Published in final edited form as:

Inj Prev. 2016 April ; 22(Suppl 1): i34–i42. doi:10.1136/injuryprev-2015-041813.

Harnessing information from injury narratives in the 'big data' era: Understanding and applying machine learning for injury surveillance

Kirsten Vallmuur[*],
Queensland University of Technology, Centre for Accident Research and Road Safety – Queensland, Queensland, Australia

Helen R Marucci-Wellman,
Center for Injury Epidemiology, Liberty Mutual Research Institute for Safety, Hopkinton, MA, USA

Jennifer A Taylor,
Department of Environmental & Occupational Health, School of Public Health, Drexel University, Philadelphia, PA, USA

Mark Lehto,
School of Industrial Engineering, Purdue University, West Lafayette, IN, USA

Helen L Corns, and
Center for Injury Epidemiology, Liberty Mutual Research Institute for Safety, Hopkinton MA, USA

Gordon S Smith
National Center for Trauma and EMS, University of Maryland School of Medicine, Baltimore, MD, USA

Abstract

Objective—Vast amounts of injury narratives are collected daily and are available electronically in real time and have great potential for use in injury surveillance and evaluation. Machine learning algorithms have been developed to assist in identifying cases and classifying mechanisms leading to injury in a much timelier manner than is possible when relying on manual coding of narratives. The aim of this paper is to describe the background, growth, value, challenges and future directions of machine learning as applied to injury surveillance.

Methods—This paper reviews key aspects of machine learning using injury narratives, providing a case study to demonstrate an application to an established human-machine learning approach.

[*] Corresponding author details: Queensland University of Technology, Centre for Accident Research and Road Safety – Queensland, 130 Victoria Park Road, Kelvin Grove 4059, Brisbane, Queensland, Australia, k.vallmuur@qut.edu.au Phone: +61 7 3138 9753 Fax: +61 7 3138 5515.

Competing Interests

No competing interests to declare.

Authors Contributions

KV planned the manuscript, drafted sections, consolidated and revised drafts from authors, and prepared the final manuscript. ML, HW and HC wrote the case study. JT, ML, HW, HC, and GS provided drafts of sections and edited and revised consecutive drafts of the paper.

Results—The range of applications and utility of narrative text has increased greatly with advancements in computing techniques over time. Practical and feasible methods exist for semi-automatic classification of injury narratives which are accurate, efficient and meaningful. The human-machine learning approach described in the case study achieved high sensitivity and positive predictive value and reduced the need for human coding to less than one-third of cases in one large occupational injury database.

Conclusion—The last 20 years have seen a dramatic change in the potential for technological advancements in injury surveillance. Machine learning of 'big injury narrative data' opens up many possibilities for expanded sources of data which can provide more comprehensive, ongoing and timely surveillance to inform future injury prevention policy and practice.

Keywords

Injury surveillance; Machine learning; Narrative text; Coding

Introduction

Injury narratives have long been recognized as valuable sources of information to understand injury circumstances and are increasingly available in the era of 'big data'. Narrative text mining and machine learning techniques have been developed that take advantage of greatly increased computing power and 'big data' to make predictions based on algorithms constructed from the data. However, along with the opportunities, challenges in adequately accessing and utilizing injury narratives for public health surveillance and prevention exist. In this paper the authors describe the background, growth and utility of machine learning of injury narratives. A case study is also provided to demonstrate the application of an established human-machine learning approach. The authors then discuss the challenges and future directions of machine learning as applied to injury surveillance.

Background

The 1990's marked the beginning of the electronic era, e-mail and the internet were surfacing and electronic records took the form of .dbf files transcribed from hard copy files. In a 1997 article Sorock and colleagues identified innovative approaches to improvements in work-related injury surveillance that reflected the utility of electronic records at this time (1). These include: (1) the use of narrative text fields from injury databases to extract useful epidemiologic data; (2) data set linkage for aiding in incidence rate calculations and (3) the development of comprehensive company-wide injury surveillance systems. Now almost 20 years later, the opportunities have expanded greatly; Large amounts of coded injury data and text descriptions of injury circumstances (injury narratives) are being collected daily and are available in real time. However, while there have been some collective efforts to standardize injury data collection and classification systems, very little has been done to develop and standardize machine learning approaches using injury narratives.

WHO guidelines specify the following requirements for injury surveillance: to facilitate ongoing data collection, in a systematic way, which enables analysis and interpretation for timely dissemination which can be applied to prevention and control (2). However, often injury information (for morbidity and mortality incidence reporting) is collected and may be classified without considering these requirements. While the data may be coded according to a standardized classification protocol (e.g. ICD coding in hospitals) the people assigning the codes are often administrative staff classifying the case for billing purposes (not for prevention), with little profession training although hospital discharge data is usually coded by a professional nosologists. In order to get these data re-coded in such a way as to satisfy the requirements of surveillance requires significant investment and resources.

On the other hand there are some national agencies such as the National Center for Health Statistics which in addition to mortality coding use their nosologists to classify medical conditions, drugs and injuries reported in their large national health surveys in the United States (e.g. the National Health and Nutrition Examination Survey and the National Health Survey). Coding systems useful to injury epidemiologists include: the International Classification of Diseases (ICD), International Classification of External Causes (ICECI) (3), and Nordic Classification of External Causes (NOMESCO) (4). Occupational injury surveillance systems however usually assign and utilize separate coding strategies aimed at identifying work exposures such as the National Institute for Occupational Safety and Health (NIOSH) Occupational Injury and Illness Classification System (OIICS) (5) and the Type Of Occurrence Classification Scheme (TOOCS) (6). These codes are often used for surveillance. However, even if the time and resources have been allotted to having trained coders assign these codes, there are still limitations in using the coded data alone. These include the limited scope, breadth and depth of injury mechanisms and scenarios captured from the codes (specifically reducing their value for injury prevention and control) and reliance on predetermined circumstances that may not capture all or the very unique case scenarios (7), nor all relevant injury factors (host, agent, vector, environment) contributing to an injury event as defined by Haddon(8).

The utility of injury narratives for surveillance

Two recent reviews (9, 10) outlined a range of benefits for using narratives as a supplement to the restrictions of coded data, including: the identification of cases not able to be detected from coded data elements alone, extracting more specific information than codes allow, extracting data fields which aren't part of the prior coding schemas, establishing chain-of-events, identifying causes without specific codes, and assessing coding accuracy.

Narrative text analyses also enables the identification of rare or emerging events usually not found using administratively assigned codes, a critical concern in injury surveillance (11–14). Incident narratives in their raw form can also be available in a more 'timely' manner than coded data and are now being used in novel applications such as syndromic surveillance (15, 16).

The range of applications and utility of narrative text has also increased with recent advancements in computing techniques. However, some of the earliest applications predate the ability to search text electronically and were simply to identify cases to overcome the lack of reported or coded data. These include using newspaper clipping services where people were paid to read newspapers and identify articles that reference any of the injury or fatality topics on a list related to clients' interests who had paid the service to look for articles containing target words about specific companies (17) (18). Now that news articles are on the web, computerized search has greatly simplified the process of searching for injury incidents using services such as Nexus.

Nowadays, with significant increases in the technological capabilities and capacity of computer systems, injury narratives which contain essential information about how the injury event occurred are more widely available in an ‘ongoing’ manner across a range of agencies [including but not limited to emergency services/first responders (ambulance, fire service, police), emergency departments/hospitals/trauma registries, coronial systems, occupational health and safety, insurance/compensation agencies (workplace/health/motor vehicle), consumer safety agencies, news services and even social networking sites (twitter/facebook) etc].

However, utilizing these data for surveillance has historically proven cost-prohibitive and fraught with human error. Bertke et al (2012) reported that it took a single researcher 10 hours (over the course of a few weeks to mitigate fatigue) to code 2,400 workers’ compensation injuries (19). Taylor et al reported 100 total hours for three coders to discern cause of injury and reconcile differences from firefighter near-miss and injury narratives (20). As a database grows, the additional resources required to code the records become increasingly labor, cost, and time prohibitive. Only recently has the use of computerized coding algorithms enabled large-scale analysis of narrative text, presenting an efficient and plausible way for individuals to code large narrative datasets with accuracies of up to 90% (19, 21). While auto-coding increases accuracy and efficiency, but it does not eliminate the need for human review entirely as humans must initially train the algorithm and conduct post-hoc quality review.

There have been some limited situations where automated classification of injury narratives has become integrated into routine processes for national statistical purposes to reduce the amount and costs of manual coding, improve coding uniformity and reduce the time taken to process records. For example, many countries use software to automatically process injury text recorded on death certificates for broad ICD cause of death coding (22) and the National Institute for Occupational Safety and Health in the USA has made available an online tool to aid state public health organizations in determining NIOSH occupation and industry codes (23). These software programs built over several decades allow a substantial subset of records to be automatically coded usually with the caveat of limited accuracy. The accuracy however can often be improved if the algorithm is able to identify those which would be more accurately coded by humans (or should be unclassifiable) or that the software cannot confidently assign a code.

Over the past two decades, several authors of this paper have completed a number of studies ((1, 24) (25) (21, 26, 27) (20)) on the utilization of computer algorithms to streamline the classification of the event (or causes) documented in injury narratives for surveillance purposes. Their focus has been to create machine learning techniques to quickly filter through hundreds of thousands of narratives to accurately and consistently classify and track high magnitude, high risk and emerging causes of injury, information which can be used to guide the development of interventions for prevention of future injury incidents (28). The results of this work has enabled the annual classification of very large batches of workers compensation (WC) claim incident narratives into Bureau of Labor Statistics (BLS) occupational injury and illness classification (OIIC) event codes for input in deriving the annual Liberty Mutual Workplace Safety Index --a surveillance metric ranking the leading causes (in terms of direct cost WC cost) of the most disabling work-related injuries in the U.S. every year (29).

Table 1 also provides examples of other studies, describing both early uses and other more complex uses of narrative text. These examples include the integration of machine learning techniques to demonstrate the changing nature of this field.

Case study

To demonstrate one successful approach to the use of machine learning to classify injury narratives, the following case study briefly summarizes a recent study by Marucci-Wellman et al (26) that accurately classifed 30,000 workers compensation (WC)narratives into injury events using a human-machine learning approach in order to match cost of claims by event category with national counts from the BLS Survey of Occupational Injury and Illness data. Coders who had been trained extensively on the BLS Occupational Injury and Illness Classification System (OIICS) read each claim accident narrative on the case and classified the event that led to work-related injury into one of approximately 40 2-digit event codes. The dataset was divided into a training set of 15,000 cases, used for model development, and a prediction dataset of 15,000 cases used for evaluating the algorithms performance on new narratives. A sample of WC claims accident narratives with BLS OIICS code assignments are shown below:

1. "STANDING UP FROM BENDING OVER STRUCK BACK ON MAID CART" -> Classified as BLS OIICS event code 63 - struck against object or equipment.
2. "FELT PAIN WHILE PULLING LOAD OF WOOD WITH PALLET JACK" -> Classified as BLS OIICS event code 71 – overexertion involving outside sources.
3. "STOPPED AT STOP SIGN WHEN REAR-ENDED BY ANOTHER VEH." -> Classified as BLS OIICS event code 26 - Roadway incidents involving motorized land vehicle.
4. "SLIPPED AND FELL ON UNK SURFACE TWISTING HIS ANKLE SPRAININGIT".-> Classified as BLS OIICS event code 42 - Falls on same level.
5. "EMPLOYEE WAS WALKING ON THE STREET WHEN HIS RIGHT KNEE POPPED" ->Classified as BLS OIICS event code 73 - Other exertions or bodily reactions.

Using the 15,000 narratives and manually assigned codes from the training set, a keyword list was created by parsing the words in each narrative (e.g., standing, up, from, bending, etc.). The occurrence or probability of each word in each category (Pn_j/C_i) was calculated as well as the marginal probability of each event category in the training data set ($P(C_i)$; These are the two parameters necessary for the reduced Naïve Bayes algorithm ((26)). These statistics calculated from the training narratives were stored in a probability table and used to train the algorithm. A similar word list and probability table was constructed for 2, 3 and 4 word sequences (each sequence considered as a keyword, e.g. standing-up, up-from, from-bending, standing-up-from etc.). The Naïve Bayes model was used to assign a probability to each event code based on the keywords present in a particular narrative. The event code with the largest estimated probability was then chosen as the prediction for the words present.

The theoretical basis for the Naïve Bayes classifier and detailed instructions on how to implement the algorithm with narrative data have been thoroughly defined previously (21, 26). Various software packages are now publically available for training (or building) the models based on the training dataset and then making subsequent predictions. Weka (39) and Python (40) are two examples of publically available, easily downloadable and easily adaptable packages for development of the Naïve Bayes Model. For this study, the Textminer software developed by one of the authors (ML) was used. The narratives were used in their raw form; although improved performance can be expected when misspellings are cleaned and words that have the same meaning are morphed into one syntax, the aim was to demonstrate what could be achieved by machine learning with little pre-processing of the narratives. However, a small list of frequently occurring “stop words” believed to have little meaning for the classification assignment (e.g. a, and, left, right) was removed from the narratives prior to calculating probabilities.

Two Naïve Bayes algorithms were run on each of the 15,000 prediction narratives using first the set of single keyword probabilities and second the sequenced keyword probabilities (stored in probability tables) from the training narratives in order to assign two independent computer generated classifications to the 15,000 prediction narratives.

The authors (26) found while the overall sensitivity of the two independent models was fairly good (0.67 naive_{sw}, 0.65 naive_{seq}), both algorithms independently predicted some categories much better than others, skewing the final distribution of the coded data (χ^2 P<0.0001), and most of the cases in the smaller categories were not found. The sequence-word model showed improved performance where word order was important for differentiating causality. Still many categories had low performance. We consequently integrated a rule where we would *only* use the computer classifications when the two models agreed and then would manually code the remaining narratives. Implementing this rule resulted in an overall sensitivity of codes for the final coded dataset of 87% with high sensitivity and positive predictive values across all categories (See Table 2 and 3 and Marucci-Wellman et al (26) for more details). Note, both high sensitivity and positive predictive value is important for resulting in a final unbiased distribution of the coded data for surveillance and targeting prevention efforts. Also using this human-machine pairing resulted in 68% of the narratives coded by the algorithm leaving only 32% to be coded by humans.

The authors found the accuracy of the human-machine system was at least as good and likely was even better than manual coding alone of all 15,000 records as the system uses consistent rules. This was demonstrated by comparing the results with inter-rater reliability data for four well trained human coders. While the evaluation of inter-rater reliability relies on different metrics, the inter -rater reliability performance of the four coders does not appear to be as systematically high and consistent as what is projected from the sensitivity and positive predictive value (PPV) values of the human-machine pairing method for the very large categories, nor the very small categories. Other readily available and easily adaptable machine learning techniques for narrative text analyses other than the Bayesian algorithms exist such as support vector machine (SVM) and logistic regression (LR) and could also be incorporated to improve accuracy. Work has begun to investigate ensembles consisting of agreement between these various algorithms with some slightly improved results over the ones presented in the case study summary (See Table 4). Overall, this case study demonstrates that a practical and feasible method exists for human-machine learning of short injury narratives. The computer was able to accurately classify many of the narratives of a large WC dataset leaving one-third for human review and resulting in a very high overall accuracy and very high accuracy across almost all categories (large and small) in the final coded dataset. Accuracy can be further improved when a percent of difficult cases, predicted by the algorithm with a low confidence, are rejected for manual coding.

Discussion: Challenges and future directions

As illustrated in the previous case study, the use of off-the-shelf machine learning methods combined with human review of weakly predicted cases is an effective, easily applied method. However, this approach still required developing a large training set of previously coded cases to develop the model and then subsequent human review of around 1/3 of the cases to attain high sensitivities across all categories in the prediction set. In practice, obtaining a good training set and the need for human review (which could be substantial if 1/3 of a very large data set still requires manual coding) may both be major application bottlenecks. Numerous strategies and approaches for tailoring methods to address this problem exist. For the most part, these strategies and approaches can be roughly divided as: focusing on obtaining more data (a larger training set), applying better learning algorithms, or going beyond the training set, using other sources of information, causal models, or human knowledge to preprocess the information used by the learning algorithm. The following discussion briefly builds on ideas generated by the case study and introduces some of these other approaches, their effectiveness, and emerging trends in their use.

Obtaining more data or applying better algorithms

The use of a larger training set and better learning algorithms are both commonly suggested strategies for improving model predictions. Previous work (32) has shown that model performance improves for short injury narratives with larger training sets. The latter study also showed that SVM algorithm performed better than Naïve Bayes and several other learning algorithms. However, the improvements were clearly slowing down as the increase of training data continued. Furthermore, smaller categories were often poorly predicted by the algorithm, just as found in the case study above for Naïve Bayes, Logistic Regression, and SVM. Some further improvements in the SVM model performance were also observed by Chen et al. (32) after model factorization using Singular Value Decomposition to map the word vectors to a lower dimensional space. The latter result was consistent with earlier studies showing improvements after feature space reduction using Singular Value Decomposition (SVD) (41, 42), and SVD approaches are likely to be especially useful in ‘big data’ applications where there is substantial training data available for mapping the lower dimensional space.

Preprocessing data

Overall though, the results using thousands of training examples across multiple studies suggest that it is doubtful that the need for human review will be completely eliminated with more data or by better learning algorithms alone for complex multi-class coding schemes and especially so when there is a need to assign rarely occurring categories (i.e. needle stick injuries in the case study). One potentially promising strategy for improving performance for smaller categories is to go beyond the training set, using other sources of information, causal models, or human knowledge to preprocess the information used by the learning algorithm. Numerous approaches have been used for preprocessing injury text prior to applying the learning algorithms such as word stemming, lemmatization, dropping infrequent or frequent words, or weighting schemes such as TF-IDF (32). One advantage of such approaches is that they provide an easy way of reducing the dimensionality of the word vector, which can speed learning of any machine learning algorithm. However, this may sacrifice accuracy, with the authors preliminary work using Naïve Bayes, Logistic Regression, and SVM showing that these pre-processing approaches have the potential to reduce the overall detection (distinguishing between categories) capability, and especially for small categories (43). Part of the problem is that such approaches do not consider the meaning of words. For example, in related as yet unpublished work, the authors found that stemming or lemmatizing the words “lifting” and “lifts” to their root “lift” reduces the ability of SVM, NB, and LR to distinguish injuries related to exertion from those caused by man lifts or fork lifts. Similarly, dropping infrequent words in this large word set of 10,000 words such as “muggers” or “rape” reduced the ability to identify assault cases.

Targeted mapping of only certain words to a common meaning, on the other hand, tended to improve performance (for example, HOT and SCALDING or bike and bicycle).The latter approach was especially useful for finding predictive word sequences (for example, “all words that mean a person” followed by the word “fell” separates struck by events from fall events). Based on the author’s preliminary results, systematic development of a lexicon mapping words, word-sequences, and word combinations that relate to important concepts can greatly improve the sensitivity across categories of any machine learning algorithm. For example, the authors found the generic concept “hit body part on” identified as a sequence of words that can mean hit, followed by words that can mean a body part, followed by either the frequent words “or” or “against”, greatly improved the ability of Naïve Bayes, SVM, and LR alike to distinguish struck against events from both falls and struck against events. The finding that a good lexicon can improve the performance of machine learning algorithms for short injury narratives is not surprising. The caveat is that manually developing a good lexicon is very time consuming, since datasets will contain thousands of unique words and words will have different meanings depending on what other words are present (really requiring topic appropriate linguist experts to do this work). Further complicating the matter, a causal model may be necessary to organize the concepts into a predictive model. Illustrating recent developments in this direction, Abdat, et al (44) developed a causal model of construction accidents using a Bayesian network to identify the probable explanation of accidents based on generic factors extracted by expert from accident scenarios. Other work in this direction included the use of automated named entity recognition techniques to automatically parse unstructured data from several databases which were then used in a Bayesian network to identify and code safety factors (35).

An interesting conjecture is that these findings suggest a lexicon or causal factors generated from one text mining project can be used to help code another project's uncoded narratives. Transfer of results would seem to be especially promising when data sets have the same focus, like occupational hazards. For example, if the results obtained using the database from the National Firefighter Near-Miss Reporting System (NFFNMRS) (20) were applied to narratives from the Fire Fighter Fatality Investigation and Prevention Program (FFFIPP), one would expect falls to be predicted with fairly good accuracy because the language firefighters use to describe their hazards is similar ("roof, spongy" are precise predictors for firefighter falls caused from a weakening roof on fire). Similarly, a multitude of terms identified as toxic chemicals (e.g. hydrogen sulfide, toluene) in one data set could be directly mapped to the concept "toxic chemical" used in a new application, rather than relying on the training set alone. Future studies might also explore how well key words and word predictors in a home and leisure injury database (25) would predict injuries in occupational narratives. If one wanted to auto code causes of injury in firefighter narratives using results obtained from a knowledge database (meaning a collection of either narratives linked to manually assigned codes or word lists with corresponding probability weights) created from a home and leisure population level database, the terms used to describe important concepts in a fire fighter database could be nodes in a Bayesian network retrained using the home and leisure injury database to estimate probability weights (Pn_j/C_i) for the new database. The new weights would adjust the original weights for terms such as "roof, spongy" used as a precise predictor for firefighter falls but unlikely to indicate a fall when at home or in leisure activities. This approach will enable the development of weighting coefficients (as adjustments) to the probabilities that comprise the knowledge database before it is transferred from population narratives to occupational narratives. This work – while currently hypothetical – would, if feasible, provide critical proof of concept: if high specificity, sensitivity, and positive predictive value are able to be attained, there would be good evidence that weighting of probabilities would be the next step in making machine learning algorithms more broadly transferrable helping to reduce resources needed for human coding.

Building an open source knowledgebase

For machine learning algorithms to be broadly utilized, they need to be accessible and refined in an open source manner. Ideally, researchers could share both data and algorithms, perhaps in a cloud-based shared-access knowledge database. Along these lines, Purdue University (ML) is in the process of creating an open source framework that can serve as a repository for shared injury coding knowledge databases. This framework would allow remote access to datasets of coded and uncoded narratives, machine learning algorithms, lexicons, and other information, enabling researchers to share their results, develop better models more quickly, and ultimately reduce the need to manually code in the traditionally resource-dependent manner. The expectation is that as the open source repository grows, new models will be developed that accurately code injury narratives within specific content areas. As more narratives are put into the knowledge database such models should perform more precisely and accurately. The end product would be an open-sourced knowledge repository that stores words and associated probabilities in order to code injury narratives, where researchers and other organizations may upload their injury narratives, select what rubric and algorithm to apply, and then run the model to obtain injury codes for their narrative data.

Providing better access to training data and cloud-based computer coding methods would enable researchers without previous access to computerized coding software and/or without a training set for the algorithm to code their data. This has global implications because health systems in the developing world have yet to move to computerized information systems and their only option may be narratives as trained coders are often scarce.

A shared knowledge database would enable injury researchers, organizations, and government health agencies to code and analyze large injury narrative datasets without the need for substantial resources as previously required, liberating these untapped data sources to be used for surveillance, policy, and implementing interventions. Ultimately, the future of injury surveillance must address who funds such a data warehouse and how it is financially sustained with appropriate technical assistance.

One of the challenges in building a knowledgebase of narratives and moving from privately used datasets to publically available datasets is the issue of confidentiality. Injury narratives may contain personally identifiable information (such as patient names) or company identifiable information (such as brands of products). To enable sharing of narratives more publically, language parsing techniques which can automatically de-identify details from narrative text (without losing the context of the narrative) will need to be incorporated into text mining methods, and there have already been significant advances in such techniques (See for example Deleger, 2013 et al (45)).

Human-directed learning

Nevertheless, algorithms do only what humans tell them. The human factors of manual review, quality assurance, and “knowing your data” will still be required especially to identify new or emerging hazards and to understand the complex interaction of contributory factors - a principle of surveillance. Text mining for injury surveillance stands apart from other data mining efforts such as that used by generic search engines. Generic search engines allow algorithms to find whatever they can, while human-directed injury surveillance through text mining is looking for *particular* outcomes – injuries, and particular features (for example, host, agent, vector environment), classifiable to specified categories defined by the end-user. The role of the human in teaching the algorithm how to behave is vital to getting it right.

It is difficult for an algorithm on its own to be able to assign classifications in all categories with the same level of confidence and very difficult to improve the accuracy of computer generated codes for the small categories or for identifying emerging hazards. Improvement beyond simply modeling of a training data set to use on a prediction dataset requires either sophisticated filtering or tailoring of the algorithm (with natural language processing) to identify small categories or other nuances of the coding protocol and the latter approach will still not allow for emerging risks to surface.

It was stated from the beginning (25) that manual coding should never be completely replaced and therefore a best practice approach should incorporate some manual coding, assigning a computer classification only for more repetitive events where the models are able to confidently predict the correct classification. This will be especially important for rare events and/or emerging hazards that appear only a very small number of times or not at all in a training dataset. For example, a new motor vehicle crash hazard (exploding magnesium steering column) would cause a human reviewer to query why steering columns explode on impact and if they represent a new material hazard to drivers and first responders. An algorithm would simply say this does not happen enough to be coded with certainty and would flag it for manual review. For large administrative datasets, incorporation of methods based on human-machine pairings such as presented in this paper utilizing readily available off the shelf machine learning techniques result in only a fraction of narratives that require manual review.

Conclusion

Machine learning of ‘big injury narrative data’ opens up many possibilities for expanded sources of data that can provide more comprehensive, ongoing and timely surveillance to inform injury prevention policy and practice in the future. This paper has demonstrated the significant value that injury narratives provide beyond structured coded datasets. It is critically important that, as an injury prevention community, we continue to advocate for the need for narratives to be included (or introduced) in routine data sources to capitalize on this potential as computing and technical capacity expands and not just rely on coded checkboxes. Secondly, the authors have argued for the need for a more systematic and incremental approach to developing machine learning approaches for the specialized purpose of injury surveillance, as distinct from other applications of machine learning more broadly. Modelling techniques (and research applications) vary in terms of levels of specificity and sensitivity, simplicity and complexity, and the building and refinement of these techniques require input from content experts and technical experts. The authors proposed future steps towards developing a ‘big injury narrative data’ platform to allow for the building, testing and refinement of machine learning algorithms. Finally, the need for human-machine pairings was reiterated to ensure machine learning approaches continue to reflect the underlying principles of injury surveillance.

The last 20 years has seen a dramatic change in the potential for technological advancements in injury surveillance and we have many examples of successful applications of such technology to injury narratives. It is now time to consolidate these learnings to build more sustainable, reliable and efficient approaches which will ensure the most robust use of the evidence-base for injury prevention.

Acknowledgments

Funding

Kirsten Vallmuur is supported by an Australian Research Council Future Fellowship under Grant FT120100202. Gordon Smith is supported by a grant from the U.S. National Institute on Alcohol Abuse and Alcoholism (R01AA18707).

References

1. Sorock GS, Smith GS, Reeve GR, Dement J, Stout N, Layne L, et al. Three perspectives on work-related injury surveillance systems. American journal of industrial medicine. 1997; 32(2):116–28. [PubMed: 9215434]
2. World Health Organisation. WHO Injury Surveillance Guidelines. Geneva: World Health Organisation; 2001.
3. World Health Organization (WHO). International Classification of External Causes of Injury (ICECI). Geneva: 2003.
4. Nordic Medico-Statistical Committee. NOMESCO Classification of External Causes of Injuries. Copenhagen: AN:sats; 2007. Fourth revised edition
5. United States Department of Labor Bureau of Labor Statistics. Occupational Injury and Illness Classification Manual, Version 2.01. USA: 2012.
6. Australian Safety and Compensation Council. Type of Occurrence Classification System (TOOCS) Third Edition Revision. Canberra, Australia: Australian Government; 2008.
7. McKenzie K, Fingerhut L, Walker S, Harrison A, Harrison J. Classifying External Causes of Injury: History, Current Approaches, and Future Directions. Epidemiologic Reviews. 2012; 34:4–16. [PubMed: 22045696]
8. Runyan C. Introduction: Back to the future - Revisiting Haddon’s conceptualization of injury epidemiology and prevention. Epidemiologic Reviews. 2003; 25:60–4. [PubMed: 12940231]
9. McKenzie K, Scott D, Campbell M, McClure R. The Use of Narrative Text for Injury Surveillance Research: A Systematic Review. Accident Analysis and Prevention. 2010; 42(2):354–63. [PubMed: 20159054]
10. Vallmuur K. Machine Learning Approaches to Analysing Textual Injury Surveillance Data: A Systematic Review. Accident Analysis and Prevention. 2015; 79:41–9. [PubMed: 25795924]
11. Stout, N. Occupational Injury. CRC Press; 1998. Analysis of narrative text fields in occupational injury data.
12. Bunn TL, Slavova S, Hall L. Narrative text analysis of Kentucky tractor fatality reports. Accident Analysis And Prevention. 2008; 40(2):419–25. [PubMed: 18329390]
13. Lipscomb HJ, Glazner J, Bondy J, Lezotte D, Guarini K. Analysis of text from injury reports improves understanding of construction falls. Journal Of Occupational And Environmental Medicine. 2004; 46(11):1166–73. [PubMed: 15534504]
14. Smith GS, Timmons RA, Lombardi DA, Mamidi DK, Matz S, Courtney TK, et al. Work-related ladder fall fractures: identification and diagnosis validation using narrative text. Accident Analysis And Prevention. 2006; 38(5):973–80. [PubMed: 16750154]
15. Chapman WW, Christensen LM, Wagner MM, Haug PJ, Ivanov O, Dowling JN, et al. Classifying free-text triage chief complaints into syndromic categories with natural language processing. Artificial Intelligence In Medicine. 2005; 33(1):31–40. [PubMed: 15617980]
16. Muscatello DJ, Churches T, Kaldor J, Zheng W, Chiu C, Correll P, et al. An automated, broad-based, near real-time public health surveillance system using presentations to hospital Emergency Departments in New South Wales, Australia. BMC Public Health. 2005; 5:141. [PubMed: 16372902]
17. Rainey D, Runyan C. Newspapers: A Source for Injury Surveillance? American Journal of Public Health. 1992; 82:746.
18. Archer P, Mallonee S, Schmidt A, Ikeda R. Oklahoma Firearm-Related Injury Surveillance. American Journal of Preventive Medicine. 1998; 15(3S):83–91. [PubMed: 9791627]
19. Bertke S, Meyers A, Wurzelbacher S, Bell J, Lampl M, Robins D. Development and evaluation of a Naïve Bayesian model for coding causation of workers’ compensation claims. Journal of safety research. 2012; 43(5):327–32. [PubMed: 23206504]

20. Taylor JA, Lacovara AV, Smith GS, Pandian R, Lehto M. Near-miss narratives from the fire service: A Bayesian analysis. Accident Analysis & Prevention. 2014; 62:119–29. [PubMed: 24144497]

21. Lehto M, Marucci-Wellman H, Corns H. Bayesian methods: a useful tool for classifying injury narratives into cause groups. Injury Prevention. 2009; 15(4):259–65. [PubMed: 19652000]

22. Ossiander E. Using Textual Cause-of-Death Data to Study Drug Poisoning Deaths. American Journal of Epidemiology. 2014; 179(7):884–94. [PubMed: 24521559]

23. Centers for Disease Control and Prevention. NIOSH Industry and Occupation Computerized Coding System (NIOCCS). 2015. Available from: http://wwwn.cdc.gov/niosh-nioccs/

24. Lehto M, Sorock G. Machine learning of motor vehicle accident categories from narrative data. Methods of Information in Medicine. 1996; 35:309–16. [PubMed: 9019094]

25. Marucci-Wellman H, Lehto MR, Sorock GS, Smith GS. Computerized coding of injury narrative data from the National Health Interview Survey. Accident; Analysis And Prevention. 2004; 36(2): 165–71.

26. Marucci-Wellman HR, Lehto MR, Corns HL. A Practical Tool for Public Health Surveillance: Semi-Automated Coding of Short Injury Narratives from Large Administrative Databases Using Naïve Bayes Algorithms. Accident Analysis and Prevention. 2015 Accepted for publication, June 29, 2015.

27. Marucci-Wellman H, Lehto M, Corns H. A combined Fuzzy and Na ve Bayesian strategy can be used to assign event codes to injury narratives. Injury Prevention. 2011; 17(6):407–14. [PubMed: 21482563]

28. Horan JM, Mallonee S. Injury Surveillance. Epidemiol Rev. 2003; 25(1):24–42. [PubMed: 12923988]

29. Marucci-Wellman HR, Courtney TK, Corns HL, Sorock GS, Webster BS, Wasiak R, et al. The direct cost burden of 13 years of disabling workplace injuries in the U.S. (1998–2010): Findings from the Liberty Mutual Workplace Safety Index. Journal of Safety Research. 2015 e-pub ahead of print.

30. Homan, C.; RJ; Liu, T.; ML; Silenzio, V.; COA, editors. Toward Macro-Insights for Suicide Prevention: Analyzing Fine-Grained Distress at Scale; Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality Proceedings of the Workshop; 2014; Baltimore, Maryland, USA.

31. Hume PA, Chalmers DJ, Wilson BD. Trampoline injury in New Zealand: emergency care. British journal of sports medicine. 1996; 30(4):327–30. [PubMed: 9015596]

32. Chen L, Vallmuur K, Nayak R. Injury Narrative Text Classification using the Factorization Model. BMC medical informatics and decision making. 2015; 15(Suppl 1):S5. [PubMed: 26043671]

33. Sorock GS, Ranney TA, Lehto MR. Motor vehicle crashes in roadway construction workzones: an analysis using narrative text from insurance claims. Accident; Analysis And Prevention. 1996; 28(1):131–8.

34. Bauer R, Sector M. Preventive product safety – monitoring accidental injuries related to consumer products in the European Union. Injury Control and Safety Promotion. 2003; 10(4):253–5. [PubMed: 14664371]

35. Pan S, Wang L, Wang K, Bi Z, Shan S, Xu B. A Knowledge Engineering Framework for Identifying Key Impact Factors-from Safety Related Accident Cases. Systems Research and Behavioral Science. 2014

36. Bondy J, Lipscomb H, Guarini K, Glazner JE. Methods for using narrative text from injury reports to identify factors contributing to construction injury. American journal of industrial medicine. 2005; 48(5):373–80. [PubMed: 16254951]

37. Zhao D, McCoy A, Kleiner B, Smith-Jackson T. Control measures of electrical hazards: An analysis of construction industry. Safety Science. 2015; 77:143–51.

38. Zhao D, McCoy A, Kleiner B, Du J, Smith-Jackson T. Decision-Making Chains in Electrical Safety for Construction Workers. Journal of Construction Engineering and Management. 201510.1061/(ASCE)CO.1943-7862.0001037

39. Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P, Witten I. The WEKA Data Mining Software: An Update. SIGKDD Explorations. 2009; 11(1)

40. Pedregosa F. Scikit-learn: Machine Learning in Python. The Journal of Machine Learning Research. 2011; 12:2825–30.

41. Noorinaeini, A.; Lehto, M. Mathematical Models of Human Text Classification. In: Duffy, V., editor. Handbook of Digital Human Modeling for Human Factors and Ergonomics. Mahwah, NJ: Lawrence Erlbaum Associates, Inc; 2009. p. 17.1-.5.

42. Noorinaeini A, Lehto M. Hybrid Singular Value Decomposition; a Model of Text Classification. International Journal of Human Factors Modeling and Simulation. 2006; 1(1):95–118.

43. Huang, H.; Lehto, M. Significance of low-frequency words in text classification of open-ended survey responses. 2nd Global Conference on Engineering and Technology Management; September 4–5, 2015; Chicago, IL, USA. 2015.

44. Abdat F, Leclercq S, Cuny X, Tissot C. Extracting recurrent scenarios from narrative texts using a Bayesian network: Application to serious occupational accidents with movement disturbance. Accident Analysis & Prevention. 2014; 70:155–66. [PubMed: 24769246]

45. Deleger L, Molnar K, Savova G, Xia F, Lingren T, Li Q, et al. Large-scale evaluation of automated clinical note de-identification and its impact on information extraction. Journal of the American Medical Informatics Association. 2013; 20:84–94. [PubMed: 22859645]

Key Messages

What is already known on this subject

- Large amounts of coded injury data and injury narratives are being collected universally daily and are available real time, yet the development and standardization of machine learning approaches using injury narratives is nascent.
- Injury narratives provide opportunities to a) identify the cases not able to be detected due to coding limitations, b) extract more specific information than codes allow, c) extract data fields which aren't part of the coding schema, d) establish chain-of-events scenarios, and e) assess coding accuracy.
- The main focus of machine learning techniques using injury narratives have been to quickly filter large numbers of narratives to accurately and consistently classify and track high magnitude, high risk and emerging causes of injury, to guide the development of interventions for prevention of future injury incidents.

What this study adds

- Reiteration of the significant value that injury narratives provide beyond structured coded datasets and evidence for the continued need to advocate for narratives to be included (or introduced) in routine data sources to capitalize on this potential as computing and technical capacity expands.
- Demonstration of a practical and feasible method for semi-automatic classification using human-machine learning of injury narratives which is accurate, efficient and meaningful and applicable to different injury domains.
- The opening of a dialogue within the injury surveillance community regarding future steps towards developing a 'big injury narrative data' knowledgebase to allow for the building, testing and refinement of machine learning algorithms.

Table 1

Examples of original and complex applications of narrative text over time

Original applications			More complex applications		
Article details	**Technique**	**Application**	**Article details**	**Technique**	**Application**
Archer et al, 1998 (18)	Newspaper clipping service used to manually identify cases of firearm-related injuries (unintentional and intentional) along with other sources obtained from hospital, police and vital statistics.	Newspaper clipping service identified almost one-third of firearm-related cases (but only 17% of suicides) and were a cheap, accessible and simple data source albeit incomplete, especially for suicide.	Homan et al, 2014 (30)	Extracted 200 tweets from 2.5 million tweets which noted suicide terms, used expert and novice ranks of tweets for distress levels, and used support vector machine approach to topic model data.	Automated tweet classification by distress levels to enable identification of individuals at risk of suicide through social media, with use of expert coders for training data and machine learning model choice important factors affecting performance of model.
Hume, Chalmers and Wilson, 1996 (31)	Free text search of emergency department data from one New Zealand hospital for one year for one product (trampoline).	Identified the number of trampoline-related incidents and allowed case identification to enable further review of text and manual coding of extra circumstance details.	Chen, Nayak and Vallmuur 2015 (32)	Automatic classification of mechanism and object categories for 15,000 emergency department cases across multiple hospitals using machine learning (matrix factorization approach).	Classified mechanism and objects quickly with accuracy of 0.93, showing potential for use to reduce need for manual coding for injury surveillance, though need for expert input into modelling required throughout process to improve algorithm performance.
Sorock, Ranney and Lehto, 1996 (33), and Lehto and Sorock 1996 (24)	Free text search of motor vehicle insurance claims database for 4 years to identify claims where road work occurring and key word categorization of pre-crash activities and crash types through word frequency count and manual grouping of similar words to prepare key word search strategy. Expanded to test a Bayesian modelling approach in second paper.	First paper identified number of incidents and categorized pre-crash activities and crash types to examine patterns of incidents. Second paper established Bayesian approach more accurately classified cases than keywords and pointed to the early potential for Bayesian approaches to be developed in this field.	Taylor et al (2014) (20)	Classified 2285 fire fighter-occupation specific narratives (longer narratives with average of 216 words), with near-misses & injury into injury mechanism and injury outcome using fuzzy and naïve Bayesian models with single word predictors.	Classified external causes with accuracy of 0.74 using fuzzy model and 0.678 using Naïve model, with increased training set size producing higher sensitivity. Showed that Bayesian methods can be used for coding long narratives for both injury incidents and near misses.
Bauer and Sector (2003) (34)	Development of a keyword based search to identify extent of product involvement in injury from emergency department based injury surveillance database, as well as use of expert panel to assess preventability and potential for product safety responses.	Ability to flag cases where high likelihood of consumer product involvement (defective, maladapted or intrinsically risky) and identify products most commonly associated with each category.	Pan et al, 2014 (35)	Use of named entity recognition techniques to automatically parse unstructured data from a range of databases (including RAPEX, CPSC and product safety databases in China and Japan). Used Bayesian network approach to identify and code safety factors pertaining to electric shock.	Automated extraction and coding of relevant cases incorporating a number of large publically available databases from different regions. Identification of the key safety factors involved in electric shock incidents (near miss and injuries), showing potential of multiple databases to extract common scenarios.
Bondy et al, 2005 (36)	Manual review of 4000 injury text reports from construction of Denver International Airport, and expert classification of case details according to Haddon's Matrix framework.	Classification of text reports according to Haddon's Matrix framework provided a more complete injury description than only coding certain injury elements, as well as providing richer data to understand injury scenario and target prevention activity.	Zhao et al 2015 (37) and Zhao et al 2015 (38)	Use of electrocution text reports in national occupational injury database to extract either key features according to hierarchy of control framework or Haddon's Matrix framework. Used narrative text analysis (such as word clusters, entity extraction, word tagging and	Automated extraction and tagging of key features of reports and grouping according to overarching injury prevention frameworks, to examine main prevention foci as well as illustrate decision making chains. Demonstrates the utility of text analysis to extract and elucidate more complex injury causation scenarios.

Original applications			**More complex applications**		
Article details	**Technique**	**Application**	**Article details**	**Technique**	**Application**
				“textual tag clouds” using NVivo qualitative software.	

Table 2

The Accuracy of the Human-Machine Classification System: Implementation of a Strategic Filter[a] Based on Agreement Between Two Naïve Bayes Algorithms

Adapted from Accident Analyses and Prevention. Marucci-Wellman, Lehto, Corns. A practical too for public health surveillance : Semi-automated coding of short injury narrative from large administrative databases using Naïve Bayes algorithms. 2015

BLS OIICS 2-Digit Event Code		Gold Standard[c]	Human-Machine System Coding of all Narratives[d]						%Agreement Between 2 Manual Coders[j]	Fleiss Kappa[k] manual coders
		(n)	n_{pred}[e]	$\%_{pred}$[f,g]	Sen[h]	95% CI	PPV[i]	95% CI		
1* Violence and other injuries by persons or animals										
11	Intentional injury by person	159	132	0.9	0.81	0.75, 0.87	0.98	0.95, 1.00	81%–97%	0.85
2* Transportation incidents										
24	Pedestrian vehicular incidents	120	117	0.8	0.78	0.71, 0.86	0.80	0.73, 0.88	57%–78%	0.65
26	Roadway incidents motorized land vehicle	650	672	4.5	0.98	0.97, 0.99	0.95	0.93, 0.97	93%–96%	0.94
27	Nonroadway incidents motorized land vehicle	136	122	0.8	0.80	0.73, 0.87	0.89	0.84, 0.95	52%–84%	0.62
4* Falls, slips, trips										
41	Slip or trip without fall	806	658	4.4	0.70	0.67, 0.73	0.86	0.83, 0.89	66%–89%	0.71
42	Falls on same level	2,148	2386	15.9	0.92	0.91, 0.93	0.83	0.81, 0.84	85%–93%	0.86
43	Falls to lower level	1,065	1176	7.8	0.89	0.87, 0.91	0.81	0.79, 0.83	78%–92%	0.81
5* Exposure to harmful substances or environments										
53	Exposure to temperature extremes	141	130	0.9	0.86	0.8, 0.92	0.93	0.89, 0.97	82%–98%	0.88
55	Exposure to other harmful substances	175	165	1.1	0.83	0.77, 0.88	0.88	0.83, 0.93	81%–96%	0.87
6* Contact with objects and equipment										
62	Struck by object or equipment	1,651	1749	11.7	0.90	0.89, 0.92	0.85	0.83, 0.87	82%–90%	0.82
63	Struck against object or equipment	466	397	2.6	0.74	0.7, 0.78	0.87	0.84, 0.91	66%–83%	0.68
64	Caught in or compressed by equipment	505	532	3.5	0.90	0.87, 0.93	0.86	0.83, 0.89	72%–83%	0.75
7* Overexertion and bodily reaction										
70	Overexertion and bodily reaction, uns	188	151	1.0	0.59	0.51, 0.66	0.73	0.66, 0.80	6%–48%	0.19
71	Overexertion involving outside sources	4,189	4334	28.9	0.95	0.95, 0.96	0.92	0.91, 0.93	87%–95%	0.87
72	Repetitive motions involving micro tasks	484	537	3.6	0.90	0.87, 0.92	0.81	0.77, 0.84	71%–83%	0.75
73	Other exertions or bodily reactions	916	827	5.5	0.79	0.76, 0.82	0.88	0.85, 0.90	56%–85%	0.64
X* All other classifiables (n<100) in training dataset										

BLS OIICS 2-Digit Event Code		Gold Standard[c]	Human-Machine System Coding of all Narratives[d]						%Agreement Between 2 Manual Coders[j]	Fleiss Kappa[k] manual coders
		(n)	n_{pred}[e]	$\%_{pred}$[f,g]	Sen[h]	95% CI	PPV[i]	95% CI		
xx	Other small (n<100 cases) classifiable categories[b]	632	467	3.1	0.68	0.64, 0.72	0.92	0.89, 0.94	-	-
Nonclassifiable										
9999	Nonclassifiable	569	448	3.0	0.70	0.66, 0.74	0.89	0.86, 0.92	69%–84%	0.72
Overall		15,000	15,000	100.0	0.87	0.87, 0.88	0.87	0.87, 0.88	77%–90%	0.78

[a] A filter is a technique to decide which narratives the computer should classify vs. which should be left for a human to read and classify.

[b] Two-digit categories with <100 cases.

[c] Gold Standard codes were assigned to each narrative by expert manual coders.

[d] Human-Machine system: The computer assigns codes to narratives that the algorithms agreed on the classification (68% of the dataset), and the remainder are manually coded (32 % of the dataset).

[e] n_{pred} = number predicted into category.

[f] $\%_{pred}$ = percent of cases in whole dataset predicted into category.

[g] The distribution of two-digit classifications will be skewed towards categories with high sensitivity, biasing the finally distribution of the coded datasets.

[h] Sen = Sensitivity: (true positives) the percentage of narratives that had been coded by the experts into each category that were also assigned correctly by the algorithm.

[i] PPV = Positive Predicted Value: the percentage of narratives correctly coded into a specific category out of all narratives placed into that category by the algorithm.

[j] Two-coder agreement, e.g. 6 total comparisons, coder 1 compared to 2,3,4, coder 2 compared to 3,4 coder 3 compared to 4.

[k] Fleiss Kappa between 0 and 1, > 0.6 considered good agreement, >.8 considered very good agreement.

$Naive_{SW}$ = Naïve Bayes Single Word Algorithm. $Naive_{seq}$ = Naïve Bayes Sequence Word Algorithm

Table 3

The Accuracy of the Human-Machine Classification System: Implementation of a Strategic Filter[a] Based on Agreement Between the Two Naïve Bayes Algorithms (Results for Small Categories Only, n< 100 Cases in Each Category)

Adapted from Accident Analyses and Prevention. Marucci-Wellman, Lehto, Corns. A practical too for public health surveillance : Semi-automated coding of short injury narrative from large administrative databases using Naïve Bayes algorithms. 2015

BLS OIICS 2-Digit Event Code		Gold Standard[b]	Human-Machine System Coding of All Narratives[c]					%Agreement Between 2 Manual Coders[g]	Fleiss Kappa[h] manual coders
		(n)	n_{pred}[d]	Sen[e]	(95% CI)	PPV[f]	95% CI		
1* Violence and other injuries by persons or animals									
12	Injury by person - intentional or intent unknown	96	78	0.66	0.56, 0.75	0.81	0.71, 0.88	47%–78%	0.57
13	Animal and insect related incidents	99	79	0.80	0.71, 0.87	1.00	1.00, 1.00	79%–94%	0.87
2* Transportation incidents									
20	Transportation incident, unspecified	3	3	1.00	1.00, 1.00	1.00	1.00, 1.00	0%–0%	0.00
21	Aircraft incidents	22	15	0.68	0.47, 0.89	1.00	1.00, 1.00	0%–75%	0.17
22	Rail vehicle incidents	6	4	0.67	0.12, 1.00	1.00	1.00, 1.00	0%–100%	0.67
23	Animal & other non-motorized vehicle transport incidents	14	13	0.86	0.65, 1.00	0.92	0.76, 1.00	0%–0%	0.00
25	Water vehicle incidents	11	5	0.45	0.1, 0.81	1.00	1.00, 1.00	0%–88%	0.25
3* Fires and explosion									
31	Fires	22	20	0.91	0.78, 1.00	1.00	1.00, 1.00	55%–88%	0.58
32	Explosions	21	18	0.86	0.69, 1.00	1.00	1.00, 1.00	44%–83%	0.46
4* Falls, slips, trips									
40	Fall, slip, trip, unspecified	4	2	0.50	0.00, 1.00	1.00	1.00, 1.00	0%–0%	0.00
44	Jumps to lower level	57	39	0.61	0.48, 0.74	0.90	0.80, 1.00	51%–90%	0.65
45	Fall or jump curtailed by personal fall arrest system	3	2	0.67	0.00, 1.00	1.00	1.00, 1.00	0%–0%	0.00
5* Exposure to harmful substances or environments									
50	Exposure to harmful substances or environ, unspecified	23	18	0.78	0.6, 0.96	1.00	1.00, 1.00	21%–88%	0.33
51	Exposure to electricity	27	18	0.67	0.48, 0.86	1.00	1.00, 1.00	65%–88%	0.81
52	Exposure to radiation and noise	38	36	0.87	0.76, 0.98	0.92	0.82, 1.00	54%–100%	0.80
54	Exposure to air and water pressure change	1	0	0.00	.	0.00	.	0%–100%	0.40
57	Exposure to traumatic or stressful even nec	32	23	0.72	0.55, 0.88	1.00	1.00, 1.00	73%–85%	0.80
59	Exposure to harmful substances or environments, nec	1	7	0.00	.	0.00	.	0%–100%	0.12

BLS OIICS 2-Digit Event Code		Gold Standard[b]	Human-Machine System Coding of All Narratives[c]					%Agreement Between 2 Manual Coders[g]	Fleiss Kappa[h] manual coders
		(n)	n_{pred}[d]	Sen[e]	(95% CI)	PPV[f]	95% CI		
6* Contact with objects and equipment									
60	Contact with objects and equipment, uns	78	43	0.54	0.43, 0.65	0.98	0.93, 1.00	12%–63%	0.25
61	Needle stick	1	1	1.00	1.00, 1.00	1.00	1.00, 1.00	-	-
65	Struck/caught/crush in collapsing structure, equip or material	5	3	0.60	0.00, 1.00	1.00	1.00, 1.00	0%–0%	0.33
66	Rubbed or abraded by friction or pressure	16	12	0.69	0.43, 0.94	0.92	0.73, 1.00	0%–50%	0.11
67	Rubbed abraded or jarred by vibration	7	4	0.57	0.08, 1.00	1.00	1.00, 1.00	0%–67%	0.14
69	Contact with objects and equipment, nec	1	1	1.00	1.00, 1.00	1.00	1.00, 1.00	-	-
7* Overexertion and bodily reaction									
74	Bodily conditions nec	20	10	0.50	0.26, 0.74	1.00	1.00, 1.00	0%–75%	0.33
78	Multiple types of overexertions and bodily reactions	23	13	0.39	0.18, 0.61	0.69	0.40, 0.98	0%–0%	0.00
79	Overexertion and bodily reaction and exertion, nec	1		0.00	.	0.00	.	-	-
Overall		437	467	0.68	0.64, 0.72	0.92	0.89, 0.94		

[a] A filter is a technique to decide which narratives the computer should classify vs. which should be left for a human to read and classify.

[b] Gold Standard codes were assigned to each narrative by expert manual coders

[c] Human-machine system consisted of human coding 32% of the dataset, machine coding 68% of the dataset.

[d] n_{pred} = number predicted into category.

[e] Sen = Sensitivity: (true positives) the percentage of narratives that had been coded by the experts into each category that were also assigned correctly by the algorithm.

[f] PPV = Positive Predicted Value: the percentage of narratives correctly coded into a specific category out of all narratives placed into that category by the algorithm.

[g] Two-coder agreement, e.g. 6 total comparisons, coder 1 compared to 2,3,4, coder 2 compared to 3,4 coder 3 compared to 4.

[h] Fleiss Kappa between 0 and 1, > 0.6 considered good agreement, >.8 considered very good agreement. $Naive_{SW}$ = Naïve Bayes Single Word Algorithm. $Naive_{seq}$ = Naïve Bayes Sequence Word Algorithm.

Table 4

The Accuracy of the Human-Machine Classification System: Implementation of a Strategic Filter[a] Based on Agreement of Predictions Between Selected Combinations of Different Algorithms (Naïve Bayes Single Word, Naïve Bayes Bi-gram, SVM, Logistic Regression)

	Two Model Agreement					Three Model Agreement	
Models	**SVM= Naïve Bayes Single Word**	**SVM = Naïve Bayes Bi-gram**	**SVM= Logistic**	**Logistic= Naïve Bayes Single Word**	**Logistic= Naïve Bayes Bi-gram**	**SVM = Naïve Bayes Single Word =Logistic**	**SVM=Naïve Bayes Single Word = Naïve Bayes Bi-gram**
Overall							
Sensitivity/PPV	87%	89%	81%	86%	88%	89%	93%
Manual Coded	28%	33%	14%	24%	29%	31%	43%

[a] A filter is a technique to decide which narratives the computer should classify vs. which should be left for a human to read and classify