categories: (1) strongly predicted, (2) moderately predicted, and (3) poorly predicted cases. The cases were assigned based on prediction strength and whether the Fuzzy and Naïve predictions agreed.

The strongly predicted cases corresponded to narratives in which the Fuzzy and Naïve predictions agreed ($n = 475$). For this category, the distribution of narratives to be included in the cross validation set matched that of the distribution of in the original sample of 475.

The poorly and moderately predicted categories corresponded to cases where the Fuzzy and Naïve predictions disagreed. The latter cases were further subdivided based on prediction strength. Prediction strength was simply the probability assigned by the respective model to its prediction (see Eqs. (1) and (2)). The poorly predicted cases were those where the Fuzzy and Naïve models disagreed on the prediction, and both had strength predictors in the top 50% of their respective distributions. For example, Fuzzy might predict "Fire-Fall" with a prediction strength of 0.99, while Naïve predicted "Fire-Burn" with a strength of 0.97. They disagree, and both predictions are strong. The moderately predicted cases were those cases where Fuzzy and Naïve disagreed, and one had a strength predictor in the top half of their distribution, and the other had a strength predictor in the bottom half of their distribution. We considered the percentile ranks of these strength predictions to build our poor and moderate samples.

One-hundred narratives were randomly selected from each of the three categories. Each narrative was assigned a mechanism of injury code by each coder. The 300 narratives were then reconciled so that each narrative received a single code. These codes were then compared to the codes predicted by the Fuzzy and Naïve algorithms.

2.6.3. Proximal cause prediction by injury outcome

Finally, we wanted to determine how well each model was able to correctly predict a mechanism of injury code, according to injury outcome. After both models had been run, the training set was separated by injury outcome (injury vs. near-miss), and sensitivity was obtained for each. The effect of increased training set size (in iterations of 100 narratives) was also evaluated by calculating sensitivity separately for each sized training set (for injury vs. near-miss).

3. Results

3.1. Characteristics of narratives

Within the fire emergency event narratives ($n = 2285$), the mean word count was 216, with a median count of 156 words and a range from 2 words to 2420 words.

3.2. Intra- and inter-rater reliability

Agreement between coders improved substantially with an overall agreement above 79%. Agreement between coders 1 and 2 improved 12% ($\kappa = 0.785$), coders 1 and 3 improved 8% ($\kappa = 0.75$), and coders 2 and 3 improved 13% ($\kappa = 0.774$). Each of the three coders had substantial agreement with their original scores when coding the same narratives a second time ($0.68 < \kappa < 0.80$).

3.3. Modification of the coding rubric

Creation of the coding rubric was an iterative process. With each narrative read, common themes occurred and thus informed the creation of specific sub-categories. For example, when there was a roof collapse, we assumed the mechanism to be "struck by/against" unless the reporter specified otherwise. We reached saturation of repetitious events after batch 3 and the rubric did not change for coding of the remaining batches.

3.4. Performance of automated coding

3.4.1. Training set

Overall, Fuzzy Bayes performed better than Naïve Bayes. Table 2 shows the top predictor words when applying the Fuzzy model. Fuzzy outperformed Naïve Bayes with a sensitivity of 0.74 compared to 0.678 (Table 3). The fire-burn category was well predicted by both Naïve and Fuzzy, though the specificity and PPV was higher with Fuzzy. For the categories of fire-fall and fire-struck-by, Fuzzy had better sensitivity while Naïve had better PPV. In general, Fuzzy performed with higher sensitivity, specificity and PPV, particularly in the larger categories. Naïve performed a bit better with the smaller categories.

Increasing the size of the training set improved the performance of the algorithm (Fig. 2). For example, using a training set of 100 narratives to predict the entire dataset of 764, the Fuzzy model

Table 2
Top 3 predictor words for Fuzzy Bayes for largest 5 cause categories.

Fire-fall	Pit (0.86)	Stories (0.83)	Spongy, waist (0.78)
Fire-struck by/against	Strut (0.90)	Cracking (0.86)	Effect (0.83)
Fire-burn	Burns (0.91)	Flashed (0.84)	Intense (0.81)
Fire-electric current	Energized (0.93)	Arcing (0.9)	Volt, arced (0.89)
Fire-CO, smoke, fumes, etc.	Inhalation (0.88)	Inhaled (0.86)	Speak (0.83)

Table 3
Fuzzy and Naïve Bayesian analyses: sensitivity, specificity, and PPV.

Mechanism of injury category	*N*	Fuzzy model			Naïve model		
		Sensitivity	Specificity	PPV	Sensitivity	Specificity	PPV
OVERALL	764	0.740	–	–	0.678	–	–
Fire-fall	196	0.745	0.887	0.695	0.561	0.995	0.973
Fire-struck by/against	184	0.728	0.933	0.775	0.342	1	1
Fire-burn	169	0.941	0.877	0.685	1	0.652	0.449
Fire-electric current	68	1	0.974	0.791	0.853	1	1
Fire-CO, smoke, fumes, etc.	48	0.521	0.997	0.926	0.917	1	1
Fire-explosion caused by fire	25	0.12	0.999	0.75	0.84	0.996	0.875
Fire-equipment/machinery	17	0	1	–	0.412	1	1
Fire-medical condition	9	0.889	1	1	1	0.999	0.9
Fire-caught in/between	6	0	1	–	1	0.992	0.5
MV-FF struck-by vehicle	26	0.577	0.999	0.938	0.692	1	1
Firearm/ammunition	9	0.778	0.999	0.875	1	0.966	0.257
Cutting/piercing instruments/objects	7	0	1	–	0.571	1	1

had a sensitivity of 43%. From the initial training set of 100 to the final training set of 764, the algorithm improved by 31% for Fuzzy and 35% for Naïve. The algorithm appeared to be learning with each additional batch of narratives added to the training set. It is possible we were approaching a threshold with the Fuzzy sensitivity, judging by the incremental gains as the training set progressed past 700 narratives. Naïve appeared to still be improving by the final training set.

3.4.2. Prediction set and cross validation

Out of the 300 narratives within the cross-validation set, the manual coders identified 7 narratives that were not sufficiently detailed, or were not fire-related, and thus not included in the final analysis. Overall, for the 293 cases examined Fuzzy had a sensitivity of 51.9%, while the sensitivity for Naïve was about half, at 24.9% (Table 4). For those narratives within the strong category, of which Fuzzy and Naïve had the same prediction, the sensitivity was 60.2%.

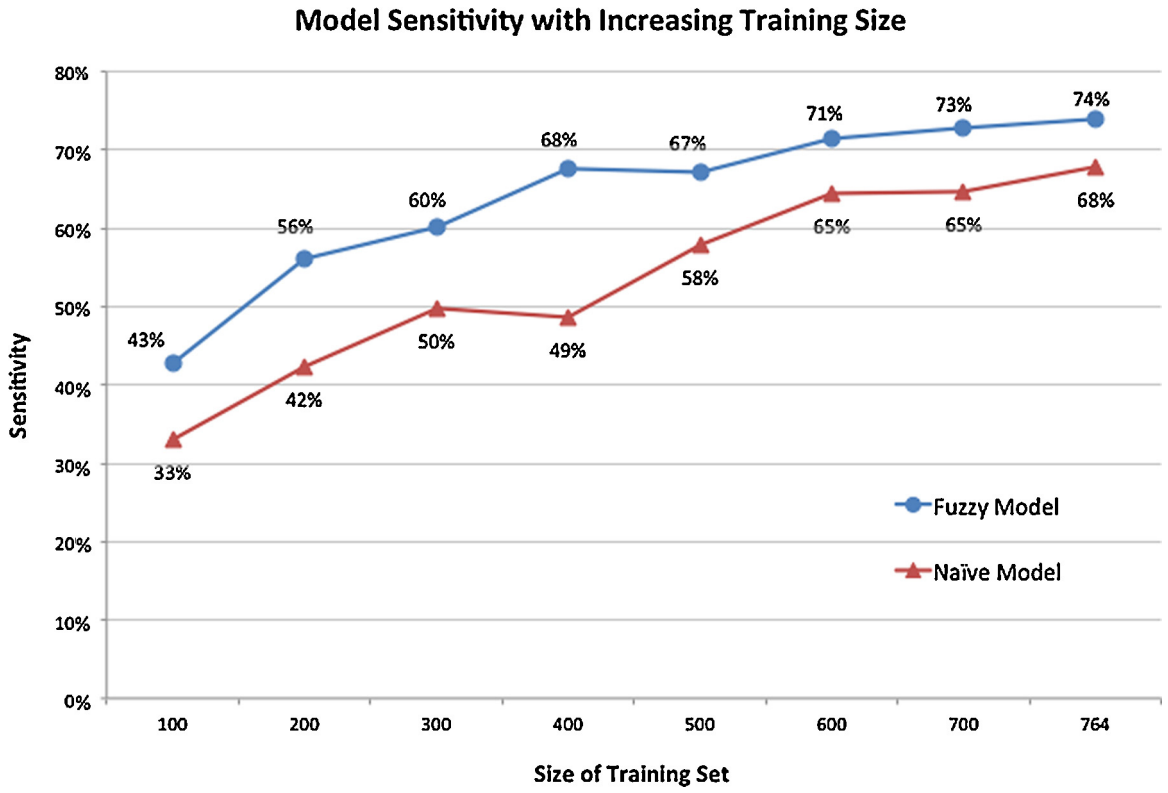


Fig. 2. Model sensitivity with increasing size of training set.

Table 4
Cross validation of the prediction set.

Prediction strength (proximal cause)	*n*	Fuzzy correct predictions (*n*)	Fuzzy sensitivity (%)	Naïve correct predictions (*n*)	Naïve sensitivity (%)
Strong—where fuzzy and naïve predicted the same category	98	59	60.2	59	60.2
Moderate—where Fuzzy and Naïve disagreed on the prediction, one had a good strength indicator, the other did not	99	40	40.4	7	7.1
Poor—where Fuzzy and Naïve disagreed on the prediction, and both had good strength indicators associated with their predictions	96	53	55.2	7	7.3
Overall	293	152	51.9	73	24.9

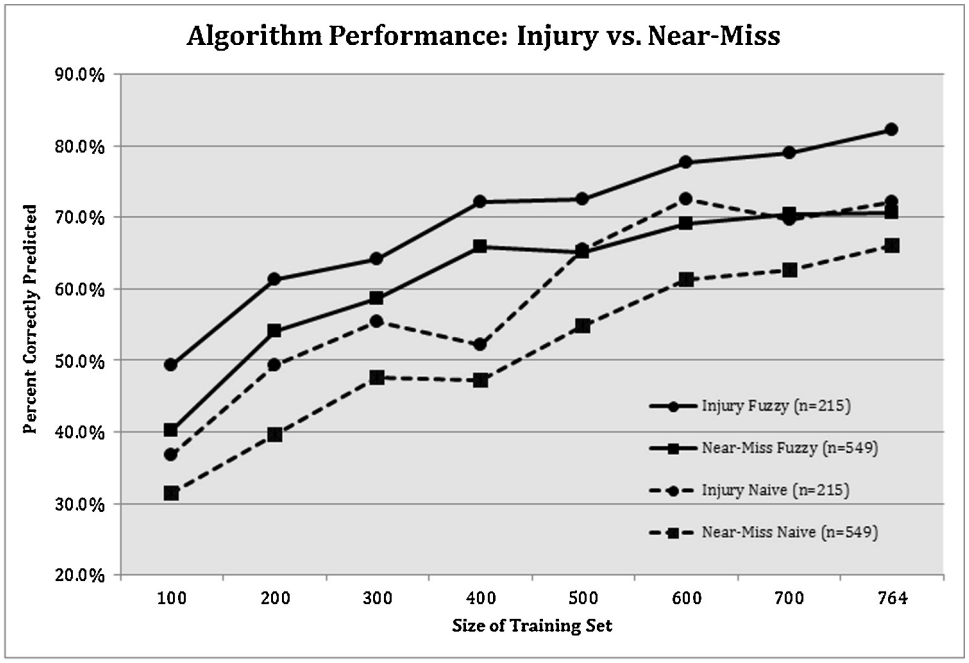
In the moderate and poor categories, Fuzzy performed much better, giving a sensitivity of 40.4% and 55.2%, respectively.

3.4.3. Model performance by injury outcome

Manually coding of the narratives for injury outcome yielded 215 injuries (28%) and 549 (72%) near-misses. Thus, we were able to create a new quantitative variable "Injury (yes/no)". Furthermore, applying the Bayesian models to the training set ($n = 764$) to predict injury outcome, Fuzzy sensitivity reached 92% (data not shown).

Using this new variable, the Fuzzy model predicts the mechanism of injury with a higher sensitivity for injury narratives (0.823) than near-miss narratives (0.707) (Fig. 3). In general, the mechanism of injury is correctly predicted more frequently for Injury narratives than for near-miss narratives. Regardless, the overall sensitivity of the algorithm improved for both models, regardless of injury outcome.

From the above results, two new quantitative data elements were created for fire events in the NFFNMRS: Injury (yes/no; "no" indicating a near-miss) and Mechanism of Injury. Table 5 exhibits near-misses and injuries by cause that were developed by analysis of the training data set which reflect persistent and emerging hazards in firefighting. The distribution of causes was



Injury Narratives — **Size of Training Set**

	100	200	300	400	500	600	700	764
Fuzzy Sensitivity	0.493	0.614	0.642	0.721	0.726	0.777	0.791	0.823
Naïve Sensitivity	0.367	0.493	0.553	0.521	0.656	0.726	0.698	0.721

*Denominator is total number of injury records, n=215

Near-Miss Narratives — **Size of Training Set**

	100	200	300	400	500	600	700	764
Fuzzy Sensitivity	0.403	0.541	0.587	0.659	0.650	0.690	0.705	0.707
Naïve Sensitivity	0.315	0.397	0.475	0.472	0.548	0.614	0.627	0.661

*Denominator is total number of near-miss records, n=549

Fig. 3. Algorithm performance by injury vs. near-miss.

Table 5
Distribution of mechanism of injury for proximal cause, fire events only.

Training set (n = 764)					
Near-miss narratives	n	%	Injury narratives	n	%
Fire-fall	141	26	Fire-burn	76	34
Fire-struck-by, against	139	26	Fire-fall	55	25
Fire-burn	93	17	Fire-struck-by/against	45	20
Fire-electric current	59	11	Fire-CO, smoke, fumes	17	8
Fire-CO, smoke, fumes	31	6	Fire-electric current	9	4
MV-FF struck-by	26	5	Fire-explosion	5	2
Fire-explosion by fire	20	4	Cutting/piercing object	5	2
Fire-equipment/machinery	17	3	Fire-caught-in, between	2	1
Fire-medical condition	9	2	Firearm/ammunition	1	0
Firearm/ammunition	8	1	Total	215	100
Fire-caught-in, between	4	1			
Cutting/piercing object	2	0			
Total	549	100			

similar between the training set and the prediction set, but as cross-validation of the prediction set (n = 300) demonstrated only 64% accuracy we are only presenting the results from the training data set.

4. Discussion

We found that TextMiner was able to correctly predict a mechanism of injury code for 74% of the narratives using the Fuzzy model and 68% of the narratives using the Naïve model. Injuries were correctly predicted at a higher rate (Fuzzy 0.82, Naïve 0.72) than near-misses (Fuzzy 0.71, Naïve 0.66). Overall, our sensitivity is comparable to the results of Lehto et al. (2009), which saw sensitivity between 70% and 80% for Naïve and between 64% and 78% sensitivity for Fuzzy when analyzing injury narratives. To our knowledge, this study is the first of its kind to successfully use machine learning algorithms to assign mechanism of injury codes to near-miss narratives. Previous research has only looked at injury narratives.

Our findings are comparable with the growing body of seminal studies on narrative autocoding (Table 6).

4.1. Manual coding of near-miss narratives

Coding near-miss narratives is not as straight-forward as coding actual injury narratives. To do so, we must look for the most likely outcome that could have occurred, recognizing that one decision must be made when multiple outcomes are possible. Such decision-making is time-consuming and therefore expensive in terms of human resources. In the Methods, we discussed the importance of and adherence to the coding order of operations: injury outcome first, then proximal cause, then precipitating cause. The challenge of coding near-miss events is that it is often difficult to determine a finite point from which to work backward because there is no injury. For this reason, starting with "Did an injury happen (yes/no)?" was invaluable in helping us determine the mechanism of that injury (or near-miss) and then assess what started the chain of events in motion (precipitating). However, there were times when the coders were often forced to speculate on the outcome and select a code. This was where the majority of disagreement between codes occurred. If too little information was provided or the report did not provide a clear understanding of the potential outcome, we coded it as NOC (not otherwise classifiable) and omitted it from the analysis. Fig. 4 presents two contrasting narratives that illustrate the challenges of coding.

Given the challenges of coding near-misses compared to injuries we were pleased with our level of substantial agreement. We obtained 79% agreement, which is comparable to research by Lehto et al. (2009) showing 75% agreement. Percentage agreement and kappa statistics were in the lower range of previous studies (Bondy et al., 2005; Lehto et al., 2009; Lombardi et al., 2009; Marucci-Wellman et al., 2011). Therefore, we conclude that this method – which has been rigorously applied to injuries – is substantiated for scenarios with less definitive outcomes like near-misses.

4.2. Structure of the data system

In NFFNMRS, reporters are asked to "Describe the event", allowing them to say anything. Narratives often begin with information about arrival and staging which are events that precede the beginning of the chain of events leading up to an injury or near-miss. It is important to note that the "Describe the event" field does not ask specific questions about any injuries that did happen or could have happened. This is different than other data systems like the National Health Interview Survey (NHIS) which asks "How did your injury on [date] happen? Please describe fully the circumstances or events leading to the injury and any objects, substances, or other people involved". In addition to asking how the injury occurred, the NHIS also asks a series of specific prompts to seek for more detailed information for certain causes such as whether the injured individual was in a motor vehicle, on a bike, scooter, skateboard, skates, skis, horse, etc., a pedestrian who was struck by a vehicle such as a car or bicycle, in a boat, train, or plane, suffered a fall, or burned or scalded by substances such as hot objects or liquids, fire, or chemicals (National Center for Health Statistics, 2009). The average length of narratives within the NFFNMRS dataset is quite long (mean word count 216), as compared to other datasets. In contrast, narratives from the NHIS contain 11 words on average (Wellman et al., 2004). Furthermore, the time required to manually code our initial 1000 narratives was approximately 25 h per coder, with an additional 25 h required for reconciliation of these 1000 narratives. Using worker's compensation narratives of approximately 20 words, Bertke et al. (2012) stated that it took them 10 h to code 2400 worker's compensation claims—which is 2.4 times the number of narratives we coded, in less than half of the time. We observed that coding of lengthy narratives – especially those without known injury outcome – is very time consuming and requires extensive human resources. Therefore the algorithm's high performance is especially welcome for narratives that emanate from generic prompts such as "Describe the event".

4.3. Performance of autocoding

Considering that coding of near-miss narratives via automated methods has not been previously described in the literature we were pleased with the performance level of the algorithm on near-miss narratives, reaching above 70% specificity.

The higher performance of the Fuzzy model as compared to the Naïve model was not too surprising, given the longer

Table 6
Comparison of results with previous auto-coding studies.

	Motor vehicle accident data (Lehto and Sorock, 1996)	NHIS (Wellman et al., 2004)	Worker's Comp (Lehto et al., 2009)	Worker's Comp (Marucci-Wellman et al., 2011)	Worker's Comp (Ohio) (Bertke et al., 2012)	NFFNMRS narratives (current study)
Narrative type	Insurance company automotive accident narratives	General population injury narratives	Worker's compensation injury narrative	Worker's compensation injury narrative	Ohio Bureau of Worker's Compensation Claims	Fire fighter-occupation specific narratives, with near-misses & injury
Narrative characteristics	Short narratives (2–3 sentences long)	Short narratives (avg. 11 words)	Short narratives (avg. 20 words)	Short narratives (avg. 20 words)		Long narratives (avg. 216 words)
#Cause categories	9 coding categories (2 main groups)	13 coding categories	21 coding categories used (out of 40 OIICS codes)	21 coding categories used (out of 40 OIICS codes)	3 broad coding categories; 8 specific coding categories	14 coding categories
Coding scheme	2 categories: Pre-crash (5 codes) and Crash (4 codes)	ICD9-CM (2-digit)	OIICS classification (2-digit)	OIICS classification (2-digit)	OIICS classification	ICD9-CM (3-digit)
Size of dataset	3686	5677 (all pre-coded)	17,000 (uncoded)	17,000 (uncoded)	10,132 (uncoded)	2280 (uncoded)
Training set size; % of dataset	3686 narratives; training set was a set of keywords, not coded narratives	5677; 100%	11,000 (manually coded)	11,000 (manually coded); Training set 367% larger than prediction set	2240 (2400, minus 160 due to coder disagreement or NOC); 22.1%	Total of 1000 manually coded with 764 used to train the algorithm; 43.4%
Coder agreement	Only 1 coder	n/a—records pre-coded	Overall 1-digit agreement of 87%; 2-digit agreement of 75%	Overall 1-digit agreement of 87%; 2-digit agreement of 75%	Overall agreement of 93.8%	Final coder agreement greater than 79% ($\kappa > 0.75$)
Prediction set	419	5677 (same as training set)	3000 (pre-coded)	3000	7732	2285 (includes training set)
Training set modifications	Keyword list of 2619 was morphed, endings removed (ing, ed), articles removed, misspellings corrected	Creation of keyword list-words occurring more than 3 times in dataset; drop word lists; synonym words	Drop word list; drop words occurring fewer than 3 times; remove punctuation and non-alphanumeric characters	List of keywords and drop words was generated; transformation of synonyms; correction of misspelling	None described.	Drop words occurring fewer than 3 times. No synonyms, or stop words
Analyses	Leave-one-out/Naïve Bayesian and Fuzzy Bayesian	Single word Fuzzy; Multiple word Fuzzy (single words, up to 4-word combos)	Naïve and Fuzzy Bayes (comparison)	Naïve and Fuzzy Bayes (combined); 1st strategy: assign cases for manual review if Fuzzy and Naïve models disagree; 2nd Strategy: selection of additional cases for manual review from Agree dataset using prediction strength to reach level of 50% computer and 50% manual coding	Assessed number of categories and size of training set on prediction set sensitivity. Assessed use of training set from one sector upon another sector	Fuzzy Bayesian and Naïve Bayesian models using Single word predictor; comparison of predictive ability as it relates to injury or near-miss
Distribution of codes	Not provided in results	Heavily weighted to falls (35%), followed by struck-by (16%), and overexertion (12%)	Weighted toward overexertion (17.8%), falls (17.4%) and struck-by (9.8%)	Weighted toward overexertion (17.8%), falls (17.4%) and struck-by (9.8%)	Weighted toward contact with object or equipment (49.3%), slips, trips and falls (23.8%), and musculoskeletal disorders (18.0%)	Weighted mostly to fire-fall (25.7%), fire-struck-by (24.1%) and fire-burn (22.1%)
Results	Keyword based classification results consistently good. Fuzzy Bayes can augment results in cases where keyword classification failed and in categories where keyword classification performed poorly	A computer program based on fuzzy Bayes logic is capable of accurately categorizing cause-of-injury codes from injury narratives. The ability to set threshold levels significantly reduced the amount of manual coding required, without sacrificing accuracy	Single-digit codes predicted better than double-digit; Naïve slightly more accurate than Fuzzy; Naïve had sensitivity of 80% and 70% (for one and two digit codes, respectively), Fuzzy Bayes had a sensitivity of 78% and 64%. Specificity and PPV was higher in Naïve than Fuzzy	1st strategy: agreement alone as filtering strategy left 36% for manual review (computer coded 64%, $n = 1928$). Overall combined sensitivity was 0.90 and PPV > 0.90 for 11 of 18 2-digit categories	Naïve Bayesian auto-coding of narrative text and injury diagnosis showed up to 90% accuracy, improvement in performance with increased training size, and training sets with broader coding performed as well or better to predict more specific sector claims	The Fuzzy model performed better than Naïve, with a sensitivity of 0.74 compared to 0.678. As the number of records in the training set increased, the models performed at a higher sensitivity. Both injuries and near-misses could be predicted, but injuries were predicted with greater sensitivity

INJURY OUTCOME: Near-miss

"On arrival, there was fire showing on the second floor "B" side of a two and one half story wood-frame residential structure. We had been operating a two and one half inch attack line for approximately ten minutes. As Division Two Commander, I felt at that time that we were beginning to lose progress. On orders of the Incident Commander, orders were given to immediately evacuate the second floor. As Division Two Commander, all crews were evacuated, excluding myself and two other crew members in a final effort, despite the Incident Commanders orders. Upon evacuating, after the Incident Commander's second order to evacuate, we observed the second floor ***flashover*** *and* ***collapse.****"*

Coders' determination:

PRECIPITATING CAUSE: Fire-Collapse

PROXIMAL CAUSE: Fire

In this narrative, the potential resultant injuries included possible struck-by or fall from the collapse, and/or burn injuries from the flashover. The collapse was the most precipitating of causes, while the generic code of "Fire" as the proximal cause indicated that too many possible outcomes were described, preventing a single cause to be identified.

INJURY OUTCOME: Injury

"This was an extremely cold night, 0400, -4 degrees. First due companies found a well involved dwelling. The Incident Commander call for a 2nd alarm. I was the lieutenant to the 4th due Engine Company. We were told to search the second floor for victims. We had cars in the driveway and kids' toys in the driveway as well. No one was coming forward to say everyone was out of the building. The main body of fire was knocked down when we made access to the second floor bathroom. One of my crew asked me if I had checked the bath tub because kids hide in tubs. ***I reach into the tub just as its weight caused the floor in the bathroom to collapse.***

I don't remember falling but I do remember hearing the Mayday from my crew and wondering to myself "Oh My god who's hurt". I had landed on the first floor on my back with my SCBA on. The crew on the first floor got me up and out of the building. I had no obvious injury and went back to work, until three hours later. I was back at the station making out reports, when a low back spasm caused me excruciating pain. I was taken to the hospital with torn muscles in my lumbar area. I was out of work for 6 months. The people that lived in the house were outside in a squad car. The patrolman never told us everyone was out of the building."

Coders' determination:

PRECIPITATING CAUSE: Fire-Collapse

PROXIMAL CAUSE: Fire-Fall

In this narrative, a clear sequence of events is presented. Working backwards from the injury (torn back muscles), he suffered a fall (proximal), from collapse of second floor bathroom (precipitating).

Fig. 4. Example narratives: near-miss and Injury comparison.

narratives in our dataset. Previous research has used much shorter narratives and seen exceptional performance by the Naïve model—particularly because with less words in each narrative, there are fewer opportunities for a strong predictive word to outweigh the other words within the narrative. Categories with fewer narratives tended to be better predicted by the Naïve model, likely because it took into account all words, rather than picking words with the single strongest predictor (as in Fuzzy). Previous research done with the TextMiner software has been applied to shorter narratives, predominately using the Naïve model.

To further elucidate why the Fuzzy model was performing better, we checked to see if there was any evidence of overfitting of the data. We analyzed the difference in correct predictions between the training set and prediction set, for both the Fuzzy and Naïve models and found no indication that the Fuzzy model was overfitting the data, suggesting that the Fuzzy model truly does perform better with this particular dataset.

The results of the Cross-validation showed that when the Fuzzy and Naïve models both predicted the same code, the agreement of the autocoding to the manual codes reached 60.2%. In using a similar technique to filter cases for manual review, Marucci-Wellman et al. (2011) reported a Fuzzy and Naïve agreement of 64%. Applying both the Fuzzy and Naïve models to a dataset could be another way of optimizing the performance and accuracy of autocoding.

The process of adding narratives to the training set in increments of 100 showed marked improvement, suggesting that the algorithm was learning with each addition. It did not appear that the algorithm had yet reached a threshold, suggesting that addition of more cases (beyond 764) will result in improved prediction rate by the models. The work of Bertke et al. (2012) showed increasing improvement in the sensitivity as the training set increased up to 1000 (with remaining cases as prediction set $n = 1240$), with marginal returns beyond that.

Using additional modifications such as paired words, word sequences, morphs, and drop words lists would likely improve the hit rate. In fact, in a preliminary analysis using paired words and 3-word sequences, we saw an increase in prediction success by the Fuzzy model (82% and 85%, respectively, data not shown). However, this indicates that with minor modifications, the predictive capability of the algorithm can improve to a significantly higher level of sensitivity, thereby reducing the amount of narratives that would need manual review.

4.4. Creation of additional quantitative data elements

Applying the Bayesian models enabled us to create two new quantitative data elements: injury outcome and mechanism of injury. This enriches the analysis of existing quantitative data in the NFFNMRS because we can look at differences between near-misses and injuries, and construct hazard scenarios.

5. Conclusion

In this study of narratives from the fire service we were able to successfully apply the Fuzzy and Naïve models to injury and near-miss narratives, which were much longer than those that have previously been investigated. While both models had relatively high sensitivity, Fuzzy proved to be the more agile model for very long narratives.

We trained the algorithm to assign a mechanism of injury and an injury outcome for each narrative. This process resulted in the creation of two new quantitative data elements that will empower more in-depth analyses of the National Fire Fighter Near Miss Reporting System.

Previous studies have the benefit of their short narratives emanating from specific questions about how the injury occurred. That the near miss narratives had fairly vague instructions to “describe the event”, and that the machine learning methods were able to assign a specific mechanism of injury code is a testament to the power of Bayesian models. An important point is that no effort was made in the current study to optimize the predictive models used. Additional steps could be taken that would be likely to improve the performance of both models, such as increasing the sample size. Other steps are also likely to lead to significant improvements, such as trimming the word set by dropping common noise words to improve performance of the Naïve Bayes model, or using word combinations and sequences to increase the sensitivity of the Fuzzy model.

Acknowledgments

We would like to thank the International Association of Fire Chiefs for the opportunity to work with these data.

This research was supported through a grant from the National Institute for Occupational Safety and Health (no. 5R03OH00984-02).

References

Alamgir, H., Yu, S., Gorman, E., Ngan, K., Guzman, J., 2009. Near miss and minor occupational injury: does it share a common causal pathway with major injury? American Journal of Industrial Medicine 52 (1), 69–75.

Aspden, P., Corrigan, J.M., Wolcott, J., Erickson, S.M. (Eds.), 2004. Near-Miss Analysis. Patient Safety: Achieving a New Standard for Care. The National Academies Press, Washington, DC, pp. 226–245.

Barach, P., Small, S.D., 2000. Reporting and preventing medical mishaps: lessons from non-medical near miss reporting systems. BMJ 320 (7237), 759–763.

Bertke, S.J., Meyers, A.R., Wurzelbacher, S.J., Bell, J., Lampl, M.L., Robins, D., 2012. Development and evaluation of a naive Bayesian model for coding causation of workers’ compensation claims. Journal of Safety Research 43 (5–6), 327–332.

Bondy, J., Lipscomb, H., Guarini, K., Glazner, J.E., 2005. Methods for using narrative text from injury reports to identify factors contributing to construction injury. American Journal Industrial Medicine 48 (5), 373–380.

Bunn, T.L., Slavova, S., Hall, L., 2008. Narrative text analysis of Kentucky tractor fatality reports. Accident Analysis and Prevention 40 (2), 419–425.

Leape, L.L., 2002. Reporting of adverse events. New England Journal of Medicine 347 (20), 1633–1638.

Lehto, M., Marucci-Wellman, H., Corns, H., 2009. Bayesian methods: a useful tool for classifying injury narratives into cause groups. Injury Prevention 15 (4), 259–265.

Lehto, M.R., Sorock, G.S., 1996. Machine learning of motor vehicle accident categories from narrative data. Methods of Information in Medicine 35 (4-5), 309–316.

Lincoln, A.E., Sorock, G.S., Courtney, T.K., Wellman, H.M., Smith, G.S., Amoroso, P.J., 2004. Using narrative text and coded data to develop hazard scenarios for occupational injury interventions. Injury Prevention 10 (4), 249–254.

Lipscomb, H.J., Glazner, J., Bondy, J., Lezotte, D., Guarini, K., 2004. Analysis of text from injury reports improves understanding of construction falls. Journal of Occupational and Environmental Medicine 46 (11), 1166–1173.

Lombardi, D.A., Matz, S., Brennan, M.J., Smith, G.S., Courtney, T.K., 2009. Etiology of work-related electrical injuries: a narrative analysis of workers’ compensation claims. Journal of Occupational and Environmental Hygiene 6 (10), 612–623.

Marucci-Wellman, H., Lehto, M., Corns, H., 2011. A combined fuzzy and naive Bayesian strategy can be used to assign event codes to injury narratives. Injury Prevention 17 (6), 407–414.

National Center for Health Statistics, 2009. Injury and Poisoning Questions on the National Health Interview Survey: 1997-Present. Centers for Disease Control and Prevention.

Patel, M.D., Rose, K.M., Owens, C.R., Bang, H., Kaufman, J.S., 2012. Performance of automated and manual coding systems for occupational data: a case study of historical records. American Journal of Industrial Medicine 55 (3), 228–231.

Rivard, P.E., Rosen, A.K., Carroll, J.S., 2006. Enhancing patient safety through organizational learning: are patient safety indicators a step in the right direction? Health Services Research 41, 1633–1653, 4p2.

Smith, G.S., Timmons, R.A., Lombardi, D.A., Mamidi, D.K., Matz, S., Courtney, T.K., Perry, M.J., 2006. Work-related ladder fall fractures: identification and diagnosis validation using narrative text. Accident Analysis and Prevention 38 (5), 973–980.

Stout, N., 1998. Analysis of narrative text fields in occupational injury data. In: Occupational Injury: Risk Prevention and Intervention. Taylor & Francis, United Kingdom, pp. 15–20.

Wellman, H.M., Lehto, M.R., Sorock, G.S., Smith, G.S., 2004. Computerized coding of injury narrative data from the national health interview survey. Accident Analysis & Prevention 36 (2), 165–171.