A practical tool for public health surveillance: Semi-automated coding of short injury narratives from large administrative databases using Naïve Bayes algorithms

Helen R. Marucci-Wellman\textsuperscript{a,\emph{*}}, Mark R. Lehto\textsuperscript{b}, Helen L. Corns\textsuperscript{a}

\textsuperscript{a}Center for Injury Epidemiology, Liberty Mutual Research Institute for Safety, Hopkinton, MA, USA
\textsuperscript{b}School of Industrial Engineering, Purdue University, West Lafayette, IN, USA

\begin{abstract}
Public health surveillance programs in the U.S. are undergoing landmark changes with the availability of electronic health records and advancements in information technology. Injury narratives gathered from hospital records, workers compensation claims or national surveys can be very useful for identifying antecedents to injury or emerging risks. However, classifying narratives manually can become prohibitive for large datasets.

The purpose of this study was to develop a human–machine system that could be relatively easily tailored to routinely and accurately classify injury narratives from large administrative databases such as workers compensation. We used a semi-automated approach based on two Naïve Bayesian algorithms to classify 15,000 workers compensation narratives into two-digit Bureau of Labor Statistics (BLS) event (leading to injury) codes. Narratives were filtered out for manual review if the algorithms disagreed or made weak predictions.

This approach resulted in an overall accuracy of 87\%, with consistently high positive predictive values across all two-digit BLS event categories including the very small categories (e.g., exposure to noise, needle sticks). The Naïve Bayes algorithms were able to identify and accurately machine code most narratives leaving only 32\% (4853) for manual review. This strategy substantially reduces the need for resources compared with manual review alone.
\end{abstract}

© 2015 Published by Elsevier Ltd.

1. Introduction

Electronic health records containing real-time medical data create the potential for vast changes and improvements in public health research and surveillance. One important goal of injury surveillance is to determine important antecedents to injury and rank them according to magnitude, risk, severity or burden (Souza et al., 2011). To do this requires accurately classifying a comprehensive and representative sample of cases. Narratives in large administrative databases such as hospital records, workers compensation (WC) claims, or national surveys, can provide useful information about potential causes, prevention and recovery alone or as a supplement to existing coded data (Sorock et al., 1996, 1997; Stutts et al., 2001; Williamson et al., 2001; Lombardi et al., 2005; Verma et al., 2008; Lombardi et al., 2009; McKenzie et al., 2010; Vallmura, 2015; Taylor et al., 2014). However, classifying narratives into groups manually can become prohibitive for large datasets.

Advances in computer technology provide one potential solution. Automated methods of text processing developed within the field of medical informatics are currently used to perform highly sophisticated and accurate automated identification of specific events for applications such as medical monitoring or syndromic surveillance and coding of chief complaints (Wagner et al., 2004; Chapman et al., 2005; Brown et al., 2010; Gerbier et al., 2011). However, tailoring software for one particular coding strategy can consume tremendous resources (including the integration of specialized linguistics and semantics).

The purpose of this study was to develop a human–machine system that could be easily tailored to routinely and accurately classify injury narratives from large administrative databases such as workers compensation. Specifically, we sought to identify a representative set of classifications at similar levels of accuracy across all potential categories, those occurring the most often or those fairly infrequent but with potential high or emerging risk, or high severity. Recent research has shown that a simple...
multinomial Naive Bayes model can outperform state-of-art methods of text classification for short snippets of text and when there are few training cases (Wang and Manning, 2012). Therefore, given the lack of detail and specificity in short administrative narratives such as those collected on the first report of injury (snippets of text typically two to 15 words long), and our interest in identifying a large number of small categories (each with very few training cases), the Naive Bayes model seemed especially appropriate for the purposes of this study.

2. Methods

Thirty thousand records were randomly extracted from claims filed with a large WC insurance provider between January 1 and December 31, 2007. Four coders, trained on the Bureau of Labor Statistics (BLS) Occupational Injury and Illness Classification system (OIICS) 2012 version, classified records into two-digit event codes using the accident (what happened, 120 character maximum) and injury (type, e.g., strain, fracture, 20 character maximum) narratives as they appeared on the first report of injury. These manual codes served as our gold standard.

The dataset was divided into a training set of 15,000 cases, for model development, and a prediction dataset of 15,000 cases for evaluation. Each record included a unique identifier, a narrative describing how the injury occurred, and a two-digit BLS OIICS event code. The distribution of the two-digit OIICS event codes did not differ between datasets ($\chi^2 P = 0.87$).

2.1. Model development

Following the approach of Lehto et al. (2009), the Naive Bayes model can be used to assign a probability to each event code based on the words present in a particular narrative. The event code with the largest estimated probability is then chosen as the prediction for the words present. The assigned probability follows from Eq. (1) below:

$$P(C_i|n) = \prod_j P(n_j|C_i) \times P(C_i)$$

where $P(C_i|n)$ = the predicted probability of event code $i$ given the set of $j$ words in the particular narrative, $P(n_j|C_i)$ = the probability of observing word $n_j$ in a training narrative corresponding to event code $i$, $P(n_j)$ = the marginal probability of observing word $n_j$ in any training narrative, $P(C_i)$ = the marginal probability of event code $i$ in the training dataset.

To estimate $P(C_i|n)$, we followed the approach of Lehto et al. (2009) and only considered the words that were actually present in a narrative. Specifically, we employed the multinomial Naive Bayes model, as opposed to the Bernoulli Naive Bayes model. The multinomial model is not only simpler because it allows the missing words to be ignored, but also in our experience and that of others (McCallum and Nigam, 1998), it gives consistently better results for text classification. Another simplification follows from the fact that each estimate of $P(C_i|n)$ in Eq. (1) above contains the constant multiplier $P(n_j)$ which, therefore, can be factored out of (dropped from) the calculations. Consequently, the model can be fit by estimating only two parameters $P(C_i)$ and $P(n_j|C_i)$ from the training data. In this study, the first parameter is simply determined:

$$P(C_i) = \frac{N_i}{N}$$

where $N_i$ is the number of times category $C_i$ occurred in the training set and $N$ is the total number of training cases.

The second parameter, $P(n_j|C_i)$, was estimated as defined above but also by adding a smoothing parameter to the observed relative frequency of word $n_j$ in each category. The way this was done corresponds to assuming a uniform prior probability density function for the frequency of word $n_j$ in each category, given a prior sample size of $\alpha \times N$, where $N$ is the total number of training cases. From a practical perspective a value of alpha greater than 0 avoids inappropriately setting the $P(n_j|C_i)$ to zero when a word does not occur in the category of interest in the training dataset (often due to a noisy or odd misspelling of word). That is,

$$P(n_j|C_i) = \frac{\alpha + N_i}{\alpha + N}$$

where $N_i$ was the number of times category $C_i$ occurred in the training set, $N_j$ was the number of times word $n_j$ occurred in the training set, and $N_{ij}$ was the number of times of the training set that word $n_j$ occurred in category $C_i$. The smoothing constant $\alpha$ was set to a value of 0.01, corresponding to strong confidence that differences in the relative frequency of words between categories were in fact predictive of the category.

To improve precision and reduce potential rounding errors, instead of multiplying through $P(n_j|C_i)$ for each word in the narrative we calculate $P(C_i|n)$ using the log transform of Eq. (1). This gives the result of

$$\ln(P(C_i|n)) = \ln(P(C_i)) + \sum_j [\ln(P(n_j|C_i))]$$

The normalized estimates of each $P(C_i|n)$ can then be obtained as

$$P(C_i|n) = \frac{\exp(\ln(P(C_i|n)))}{\sum_j \exp(\ln(P(C_i|n)))}$$

Note: words never found in the training dataset are ignored (e.g., have no effect) when calculating $P(C_i|n)$.

Normalizing Eq. (1), using Eq. (5), results in an estimate of the true value of $P(C_i|n)$, which will be optimal if the words are conditionally independent. Our earlier studies have found this technique results in a well calibrated estimate of $P(C_i|n)$ strongly related to prediction accuracy (Lehto et al., 2009; Choe et al., 2013). This estimate contains no free-weighting parameters that are being adjusted to “fit” the model to the data. As such, this technique differs from methods such as neural networks, weighted Bayesian models, or logistic regression which have a large number of free parameters that are being adjusted to fit the data to minimize some objective function.

In summary, the algorithm is trained by simply calculating weights for each category (Eq. (2)) and the predictive relation of words to each category (Eq. (3)) using the training narratives. To generate the word counts used in Eqs. (2) and (3), first parse each training narrative to obtain a list of all the words and word sequences, and then count how often each word occurs in the training dataset for each category.

These probabilities (Eqs. (2) and (3)) are then used to assign classifications to the 15,000 prediction narratives. The probability of each category, given the set of words in a new prediction narrative $P(C_i|n)$, is calculated by adding up the log transformed weights (Eq. (4)) for each single word, or word sequence in the narrative, then normalizing (Eq. (5)). The category with the highest value of $P(C_i|n)$ is chosen as the prediction.

Sample calculations of $P(n_j|C_i)$, $P(C_i)$ and the category prediction strengths, $P(C_i|n)$ for the following two examples are provided in Appendix A, Tables A1 and A2. The first example, “Glove got caught in conveyor” evaluated using the single-word model (naive $\text{sw}$)
Table 1
The accuracy of two independent Naïve Bayes algorithms for classifying the events leading to injury of N = 15,000 workers compensation narratives (results for categories n \geq 100).

<table>
<thead>
<tr>
<th>BLS OIICS 2-digit event code</th>
<th>Gold standard(b)</th>
<th>Naïve Bayes single word model naïve(_{sw})</th>
<th>Naïve Bayes sequence word model ( naïve(_{seq}))</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(n) (%) (n_{pred}a) (%<em>{pred}a) Sen(c) 95% CI PPV(d) 95% CI (n</em>{pred}e) (%_{pred}e) Sen(f) 95% CI PPV(g) 95% CI</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1* Violence and other injuries by persons or animals</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11 Intentional injury by person</td>
<td>159 1.1 92 0.6 0.38 0.31, 0.46 0.66 0.56, 0.76 97 0.6 0.30 0.23, 0.37 0.49 0.39, 0.49</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2* Transportation incidents</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>24 Pedestrian vehicular incidents</td>
<td>120 0.8 116 0.8 0.51 0.42, 0.60 0.53 0.43, 0.62 143 1.0 0.45 0.36, 0.54 0.38 0.30, 0.46</td>
<td></td>
<td></td>
</tr>
<tr>
<td>26 Roadway incidents involv motorized land vehicle</td>
<td>650 4.3 669 4.5 0.90 0.88, 0.93 0.88 0.85, 0.90 670 4.5 0.86 0.83, 0.88 0.83 0.80, 0.86</td>
<td></td>
<td></td>
</tr>
<tr>
<td>27 Nonroadway incidents involv motorized land vehicles</td>
<td>136 0.9 177 1.2 0.42 0.34, 0.50 0.32 0.25, 0.39 93 0.6 0.32 0.24, 0.40 0.46 0.36, 0.57</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4* Falls, slips, trips</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>41 Slip or trip without fall</td>
<td>806 5.4 608 4.1 0.43 0.39, 0.46 0.56 0.52, 0.60 701 4.7 0.48 0.45, 0.52 0.56 0.52, 0.59</td>
<td></td>
<td></td>
</tr>
<tr>
<td>42 Falls on same level</td>
<td>2,148 14.3 2,554 17.0 0.77 0.75, 0.78 0.64 0.63, 0.66 2,524 16.8 0.78 0.76, 0.80 0.66 0.65, 0.68</td>
<td></td>
<td></td>
</tr>
<tr>
<td>43 Falls to lower level</td>
<td>1,065 7.1 1,338 8.9 0.70 0.67, 0.73 0.56 0.53, 0.58 1,383 9.2 0.74 0.72, 0.77 0.57 0.55, 0.60</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5* Exposure to harmful substances or environments</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>53 Exposure to temperature extremes</td>
<td>141 0.9 134 0.9 0.62 0.54, 0.71 0.66 0.58, 0.74 76 0.5 0.33 0.25, 0.40 0.61 0.49, 0.72</td>
<td></td>
<td></td>
</tr>
<tr>
<td>55 Exposure to other harmful substances</td>
<td>175 1.2 123 0.8 0.48 0.41, 0.55 0.68 0.60, 0.77 86 0.6 0.27 0.21, 0.34 0.56 0.45, 0.67</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6* Contact with objects and equipment</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>62 Struck by object or equipment</td>
<td>1,651 11.0 1,988 13.3 0.67 0.65, 0.70 0.56 0.54, 0.58 1,732 11.5 0.66 0.64, 0.69 0.63 0.61, 0.66</td>
<td></td>
<td></td>
</tr>
<tr>
<td>63 Struck against object or equipment</td>
<td>466 3.1 315 2.1 0.28 0.24, 0.32 0.42 0.36, 0.47 343 2.3 0.34 0.30, 0.38 0.46 0.41, 0.51</td>
<td></td>
<td></td>
</tr>
<tr>
<td>64 Caught in or compressed by equipment</td>
<td>505 3.4 531 3.5 0.62 0.58, 0.67 0.59 0.55, 0.64 547 3.6 0.66 0.62, 0.70 0.61 0.57, 0.65</td>
<td></td>
<td></td>
</tr>
<tr>
<td>7* Overexertion and bodily reaction</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>70 Overexertion and bodily reaction, uns</td>
<td>188 1.3 184 1.2 0.29 0.22, 0.35 0.29 0.23, 0.36 143 1.0 0.20 0.14, 0.25 0.26 0.19, 0.33</td>
<td></td>
<td></td>
</tr>
<tr>
<td>71 Overexertion involving outside sources</td>
<td>4,189 27.9 4,217 28.1 0.85 0.83, 0.86 0.84 0.83, 0.85 4,146 27.6 0.80 0.79, 0.82 0.81 0.80, 0.83</td>
<td></td>
<td></td>
</tr>
<tr>
<td>72 Repetitive motions involving micro tasks</td>
<td>484 3.2 524 3.5 0.70 0.66, 0.75 0.65 0.61, 0.69 526 3.5 0.64 0.60, 0.68 0.59 0.55, 0.63</td>
<td></td>
<td></td>
</tr>
<tr>
<td>73 Other exertions or bodily reactions</td>
<td>916 6.1 781 5.2 0.54 0.51, 0.57 0.64 0.60, 0.67 792 5.3 0.51 0.47, 0.54 0.59 0.55, 0.62</td>
<td></td>
<td></td>
</tr>
<tr>
<td>X* All other classifiables (n &lt; 100) in training dataset</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>xx Other small (n &lt; 100) classifiable categories(h)</td>
<td>632 4.2 326 2.2 0.26 0.23, 0.30 0.51 0.45, 0.56 340 2.3 0.14 0.11, 0.17 0.26 0.21, 0.31</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Nonclassifiable</td>
<td>9999 Nonclassifiable</td>
<td>569 3.8 323 2.2 0.41 0.37, 0.46 0.73 0.68, 0.78 658 4.4 0.44 0.40, 0.48 0.38 0.34, 0.42</td>
<td></td>
</tr>
<tr>
<td>Overall</td>
<td>15,000 100.0 15,000 100.0 0.67 0.66, 0.68 0.67 0.66, 0.68 15,000 100.0 0.65 0.64, 0.66 0.65 0.64, 0.66</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

\(a\) Two-digit categories with < 100 cases.

\(b\) Gold standard codes were assigned to each narrative by expert manual coders.

\(c\) \(n_{pred}\) = number predicted into category.

\(d\) \(\%_{pred}\) = percent of cases in whole dataset predicted into category.

\(e\) The distribution of two-digit classifications will be skewed towards categories with high sensitivity, biasing the final distribution of the coded datasets.

\(f\) 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.

\(g\) 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.

\(h\) Nonclassifiable

Table 2a
The accuracy of the human–machine classification system: implementation of a strategic filter based on agreement between two Naïve Bayes algorithms (results for categories \( n \geq 100 \)).

<table>
<thead>
<tr>
<th>BLS OIlCs 2-digit event code</th>
<th>The subset of narratives where the naivesw and naiveseq algorithms independently assigned the same classification(^a)</th>
<th>Human–machine system coding of all narratives(^i)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>( n_\text{pred}^d )</td>
<td>( %_\text{pred}^e )</td>
</tr>
<tr>
<td>1* Violence and other injuries by persons or animals</td>
<td></td>
<td></td>
</tr>
<tr>
<td>11 Intentional injury by person</td>
<td>34</td>
<td>0.3</td>
</tr>
<tr>
<td>2* Transportation incidents</td>
<td></td>
<td></td>
</tr>
<tr>
<td>24 Pedestrian vehicular incidents</td>
<td>64</td>
<td>0.6</td>
</tr>
<tr>
<td>26 Roadway incidents involving motorized land vehicle</td>
<td>564</td>
<td>5.6</td>
</tr>
<tr>
<td>27 Nonroadway incidents involving motorized land vehicle</td>
<td>41</td>
<td>0.4</td>
</tr>
<tr>
<td>4* Falls, slips, trips</td>
<td></td>
<td></td>
</tr>
<tr>
<td>41 Slip or trip without fall</td>
<td>338</td>
<td>3.3</td>
</tr>
<tr>
<td>42 Falls on same level</td>
<td>1849</td>
<td>18.2</td>
</tr>
<tr>
<td>43 Falls to lower level</td>
<td>869</td>
<td>8.6</td>
</tr>
<tr>
<td>5* Exposure to harmful substances or environments</td>
<td></td>
<td></td>
</tr>
<tr>
<td>53 Exposure to temperature extremes</td>
<td>50</td>
<td>0.5</td>
</tr>
<tr>
<td>55 Exposure to other harmful substances</td>
<td>63</td>
<td>0.6</td>
</tr>
<tr>
<td>6* Contact with objects and equipment</td>
<td></td>
<td></td>
</tr>
<tr>
<td>62 Struck by object or equipment</td>
<td>1122</td>
<td>11.1</td>
</tr>
<tr>
<td>63 Struck against object or equipment</td>
<td>124</td>
<td>1.2</td>
</tr>
<tr>
<td>64 Caught in or compressed by equipment</td>
<td>331</td>
<td>3.3</td>
</tr>
<tr>
<td>7* Overexertion and bodily reaction</td>
<td></td>
<td></td>
</tr>
<tr>
<td>70 Overexertion and bodily reaction, uns</td>
<td>68</td>
<td>0.7</td>
</tr>
<tr>
<td>71 Overexertion involving outside sources</td>
<td>3462</td>
<td>34.1</td>
</tr>
<tr>
<td>72 Repetitive motions involving micro tasks</td>
<td>378</td>
<td>3.7</td>
</tr>
<tr>
<td>73 Other exertions or bodily reactions</td>
<td>453</td>
<td>4.5</td>
</tr>
<tr>
<td>X* All other classifiable ((n&lt;100)) in training dataset</td>
<td></td>
<td></td>
</tr>
<tr>
<td>XX Other small ((n&lt;100)) cases classifiable categories(^b)</td>
<td>89</td>
<td>0.9</td>
</tr>
<tr>
<td>Nonclassifiable</td>
<td>9999</td>
<td></td>
</tr>
<tr>
<td>Overall</td>
<td>10,147</td>
<td>100.0</td>
</tr>
</tbody>
</table>

**Notes:**
- Naivesw = Naïve Bayes single word algorithm; Naiveseq = Naïve Bayes sequence word algorithm.
- 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.
- Two-digit categories with < 100 cases, detailed results for these categories are shown below.
- Distribution and accuracy of codes for just those narratives where the two models agreed (68% of complete dataset).
- %pred\(^d\) = number predicted into category.
- \( \%_\text{pred}^e \) = percent of cases in whole dataset predicted into category.
- Distribution of two-digit classifications will be skewed towards categories with high sensitivity, biasing the finally distribution of the coded datasets.
- 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.
- 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.
- 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).

\(^a\) 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, detailed results for these categories are shown below.
\(^c\) Distribution and accuracy of codes for just those narratives where the two models agreed, (68% of complete dataset).
\(^d\) \( n_\text{pred} \) = number predicted into category.
\(^e\) \( \%_\text{pred} \) = percent of cases in whole dataset predicted into category.
\(^f\) 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.
\(^h\) 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.
\(^i\) 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).
Table 2b
The accuracy of the human-machine classification system: implementation of a strategic filter based on agreement between the two Naïve Bayes algorithms (results for small categories only, n < 100 cases in each category).

<table>
<thead>
<tr>
<th>BLS OIICS 2-digit event code</th>
<th>Gold standard</th>
<th>The subset of narratives where the naive&lt;sub&gt;sw&lt;/sub&gt; and naive&lt;sub&gt;seq&lt;/sub&gt; algorithms independently assigned the same classification</th>
<th>Human-machine system coding of all narratives</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(n)</td>
<td>n&lt;sub&gt;pred&lt;/sub&gt;&lt;sup&gt;a&lt;/sup&gt;</td>
<td>Sen&lt;sup&gt;b&lt;/sup&gt;</td>
</tr>
<tr>
<td>1* Violence and other injuries by persons or animals</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>12 Injury by person—intentional or intent unknown</td>
<td>96</td>
<td>22</td>
<td>0.18</td>
</tr>
<tr>
<td>13 Animal and insect related incidents</td>
<td>99</td>
<td>19</td>
<td>0.49</td>
</tr>
<tr>
<td>2* Transportation incidents</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>20 Transportation incident, unspecified</td>
<td>3</td>
<td>0</td>
<td>0.00</td>
</tr>
<tr>
<td>21 Aircraft incidents</td>
<td>22</td>
<td>3</td>
<td>0.30</td>
</tr>
<tr>
<td>22 Rail vehicle incidents</td>
<td>6</td>
<td>0</td>
<td>0.00</td>
</tr>
<tr>
<td>23 Animal &amp; other non-motorized vehicle transport incidents</td>
<td>14</td>
<td>3</td>
<td>0.50</td>
</tr>
<tr>
<td>25 Water vehicle incidents</td>
<td>11</td>
<td>0</td>
<td>0.00</td>
</tr>
<tr>
<td>3* Fires and explosion</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>31 Fires</td>
<td>22</td>
<td>0</td>
<td>0.00</td>
</tr>
<tr>
<td>32 Explosions</td>
<td>21</td>
<td>1</td>
<td>0.25</td>
</tr>
<tr>
<td>4* Falls, slips, trips</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>40 Fall, slip, trip, unspecified</td>
<td>4</td>
<td>0</td>
<td>0.00</td>
</tr>
<tr>
<td>44 Jumps to lower level</td>
<td>57</td>
<td>8</td>
<td>0.15</td>
</tr>
<tr>
<td>45 Fall or jump curtailed by personal fall arrest system</td>
<td>3</td>
<td>0</td>
<td>0.00</td>
</tr>
<tr>
<td>5* Exposure to harmful substances or environments</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>50 Exposure to harmful substances or environ, unspecified</td>
<td>23</td>
<td>2</td>
<td>0.29</td>
</tr>
<tr>
<td>51 Exposure to electricity</td>
<td>27</td>
<td>1</td>
<td>0.10</td>
</tr>
<tr>
<td>52 Exposure to radiation and noise</td>
<td>38</td>
<td>14</td>
<td>0.69</td>
</tr>
<tr>
<td>54 Exposure to air and water pressure change</td>
<td>1</td>
<td>0.00</td>
<td>.</td>
</tr>
<tr>
<td>57 Exposure to traumatic or stressful event nec&lt;sup&gt;h&lt;/sup&gt;</td>
<td>32</td>
<td>1</td>
<td>0.10</td>
</tr>
<tr>
<td>59 Exposure to harmful substances or environments, nec&lt;sup&gt;h&lt;/sup&gt;</td>
<td>1</td>
<td>0</td>
<td>0.00</td>
</tr>
<tr>
<td>6* Contact with objects and equipment</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>60 Contact with objects and equipment, unspecified</td>
<td>78</td>
<td>1</td>
<td>0.00</td>
</tr>
<tr>
<td>61 Needle stick</td>
<td>1</td>
<td>0</td>
<td>0.00</td>
</tr>
<tr>
<td>65 Struck/caught/crush in collapsing structure, equip or material</td>
<td>5</td>
<td>0</td>
<td>0.00</td>
</tr>
<tr>
<td>66 Rubbed or abraded by friction or pressure</td>
<td>16</td>
<td>2</td>
<td>0.17</td>
</tr>
<tr>
<td>67 Rubbed abraded or jarred by vibration</td>
<td>7</td>
<td>0</td>
<td>0.00</td>
</tr>
<tr>
<td>69 Contact with objects and equipment, nec&lt;sup&gt;h&lt;/sup&gt;</td>
<td>1</td>
<td>0</td>
<td>0.00</td>
</tr>
<tr>
<td>7* Overexertion and bodily reaction</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>74 Bodily conditions nec&lt;sup&gt;h&lt;/sup&gt;</td>
<td>20</td>
<td>1</td>
<td>0.09</td>
</tr>
<tr>
<td>78 Multiple types of overexertions and bodily reactions</td>
<td>23</td>
<td>4</td>
<td>0.00</td>
</tr>
<tr>
<td>Overall</td>
<td>437</td>
<td>89</td>
<td>0.21</td>
</tr>
</tbody>
</table>

Naive<sub>sw</sub> = Naïve Bayes single word algorithm; Naive<sub>seq</sub> = Naïve Bayes sequence word algorithm.

<sup>a</sup>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.

<sup>b</sup>Gold Standard codes were assigned to each narrative by expert manual coders.

<sup>c</sup>Distribution and accuracy of codes for just those narratives where the two models agreed, (68% of complete dataset).

<sup>d</sup>n<sub>pred</sub>= number predicted into category.

<sup>e</sup>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.

<sup>f</sup>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.

<sup>g</sup>Human-machine system consisted of human coding 32% of the dataset, machine coding 68% of the dataset.

<sup>h</sup>nec = not elsewhere classifiable.
maximizes the probability of the *caught-in-or-compressed-by* category where \( P(C|n) = 0.999 \). The sequence-word model (naïve\textsubscript{seq}) predicts the same category, given the parsed sequences: “glove-got,” “got-caught,” “caught-in,” “in-conveyor,” “glove-got-caught,” “got-caught-in” with a probability of \( P(C|n) = 0.999 \). “Caught” and “caught-in” are both very strong predictors in the models (e.g., the word is much more commonly found in the *caught-in-or-compressed-by* category.)

In contrast, for the narrative “Box fell against ankle,” the single-word model predicts the *fall-on-same-level* category incorrectly, with a relatively low prediction strength of \( P(C|n) = 0.576 \). The sequence-word model uses the predictors “box-fell,” “fell-against,” “against-his,” “his-ankle” and “fell-against-his” to correctly predict the *struck-by-object* category with a much higher prediction strength \( P(C|n) = 1.0. \) The single-word model was incorrect because the word “fell” alone occurred much more often in the *fall-on-same-level* category than the *struck-by-object* category and incorrectly skewed the prediction strength for *fall-on-same-level* category above the *struck-by-object* category. When the word “fell” is only considered in sequence with other words, “fell-against” and “box-fell,” the prediction strength becomes much lower for the *fall-on-same-level* category and higher for the *struck-by-object* category.

The Textminer program developed by one of the authors (ML) was used to generate word counts and estimate their predictive weights for all 15,000 training narratives. This was done independently for single words and word sequences, resulting in two different sets of predictions. (Note: others have also implemented these steps in SAS (Bertke et al., 2012). Software for performing such analysis is available from many other sources, for example the WEKA platform includes numerous versions of Naïve Bayes and other classifiers. The sequence-word model uses the exact same procedure described above for the single-word model but considers each two, three or four words in sequence as an individual “keyword” as demonstrated in the examples above. Some (minimal) data cleaning was performed during this process to reduce noise in the calculations. This included removing a short list of frequently occurring “stop words” thought not to be meaningful from all narratives prior to word counting (e.g., the words A, AN, AND, ETC, HE, HER, HIM, HIS, I, LEFT, LT, MY, OF, RT, RIGHT, SHE, THE, R, L). We chose not to edit misspelled words or assign morphs to words with the same meaning although this may help if you have only a small number of training narratives (e.g., the words “CARS VEHS AUTO VEHICLE VEH VEHICLE AUTOMOBILE VEHICLES VECHILE VECHILES VEHICL” could be changed to be spelled the same, “automobile”). Assigning morphs can be very time consuming, and our objective was to test how well a model with minimal human editing of the words would work.

2.2. Evaluation of the human–machine coded approach

Our primary objective was to classify 15,000 new WC narratives as efficiently and accurately as possible. Our focus was not to find one particular subcategory or type of claim (i.e., data mining) but rather classify all narratives into categories for understanding the magnitude and ranking of root causes of injuries.

We know from prior results that relying on Bayesian algorithms alone will not result in consistently high positive predictive values (PPV) across all categories (Lehto et al., 2009; Marucci-Wellman et al., 2011). Also, Bayesian models can assign predictions for short narratives (text snippets such as those in WC claims narratives) with a degree of confidence (prediction strength) that is strongly related to the actual probability of being correct and which can be utilized as a filter for the human-machine approach. A filter is just a technique to decide which narratives the computer should classify vs. which should be left for a human to read and classify. An assortment of filters are possible and make this method versatile for many different applications; the selection of filter(s) should consider both accuracy and resource constraints for the individual classification task at hand. For the current study, we were interested in achieving high PPV across all categories, including the small ones. Therefore, we used the following filters: 1) if the predictions from the single-word model and sequence-word models agreed (naïve\textsubscript{sw} = naïve\textsubscript{seq}), this indicated high confidence in the prediction and the remaining disagree cases were manually reviewed and 2) if the predictive strength of the single-word model was very high, this also indicated high confidence in the prediction and the remaining cases below a probability threshold value were manually reviewed. All evaluation statistics were determined using SAS version 9.3 (SAS Institute, Cary, NC).

We evaluated both sensitivity and PPV for all 15,000 prediction narratives by comparing the machine classifications with gold standard codes. Sensitivity (true positives) is the percentage of narratives coded by the experts into each category (gold standard codes) that were also coded by the program; PPV is the percentage of narratives correctly coded into a specific category out of all narratives placed into that category.

Using a chi-square goodness of fit test, we evaluated whether the proportionate distributions by two-digit event category of the simulated resulting datasets were the same as the proportionate distribution in the gold standard dataset. We also evaluated the capability of each approach to identify small categories.

Since we also wanted to evaluate whether we could expect any improvement in accuracy compared to all manual coding, we present inter-reliability results for four manual coders who independently assigned classifications to a separate set of 4000 workers compensation narratives. Reliability metrics include the range in the percent agreement between each set of two coders from the four and the Kappa statistic (Viera and Garrett, 2005) for each two-digit category.

3. Results

The individual performance of the two independent models, naïve\textsubscript{sw} and naïve\textsubscript{seq}, are shown in Table 1. Similar to prior reports (Lehto et al., 2009), while the overall sensitivity was fairly good (0.67 naïve\textsubscript{sw}, 0.65 naïve\textsubscript{seq}), both algorithms independently predicted some categories much better than others, skewing the final distribution of the coded data \( (x^2 P < 0.0001) \), 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. For example, Table 1 shows a higher sensitivity for the *struck-against* category and higher PPV for the *struck-by* category using the sequence-word model. Still, many categories had low performance.

3.1. Addition of strategic filters for human–machine coded narratives

3.1.1. Filter 1: agreement in predictions between the single-word and sequence-word models

Using this filter, 68% of the narratives were coded by machine with an overall sensitivity and PPV of 81% and fairly high PPV across categories with more than 100 cases (Table 2a). However, using only these narratives \( (n = 10,147) \) would result in a biased distribution of the dataset with proportionally more overexertion-from-outside-sources injuries, 34% vs. 28%, and proportionally fewer struck-against injuries, 1% vs. 3% \( (x^2, P < 0.0001) \) than represented in the gold standard codes \( (n = 15,000) \). Additionally, most cases in the very small unique categories would never be found (Table 2b).

However, if the remaining 32% of records where the naïve\textsubscript{sw} model had a different classification than the naïve\textsubscript{seq} are identified,
pulled out and manually coded, the final human–machine dataset becomes distributed very similar to the original gold standard coded dataset (is representative), with consistently high sensitivity and PPV across all categories (Table 2a). Also, since many of the narratives included in the smaller categories ended up being pulled out for manual review, the performance of the small categories improved markedly (Table 2b). Note–worthy is the high sensitivity for these small categories, indicating that many were found because of the filter.

3.1.2. Filter 2: prediction strength

Filtering simply on prediction strength alone is another option. For example, filtering on the Naive\textsubscript{sw} algorithm where those narratives with a prediction strength lower than 0.90 are pulled out for manual review (40% of the narratives) would result in approximately 88% sensitivity of the final machine-human coded data. Fig. 1 demonstrates how overall PPV is affected by setting different threshold levels. The results across all categories however was not as good as the filter based on agreement between the two algorithms.

3.2. Tailoring the approach

Different “filtering” strategies can be followed depending on human resource and accuracy trade–offs: (1) filtering more narratives for manual review from the dataset where the two models agree will improve results (if we need higher sensitivity or PPV results across categories), or (2) filtering from the disagree dataset strategically adds additional machine classifications and reduces human coding requirements (if we cannot afford to manually classify 32%, Fig. 1). Using a combination of filters 1 and 2 we can: (1) filter more narratives for manual review from the dataset where the two models agree in order to improve results (i.e., need higher sensitivity), or (2) filter from the disagree dataset to strategically add additional computer classifications and further reduce human coding, Fig. 1.

Since our objective was to apply the most efficient, yet highly representative and accurate approach across all categories, we created an additional filter on the agree dataset (where naive\textsubscript{sw}=naive\textsubscript{seq}, n=10,147) using a naive\textsubscript{sw} prediction strength of 0.90. The overall sensitivity for the combined filters improved to

---

**Fig. 1.** Accuracy expected when implementing a human–machine approach for coding 15,000 workers compensation narratives: alternative methods based on various filters. Footnotes to figure: 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.

Filter 1: If the classification predicted by the single–word model and sequence–word models agree (naive\textsubscript{sw} = naive\textsubscript{seq}), this indicates high confidence in the classification. Therefore, for filter 1 we accept the computer classifications where the algorithms agreed, then manually review the remaining disagree cases.

Filter 2: If the predictive strength of an algorithm is very high, this indicates high confidence in the prediction. Therefore, for filter 2 we accept the computer classification when the prediction strength, \( P(C|n) \) is above a probability threshold value, then manually review the remaining cases.

Note: The numbers on the plots in parenthesis are the predictive strengths \( P(C|n) \) calculated by the algorithm and used to determine whether to accept or reject the computer code.

naive\textsubscript{sw} = Naive Bayes single–word algorithm.

naive\textsubscript{seq} = Naive Bayes sequence–word algorithm.
0.94 (Fig. 1), and the sensitivity of the struck-against category (which previously had the lowest performance in Table 2) improved to 0.87 (PPV 0.95, Appendix A, Table A3). 20% additional narratives were pulled out for human review (Fig. 1).

### 3.3. Inter-rater reliability

Using a separate set of narratives to test inter-rater reliability (n = 4,000), we found 77–90% agreement overall between two coders, 0.78 Kappa (Table 3). Some categories were assigned consistently among all coders, such as the roadway-vehicle category (93–96% range in inter-rater agreement between each set of two coders, 0.94 Kappa), yet some categories had very low agreement, such as aircraft (0–75% agreement, 0.17 Kappa, Table 3), water-vehicle-incidents (0–88% agreement, 0.25 Kappa) and non-roadway-incidents-involving-motorized-land-vehicle (52–84% agreement, 0.62 Kappa).
4. Discussion

In this study, we demonstrate that a strategic human-machine approach based on Naive Bayesian algorithms was able to comprehensively and accurately classify the events leading to injury. Accurately categorizing the cause, source and location of injuries from hospital, trauma center records, or WC claims is an essential part of the analytic process of epidemiology and surveillance and provides critical information for preventing future events such as amputations (Friedman et al., 2013), motor vehicle crashes (Pollack et al., 2013) or childhood injuries (Chan et al., 2001; McGeehan et al., 2006). For large administrative data sources, however, costs of personnel necessary to manually classify each narrative limits the utility. This study demonstrates the feasibility of a human–machine classification method to reduce such costs.

We demonstrate the importance of utilizing the complete sample for surveillance vs. a subset of cases. Utilizing only the cases where the two models agreed would result in a dataset that contained many accurate codes; however, that dataset would not, in fact, be representative of the population of cases we began with and would include very few of the cases in the small unique categories. In fact, we believe any approach solely using machine codes will have low sensitivity in finding the small categories because there can only be relatively very few training cases for these categories. A semantic model to find cases from the small unique categories may be possible but would take a very large amount of effort to develop. We further note that if only very few of a certain category are used in the final analyses, such estimates will be unreliable with large standard errors and, effectively, be unusable for surveillance.

A particularly striking benefit of the Naive Bayes algorithms for short narratives is the ability to filter out cases for manual review where there is low confidence in the algorithm assigned code. We believe that reserving human resources to focus on a smaller pool will improve the efficiency of the machine–human method and allow personnel to focus on the rarer narratives and those that the

Table A1
Sample calculations: “Glove got caught in conveyor”: calculation of \( P(C|n) \) using the naive\(_{\text{wr}}\) and naive\(_{\text{wr}}\) algorithms independently.

*Example - "GLOVE GOT CAUGHT IN CONVEYOR"

<table>
<thead>
<tr>
<th>Words()</th>
<th>Intentional Injuries by Persons</th>
<th>Intentional Motorized Veh.</th>
<th>Incident Injuries by Persons</th>
<th>Incident Motorized Veh.</th>
<th>Slp, Trip, No Fall</th>
<th>Fall Same Level</th>
<th>Fall Different Level</th>
<th>Exposure to Electricity</th>
<th>Struck By Object</th>
<th>Struck Against Object</th>
<th>Overexpression</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>GLOVE</strong></td>
<td>0.1%</td>
<td>0.1%</td>
<td>0.2%</td>
<td>4.3E-02</td>
<td>1.8E-02</td>
<td>0.2%</td>
<td>0.4%</td>
<td>0.1%</td>
<td>4.7%</td>
<td>3.8E-02</td>
<td></td>
</tr>
<tr>
<td><strong>GOT</strong></td>
<td>1.7%</td>
<td>0.8%</td>
<td>3.4%</td>
<td>2.9%</td>
<td>2.2%</td>
<td>2.2%</td>
<td>2.8%</td>
<td>3.7%</td>
<td>2.8%</td>
<td>19.5%</td>
<td></td>
</tr>
<tr>
<td><strong>CAUGHT</strong></td>
<td>2.5%</td>
<td>1.1%</td>
<td>6.0%</td>
<td>3.7%</td>
<td>4.8%</td>
<td>3.7%</td>
<td>5.3%</td>
<td>5.9%</td>
<td>43.4%</td>
<td>1.5%</td>
<td></td>
</tr>
<tr>
<td><strong>IN</strong></td>
<td>33.2%</td>
<td>40.4%</td>
<td>28.5%</td>
<td>28.5%</td>
<td>26.8%</td>
<td>22.2%</td>
<td>29.2%</td>
<td>20.3%</td>
<td>17.9%</td>
<td>34.3%</td>
<td></td>
</tr>
<tr>
<td><strong>CONVEYOR</strong></td>
<td>0.3%</td>
<td>0.1%</td>
<td>0.5%</td>
<td>0.3%</td>
<td>0.3%</td>
<td>0.5%</td>
<td>0.5%</td>
<td>0.3%</td>
<td>4.3%</td>
<td>0.7%</td>
<td></td>
</tr>
<tr>
<td>Category Overall Prevalence</td>
<td>1.0%</td>
<td>4.3%</td>
<td>0.1%</td>
<td>0.1%</td>
<td>5.3%</td>
<td>14.1%</td>
<td>0.2%</td>
<td>10.7%</td>
<td>3.0%</td>
<td>2.9%</td>
<td></td>
</tr>
<tr>
<td>Sum of Log transformed weights**</td>
<td>-25.9</td>
<td>-27.8</td>
<td>-25.2</td>
<td>-25.9</td>
<td>-24.5</td>
<td>-25.8</td>
<td>-21.2</td>
<td>-23.4</td>
<td>-13.3</td>
<td>-23.5</td>
<td></td>
</tr>
<tr>
<td>Normalized Probability ( P(C</td>
<td>n) )</td>
<td>3.3E-04</td>
<td>4.5E-04</td>
<td>6.5E-04</td>
<td>3.3E-04</td>
<td>1.4E-03</td>
<td>1.7E-03</td>
<td>3.6E-04</td>
<td>3.6E-02</td>
<td>4.1E-03</td>
<td>99.93%</td>
</tr>
</tbody>
</table>

**Category weights: \( p(C) = n_i / N \) the marginal probability of event code \( i \) (Category Prevalence, \( N \)) where \( N_i \) is the number of times category \( C \) occurred in the training set and \( N \) is the total number of training cases.

Finish the log transformed weights within each category to reach value below the double line. \( \ln (P(C|n) / n) = \ln (P(C|n)) + \sum (\ln(n_i / n_i)) \)

Exponentiate (reverse transformation) and normalize estimates (scaling probabilities to sum to 1 (100%))

\[ P(C|n) = \text{cosh}(\ln(P(C|n)) + \sum(\ln(n_i / n_i))) \]

**Highest normalized probability category selected as the classification

** This is done instead of taking the product of the weights to simply improve precision of the final estimates and reduce rounding errors.

Naive Bayes Sequences Word Calculations

<table>
<thead>
<tr>
<th>Words()</th>
<th>Intentional Injuries by Persons</th>
<th>Intentional Motorized Veh.</th>
<th>Incident Injuries by Persons</th>
<th>Incident Motorized Veh.</th>
<th>Slp, Trip, No Fall</th>
<th>Fall Same Level</th>
<th>Fall Different Level</th>
<th>Exposure to Electricity</th>
<th>Struck By Object</th>
<th>Struck Against Object</th>
<th>Overexpression</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>GLOVE-GOT</strong></td>
<td>3.0E-02</td>
<td>1.2E-02</td>
<td>0.1%</td>
<td>9.5E-03</td>
<td>4.0E-03</td>
<td>0.1%</td>
<td>5.1E-03</td>
<td>1.5E-02</td>
<td>1.6%</td>
<td>2.1E-03</td>
<td></td>
</tr>
<tr>
<td><strong>GOT-CAUGHT</strong></td>
<td>0.4%</td>
<td>0.2%</td>
<td>1.3%</td>
<td>0.8%</td>
<td>1.2%</td>
<td>0.7%</td>
<td>0.7%</td>
<td>1.0%</td>
<td>1.2%</td>
<td>9.5%</td>
<td></td>
</tr>
<tr>
<td><strong>CAUGHT-IN</strong></td>
<td>0.4%</td>
<td>0.1%</td>
<td>0.7%</td>
<td>0.7%</td>
<td>0.5%</td>
<td>0.7%</td>
<td>0.7%</td>
<td>0.4%</td>
<td>13.0%</td>
<td>0.3%</td>
<td></td>
</tr>
<tr>
<td><strong>IN-CONVEYOR</strong></td>
<td>1.3E-02</td>
<td>5.1E-03</td>
<td>2.3E-02</td>
<td>2.9E-02</td>
<td>4.2E-03</td>
<td>1.8E-03</td>
<td>2.3E-02</td>
<td>6.6E-03</td>
<td>0.7%</td>
<td>9.4E-04</td>
<td></td>
</tr>
<tr>
<td><strong>GLOVE-GOT-CAUGHT</strong></td>
<td>3.0E-02</td>
<td>1.2E-02</td>
<td>0.1%</td>
<td>9.5E-03</td>
<td>4.0E-03</td>
<td>0.1%</td>
<td>5.1E-03</td>
<td>1.5E-02</td>
<td>1.6%</td>
<td>2.1E-03</td>
<td></td>
</tr>
<tr>
<td><strong>GOT-CAUGHT-IN</strong></td>
<td>0.2%</td>
<td>0.1%</td>
<td>0.3%</td>
<td>0.3%</td>
<td>0.5%</td>
<td>0.2%</td>
<td>0.2%</td>
<td>0.2%</td>
<td>0.2%</td>
<td>4.0%</td>
<td></td>
</tr>
<tr>
<td><strong>CAUGHT-IN-CONVEYOR</strong></td>
<td>1.3E-02</td>
<td>5.1E-03</td>
<td>2.3E-02</td>
<td>2.9E-02</td>
<td>4.2E-03</td>
<td>1.8E-03</td>
<td>2.3E-02</td>
<td>6.6E-03</td>
<td>0.7%</td>
<td>9.4E-04</td>
<td></td>
</tr>
<tr>
<td>Category Overall Prevalence</td>
<td>1.0%</td>
<td>4.3%</td>
<td>0.1%</td>
<td>0.1%</td>
<td>5.3%</td>
<td>14.1%</td>
<td>0.2%</td>
<td>10.7%</td>
<td>3.0%</td>
<td>2.9%</td>
<td></td>
</tr>
<tr>
<td>Sum of Log transformed weights**</td>
<td>-56.1</td>
<td>-61.5</td>
<td>-53.6</td>
<td>-54.2</td>
<td>-56.6</td>
<td>-60.3</td>
<td>-54.2</td>
<td>-59.6</td>
<td>-56.4</td>
<td>-29.4</td>
<td></td>
</tr>
<tr>
<td>Normalized Probability ( P(C</td>
<td>n) )</td>
<td>2.5E-10</td>
<td>1.2E-12</td>
<td>3.1E-09</td>
<td>1.8E-09</td>
<td>1.6E-10</td>
<td>3.9E-12</td>
<td>1.8E-09</td>
<td>8.2E-12</td>
<td>1.8E-10</td>
<td>99.99%</td>
</tr>
</tbody>
</table>
algorithm classifies poorly. Overall sensitivity and PPV improved substantially using the integrated human–machine approaches.

In our example, applying a filter using agreement between the Naïve single-word and sequence-word models, resulted in an overall sensitivity of 87% with consistently high PPV across all two-digit BLS event categories. The combined human–machine approach was able to identify and code many narratives (68%) leaving only 32% (4800) for manual review, and was able to find many of the narratives (68%) in the very small categories. Further manual review of cases using the additional filter allowed for even higher accuracy (PPVs ranged from 0.88 to 1.0 across all categories).

The design and implementation of this human–machine coding system required the following resources: (1) training coders over a four-week period, (2) manually classifying 30,000 narratives for training and evaluation, (3) pre-processing narratives by removing some non-meaningful words using global re-assignment methods, (4) computing word probabilities for training the algorithm, (5) applying the algorithm to prediction narratives, and (6) identifying the optimal filtering potential of the data. This total investment included approximately 4 months’ effort and provided an algorithm capable of classifying the bulk of future WC narratives. The Naïve algorithms easily classified 48–68% of the WC narratives (freeing up human resources) and resulted in a higher accuracy of the ensuing combined human-machine assigned dataset, likely beyond what either a human or computer alone could accomplish.

Relying on human coders alone to manually classify injury narratives for surveillance has other disadvantages beyond consuming large amounts of resources: human coders require periodic re-training to code accurately and systematically (CDC, 2013), experienced coders cannot be replaced without investing in training of new coders, human coders can become bored with or just indifferent to repetitive and mundane narratives which may lead to coding inconsistencies over time. Conversely, an algorithm can code systematically and consistently for a limitless amount of repetitive, mundane narratives without experiencing fatigue.

Other computer algorithms have been developed resulting in similar or higher accuracy (e.g., natural language processing or optimal classifiers, such as those based on logistic regression); yet these require either advanced modeling or tailored programing to contain all the necessary words, phrases, and ontologies for a specific classification or type of narrative. Naïve Bayes algorithms are less complex and require fewer computations compared with Fuzzy Bayes (prior analyses, see (Lehto et al., 2009; Marucci-Wellman et al., 2011)) or other more sophisticated models (e.g., support vector machine (SVM), logistic regression). The training of a probability-based Naïve Bayes algorithm relies simply on pre-coded narratives and results in intuitive evidence supporting each prediction.

Optimization using methods such as Regularized Logistic Regression or SVM will almost always moderately improve overall results (prior to filtering) over the simple multinomial Naïve Bayes model, but studies have shown a tendency for the Naïve Bayes model to outperform optimized models when there are few training cases (Ng and Jordan, 2002) such as for small categories or for shorter snippets of text (Wang and Manning, 2012). Interestingly, the performance of multinomial Naïve Bayes can also be improved by simple data transforms that retain the large
advantage Naïve Bayes has over optimization in terms of computation requirements (Rennie et al., 2003).

Previously, we demonstrated similar potential for high accuracy using agreement between Fuzzy and Naïve Bayesian models (Marucci-Wellman et al., 2011). However, Fuzzy Bayes is computationally cumbersome and better suited for long narratives (Taylor et al., 2014). Use of the two Naïve Bayesian models shown here required much less model development time and had similar results. The present strategy is more feasible for resource-constrained public health organizations when utilizing short narratives from large administrative databases for many different applications.

4.1. Limitations

There are several limitations to this approach. Currently, we have evaluated the accuracy using only one classification scheme and with fairly short WC narratives. More research will be required to understand how the algorithm performs with more-detailed classifications, on other types of classification protocols or with different types of narratives. Sometimes additional variables may be required (e.g., gender, age) for accurate coding—it is very important to provide the algorithm with the same variables used by humans to do classifications. We note, however, that the Naïve algorithm accommodates these other variable as predictors (or additional keywords) if they do not add redundancy with words already in the narrative (e.g., the nature of injury included as a “fall”or “burn” improves the machine classification).

5. Conclusion

These results suggest that large amounts of resources could be saved by public health organizations which routinely classify short narratives in large administrative databases for surveillance purposes by strategically deploying computer coding for the majority of cases. Noteworthy, is the ability of the Naïve algorithm to pull out narratives from the small and novel categories for manual assignment, improving results substantially.

More sophisticated classifiers than Naïve Bayes exist. However, Naïve Bayes algorithms have good performance for classifying short, noisy snippets of text and are a simple, practical, easily implemented approach. This study shows that filtering with two Naïve Bayes models to selectively guide manual review successfully generated an unbiased estimate of the frequency of injury cause/events. In future studies, we will look at whether SVM or other models can also be used to selectively filter out the small categories for review. Simply put, we have observed that this simple approach is effective, but further improvement in overall performance may be attainable with other classifiers.

Table A3

Accuracy of the human-machine classification system: implementation of two strategic filters* (1) where the two Naïve Models agree on the classification and (2) where the classification strength is above 0.90 for the naïve-naïve algorithm.

<table>
<thead>
<tr>
<th>BLS ORICS 2-Digit Event Code</th>
<th>The Subset of Narratives Where the NaïveBayes and NaïveSVM Algorithms Independently Assigned the Same Classification AND the Classification Strength Was Above 0.90 for NaïveBayes Algorithm1</th>
<th>Human-Machine System Coding of all Narratives2</th>
</tr>
</thead>
<tbody>
<tr>
<td>1st Violence and other injuries by persons or animals</td>
<td>159.9 1.1 24.0 0.3 0.53 0.38 0.68 1.00 1.00 1.10</td>
<td>138.9 0.9 0.87 0.81 0.92 1.00 1.00 1.00</td>
</tr>
<tr>
<td>2nd Transportation incidents</td>
<td>120.8 0.8 49.0 0.5 0.64 0.48 0.80 0.63 0.47 0.78</td>
<td>121.8 0.8 0.88 0.83 0.94 0.89 0.82 0.94</td>
</tr>
<tr>
<td>26 Roadway incidents involving motorized land vehicle</td>
<td>650.4 4.3 329.7 7.2 0.99 0.98 1.00 0.97 0.95 0.98</td>
<td>662.4 4.4 0.99 0.98 1.00 0.97 0.96 0.99</td>
</tr>
<tr>
<td>27 Normandy incident involvd motorized land vehicle</td>
<td>136.0 0.9 35.0 0.5 0.59 0.43 0.74 0.69 0.52 0.85</td>
<td>130.9 0.9 0.88 0.82 0.93 0.92 0.87 0.96</td>
</tr>
<tr>
<td>4th Falls, Slips, Tries</td>
<td>806.5 5.4 154.2 2.1 0.60 0.53 0.66 0.82 0.76 0.89</td>
<td>747.0 0.9 0.87 0.87 0.91 0.96 0.95 0.98</td>
</tr>
<tr>
<td>42 Falls on level</td>
<td>2,148.14 1283.17 15.7 0.94 0.92 0.95 0.87 0.85 0.99</td>
<td>2241.14 0.9 0.97 0.96 0.97 0.92 0.91 0.94</td>
</tr>
<tr>
<td>43 Falls to lower level</td>
<td>1,005.7 11.1 537.7 7.3 0.89 0.86 0.92 0.82 0.79 0.85</td>
<td>1105.7 1.4 0.95 0.93 0.96 0.91 0.96 0.93</td>
</tr>
<tr>
<td>5th Exposure to harmful substances or environments</td>
<td>141.9 0.9 48.0 0.7 0.77 0.66 0.89 0.85 0.75 0.96</td>
<td>136.9 0.9 0.91 0.87 0.96 0.93 0.91 0.99</td>
</tr>
<tr>
<td>55 Exposure to other harmful substances</td>
<td>175.1 0.6 25.6 0.8 0.70 0.57 0.82 0.70 0.57 0.82</td>
<td>175.1 0.2 0.90 0.86 0.95 0.90 0.86 0.95</td>
</tr>
<tr>
<td>6th Contact with objects and equipment</td>
<td>1,651.11 0.0 727.9 0.9 0.87 0.84 0.89 0.81 0.79 0.84</td>
<td>1695.1.1 0.3 0.96 0.93 0.96 0.92 0.91 0.93</td>
</tr>
<tr>
<td>62 Struck by object or equipment</td>
<td>466.3 1.0 60.8 0.8 0.38 0.28 0.48 0.62 0.49 0.74</td>
<td>428.2 0.9 0.87 0.84 0.90 0.95 0.92 0.97</td>
</tr>
<tr>
<td>63 Struck against object or equipment</td>
<td>505.3 0.4 252.3 0.4 0.90 0.86 0.94 0.81 0.77 0.86</td>
<td>526.9 0.3 0.95 0.93 0.97 0.91 0.89 0.94</td>
</tr>
<tr>
<td>7th Overexertion and Bodily Reaction</td>
<td>188.1 0.3 37.0 0.5 0.30 0.17 0.42 0.43 0.26 0.60</td>
<td>171.1 0.4 0.74 0.74 0.86 0.88 0.83 0.93</td>
</tr>
<tr>
<td>70 Overexertion and bodily reaction, uns</td>
<td>4,189.27 3,183.83 0.97 0.96 0.97 0.94 0.95 0.95</td>
<td>4265.28 0.8 0.97 0.98 0.97 0.96 0.96 0.97</td>
</tr>
<tr>
<td>72 Repetitive motions involving microtasks</td>
<td>804.32 0.7 272.17 3.7 0.88 0.84 0.92 0.82 0.77 0.86</td>
<td>805.3 0.4 0.94 0.92 0.96 0.90 0.87 0.93</td>
</tr>
<tr>
<td>73 Other exertions or bodily reactions</td>
<td>916.6 0.1 275.7 0.7 0.75 0.70 0.80 0.83 0.78 0.87</td>
<td>887.9 0.2 0.92 0.90 0.93 0.95 0.93 0.96</td>
</tr>
<tr>
<td>10 All other classifiable (n=100 in training dataset)</td>
<td>632.4 0.2 64.0 0.4 0.30 0.23 0.38 0.66 0.53 0.76</td>
<td>554.3 0.3 0.84 0.81 0.87 0.96 0.94 0.97</td>
</tr>
<tr>
<td>85 Other small n=100 cases classifiable categories</td>
<td>9999 Nonclassifiable</td>
<td>569.3 0.8 110.1 0.5 0.58 0.51 0.65 0.82 0.76 0.89</td>
</tr>
<tr>
<td>Overall</td>
<td>15000.1000 7343.1000 0.88 0.87 0.89 0.88 0.87 0.89</td>
<td>15000.1500 0.94 0.94 0.94 0.94 0.94 0.94</td>
</tr>
</tbody>
</table>

* 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

1Two-digit classifications with 100+ cases

2Distribution of accuracy for just those narratives where the two models agreed and have a naive-naive strength value above 0.9, 7343 (49%) of complete dataset.

3n=numerical predictor into category

4n=percentage of cases in the full dataset protected into category

5The distribution of two-digit classifications will be skewed towards categories with high sensitivity, biasing the final distribution of the coded datasets

6SVM = 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

7SVM = Positive Predictive Value; the percentage of narratives correctly coded into a specific category out of all narratives placed into that category by the algorithm

8Human-machine system consisted of human coding 51% of the dataset, machine coding 49% of the dataset

NaïveBayes = Naïve Bayes Single Word Algorithm

NaïveSVM = Naïve Bayes Sequence Word Algorithm
Acknowledgements

The authors would like to express their gratitude to Dr. David Swedler, Dr. Santosh Verma and Ted Courtney for reviewing the manuscript and Ms Margaret Rothwell for editing on the final manuscript.

Appendix A.

See Tables A1–A3.

References


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. Inj. Prev. 17 (6), 407–414.


