

Bayesian methods: a useful tool for classifying injury narratives into cause groups

M Lehto,¹ H Marucci-Wellman,² H Corns²

¹ School of Industrial Engineering, Purdue University, West Lafayette, Indiana, USA; ² Liberty Mutual Research Institute for Safety, Hopkinton, Massachusetts, USA

Correspondence to: Professor M Lehto, School of Industrial Engineering, Purdue University, 1287 Grissom Hall, West Lafayette, IN 47907, USA; lehto@purdue.edu

Accepted 17 April 2009

ABSTRACT

To compare two Bayesian methods (Fuzzy and Naïve) for classifying injury narratives in large administrative databases into event cause groups, a dataset of 14 000 narratives was randomly extracted from claims filed with a worker's compensation insurance provider. Two expert coders assigned one-digit and two-digit Bureau of Labor Statistics (BLS) Occupational Injury and Illness Classification event codes to each narrative. The narratives were separated into a training set of 11 000 cases and a prediction set of 3000 cases. The training set was used to develop two Bayesian classifiers that assigned BLS codes to narratives. Each model was then evaluated for the prediction set. Both models performed well and tended to predict one-digit BLS codes more accurately than two-digit codes. The overall sensitivity of the Fuzzy method was, respectively, 78% and 64% for one-digit and two-digit codes, specificity was 93% and 95%, and positive predictive value (PPV) was 78% and 65%. The Naïve method showed similar accuracy: a sensitivity of 80% and 70%, specificity of 96% and 97%, and PPV of 80% and 70%. For large administrative databases, Bayesian methods show significant promise as a means of classifying injury narratives into cause groups. Overall, Naïve Bayes provided slightly more accurate predictions than Fuzzy Bayes.

In the USA, most hospitals and trauma centres classify injury causes into International Classification of Disease (ICD) E-code categories. The same coding scheme is used in large health surveys such as the National Health Interview Survey (NHIS) conducted by the National Center for Health Statistics (NCHS). Public health authorities, government agencies and researchers use the coded data to identify patterns and trends in external causes of injury.¹⁻⁴ The Bureau of Labor Statistics (BLS) uses a similar approach to identify events leading to workplace injuries as defined by the occupational injury and illness classification system, OIICS.⁵

For the most part, the injury causes are assigned manually in these systems by humans based on a narrative description. This process is resource intensive. The accuracy and completeness of manual coding is another important issue. One recent study evaluating the accuracy of E-codes assigned in emergency department data found an accuracy of 65% for work-related injuries and 57% for non-work-related injuries.⁶ Other studies reviewed in a recent report prepared by the Centers for Disease Control (CDC) showed levels of inaccurate E-coding from 13% to 18%.⁷ Recently, the CDC has been actively promoting strategies for improving the quality and completeness of exter-

nal cause of injury coding in the USA, including the use of automated systems that assist coders in assigning event codes.⁷

The simplest automated systems assign event codes to cases on the basis of presence or absence of keywords in injury narratives. For example, a case might be assigned the code "contact with electricity" if it contains the word "electricity" or "volt." Previous work has shown that keyword-based approaches can classify injury narratives with a high level of specificity,¹ but low sensitivity for many categories. Machine learning algorithms offer an alternative approach, in which the system learns how to classify injury text without human intervention from the typically massive amounts of previously coded narratives in administrative databases.^{3,8} The potential of this approach was demonstrated in a study in which a multiple-word Fuzzy Bayes model was used to assign event codes to narrative descriptions of injury incidents from the NHIS.³ The results had high sensitivity (83%) and also suggested the potential to filter out targeted records for manual review. However, owing to dataset constraints (ie, only ~5000 total coded narratives), the same dataset was used to train the algorithm and make predictions, which may have resulted in an optimistic bias.³

The primary objective of this study was to demonstrate that Bayesian methods of computer classification can accurately classify injury narratives from large administrative databases into cause groups, and that this can be done (1) with little or no human intervention, (2) for new injury narratives not used to train (or develop) the classification algorithm, and (3) in a way that estimates how likely each prediction is to be correct. Attaining this objective would be an important first step, in implementation of automatic coding methods. Two different Bayesian models (Fuzzy and Naïve) found to be promising in our previous studies were compared from this perspective, as expanded upon below.

METHODS

Training and prediction datasets

Over 17 000 records were randomly extracted from claims filed between January 2002 and December 2004 with a workers' compensation insurance provider. Each record included a unique identifier, a narrative describing how the injury occurred, and a two-digit BLS OIICS event code manually assigned and agreed upon by two expert coders. The OIICS scheme includes ~40 mutually exclusive injury and event categories. The two coders who classified these narratives went through an intensive training process and practised with each

Special feature

other. At the end of the training session, the two expert coders were tested on ~2000 narratives. Overall, they agreed on classifications at the one-digit level 87% of the time, and at the two-digit level 75% of the time, but the level of agreement varied by category.

The observed lack of agreement between manual raters was expected, given that the claims narratives were fairly short (usually between nine and 20 words long, with a mean length of 14 words), which obviously limited the amount of information provided. An example narrative, in this case for a transportation accident (BLS OIICS two-digit code of 41), is given below.

“HUSB. & SON WERE REARENDED AT RED TRAFFIC LIGHT
BY DRUNK DRIVER DRIV- ING AT LEAST 45 MPH INFULL
SIZE PICK-UP TRUCK//N”

As illustrated by the example, narratives tended to be noisy, with many misspellings, run-on words, abbreviations, and inconsistent or missing punctuation. After elimination of cases on which the raters disagreed, the data were then divided into a training set of 11 000 cases used for model development and a prediction dataset of 3000 cases for model evaluation.

Model development

Two different Bayesian models, referred to as Naïve Bayes and Fuzzy Bayes, were developed. (Both Bayes models were developed and evaluated using the Textminer program developed by one of the authors (ML). The Textminer program is implemented in Visual Basic and provides an interface for analysing any dataset format readable by Microsoft Access. Additional information on the program can be obtained by contacting ML.) Both models used the statistical relationship between terms in the 11 000 injury narratives in the training set and the manually assigned one-digit and two-digit BLS OIICS codes to estimate the probability that a human coder would assign a particular code to a new narrative, given the words that were present in the narrative. The training set was a random sample of cases from the 17 000 manually coded records. A large training set was used, because, as in any implementation of statistical models, a larger sample size reduces the effect of noise. (In field applications, the entire set of previously coded narratives would normally be used as the training set. Then, if narratives known to be coded correctly became available, they could be added to the training set to improve accuracy when convenient. Note that this study used a completely separate prediction set to avoid potential bias due to overfitting the data and to more accurately model the situation where predictions are made for new unclassified cases.)

After the cases had been divided into a training set and prediction set, the next step in model development was to extract the words used in each narrative. This resulted in a list showing which words were present for each narrative. The extracted words were then cleaned up, by removing punctuation marks and non-alphabetical characters. A small set of “stop” words or words that have no predictive value but create burdensome processing time (eg, “and”, “a”, “the”) were also deleted. Words occurring fewer than three times in the entire set of narratives were also dropped. As one objective of the study was to minimise the degree of human input required during model development, no additional steps were performed during this process. After identification of the words present in the narratives, the next step was to tabulate their frequency of occurrence for each of the assigned categories in the training set,

which corresponded to training the models. The two Bayesian models were then implemented, as expanded upon below.

Naïve Bayes model

The Naïve Bayes model is a commonly applied method of text classification which has been used for years in the field of information retrieval.⁹ To see how the model works, let us assume that a given narrative consists of a vector of j words, $n = \{n_1, n_2, \dots, n_j\}$. Also, assume that i possible event codes can be assigned resulting in a second vector $E = \{E_1, E_2, \dots, E_i\}$. By making what is called the conditional independence assumption,¹⁰ the probability of assigning a particular event code category can then be calculated using the expression:

$$P(E_i|n) = \prod_j \frac{P(n_j|E_i) P(E_i)}{P(n_j)}$$

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

In application, $P(n_j|E_i)$, $P(E_i)$ and $P(n_j)$ are all normally estimated on the basis of their frequency in a training set. Also, $P(n_j|E_i)$ is normally smoothed to reduce the effects of noise. The approach we implemented was to add a small constant to the number of times a particular word occurred in a category, as shown below:

$$P(n_j|E_i) = \frac{\text{count}(n_j|E_i) + \alpha \times \text{count}(n_j)}{\text{count}(E_i) + \alpha \times N}$$

where $\text{count}(n_j|E_i)$ is the number of times word n_j occurs in category E_i , $\text{count}(n_j)$ is the number of times word n_j occurs, $\text{count}(E_i)$ is the number of times category E_i occurs, and α is a smoothing constant. Larger values of α reduce the weight given to the evidence provided by each term. We chose to use a value of $\alpha = 0.05$, which corresponds to a small level of smoothing.

The conditional independence assumption is perhaps the most controversial aspect of the Naïve Bayes model. Informally, for the purposes of text classification, when this assumption holds, the probability of each index term (eg, word or word sequence) being present depends on only the event code considered and is independent of the remaining terms in the narrative. The conditional independence assumption is almost always violated in practice. However, a long history of application shows that Naïve Bayes tends to work remarkably well even when this assumption is violated.¹⁰

Fuzzy Bayes model

The Fuzzy Bayes approach avoids making the conditional independence assumption by calculating $P(A_i|n)$ using the expression:

$$P(E_i|n) = \text{MAX}_j \frac{P(n_j|E_i) P(E_i)}{P(n_j)}$$

where each term i is assigned a value as explained in equation 2 above. The primary difference from Naïve Bayes is that instead of multiplying the conditional probabilities, Fuzzy Bayes estimates $P(E_i|n)$ using the “index term” most strongly predictive of the category. In practice, n_j in the above expression can be a combination or sequence of words, which allows Fuzzy

Bayes to consider multiple pieces of evidence when calculating $P(E_i|n)$. Many word combinations and word sequences (such as “fell-off”) are accurate and intuitive predictors of event codes.^{3 11 12} Consequently, we included combinations of up to four words and sequences of up to three words as predictors in this study.

Model evaluation

Two independent trials were conducted, in which predictions were made for the 3000 narratives in the prediction set using the Naïve and Fuzzy Bayes algorithms, respectively. For each approach, the category calculated by the algorithm as having the highest prediction strength,

$$P(E_i|n)$$

for the terms in the narrative, was chosen as the “predicted code.” The obtained results were then evaluated by comparing the predictions with the manually assigned “gold standard” codes, for both one-digit and two-digit classifications. The evaluation included calculations of sensitivity, specificity, positive predictive value (PPV) and negative predictive value (NPV) for each method.

In addition to these four evaluation metrics, we tested how well calibrated the model predictions were, by plotting the computer-assigned prediction strength of the predicted categories against the observed relative frequency of the prediction being correct. We chose to do so because, if the two quantities are closely related, the prediction strength could be useful for filtering purposes, as expanded upon below.

RESULTS

Table 1 shows the frequency distribution of cases in the prediction set that were assigned one-digit and two-digit codes by both manual coders. As shown there, the number of cases varied from only 17 in the fire or explosion one-digit category, to 1013 in the bodily reaction and exertion category. Table 1 also gives the predicted frequencies by the Naïve and Fuzzy Bayes models for the same one-digit and two-digit codes. Statistics on the sensitivity, specificity, PPV and NPV of the model predictions for both one-digit and two-digit codes are also included. These statistics were also measured for the training set, revealing that sensitivity and PPV were 10–15% higher on the training set, for predictions at both the one-digit and two-digit levels, and the specificity and NPV were similar. As our focus was on the ability to predict new unclassified cases, the results presented in this paper are for the prediction set only.

Prediction of one-digit codes

The accuracy of predictions at the one-digit level was quite high for both models. A quick comparison of Naïve Bayes with Fuzzy Bayes reveals (mean) sensitivities averaged over all categories of 0.80 vs 0.78, specificities of 0.96 vs 0.93, and PPV of 0.80 vs 0.78. The predictions of both models were also well calibrated. That is, the computer-assigned prediction strength tended to be close to the observed relative frequency of the prediction being correct (fig 1). However, the fuzzy predictions appeared to be slightly better calibrated for prediction strengths of 0.8 or higher, in that Naïve prediction strengths above 0.8 tended to be greater than the observed accuracy.

Within categories, both models consistently showed a high specificity (table 1). The lowest mean specificity was a value of 0.87 for the Fuzzy Bayes predictions on the bodily motion

category (table 1). However, some variability, related to the size of the category, was observed in the mean sensitivity and PPV for both models. Mean sensitivity and PPV were both consistently high for each model on the five largest categories (contact, fall, exposure, bodily motion and transportation). On the smaller categories (fire and explosion, assaults, non-classifiable), the mean PPV continued to be high, but the mean sensitivity dropped for both models. However, the relatively large 95% CI for sensitivity and PPV on the smaller categories often overlapped that for the larger categories (ie, for Naïve Bayes a 95% CI for sensitivity of 0.68 to 0.86 for assaults versus 0.74 to 0.79 for bodily motion).

For Naïve Bayes, mean sensitivity varied from a low of 0.47 in the fire and explosion category to a high of 0.91 in the transportation category, and was greater than 0.75 for six of the eight categories (table 1). The mean PPV, ranged from a low of 0.68 for the assaults category to a high of 0.90 for bodily motion. For Fuzzy Bayes, the mean sensitivity was again the highest on the transportation category (0.90), the lowest on the non-classifiable category (0.37), and greater than 0.75 for four of the eight categories. The lowest mean sensitivity was observed for the three smaller categories (fire and explosion, assaults, non-classifiable) with mean sensitivities of 0.41, 0.54 and 0.37, respectively, but with quite large confidence intervals, because of the small sample size. Although the mean sensitivity of Fuzzy Bayes was lower than for Naïve Bayes (table 1), for the latter categories, the mean PPVs were higher. The mean PPV of Fuzzy Bayes predictions was above 0.75 for each category, and less than that of Naïve Bayes only for bodily motion and transportation.

In terms of mean sensitivity and PPV, the Naïve and Fuzzy Bayes models both performed the best for the transportation category and had difficulty with the non-classifiable category. The observed differences in model performance for the remaining categories seem to reflect a trade-off between specificity and PPV. For example, although the mean sensitivity of the Naïve model was higher than for Fuzzy on the contact category (0.80 vs 0.62), the mean PPV of the Fuzzy model was higher (0.74 vs 0.77). Similarly, the mean sensitivity of the Naïve model was higher for the fall category (0.85 vs 0.83), but mean specificity and PPV were both lower (specificity 0.92 vs 0.95, PPV 0.70 and 0.76). The opposite was true for the bodily motion category, where Fuzzy was more sensitive (mean of 0.87 vs 0.76), but had a much lower PPV (mean of 0.77 vs 0.90).

Prediction of two-digit codes

Both models had more difficulty predicting two-digit codes than one-digit BLS event codes, which was expected given the larger set of possible predictions and the similarity of two-digit codes within the same one-digit category. At the two-digit level, averaged over all categories, the Naïve Bayes model had a mean sensitivity of 0.70, specificity of 0.97 and PPV of 0.70. The lowest mean sensitivity for categories with more than 20 cases in the prediction set (table 1) was for the struck against and bodily reaction (0.46 and 0.52 respectively) categories. The lowest mean PPV was in the non-highway accidents category (0.48). The highest accuracy was in the highway accident category, with a mean sensitivity of 0.97, specificity of 0.98 and PPV of 0.83.

Analysis of the Fuzzy Bayes two-digit predictions (table 1) revealed an overall (mean) sensitivity of 0.64, specificity of 0.95 and PPV of 0.65, which were all lower than observed for the Naïve Bayes model. For eight different two-digit codes, the mean sensitivity and PPV of the Naïve Bayes model were both

Special feature

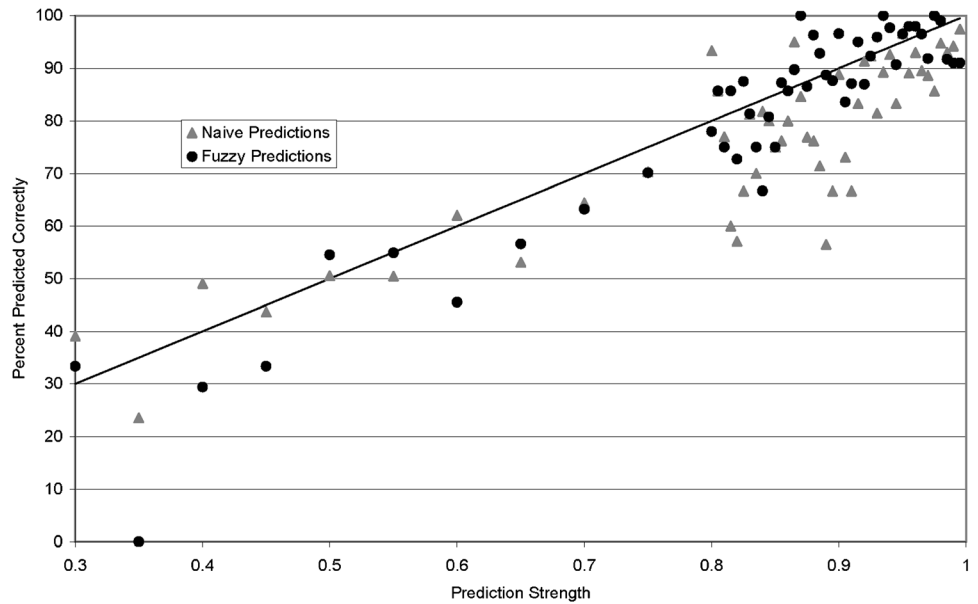
Table 1 Evaluation of Naïve and Fuzzy Bayes prediction of one-digit and two-digit BLS OIICS classifications

| BLS OIICS | Description | Gold standard | | Naïve Bayes model | | | | Fuzzy Bayes model | | | | NPV |
|--|----------------------------------|----------------------|--------------|---------------------|------|---------------------|------|-------------------|---------------------|------|---------------------|------|
| | | N _{ges} (%) | n (%) | Sen (95% CI) | Spec | PPV (95% CI) | NPV | n (%) | Sen (95% CI) | Spec | PPV (95% CI) | |
| Contact: Group 0 | | 523 (17.4) | 563 (18.8) | 0.80 (0.76 to 0.83) | 0.94 | 0.74 (0.71 to 0.78) | 0.96 | 417 (13.9) | 0.62 (0.57 to 0.66) | 0.96 | 0.77 (0.73 to 0.81) | 0.92 |
| 01 | Struck against | 145 (4.8) | 103 (3.4) | 0.46 (0.38 to 0.54) | 0.99 | 0.65 (0.56 to 0.74) | 0.97 | 61 (2.0) | 0.28 (0.20 to 0.35) | 0.99 | 0.66 (0.54 to 0.77) | 0.96 |
| 02 | Struck by | 294 (9.8) | 369 (12.3) | 0.72 (0.67 to 0.77) | 0.94 | 0.57 (0.52 to 0.62) | 0.97 | 257 (8.6) | 0.50 (0.44 to 0.56) | 0.96 | 0.57 (0.51 to 0.63) | 0.95 |
| 03 | Caught/compressed | 73 (2.4) | 76 (2.5) | 0.60 (0.49 to 0.71) | 0.99 | 0.58 (0.47 to 0.69) | 0.99 | 61 (2.0) | 0.51 (0.39 to 0.62) | 0.99 | 0.61 (0.48 to 0.73) | 0.99 |
| Fall: Group 1 | | 521 (17.4) | 632 (21.1) | 0.85 (0.82 to 0.88) | 0.92 | 0.70 (0.67 to 0.74) | 0.97 | 568 (18.9) | 0.83 (0.80 to 0.86) | 0.95 | 0.76 (0.73 to 0.80) | 0.96 |
| 11 | Fall to lower level | 185 (6.2) | 210 (7.0) | 0.70 (0.64 to 0.77) | 0.97 | 0.62 (0.55 to 0.68) | 0.98 | 180 (6.0) | 0.62 (0.55 to 0.69) | 0.98 | 0.64 (0.57 to 0.71) | 0.98 |
| 13 | Fall on same level | 322 (10.7) | 361 (12.0) | 0.73 (0.68 to 0.78) | 0.95 | 0.65 (0.60 to 0.70) | 0.97 | 365 (12.2) | 0.71 (0.66 to 0.76) | 0.95 | 0.63 (0.58 to 0.68) | 0.97 |
| Bodily motion: Group 2 | | 1013 (33.8) | 860 (28.7) | 0.76 (0.74 to 0.79) | 0.96 | 0.90 (0.88 to 0.92) | 0.89 | 1151 (38.4) | 0.87 (0.85 to 0.89) | 0.87 | 0.77 (0.74 to 0.79) | 0.93 |
| 21 | Bodily reaction | 124 (4.1) | 261 (8.7) | 0.52 (0.47 to 0.57) | 0.98 | 0.63 (0.69 to 0.79) | 0.95 | 183 (4.5) | 0.37 (0.32 to 0.42) | 0.98 | 0.74 (0.68 to 0.81) | 0.92 |
| 22 | Overexertion | 535 (17.8) | 575 (19.2) | 0.85 (0.82 to 0.88) | 0.95 | 0.79 (0.76 to 0.82) | 0.97 | 862 (28.7) | 0.92 (0.90 to 0.95) | 0.85 | 0.57 (0.54 to 0.61) | 0.98 |
| 23 | Repetitive motion | 76 (2.5) | 92 (3.1) | 0.86 (0.78 to 0.93) | 0.99 | 0.71 (0.61 to 0.80) | 1.00 | 81 (2.7) | 0.80 (0.71 to 0.89) | 0.99 | 0.75 (0.67 to 0.83) | 0.99 |
| Exposure to harmful substances or environment: Group 3 | | 303 (10.1) | 333 (11.1) | 0.88 (0.84 to 0.91) | 0.98 | 0.80 (0.76 to 0.84) | 0.99 | 318 (10.6) | 0.86 (0.82 to 0.90) | 0.98 | 0.82 (0.78 to 0.86) | 0.98 |
| 31 | Contact with electric | 28 (0.9) | 26 (0.9) | 0.75 (0.59 to 0.91) | 1.00 | 0.81 (0.66 to 0.96) | 1.00 | 30 (1.0) | 0.68 (0.51 to 0.85) | 1.00 | 0.63 (0.46 to 0.81) | 1.00 |
| 32 | Contact with temperature extreme | 93 (3.1) | 105 (3.5) | 0.88 (0.82 to 0.95) | 0.99 | 0.78 (0.70 to 0.86) | 1.00 | 110 (3.7) | 0.89 (0.83 to 0.96) | 0.99 | 0.75 (0.67 to 0.83) | 1.00 |
| 34 | Exposure to caustic substance | 110 (3.7) | 100 (3.3) | 0.76 (0.68 to 0.84) | 0.99 | 0.84 (0.77 to 0.91) | 0.99 | 106 (3.5) | 0.75 (0.67 to 0.83) | 0.99 | 0.78 (0.70 to 0.86) | 0.99 |
| 35 | Exposure to noise | 37 (1.2) | 38 (1.3) | 0.89 (0.79 to 0.99) | 1.00 | 0.87 (0.76 to 0.98) | 1.00 | 40 (1.3) | 0.97 (0.92 to 1.00) | 1.00 | 0.90 (0.81 to 0.99) | 1.00 |
| 37 | Exposure to stress | 33 (1.1) | 46 (1.5) | 0.85 (0.73 to 0.97) | 0.99 | 0.61 (0.47 to 0.75) | 1.00 | 50 (1.7) | 0.88 (0.77 to 0.99) | 0.99 | 0.58 (0.44 to 0.72) | 1.00 |
| Transportation: Group 4 | | 384 (12.8) | 407 (13.6) | 0.91 (0.89 to 0.94) | 0.98 | 0.86 (0.83 to 0.90) | 0.99 | 433 (14.4) | 0.90 (0.87 to 0.93) | 0.97 | 0.80 (0.76 to 0.84) | 0.99 |
| 41 | Highway accident | 220 (7.3) | 259 (8.6) | 0.97 (0.95 to 0.99) | 0.98 | 0.83 (0.78 to 0.87) | 1.00 | 329 (11.0) | 0.97 (0.95 to 0.99) | 0.96 | 0.65 (0.60 to 0.70) | 1.00 |
| 42 | Non-highway accident | 56 (1.9) | 86 (2.9) | 0.73 (0.62 to 0.85) | 0.98 | 0.48 (0.37 to 0.58) | 0.99 | 37 (1.2) | 0.30 (0.18 to 0.42) | 0.99 | 0.46 (0.30 to 0.62) | 0.99 |
| 43 | Pedestrian struck by vehicle | 104 (3.5) | 85 (2.8) | 0.63 (0.54 to 0.73) | 0.99 | 0.78 (0.69 to 0.87) | 0.99 | 76 (2.5) | 0.44 (0.35 to 0.54) | 0.99 | 0.61 (0.50 to 0.72) | 0.98 |
| Fire or explosion: Group 5 | | 17 (0.6) | 11 (0.4) | 0.47 (0.23 to 0.71) | 1.00 | 0.73 (0.46 to 0.99) | 1.00 | 9 (0.3) | 0.41 (0.18 to 0.65) | 1.00 | 0.78 (0.51 to 1.00) | 1.00 |
| 52 | Explosion | 11 (0.4) | 6 (0.2) | 0.45 (0.16 to 0.75) | 1.00 | 0.83 (0.54 to 1.00) | 1.00 | 9 (0.3) | 0.55 (0.25 to 0.84) | 1.00 | 0.67 (0.36 to 0.97) | 1.00 |
| Assaults and violent acts: Group 6 | | 87 (2.9) | 97 (3.2) | 0.76 (0.67 to 0.85) | 0.99 | 0.68 (0.59 to 0.77) | 0.99 | 61 (2.0) | 0.54 (0.44 to 0.64) | 1.00 | 0.77 (0.66 to 0.88) | 0.99 |
| 61 | Assaults | 82 (2.7) | 91 (3.0) | 0.77 (0.68 to 0.86) | 0.99 | 0.69 (0.60 to 0.79) | 0.99 | 82 (2.7) | 0.70 (0.60 to 0.79) | 0.99 | 0.70 (0.60 to 0.79) | 0.99 |
| Non-classifiable: Group 9 | | 152 (5.1) | 105 (3.5) | 0.49 (0.41 to 0.57) | 0.99 | 0.70 (0.62 to 0.79) | 0.97 | 69 (2.3) | 0.37 (0.29 to 0.45) | 1.00 | 0.81 (0.72 to 0.90) | 0.97 |
| 99 | Non-classifiable | 52 (1.7) | 4 (0.1) | 0.04 (0 to 0.09) | 1.00 | 0.50 (0.01 to 0.99) | 0.98 | 5 (0.2) | 0.08 (0.00-0.15) | 1.00 | 0.67 (0.13 to 1.00) | 1.00 |
| | General Unclassifiable* | 22 (0.7) | 2 (0.1) | 0.04 (0 to 0.13) | - | - | - | 7 (0.2) | 0.23 (0.05-0.40) | - | 0.01 (0 to 0.01) | 0.99 |
| | Other categories <10 | 3000 (100.0) | 3000 (100.0) | 0.70 (0.69 to 0.72) | 0.97 | 0.70 (0.68 to 0.71) | 0.97 | 3000 (100.0) | 0.64 (0.62 to 0.66) | 0.95 | 0.65 (0.63 to 0.66) | 0.98 |

*Unspecified and unclassifiable within category, ie, 10, contact unspecified.

BLS OIICS, Bureau of Labor Statistics Occupational Injury and Illness Classification System; N_{ges}, gold standard classifications; NPV, negative predictive value; PPV, positive predictive value; Sen, sensitivity; Spec, specificity.

Figure 1 Calibration curve for Naive and Fuzzy Bayes models.



as high or higher than for Fuzzy Bayes. The opposite was true only for the exposure to noise category, in which Fuzzy Bayes showed a higher mean sensitivity (0.97 vs 0.89) and PPV (0.90 vs 0.87). The Fuzzy model also did well in the overexertion category, where it showed a higher mean sensitivity (0.85 vs 0.92), but at the cost of a much lower mean PPV (0.57 vs 0.79). For the seven remaining two-digit BLS event codes, one model was slightly better on one criteria (ie, higher mean sensitivity or PPV), and not quite as good on the other, reflecting a trade-off similar to that observed at the one-digit level.

DISCUSSION

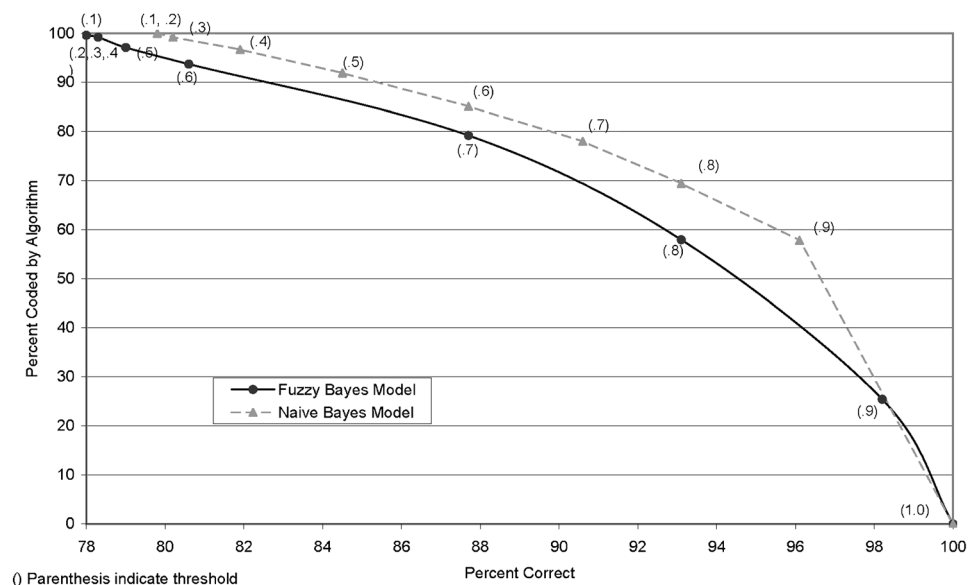
Using two Bayesian models, we were able to assign one-digit BLS event classifications to administrative workers' compensation injury narratives with high accuracy, almost matching the agreement of two manual classifiers (80% vs 87% agreement, respectively). The models learned how make these predictions with little or no human intervention, for new narratives not

used in the training set. We also demonstrated that this methodology could be used to predict more specific classifications (such as two-digit BLS OIICS event classifications). For some groups, such as highway accident, the Naïve algorithm was able to classify with 0.97 mean sensitivity, complimented with a 0.83 mean PPV, of those classifications. The current version of the algorithm was not able to detect other groups as well, such as struck against and bodily reaction (0.46 and 0.52 mean sensitivity, respectively), suggesting that more work is necessary to understand and incorporate changes to the current algorithm. With a semi-autonomous approach, such narratives would likely be the cases filtered out for manual review, giving the current model high utility even in this stage of development.

One of the significant advantages of the Bayesian models is the capability to have the computer assign classifications to narratives with almost complete confidence and then be able to filter out narratives below a certain threshold of confidence (intelligent selection) for manual review.¹³ The feasibility of

1

Figure 2 Trade-off between accuracy and proportion manually coded for semi-autonomous classification using Naive or Fuzzy Bayes models (one-digit classification). Parentheses indicate threshold.



Special feature

Table 2 Predictors used in Fuzzy Bayes model

| | Frequency of occurrence in prediction set | Frequency of times used to predict | Prediction strength |
|---------------------|---|------------------------------------|---------------------|
| Fall to lower level | | | |
| From-ladder | 7 | 4 | 0.71 |
| Off-ladder | 7 | 4 | 0.68 |
| Fell-off-ladder | 6 | 6 | 0.97 |
| Fall on same level | | | |
| Slipped-fell | 24 | 21 | 0.79 |
| Slipped-on-ice-fell | 12 | 12 | 0.93 |
| Slipped-fell-on-ice | 9 | 9 | 0.95 |
| Struck by | | | |
| Came-down | 9 | 7 | 0.74 |
| Fell-on-foot | 6 | 6 | 0.94 |
| Fell-onto-foot | 4 | 3 | 0.78 |
| Fell-struck | 3 | 3 | 0.86 |
| Highway | | | |
| Rear-ended | 10 | 10 | 0.97 |
| Was-rear-ended-by | 10 | 10 | 0.97 |
| Was-rear-ended | 10 | 10 | 0.96 |
| Struck against | | | |
| Finger-on | 6 | 4 | 0.75 |
| Struck-knee-on | 4 | 4 | 0.87 |
| Bruise&causing | 3 | 3 | 0.67 |
| Overexertion | | | |
| Felt-pain-to-lower | 5 | 5 | 0.93 |
| Lifting-boxes | 4 | 4 | 0.94 |
| Pushing-pulling | 4 | 4 | 0.94 |
| Bodily reaction | | | |
| Kneeling | 5 | 2 | 0.71 |
| Bent-over | 4 | 2 | 0.75 |
| Down&kneeling | 3 | 3 | 0.75 |
| Repetitive motion | | | |
| Typing | 9 | 9 | 0.96 |
| Keyboard | 5 | 4 | 0.76 |
| Carpal-tunnel | 3 | 3 | 0.93 |

performing such filtering is supported in this study by the finding that both Bayesian models assigned their predictions with a degree of confidence (strength) that was strongly related to the actual probability of being correct (fig 1). Figure 2 shows for both Bayesian models how the proportion of correctly coded cases at the one-digit level changes in the prediction set, when cases with a prediction strength below different threshold values are manually reviewed and then correctly coded. Threshold values are indicated in the figure as points on the separate curves shown for the two Bayesian models.

To better understand the relationship shown, if we choose a small threshold, of 0.1 for example, corresponding to a location at the top left corner of fig 2, we see that, for both models, 100% of the narratives would be coded by the algorithm, giving a 0.78–0.80 accuracy level without manual coding of any narratives (relying only on the algorithm's predictions). As the threshold increases, the percentage coded by the algorithm decreases for both models, and for the extreme case where the threshold is 1, 0% would be coded by the algorithm, but the percentage correct would go to 100% as all the cases would be manually coded (assuming that people can code them correctly). At intermediate threshold values, some narratives will have a prediction strength below the threshold. If we assume that only those cases will be manually coded, we can obtain an overall team performance that is better than for the model alone, but without requiring all of the cases to be manually coded. For

What is already known on this topic

- ▶ There is a recognised need to improve the accuracy of manually assigned E-codes.
- ▶ Computerised automated systems have been recognised as a potential solution to improve accuracy and reduce resource requirements necessary for manual classification of narratives in large administrative databases.
- ▶ The simplest form of automated systems use keyword-based search algorithms to assign event codes to cases on the basis of presence or absence of keywords in injury narratives. However, this method has been shown to have low accuracy.

What this study adds

- ▶ This study shows that a computer model based on Bayesian theory with minimal text processing can be implemented. The computer learns on its own from examples of previous work and predicts new classifications with high accuracy.
- ▶ This study also shows high sensitivity and specificity of a Naïve Bayesian model in predicting detailed two-digit Bureau of Labor statistics event classifications of injury narratives.
- ▶ A combined computer/manual coding approach allows the computer to code, in a consistent manner, narratives that use the same words over and over and has a strength parameter assigned to each prediction enabling filtering of more difficult narratives for manual review.

example, if a threshold of 0.70 was used to screen the Fuzzy Bayes predictions (see the 0.7 value on the solid line), manually coding only 20% (ie, 80% are coded by the algorithm) of the 3000 narratives would result in an overall accuracy of 0.88 (fig 2). The Naïve Bayes model does even better. That is, we can obtain the same accuracy of 0.88, using a threshold of 0.60 (now the 0.6 value on the dashed line), by manually classifying only about 14% of the narratives.

Such a reduction in the amount of manually coded cases could be useful when manual coding resources are limited, as this filtering approach would suggest which cases should be chosen to be manually coded. This capability goes well beyond traditional keyword-based systems, which generate their predictions in an all-or-nothing manner, and is a necessary step to developing systems that can filter out records needing manual review. This point is important, because classifying narratives in large administrative datasets becomes more feasible. In addition, the combination of some computer-generated codes, as well as manual codes, will likely provide accuracy above either all-computer or all-manual classifications, as human coders are not able to classify as systematically as the computer.

Another finding is that the Naïve Bayes model generally outperformed the Fuzzy Bayes model, especially for many two-digit codes. This performance advantage may reflect the fact that the word combinations and sequences needed for Fuzzy Bayes to make accurate predictions did not occur often enough in the smaller categories. By considering all the words, Naïve Bayes may have had an advantage, despite the fact that the conditional independence assumption is questionable, and was clearly violated in many cases. Another strength of Naïve Bayes is that it generated its predictions without requiring consideration of the many possible word combinations and sequences in

the injury narratives. The number of possible combinations and sequences of terms in injury narratives is immense. Reducing the items reduces computation time. However, the strengths of each model seemed to compliment each other. Therefore, it may be possible to develop an integrated approach using both the Naïve and Fuzzy classifiers and then, based on the predictive strengths of both, determine if one classification should be used over another or if manual review is required (eg, if both models predict with low strength). These methods will only improve accuracy over what we have demonstrated is possible by using the methods individually and with no filtering and very little preprocessing of the data.

Perhaps the primary advantage of the Fuzzy Bayes model was that the predictors used tended to be highly intuitive, which increased the face validity of the method. This tendency is illustrated in the examples shown in table 2. It also can be easily seen from the examples that some predictors correspond to very specific subcategories and injury causes of potential interest to the analyst (eg, fall off ladder, rear-ended by). As such, the Fuzzy Bayes method is very similar to traditional keyword-based methods, differing primarily in that the predictors are generated automatically instead of by humans by searching for words, word combinations or word sequences that are strongly related to particular categories. Application of the Fuzzy Bayes model can guide efforts to improve the sensitivity of keyword-based classification methods, by suggesting new terms not included in existing systems, or corresponding to specific subcategories of interest to the analyst.

CONCLUSIONS

Through this research we were able to demonstrate that Bayesian models can be used to automatically classify injury narratives from large administrative datasets (such as workers' compensation claims) into one-digit classifications with high accuracy and with fair accuracy at the two-digit level. This finding is substantial given the need for such classifiers in health statistics research. There are many organisations in the USA that are using a large amount of resources to manually classify large datasets. The CDC has determined the need to improve the quality and completeness of external cause of injury coding. Bayesian approaches could be used to help in this effort. We have demonstrated the potential for a semi-automatic/manual approach to classifying injury narratives which would likely improve accuracy beyond what manual coders or the computer alone can achieve, especially with regard to the two-digit classifications.

The accuracy is surprising considering the level of noise expected from workers' compensation claims narratives. We demonstrated that this approach is both feasible and accurate and requires little preprocessing of the original narratives. One of the more encouraging results is that the high levels of performance we observed in this study were obtained with a minimum level of human input. From one perspective, this level of performance implies human inputs are not necessary. Arguably, the need for developing synonym lists, spell-checking, word-stemming, pattern specification, natural language phrase generation, and other commonly applied methods may have been reduced simply because of the large size of the narrative set used in this study. On the other hand, it seems reasonable that these approaches could further improve model performance. It also seems reasonable that accuracy could be improved by increasing the size of the training set, especially for the smaller categories.

A Bayesian approach also provides a means (without using large resources) for re-evaluating narratives with different classification

protocols depending on the purpose of the investigation. Although the Naïve model had higher accuracy than the Fuzzy model, it could be seen that they complimented each other, and a combined approach may be best. Taken together, the overall conclusion that can be drawn from this research is that Bayesian approaches are clearly promising for applications requiring injury narratives to be classified. However, there are a number of practical issues related to implementing the methods that must be addressed before this can become a reality. For example, the integration of these methods with existing administrative databases will require careful consideration of complex issues related to software compatibility, record security and privacy concerns. Other important issues will have to be addressed related to the design of interfaces or modes of interaction to best assist human coders.

A key advantage of the Bayesian methods follows from the fact that the predictions are made on the basis of how often particular terms are found in particular categories. Consequently, the models can be easily updated (ie, learnt) when newly coded narratives are added to the training set by simply recalculating the relative frequencies of terms in particular categories. Such updating might be combined with manual filtering, and occur immediately every time a human confirms a prediction. In other cases, the additions to the training set might be newly coded cases accumulated over some preset interval of time. Deciding on the most appropriate method of updating the Bayesian models and how often this should be done is another important issue that will need to be addressed to implement these methods.

Competing interests: None.

REFERENCES

1. **Williamson A**, Feyer AM, Stout N, *et al*. Use of narrative analysis for comparisons of the causes of fatal accidents in three countries: New Zealand, Australia, and the United States. *Inj Prev* 2001;**7**(Suppl 1):i15–20.
2. **Lombardi DA**, Pannala R, Sorock GS, *et al*. Welding related occupational eye injuries: a narrative analysis. *Inj Prev* 2005;**11**:174–9.
3. **Wellman HM**, Lehto MR, Sorock GS. Computerized coding of injury narrative data from the National Health Interview Survey. *Accid Anal Prev* 2004;**36**:165–71.
4. **Lincoln AE**, Sorock GS, Courtney TK, *et al*. Using narrative text and coded data to develop hazard scenarios for occupational injury interventions. *Inj Prev* 2004;**10**:249–54.
5. **Bureau of Labor Statistics**. *Occupational injury and illness classification manual*. Washington, DC: US Department of Labor, December 1992.
6. **Hunt PR**, Hackman H, Berenholz G, *et al*. Completeness and accuracy of International Classification of Disease (ICD) external cause of injury codes in emergency department electronic data. *Inj Prev* 2007;**13**:422–5.
7. **Annest JL**, Fingerhut LA, Gallagher SS, *et al*. Centers for Disease Control and Prevention (CDC). Strategies to improve external cause-of-injury coding in state-based hospital discharge and emergency department data systems: recommendations of the CDC Workgroup for Improvement of External Cause-of-Injury Coding. *MMWR Recomm Rep* 2008;**57**(RR-1):1–15.
8. **Noorinaeini A**, Lehto MR. Hybrid singular value decomposition: a model of text classification. *International Journal of Human Factors Modeling and Simulation* 2006;**1**:95–118.
9. **Sebastiani F**. Machine learning in automated text categorization. *ACM Computing Surveys (CSUR)* 2002;**34**:1–47.
10. **Lewis DD**. Naive Bayes at forty: the independence assumption in information retrieval. In: *Proceedings of the 15th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1992:37–50.
11. **Corns HL**, Marucci HR, Lehto MR. Development of an approach for optimizing the accuracy of classifying claims narratives using a machine learning tool (TEXTMINER). In: *Proceedings of HCI International 2007, 12th International Conference on Human-Computer Interaction(8)*. 2007:411–16.
12. **Marucci HR**, Lehto MR, Corns HL. Computer classification of injury narratives using a Fuzzy Bayes approach: improving the model. In: *Proceedings of HCI International 2007, 12th International Conference on Human-Computer Interaction(8)*. 2007:500–6.
13. **Lehto MR**, Sorock G. Machine learning of motor vehicle accident categories from narrative data. *Methods Inf Med* 1996;**35**:1–8.
14. **Sorock G**, Ranney T, Lehto M. Motor vehicle crashes in roadway construction work zones: an analysis using narrative text from insurance claims. *Accid Anal Prev* 1996;**28**:131–8.

Bayesian methods: a useful tool for classifying injury narratives into cause groups

M Lehto,¹ H Marucci-Wellman,² H Corns²

¹ School of Industrial Engineering, Purdue University, West Lafayette, Indiana, USA; ² Liberty Mutual Research Institute for Safety, Hopkinton, Massachusetts, USA

Correspondence to: Professor M Lehto, School of Industrial Engineering, Purdue University, 1287 Grissom Hall, West Lafayette, IN 47907, USA; lehto@purdue.edu

Accepted 17 April 2009

ABSTRACT

To compare two Bayesian methods (Fuzzy and Naïve) for classifying injury narratives in large administrative databases into event cause groups, a dataset of 14 000 narratives was randomly extracted from claims filed with a worker's compensation insurance provider. Two expert coders assigned one-digit and two-digit Bureau of Labor Statistics (BLS) Occupational Injury and Illness Classification event codes to each narrative. The narratives were separated into a training set of 11 000 cases and a prediction set of 3000 cases. The training set was used to develop two Bayesian classifiers that assigned BLS codes to narratives. Each model was then evaluated for the prediction set. Both models performed well and tended to predict one-digit BLS codes more accurately than two-digit codes. The overall sensitivity of the Fuzzy method was, respectively, 78% and 64% for one-digit and two-digit codes, specificity was 93% and 95%, and positive predictive value (PPV) was 78% and 65%. The Naïve method showed similar accuracy: a sensitivity of 80% and 70%, specificity of 96% and 97%, and PPV of 80% and 70%. For large administrative databases, Bayesian methods show significant promise as a means of classifying injury narratives into cause groups. Overall, Naïve Bayes provided slightly more accurate predictions than Fuzzy Bayes.

In the USA, most hospitals and trauma centres classify injury causes into International Classification of Disease (ICD) E-code categories. The same coding scheme is used in large health surveys such as the National Health Interview Survey (NHIS) conducted by the National Center for Health Statistics (NCHS). Public health authorities, government agencies and researchers use the coded data to identify patterns and trends in external causes of injury.¹⁻⁴ The Bureau of Labor Statistics (BLS) uses a similar approach to identify events leading to workplace injuries as defined by the occupational injury and illness classification system, OIICS.⁵

For the most part, the injury causes are assigned manually in these systems by humans based on a narrative description. This process is resource intensive. The accuracy and completeness of manual coding is another important issue. One recent study evaluating the accuracy of E-codes assigned in emergency department data found an accuracy of 65% for work-related injuries and 57% for non-work-related injuries.⁶ Other studies reviewed in a recent report prepared by the Centers for Disease Control (CDC) showed levels of inaccurate E-coding from 13% to 18%.⁷ Recently, the CDC has been actively promoting strategies for improving the quality and completeness of exter-

nal cause of injury coding in the USA, including the use of automated systems that assist coders in assigning event codes.⁷

The simplest automated systems assign event codes to cases on the basis of presence or absence of keywords in injury narratives. For example, a case might be assigned the code "contact with electricity" if it contains the word "electricity" or "volt." Previous work has shown that keyword-based approaches can classify injury narratives with a high level of specificity,¹ but low sensitivity for many categories. Machine learning algorithms offer an alternative approach, in which the system learns how to classify injury text without human intervention from the typically massive amounts of previously coded narratives in administrative databases.^{3,8} The potential of this approach was demonstrated in a study in which a multiple-word Fuzzy Bayes model was used to assign event codes to narrative descriptions of injury incidents from the NHIS.³ The results had high sensitivity (83%) and also suggested the potential to filter out targeted records for manual review. However, owing to dataset constraints (ie, only ~5000 total coded narratives), the same dataset was used to train the algorithm and make predictions, which may have resulted in an optimistic bias.³

The primary objective of this study was to demonstrate that Bayesian methods of computer classification can accurately classify injury narratives from large administrative databases into cause groups, and that this can be done (1) with little or no human intervention, (2) for new injury narratives not used to train (or develop) the classification algorithm, and (3) in a way that estimates how likely each prediction is to be correct. Attaining this objective would be an important first step, in implementation of automatic coding methods. Two different Bayesian models (Fuzzy and Naïve) found to be promising in our previous studies were compared from this perspective, as expanded upon below.

METHODS

Training and prediction datasets

Over 17 000 records were randomly extracted from claims filed between January 2002 and December 2004 with a workers' compensation insurance provider. Each record included a unique identifier, a narrative describing how the injury occurred, and a two-digit BLS OIICS event code manually assigned and agreed upon by two expert coders. The OIICS scheme includes ~40 mutually exclusive injury and event categories. The two coders who classified these narratives went through an intensive training process and practised with each

Special feature

other. At the end of the training session, the two expert coders were tested on ~2000 narratives. Overall, they agreed on classifications at the one-digit level 87% of the time, and at the two-digit level 75% of the time, but the level of agreement varied by category.

The observed lack of agreement between manual raters was expected, given that the claims narratives were fairly short (usually between nine and 20 words long, with a mean length of 14 words), which obviously limited the amount of information provided. An example narrative, in this case for a transportation accident (BLS OIICS two-digit code of 41), is given below.

“HUSB. & SON WERE REARENDED AT RED TRAFFIC LIGHT
BY DRUNK DRIVER DRIV- ING AT LEAST 45 MPH INFULL
SIZE PICK-UP TRUCK//N”

As illustrated by the example, narratives tended to be noisy, with many misspellings, run-on words, abbreviations, and inconsistent or missing punctuation. After elimination of cases on which the raters disagreed, the data were then divided into a training set of 11 000 cases used for model development and a prediction dataset of 3000 cases for model evaluation.

Model development

Two different Bayesian models, referred to as Naïve Bayes and Fuzzy Bayes, were developed. (Both Bayes models were developed and evaluated using the Textminer program developed by one of the authors (ML). The Textminer program is implemented in Visual Basic and provides an interface for analysing any dataset format readable by Microsoft Access. Additional information on the program can be obtained by contacting ML.) Both models used the statistical relationship between terms in the 11 000 injury narratives in the training set and the manually assigned one-digit and two-digit BLS OIICS codes to estimate the probability that a human coder would assign a particular code to a new narrative, given the words that were present in the narrative. The training set was a random sample of cases from the 17 000 manually coded records. A large training set was used, because, as in any implementation of statistical models, a larger sample size reduces the effect of noise. (In field applications, the entire set of previously coded narratives would normally be used as the training set. Then, if narratives known to be coded correctly became available, they could be added to the training set to improve accuracy when convenient. Note that this study used a completely separate prediction set to avoid potential bias due to overfitting the data and to more accurately model the situation where predictions are made for new unclassified cases.)

After the cases had been divided into a training set and prediction set, the next step in model development was to extract the words used in each narrative. This resulted in a list showing which words were present for each narrative. The extracted words were then cleaned up, by removing punctuation marks and non-alphabetical characters. A small set of “stop” words or words that have no predictive value but create burdensome processing time (eg, “and”, “a”, “the”) were also deleted. Words occurring fewer than three times in the entire set of narratives were also dropped. As one objective of the study was to minimise the degree of human input required during model development, no additional steps were performed during this process. After identification of the words present in the narratives, the next step was to tabulate their frequency of occurrence for each of the assigned categories in the training set,

which corresponded to training the models. The two Bayesian models were then implemented, as expanded upon below.

Naïve Bayes model

The Naïve Bayes model is a commonly applied method of text classification which has been used for years in the field of information retrieval.⁹ To see how the model works, let us assume that a given narrative consists of a vector of j words, $n = \{n_1, n_2, \dots, n_j\}$. Also, assume that i possible event codes can be assigned resulting in a second vector $E = \{E_1, E_2, \dots, E_i\}$. By making what is called the conditional independence assumption,¹⁰ the probability of assigning a particular event code category can then be calculated using the expression:

$$P(E_i|n) = \prod_j \frac{P(n_j|E_i) P(E_i)}{P(n_j)}$$

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

In application, $P(n_j|E_i)$, $P(E_i)$ and $P(n_j)$ are all normally estimated on the basis of their frequency in a training set. Also, $P(n_j|E_i)$ is normally smoothed to reduce the effects of noise. The approach we implemented was to add a small constant to the number of times a particular word occurred in a category, as shown below:

$$P(n_j|E_i) = \frac{\text{count}(n_j|E_i) + \alpha \times \text{count}(n_j)}{\text{count}(E_i) + \alpha \times N}$$

where $\text{count}(n_j|E_i)$ is the number of times word n_j occurs in category E_i , $\text{count}(n_j)$ is the number of times word n_j occurs, $\text{count}(E_i)$ is the number of times category E_i occurs, and α is a smoothing constant. Larger values of α reduce the weight given to the evidence provided by each term. We chose to use a value of $\alpha = 0.05$, which corresponds to a small level of smoothing.

The conditional independence assumption is perhaps the most controversial aspect of the Naïve Bayes model. Informally, for the purposes of text classification, when this assumption holds, the probability of each index term (eg, word or word sequence) being present depends on only the event code considered and is independent of the remaining terms in the narrative. The conditional independence assumption is almost always violated in practice. However, a long history of application shows that Naïve Bayes tends to work remarkably well even when this assumption is violated.¹⁰

Fuzzy Bayes model

The Fuzzy Bayes approach avoids making the conditional independence assumption by calculating $P(A_i|n)$ using the expression:

$$P(E_i|n) = \text{MAX}_j \frac{P(n_j|E_i) P(E_i)}{P(n_j)}$$

where each term i is assigned a value as explained in equation 2 above. The primary difference from Naïve Bayes is that instead of multiplying the conditional probabilities, Fuzzy Bayes estimates $P(E_i|n)$ using the “index term” most strongly predictive of the category. In practice, n_j in the above expression can be a combination or sequence of words, which allows Fuzzy

Bayes to consider multiple pieces of evidence when calculating $P(E_i|n)$. Many word combinations and word sequences (such as “fell-off”) are accurate and intuitive predictors of event codes.^{3 11 12} Consequently, we included combinations of up to four words and sequences of up to three words as predictors in this study.

Model evaluation

Two independent trials were conducted, in which predictions were made for the 3000 narratives in the prediction set using the Naïve and Fuzzy Bayes algorithms, respectively. For each approach, the category calculated by the algorithm as having the highest prediction strength,

$$P(E_i|n)$$

for the terms in the narrative, was chosen as the “predicted code.” The obtained results were then evaluated by comparing the predictions with the manually assigned “gold standard” codes, for both one-digit and two-digit classifications. The evaluation included calculations of sensitivity, specificity, positive predictive value (PPV) and negative predictive value (NPV) for each method.

In addition to these four evaluation metrics, we tested how well calibrated the model predictions were, by plotting the computer-assigned prediction strength of the predicted categories against the observed relative frequency of the prediction being correct. We chose to do so because, if the two quantities are closely related, the prediction strength could be useful for filtering purposes, as expanded upon below.

RESULTS

Table 1 shows the frequency distribution of cases in the prediction set that were assigned one-digit and two-digit codes by both manual coders. As shown there, the number of cases varied from only 17 in the fire or explosion one-digit category, to 1013 in the bodily reaction and exertion category. Table 1 also gives the predicted frequencies by the Naïve and Fuzzy Bayes models for the same one-digit and two-digit codes. Statistics on the sensitivity, specificity, PPV and NPV of the model predictions for both one-digit and two-digit codes are also included. These statistics were also measured for the training set, revealing that sensitivity and PPV were 10–15% higher on the training set, for predictions at both the one-digit and two-digit levels, and the specificity and NPV were similar. As our focus was on the ability to predict new unclassified cases, the results presented in this paper are for the prediction set only.

Prediction of one-digit codes

The accuracy of predictions at the one-digit level was quite high for both models. A quick comparison of Naïve Bayes with Fuzzy Bayes reveals (mean) sensitivities averaged over all categories of 0.80 vs 0.78, specificities of 0.96 vs 0.93, and PPV of 0.80 vs 0.78. The predictions of both models were also well calibrated. That is, the computer-assigned prediction strength tended to be close to the observed relative frequency of the prediction being correct (fig 1). However, the fuzzy predictions appeared to be slightly better calibrated for prediction strengths of 0.8 or higher, in that Naïve prediction strengths above 0.8 tended to be greater than the observed accuracy.

Within categories, both models consistently showed a high specificity (table 1). The lowest mean specificity was a value of 0.87 for the Fuzzy Bayes predictions on the bodily motion

category (table 1). However, some variability, related to the size of the category, was observed in the mean sensitivity and PPV for both models. Mean sensitivity and PPV were both consistently high for each model on the five largest categories (contact, fall, exposure, bodily motion and transportation). On the smaller categories (fire and explosion, assaults, non-classifiable), the mean PPV continued to be high, but the mean sensitivity dropped for both models. However, the relatively large 95% CI for sensitivity and PPV on the smaller categories often overlapped that for the larger categories (ie, for Naïve Bayes a 95% CI for sensitivity of 0.68 to 0.86 for assaults versus 0.74 to 0.79 for bodily motion).

For Naïve Bayes, mean sensitivity varied from a low of 0.47 in the fire and explosion category to a high of 0.91 in the transportation category, and was greater than 0.75 for six of the eight categories (table 1). The mean PPV, ranged from a low of 0.68 for the assaults category to a high of 0.90 for bodily motion. For Fuzzy Bayes, the mean sensitivity was again the highest on the transportation category (0.90), the lowest on the non-classifiable category (0.37), and greater than 0.75 for four of the eight categories. The lowest mean sensitivity was observed for the three smaller categories (fire and explosion, assaults, non-classifiable) with mean sensitivities of 0.41, 0.54 and 0.37, respectively, but with quite large confidence intervals, because of the small sample size. Although the mean sensitivity of Fuzzy Bayes was lower than for Naïve Bayes (table 1), for the latter categories, the mean PPVs were higher. The mean PPV of Fuzzy Bayes predictions was above 0.75 for each category, and less than that of Naïve Bayes only for bodily motion and transportation.

In terms of mean sensitivity and PPV, the Naïve and Fuzzy Bayes models both performed the best for the transportation category and had difficulty with the non-classifiable category. The observed differences in model performance for the remaining categories seem to reflect a trade-off between specificity and PPV. For example, although the mean sensitivity of the Naïve model was higher than for Fuzzy on the contact category (0.80 vs 0.62), the mean PPV of the Fuzzy model was higher (0.74 vs 0.77). Similarly, the mean sensitivity of the Naïve model was higher for the fall category (0.85 vs 0.83), but mean specificity and PPV were both lower (specificity 0.92 vs 0.95, PPV 0.70 and 0.76). The opposite was true for the bodily motion category, where Fuzzy was more sensitive (mean of 0.87 vs 0.76), but had a much lower PPV (mean of 0.77 vs 0.90).

Prediction of two-digit codes

Both models had more difficulty predicting two-digit codes than one-digit BLS event codes, which was expected given the larger set of possible predictions and the similarity of two-digit codes within the same one-digit category. At the two-digit level, averaged over all categories, the Naïve Bayes model had a mean sensitivity of 0.70, specificity of 0.97 and PPV of 0.70. The lowest mean sensitivity for categories with more than 20 cases in the prediction set (table 1) was for the struck against and bodily reaction (0.46 and 0.52 respectively) categories. The lowest mean PPV was in the non-highway accidents category (0.48). The highest accuracy was in the highway accident category, with a mean sensitivity of 0.97, specificity of 0.98 and PPV of 0.83.

Analysis of the Fuzzy Bayes two-digit predictions (table 1) revealed an overall (mean) sensitivity of 0.64, specificity of 0.95 and PPV of 0.65, which were all lower than observed for the Naïve Bayes model. For eight different two-digit codes, the mean sensitivity and PPV of the Naïve Bayes model were both

Special feature

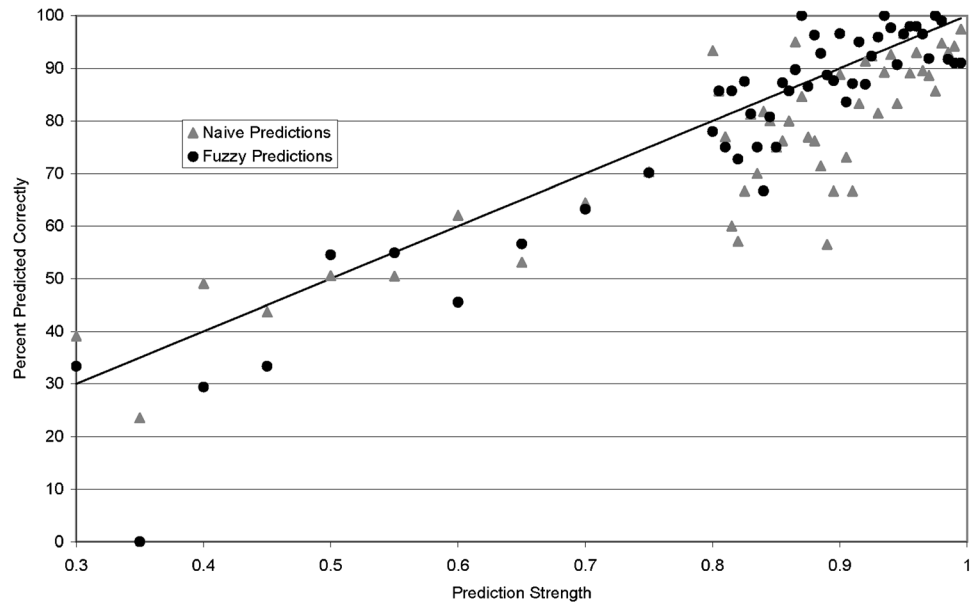
Table 1 Evaluation of Naïve and Fuzzy Bayes prediction of one-digit and two-digit BLS OIICS classifications

| BLS OIICS | Description | Gold standard | | Naïve Bayes model | | | | Fuzzy Bayes model | | | | NPV |
|--|----------------------------------|----------------------|--------------|---------------------|------|---------------------|------|-------------------|---------------------|------|---------------------|------|
| | | N _{ges} (%) | n (%) | Sen (95% CI) | Spec | PPV (95% CI) | NPV | n (%) | Sen (95% CI) | Spec | PPV (95% CI) | |
| Contact: Group 0 | | 523 (17.4) | 563 (18.8) | 0.80 (0.76 to 0.83) | 0.94 | 0.74 (0.71 to 0.78) | 0.96 | 417 (13.9) | 0.62 (0.57 to 0.66) | 0.96 | 0.77 (0.73 to 0.81) | 0.92 |
| 01 | Struck against | 145 (4.8) | 103 (3.4) | 0.46 (0.38 to 0.54) | 0.99 | 0.65 (0.56 to 0.74) | 0.97 | 61 (2.0) | 0.28 (0.20 to 0.35) | 0.99 | 0.66 (0.54 to 0.77) | 0.96 |
| 02 | Struck by | 294 (9.8) | 369 (12.3) | 0.72 (0.67 to 0.77) | 0.94 | 0.57 (0.52 to 0.62) | 0.97 | 257 (8.6) | 0.50 (0.44 to 0.56) | 0.96 | 0.57 (0.51 to 0.63) | 0.95 |
| 03 | Caught/compressed | 73 (2.4) | 76 (2.5) | 0.60 (0.49 to 0.71) | 0.99 | 0.58 (0.47 to 0.69) | 0.99 | 61 (2.0) | 0.51 (0.39 to 0.62) | 0.99 | 0.61 (0.48 to 0.73) | 0.99 |
| Fall: Group 1 | | 521 (17.4) | 632 (21.1) | 0.85 (0.82 to 0.88) | 0.92 | 0.70 (0.67 to 0.74) | 0.97 | 568 (18.9) | 0.83 (0.80 to 0.86) | 0.95 | 0.76 (0.73 to 0.80) | 0.96 |
| 11 | Fall to lower level | 185 (6.2) | 210 (7.0) | 0.70 (0.64 to 0.77) | 0.97 | 0.62 (0.55 to 0.68) | 0.98 | 180 (6.0) | 0.62 (0.55 to 0.69) | 0.98 | 0.64 (0.57 to 0.71) | 0.98 |
| 13 | Fall on same level | 322 (10.7) | 361 (12.0) | 0.73 (0.68 to 0.78) | 0.95 | 0.65 (0.60 to 0.70) | 0.97 | 365 (12.2) | 0.71 (0.66 to 0.76) | 0.95 | 0.63 (0.58 to 0.68) | 0.97 |
| Bodily motion: Group 2 | | 1013 (33.8) | 860 (28.7) | 0.76 (0.74 to 0.79) | 0.96 | 0.90 (0.88 to 0.92) | 0.89 | 1151 (38.4) | 0.87 (0.85 to 0.89) | 0.87 | 0.77 (0.74 to 0.79) | 0.93 |
| 21 | Bodily reaction | 124 (4.1) | 261 (8.7) | 0.52 (0.47 to 0.57) | 0.98 | 0.63 (0.69 to 0.79) | 0.95 | 183 (4.5) | 0.37 (0.32 to 0.42) | 0.98 | 0.74 (0.68 to 0.81) | 0.92 |
| 22 | Overexertion | 535 (17.8) | 575 (19.2) | 0.85 (0.82 to 0.88) | 0.95 | 0.79 (0.76 to 0.82) | 0.97 | 862 (28.7) | 0.92 (0.90 to 0.95) | 0.85 | 0.57 (0.54 to 0.61) | 0.98 |
| 23 | Repetitive motion | 76 (2.5) | 92 (3.1) | 0.86 (0.78 to 0.93) | 0.99 | 0.71 (0.61 to 0.80) | 1.00 | 81 (2.7) | 0.80 (0.71 to 0.89) | 0.99 | 0.75 (0.67 to 0.83) | 0.99 |
| Exposure to harmful substances or environment: Group 3 | | 303 (10.1) | 333 (11.1) | 0.88 (0.84 to 0.91) | 0.98 | 0.80 (0.76 to 0.84) | 0.99 | 318 (10.6) | 0.86 (0.82 to 0.90) | 0.98 | 0.82 (0.78 to 0.86) | 0.98 |
| 31 | Contact with electric | 28 (0.9) | 26 (0.9) | 0.75 (0.59 to 0.91) | 1.00 | 0.81 (0.66 to 0.96) | 1.00 | 30 (1.0) | 0.68 (0.51 to 0.85) | 1.00 | 0.63 (0.46 to 0.81) | 1.00 |
| 32 | Contact with temperature extreme | 93 (3.1) | 105 (3.5) | 0.88 (0.82 to 0.95) | 0.99 | 0.78 (0.70 to 0.86) | 1.00 | 110 (3.7) | 0.89 (0.83 to 0.96) | 0.99 | 0.75 (0.67 to 0.83) | 1.00 |
| 34 | Exposure to caustic substance | 110 (3.7) | 100 (3.3) | 0.76 (0.68 to 0.84) | 0.99 | 0.84 (0.77 to 0.91) | 0.99 | 106 (3.5) | 0.75 (0.67 to 0.83) | 0.99 | 0.78 (0.70 to 0.86) | 0.99 |
| 35 | Exposure to noise | 37 (1.2) | 38 (1.3) | 0.89 (0.79 to 0.99) | 1.00 | 0.87 (0.76 to 0.98) | 1.00 | 40 (1.3) | 0.97 (0.92 to 1.00) | 1.00 | 0.90 (0.81 to 0.99) | 1.00 |
| 37 | Exposure to stress | 33 (1.1) | 46 (1.5) | 0.85 (0.73 to 0.97) | 0.99 | 0.61 (0.47 to 0.75) | 1.00 | 50 (1.7) | 0.88 (0.77 to 0.99) | 0.99 | 0.58 (0.44 to 0.72) | 1.00 |
| Transportation: Group 4 | | 384 (12.8) | 407 (13.6) | 0.91 (0.89 to 0.94) | 0.98 | 0.86 (0.83 to 0.90) | 0.99 | 433 (14.4) | 0.90 (0.87 to 0.93) | 0.97 | 0.80 (0.76 to 0.84) | 0.99 |
| 41 | Highway accident | 220 (7.3) | 259 (8.6) | 0.97 (0.95 to 0.99) | 0.98 | 0.83 (0.78 to 0.87) | 1.00 | 329 (11.0) | 0.97 (0.95 to 0.99) | 0.96 | 0.65 (0.60 to 0.70) | 1.00 |
| 42 | Non-highway accident | 56 (1.9) | 86 (2.9) | 0.73 (0.62 to 0.85) | 0.98 | 0.48 (0.37 to 0.58) | 0.99 | 37 (1.2) | 0.30 (0.18 to 0.42) | 0.99 | 0.46 (0.30 to 0.62) | 0.99 |
| 43 | Pedestrian struck by vehicle | 104 (3.5) | 85 (2.8) | 0.63 (0.54 to 0.73) | 0.99 | 0.78 (0.69 to 0.87) | 0.99 | 76 (2.5) | 0.44 (0.35 to 0.54) | 0.99 | 0.61 (0.50 to 0.72) | 0.98 |
| Fire or explosion: Group 5 | | 17 (0.6) | 11 (0.4) | 0.47 (0.23 to 0.71) | 1.00 | 0.73 (0.46 to 0.99) | 1.00 | 9 (0.3) | 0.41 (0.18 to 0.65) | 1.00 | 0.78 (0.51 to 1.00) | 1.00 |
| 52 | Explosion | 11 (0.4) | 6 (0.2) | 0.45 (0.16 to 0.75) | 1.00 | 0.83 (0.54 to 1.00) | 1.00 | 9 (0.3) | 0.55 (0.25 to 0.84) | 1.00 | 0.67 (0.36 to 0.97) | 1.00 |
| Assaults and violent acts: Group 6 | | 87 (2.9) | 97 (3.2) | 0.76 (0.67 to 0.85) | 0.99 | 0.68 (0.59 to 0.77) | 0.99 | 61 (2.0) | 0.54 (0.44 to 0.64) | 1.00 | 0.77 (0.66 to 0.88) | 0.99 |
| 61 | Assaults | 82 (2.7) | 91 (3.0) | 0.77 (0.68 to 0.86) | 0.99 | 0.69 (0.60 to 0.79) | 0.99 | 82 (2.7) | 0.70 (0.60 to 0.79) | 0.99 | 0.70 (0.60 to 0.79) | 0.99 |
| Non-classifiable: Group 9 | | 152 (5.1) | 105 (3.5) | 0.49 (0.41 to 0.57) | 0.99 | 0.70 (0.62 to 0.79) | 0.97 | 69 (2.3) | 0.37 (0.29 to 0.45) | 1.00 | 0.81 (0.72 to 0.90) | 0.97 |
| 99 | Non-classifiable | 52 (1.7) | 4 (0.1) | 0.04 (0 to 0.09) | 1.00 | 0.50 (0.01 to 0.99) | 0.98 | 5 (0.2) | 0.08 (0.00-0.15) | 1.00 | 0.67 (0.13 to 1.00) | 1.00 |
| General Unclassifiable* | | 22 (0.7) | 2 (0.1) | 0.04 (0 to 0.13) | - | - | - | 7 (0.2) | 0.23 (0.05-0.40) | - | 0.01 (0 to 0.01) | 0.99 |
| Other categories <10 | | 3000 (100.0) | 3000 (100.0) | 0.70 (0.69 to 0.72) | 0.97 | 0.70 (0.68 to 0.71) | 0.97 | 3000 (100.0) | 0.64 (0.62 to 0.66) | 0.95 | 0.65 (0.63 to 0.66) | 0.98 |

*Unspecified and unclassifiable within category, ie, 10, contact unspecified.

BLS OIICS, Bureau of Labor Statistics Occupational Injury and Illness Classification System; N_{ges}, gold standard classifications; NPV, negative predictive value; PPV, positive predictive value; Sen, sensitivity; Spec, specificity.

Figure 1 Calibration curve for Naive and Fuzzy Bayes models.



as high or higher than for Fuzzy Bayes. The opposite was true only for the exposure to noise category, in which Fuzzy Bayes showed a higher mean sensitivity (0.97 vs 0.89) and PPV (0.90 vs 0.87). The Fuzzy model also did well in the overexertion category, where it showed a higher mean sensitivity (0.85 vs 0.92), but at the cost of a much lower mean PPV (0.57 vs 0.79). For the seven remaining two-digit BLS event codes, one model was slightly better on one criteria (ie, higher mean sensitivity or PPV), and not quite as good on the other, reflecting a trade-off similar to that observed at the one-digit level.

DISCUSSION

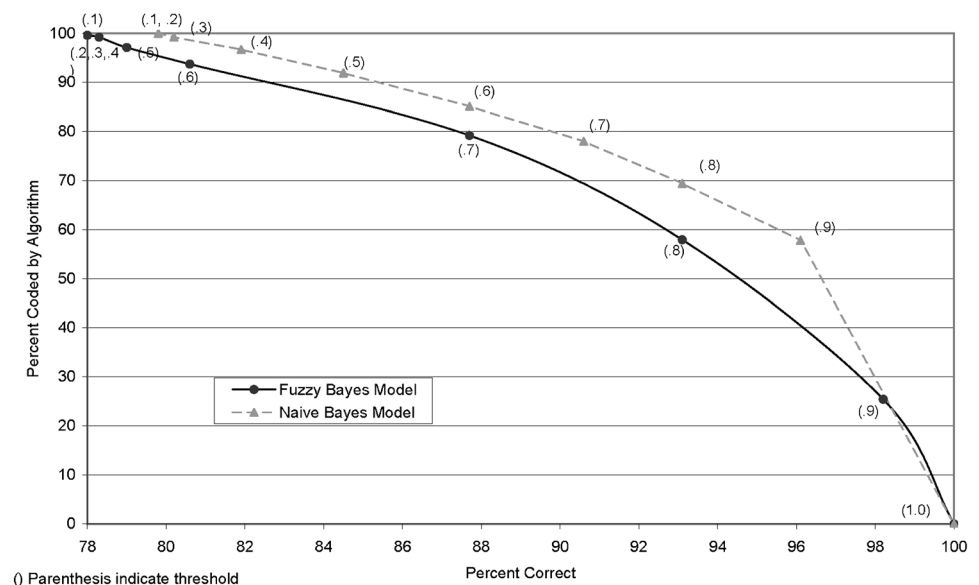
Using two Bayesian models, we were able to assign one-digit BLS event classifications to administrative workers' compensation injury narratives with high accuracy, almost matching the agreement of two manual classifiers (80% vs 87% agreement, respectively). The models learned how make these predictions with little or no human intervention, for new narratives not

used in the training set. We also demonstrated that this methodology could be used to predict more specific classifications (such as two-digit BLS OIICS event classifications). For some groups, such as highway accident, the Naïve algorithm was able to classify with 0.97 mean sensitivity, complimented with a 0.83 mean PPV, of those classifications. The current version of the algorithm was not able to detect other groups as well, such as struck against and bodily reaction (0.46 and 0.52 mean sensitivity, respectively), suggesting that more work is necessary to understand and incorporate changes to the current algorithm. With a semi-autonomous approach, such narratives would likely be the cases filtered out for manual review, giving the current model high utility even in this stage of development.

One of the significant advantages of the Bayesian models is the capability to have the computer assign classifications to narratives with almost complete confidence and then be able to filter out narratives below a certain threshold of confidence (intelligent selection) for manual review.¹³ The feasibility of

1

Figure 2 Trade-off between accuracy and proportion manually coded for semi-autonomous classification using Naive or Fuzzy Bayes models (one-digit classification). Parentheses indicate threshold.



Special feature

Table 2 Predictors used in Fuzzy Bayes model

| | Frequency of occurrence in prediction set | Frequency of times used to predict | Prediction strength |
|---------------------|---|------------------------------------|---------------------|
| Fall to lower level | | | |
| From-ladder | 7 | 4 | 0.71 |
| Off-ladder | 7 | 4 | 0.68 |
| Fell-off-ladder | 6 | 6 | 0.97 |
| Fall on same level | | | |
| Slipped-fell | 24 | 21 | 0.79 |
| Slipped-on-ice-fell | 12 | 12 | 0.93 |
| Slipped-fell-on-ice | 9 | 9 | 0.95 |
| Struck by | | | |
| Came-down | 9 | 7 | 0.74 |
| Fell-on-foot | 6 | 6 | 0.94 |
| Fell-onto-foot | 4 | 3 | 0.78 |
| Fell-struck | 3 | 3 | 0.86 |
| Highway | | | |
| Rear-ended | 10 | 10 | 0.97 |
| Was-rear-ended-by | 10 | 10 | 0.97 |
| Was-rear-ended | 10 | 10 | 0.96 |
| Struck against | | | |
| Finger-on | 6 | 4 | 0.75 |
| Struck-knee-on | 4 | 4 | 0.87 |
| Bruise&causing | 3 | 3 | 0.67 |
| Overexertion | | | |
| Felt-pain-to-lower | 5 | 5 | 0.93 |
| Lifting-boxes | 4 | 4 | 0.94 |
| Pushing-pulling | 4 | 4 | 0.94 |
| Bodily reaction | | | |
| Kneeling | 5 | 2 | 0.71 |
| Bent-over | 4 | 2 | 0.75 |
| Down&kneeling | 3 | 3 | 0.75 |
| Repetitive motion | | | |
| Typing | 9 | 9 | 0.96 |
| Keyboard | 5 | 4 | 0.76 |
| Carpal-tunnel | 3 | 3 | 0.93 |

performing such filtering is supported in this study by the finding that both Bayesian models assigned their predictions with a degree of confidence (strength) that was strongly related to the actual probability of being correct (fig 1). Figure 2 shows for both Bayesian models how the proportion of correctly coded cases at the one-digit level changes in the prediction set, when cases with a prediction strength below different threshold values are manually reviewed and then correctly coded. Threshold values are indicated in the figure as points on the separate curves shown for the two Bayesian models.

To better understand the relationship shown, if we choose a small threshold, of 0.1 for example, corresponding to a location at the top left corner of fig 2, we see that, for both models, 100% of the narratives would be coded by the algorithm, giving a 0.78–0.80 accuracy level without manual coding of any narratives (relying only on the algorithm's predictions). As the threshold increases, the percentage coded by the algorithm decreases for both models, and for the extreme case where the threshold is 1, 0% would be coded by the algorithm, but the percentage correct would go to 100% as all the cases would be manually coded (assuming that people can code them correctly). At intermediate threshold values, some narratives will have a prediction strength below the threshold. If we assume that only those cases will be manually coded, we can obtain an overall team performance that is better than for the model alone, but without requiring all of the cases to be manually coded. For

What is already known on this topic

- ▶ There is a recognised need to improve the accuracy of manually assigned E-codes.
- ▶ Computerised automated systems have been recognised as a potential solution to improve accuracy and reduce resource requirements necessary for manual classification of narratives in large administrative databases.
- ▶ The simplest form of automated systems use keyword-based search algorithms to assign event codes to cases on the basis of presence or absence of keywords in injury narratives. However, this method has been shown to have low accuracy.

What this study adds

- ▶ This study shows that a computer model based on Bayesian theory with minimal text processing can be implemented. The computer learns on its own from examples of previous work and predicts new classifications with high accuracy.
- ▶ This study also shows high sensitivity and specificity of a Naïve Bayesian model in predicting detailed two-digit Bureau of Labor statistics event classifications of injury narratives.
- ▶ A combined computer/manual coding approach allows the computer to code, in a consistent manner, narratives that use the same words over and over and has a strength parameter assigned to each prediction enabling filtering of more difficult narratives for manual review.

example, if a threshold of 0.70 was used to screen the Fuzzy Bayes predictions (see the 0.7 value on the solid line), manually coding only 20% (ie, 80% are coded by the algorithm) of the 3000 narratives would result in an overall accuracy of 0.88 (fig 2). The Naïve Bayes model does even better. That is, we can obtain the same accuracy of 0.88, using a threshold of 0.60 (now the 0.6 value on the dashed line), by manually classifying only about 14% of the narratives.

Such a reduction in the amount of manually coded cases could be useful when manual coding resources are limited, as this filtering approach would suggest which cases should be chosen to be manually coded. This capability goes well beyond traditional keyword-based systems, which generate their predictions in an all-or-nothing manner, and is a necessary step to developing systems that can filter out records needing manual review. This point is important, because classifying narratives in large administrative datasets becomes more feasible. In addition, the combination of some computer-generated codes, as well as manual codes, will likely provide accuracy above either all-computer or all-manual classifications, as human coders are not able to classify as systematically as the computer.

Another finding is that the Naïve Bayes model generally outperformed the Fuzzy Bayes model, especially for many two-digit codes. This performance advantage may reflect the fact that the word combinations and sequences needed for Fuzzy Bayes to make accurate predictions did not occur often enough in the smaller categories. By considering all the words, Naïve Bayes may have had an advantage, despite the fact that the conditional independence assumption is questionable, and was clearly violated in many cases. Another strength of Naïve Bayes is that it generated its predictions without requiring consideration of the many possible word combinations and sequences in

the injury narratives. The number of possible combinations and sequences of terms in injury narratives is immense. Reducing the items reduces computation time. However, the strengths of each model seemed to compliment each other. Therefore, it may be possible to develop an integrated approach using both the Naïve and Fuzzy classifiers and then, based on the predictive strengths of both, determine if one classification should be used over another or if manual review is required (eg, if both models predict with low strength). These methods will only improve accuracy over what we have demonstrated is possible by using the methods individually and with no filtering and very little preprocessing of the data.

Perhaps the primary advantage of the Fuzzy Bayes model was that the predictors used tended to be highly intuitive, which increased the face validity of the method. This tendency is illustrated in the examples shown in table 2. It also can be easily seen from the examples that some predictors correspond to very specific subcategories and injury causes of potential interest to the analyst (eg, fall off ladder, rear-ended by). As such, the Fuzzy Bayes method is very similar to traditional keyword-based methods, differing primarily in that the predictors are generated automatically instead of by humans by searching for words, word combinations or word sequences that are strongly related to particular categories. Application of the Fuzzy Bayes model can guide efforts to improve the sensitivity of keyword-based classification methods, by suggesting new terms not included in existing systems, or corresponding to specific subcategories of interest to the analyst.

CONCLUSIONS

Through this research we were able to demonstrate that Bayesian models can be used to automatically classify injury narratives from large administrative datasets (such as workers' compensation claims) into one-digit classifications with high accuracy and with fair accuracy at the two-digit level. This finding is substantial given the need for such classifiers in health statistics research. There are many organisations in the USA that are using a large amount of resources to manually classify large datasets. The CDC has determined the need to improve the quality and completeness of external cause of injury coding. Bayesian approaches could be used to help in this effort. We have demonstrated the potential for a semi-automatic/manual approach to classifying injury narratives which would likely improve accuracy beyond what manual coders or the computer alone can achieve, especially with regard to the two-digit classifications.

The accuracy is surprising considering the level of noise expected from workers' compensation claims narratives. We demonstrated that this approach is both feasible and accurate and requires little preprocessing of the original narratives. One of the more encouraging results is that the high levels of performance we observed in this study were obtained with a minimum level of human input. From one perspective, this level of performance implies human inputs are not necessary. Arguably, the need for developing synonym lists, spell-checking, word-stemming, pattern specification, natural language phrase generation, and other commonly applied methods may have been reduced simply because of the large size of the narrative set used in this study. On the other hand, it seems reasonable that these approaches could further improve model performance. It also seems reasonable that accuracy could be improved by increasing the size of the training set, especially for the smaller categories.

A Bayesian approach also provides a means (without using large resources) for re-evaluating narratives with different classification

protocols depending on the purpose of the investigation. Although the Naïve model had higher accuracy than the Fuzzy model, it could be seen that they complimented each other, and a combined approach may be best. Taken together, the overall conclusion that can be drawn from this research is that Bayesian approaches are clearly promising for applications requiring injury narratives to be classified. However, there are a number of practical issues related to implementing the methods that must be addressed before this can become a reality. For example, the integration of these methods with existing administrative databases will require careful consideration of complex issues related to software compatibility, record security and privacy concerns. Other important issues will have to be addressed related to the design of interfaces or modes of interaction to best assist human coders.

A key advantage of the Bayesian methods follows from the fact that the predictions are made on the basis of how often particular terms are found in particular categories. Consequently, the models can be easily updated (ie, learnt) when newly coded narratives are added to the training set by simply recalculating the relative frequencies of terms in particular categories. Such updating might be combined with manual filtering, and occur immediately every time a human confirms a prediction. In other cases, the additions to the training set might be newly coded cases accumulated over some preset interval of time. Deciding on the most appropriate method of updating the Bayesian models and how often this should be done is another important issue that will need to be addressed to implement these methods.

Competing interests: None.

REFERENCES

1. **Williamson A**, Feyer AM, Stout N, *et al*. Use of narrative analysis for comparisons of the causes of fatal accidents in three countries: New Zealand, Australia, and the United States. *Inj Prev* 2001;**7**(Suppl 1):i15–20.
2. **Lombardi DA**, Pannala R, Sorock GS, *et al*. Welding related occupational eye injuries: a narrative analysis. *Inj Prev* 2005;**11**:174–9.
3. **Wellman HM**, Lehto MR, Sorock GS. Computerized coding of injury narrative data from the National Health Interview Survey. *Accid Anal Prev* 2004;**36**:165–71.
4. **Lincoln AE**, Sorock GS, Courtney TK, *et al*. Using narrative text and coded data to develop hazard scenarios for occupational injury interventions. *Inj Prev* 2004;**10**:249–54.
5. **Bureau of Labor Statistics**. *Occupational injury and illness classification manual*. Washington, DC: US Department of Labor, December 1992.
6. **Hunt PR**, Hackman H, Berenholz G, *et al*. Completeness and accuracy of International Classification of Disease (ICD) external cause of injury codes in emergency department electronic data. *Inj Prev* 2007;**13**:422–5.
7. **Annest JL**, Fingerhut LA, Gallagher SS, *et al*. Centers for Disease Control and Prevention (CDC). Strategies to improve external cause-of-injury coding in state-based hospital discharge and emergency department data systems: recommendations of the CDC Workgroup for Improvement of External Cause-of-Injury Coding. *MMWR Recomm Rep* 2008;**57**(RR-1):1–15.
8. **Noorinaeini A**, Lehto MR. Hybrid singular value decomposition: a model of text classification. *International Journal of Human Factors Modeling and Simulation* 2006;**1**:95–118.
9. **Sebastiani F**. Machine learning in automated text categorization. *ACM Computing Surveys (CSUR)* 2002;**34**:1–47.
10. **Lewis DD**. Naive Bayes at forty: the independence assumption in information retrieval. In: *Proceedings of the 15th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1992:37–50.
11. **Corns HL**, Marucci HR, Lehto MR. Development of an approach for optimizing the accuracy of classifying claims narratives using a machine learning tool (TEXTMINER). In: *Proceedings of HCI International 2007, 12th International Conference on Human-Computer Interaction(8)*. 2007:411–16.
12. **Marucci HR**, Lehto MR, Corns HL. Computer classification of injury narratives using a Fuzzy Bayes approach: improving the model. In: *Proceedings of HCI International 2007, 12th International Conference on Human-Computer Interaction(8)*. 2007:500–6.
13. **Lehto MR**, Sorock G. Machine learning of motor vehicle accident categories from narrative data. *Methods Inf Med* 1996;**35**:1–8.
14. **Sorock G**, Ranney T, Lehto M. Motor vehicle crashes in roadway construction work zones: an analysis using narrative text from insurance claims. *Accid Anal Prev* 1996;**28**:131–8.