A combined Fuzzy and Naïve Bayesian strategy can be used to assign event codes to injury narratives

H Marucci-Wellman,[1] M Lehto,[2] H Corns[1]

[1]Center for Injury Epidemiology, Liberty Mutual Research Institute for Safety, 71 Frankland Road, Hopkinton, Massachusetts, USA
[2]School of Industrial Engineering, Purdue University, 1287 Grissom Hall, West Lafayette, Indiana, USA/ School of Management/Center for Global Innovation & Entrepreneurship Kyunghee University Seoul 130-701, Korea

Correspondence to
Dr Helen Marucci-Wellman, Center for Injury Epidemiology, Liberty Mutual Research Institute for Safety, 71 Frankland Road, Hopkinton, MA 01748, USA; helen.wellman@libertymutual.com

Accepted 11 March 2011
Published Online First
11 April 2011

ABSTRACT

Background Bayesian methods show promise for classifying injury narratives from large administrative datasets into cause groups. This study examined a combined approach where two Bayesian models (Fuzzy and Naïve) were used to either classify a narrative or select it for manual review.

Methods Injury narratives were extracted from claims filed with a worker's compensation insurance provider between January 2002 and December 2004. Narratives were separated into a training set (n=11,000) and prediction set (n=3,000). Expert coders assigned two-digit Bureau of Labor Statistics Occupational Injury and Illness Classification event codes to each narrative. Fuzzy and Naïve Bayesian models were developed using manually classified cases in the training set. Two semi-automatic machine coding strategies were evaluated. The first strategy assigned cases for manual review if the Fuzzy and Naïve models disagreed on the classification. The second strategy selected additional cases for manual review from the Agree dataset using prediction strength to reach a level of 50% computer coding and 50% manual coding.

Results When agreement alone was used as the filtering strategy, the majority were coded by the computer (n=1,928, 64%) leaving 36% for manual review. The overall combined (human plus computer) sensitivity was 0.90 and positive predictive value (PPV) was >0.90 for 11 of 18 2-digit event categories. Implementing the 2nd strategy improved results with an overall sensitivity of 0.95 and PPV >0.90 for 17 of 18 categories.

Conclusions A combined Naïve-Fuzzy Bayesian approach can classify some narratives with high accuracy and identify others most beneficial for manual review, reducing the burden on human coders.

Injury narratives contain useful information for the prevention of injuries.[1–3] Using the computer to help classify narratives has the potential for reducing the burden implicit in manually reviewing and classifying large numbers of narratives from administrative injury databases. The accuracy and completeness of manual coding is another important issue.[4 5] The Centers for Disease Control and Prevention have recently been actively promoting strategies for improving the quality and completeness of external cause of injury coding in the USA, including the use of automated systems that assist coders in assigning event codes.[5]

Machine learning algorithms offer a way potentially to learn how to classify injury text systematically from the typically massive amounts of previously manually coded narratives in administrative databases.[6–9] We recently showed that two different Bayesian models (naive vs fuzzy) classified injury narratives from a large administrative dataset into broad (one-digit) Bureau of Labor Statistic (BLS) Occupational Injury and Illness Classification System (OIICS) event categories (eg, fall, struck by, transportation, overexertion) with high sensitivity (0.78–0.80). Each model also did fairly well (sensitivity 0.64–0.70) at a more detailed (two-digit) level.[10] The naive Bayesian model performed slightly better than the fuzzy Bayesian model. However, the predictions of the fuzzy Bayesian model were more intuitive because they were based on the single word or word combination most strongly related to the assigned category.

An important result of our previous study was that the prediction strengths assigned by both models were strongly related to the actual probability that the prediction was correct. This suggested that prediction strength could be used effectively to filter out narratives for manual review in a semicomputerised approach in which part of the narratives are computer coded. The objective of this follow-on study was to develop and test methodologies for combining the fuzzy and naive Bayesian approaches along with selected manual coding. We hypothesised that a combined naive–fuzzy semicomputerised approach could both improve computer classification accuracy, and guide strategic assignment of narratives for manual review to optimise the accuracy of the final combined (human plus computer) coded dataset while minimising the number manually coded.

METHODS

Data collection

Over 17 000 records were randomly extracted from claims filed between January 2002 and December 2004 with a workers' compensation insurance provider.[9] The OIICS scheme includes approximately 40 mutually exclusive event categories (September 2007 version). The two coders who classified these narratives went through an intensive training process and had over 7 years of coding experience. After eliminating the cases on which the two coders disagreed, the remaining 14 000 records were considered 'gold standard' classifications. The data were then divided into a training set of 11 000 cases, which was used for model development, and a prediction dataset of 3000 cases that was used for evaluation of the model. 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 gold standard classifications for the training and prediction datasets were similar (p=0.25).

Model development

Two Bayesian models, referred to as naive Bayes and fuzzy Bayes, were developed to generate two independent sets of predictions.[i]

The naive Bayes model calculates the probability of a particular event code category 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. Essentially, the naive algorithm calculates the probability of an event category by multiplying the likelihood ratios for each word in the narrative and prior probabilities.

The fuzzy Bayes model calculates the probability of a particular event code using the expression:

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

The primary difference from naive 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.

These two different models were developed and evaluated using the Textminer program developed by one of the authors (ML). Both models used the statistical relationship between terms in the 11 000 injury narratives in the training set and the manually assigned two-digit BLS OIICS event codes to estimate, in the predictive dataset of 3000, the probability a human coder would assign a particular code to a new narrative, given the words that were present in the narrative.

Model evaluation

Two semicomputerised strategies were evaluated. Our earlier results suggested that predictions would be more likely to be correct when the fuzzy and naive algorithms predicted the same classification, because both models showed good performance on their own.[10] Therefore, our first strategy was to accept the computer-assigned codes if the fuzzy and naive algorithms agreed (*agree* dataset) and manually review the remaining narratives in which fuzzy and naive disagreed (*disagree* dataset). Our earlier results also suggested that narratives could be effectively filtered out for manual review using prediction strength. Therefore, our second strategy (which would be desirable if a higher positive predictive value (PPV) were desired and additional coding resources were available) was to filter the *agree* dataset further using the prediction strengths assigned by the naive Bayes model.

To test the first hypothesis, working with the predictive dataset of 3000 cases, we separately analysed prediction accuracy for the cases in which the models agreed (*agree* dataset) or disagreed (*disagree* dataset), in terms of sensitivity and PPV. Sensitivity (true positives) was the percentage of gold standard (human-coded) narratives in each category also coded by the algorithm and PPV was the percentage of narratives correctly coded into a specific category out of all narratives coded by the algorithm into that category. We did not evaluate specificity and negative predictive value because they were high (nearing 1.0) with little differentiation across categories (see earlier results).[10]

To test the second hypothesis, we took the *agree* dataset, and filtered out enough of the weakest predictions (eg, low naive prediction strength) to result in half of the 3000 prediction narratives being manually coded, because we had enough human resources to classify half of the narratives manually. We then evaluated the prediction accuracy for the refined subset of computer-coded narratives in terms of sensitivity and PPV.

A follow-on set of analyses were also conducted to evaluate prediction accuracy for this combined computer—human-coded dataset (all 3000 prediction narratives) resulting from strategies 1 and 2. We measured the sensitivity and PPV, assuming that the cases filtered out for each strategy to be coded by a human were coded correctly (eg, the *disagree* dataset for strategy 1 and additional cases from the *agree* dataset for strategy 2).

To illustrate the trade-off between the sensitivity of the computer-coded cases and the number of cases that would need to be manually classified, we repeated this process using different prediction thresholds and generated a plot showing how prediction sensitivity of the computer codes improved as the number being strategically filtered out for manual coding increased.

RESULTS

We first briefly compare model performance for cases in which the fuzzy and naive algorithms agreed versus disagreed. We then present results showing the influence of additional selections for manual review of the weakly predicted subset of cases based on the prediction strength of the cases in which the fuzzy and naive algorithms agreed and, finally, the effects of adding in a manual review of the selected out weakly predicted subset of cases for each case selection scenario. Figure 1 is a flow diagram of the strategic filtering process for assignment of narratives for computer or manual coding.

Classifications in which fuzzy and naive agreed

The fuzzy and naive algorithms agreed on 1928 (64%) of the 3000 classifications in the prediction set (table 1). The overall sensitivity for these coded narratives was 0.85, substantially higher than the sensitivity for the entire dataset using the naive algorithm (0.70). The lowest sensitivities were recorded for the struck against (0.42), bodily reaction (0.63) and explosion categories (0.60), but these categories had high PPV (0.88, 0.90 and 1.0, respectively). Very high sensitivities, but lower PPV were found for the overexertion (0.99 sensitivity, 0.83 PPV) and highway categories (1.0 sensitivity, 0.94 PPV). The two categories with the lowest PPV were non-highway (0.67) and exposure to stress (0.76).

When naive prediction strength was used to obtain the 1500 most strongly predicted computer-coded narratives (or half of the dataset), 428 narratives from the *agree* dataset with naive prediction strength values less than 0.89 were screened out for manual coding in addition to the *disagree* dataset. Dropping these cases from the *agree* dataset (most strongly predicted computer-coded narratives) improved the accuracy of the computer predictions (table 1). The PPV improved to above 0.80 across most categories, except for those with less than 10 observations and the non-highway accident category, which improved from 0.67 to 0.75 (table 1).

[i] We refer the reader to our earlier article for a comprehensive and detailed explanation of the Bayesian models used in the study.[10] The same training and prediction datasets, algorithms, gold standard classifications, machine-assigned codes and prediction strengths used in that study were used for this research.[10]

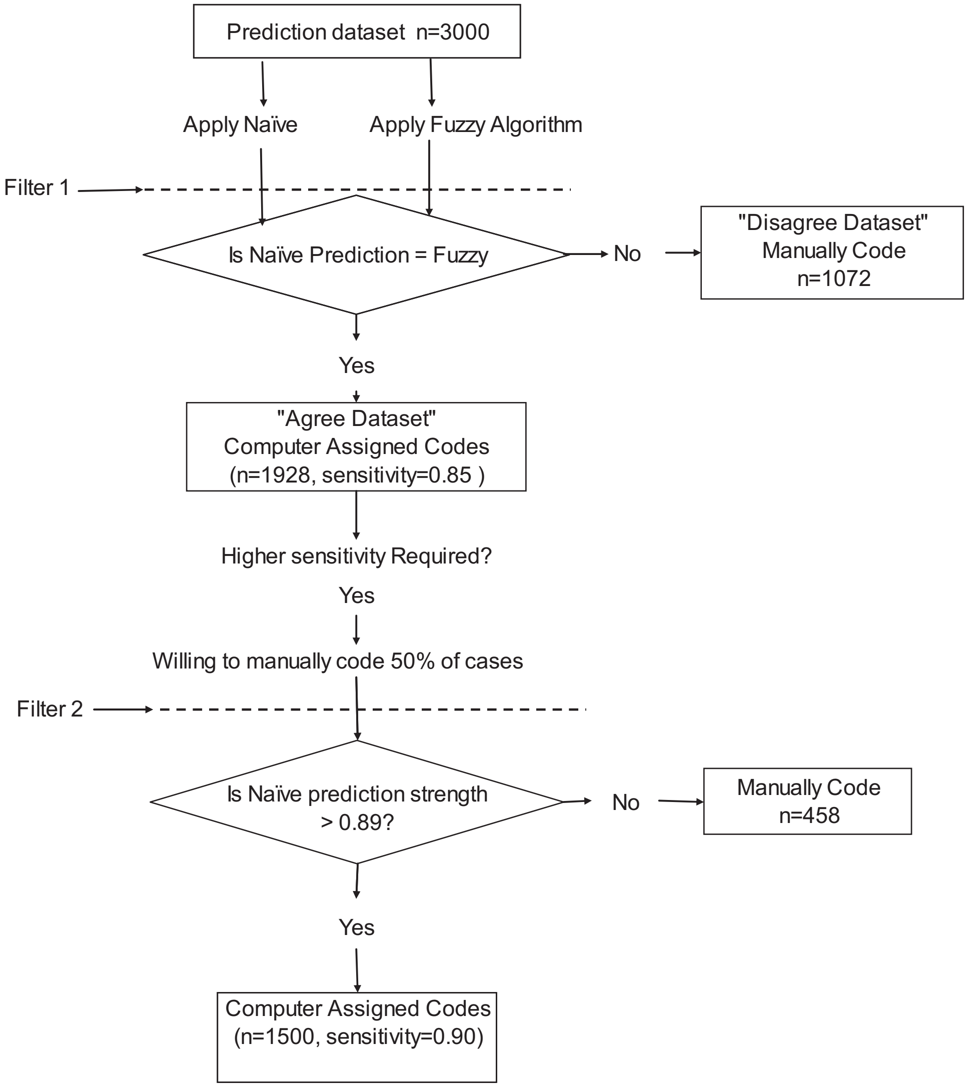

Prediction dataset n=3000
Apply Naïve
Apply Fuzzy Algorithm
Filter 1
Is Naïve Prediction = Fuzzy
No
"Disagree Dataset" Manually Code n=1072
Yes
"Agree Dataset" Computer Assigned Codes (n=1928, sensitivity=0.85)
Higher sensitivity Required?
Yes
Willing to manually code 50% of cases
Filter 2
Is Naïve prediction strength > 0.89?
No
Manually Code n=458
Yes
Computer Assigned Codes (n=1500, sensitivity=0.90)

Figure 1 Flow diagram of the strategic filtering process for assignment of narratives for computer or manual coding.

Classifications in which fuzzy and naive disagreed

A basic principal of probability theory indicates if two independent predictions (fuzzy and naive) are the same, you should be more confident that the prediction is correct. This can be seen in this example narrative: 'STRK BY NAIL HIT IN THE EYE WITH NAIL FROM NAIL GUN, LOSS OF SIGHT IN LEFT EYE'. For this narrative the fuzzy Bayesian algorithm multiple word model maximises the 'struck by' category given the sequence words 'in-eye' ($P(E_i|n)$=0.79). The naive Bayesian algorithm also found the highest strength category to be 'struck by' using all the words in the narrative ($P(E_i|n)$=0.99). In this example, all of the words in the narrative, along with the maximum evidence words in the narrative, support the same classification, the 'struck by' category. (Note: $P(E_i|n)$ is the probability of event category E_i given the set of n words in the narrative).

On the other hand, when the fuzzy and naive algorithms disagreed, there was ambiguity in the narrative. An example follows:

> 'LIFTING A PIECE OF DECK TO THROW AWAY, SLIPPED ON ICE & TWISTED BACK'

For this narrative the sequence 'lifting-piece' classified the narrative into the 'overexertion' category using the fuzzy algorithm ($P(E_i|n)$=0.75). The naive algorithm, which uses all the words in the narrative, classified the narrative into the 'bodily reaction' category ($P(E_i|n)$=0.69). There were several words that were strong predictors of the bodily reaction category 'slipped on ice' and 'twisted back', while the multiple word model maximised the 'overexertion' category given the words 'lifting-piece'.

Analysis of the 1072 (36%) cases in which the fuzzy and naive algorithms disagreed (table 2) showed, with a few exceptions, that the PPV was low for both algorithms for nearly every category, as was the overall sensitivity (fuzzy 0.29, naive 0.40, table 2). A quick comparison with table 1 reveals that the overall performance for the *agree* cases is dramatically better. This is illustrated for the 'struck by' category using a Venn diagram (figure 2). It can be seen there that few cases were predicted correctly when naive and fuzzy disagreed.

Semi-computerised coding strategies

In many real-world applications, for research, legal, administrative or other reasons, it is necessary to code all the cases. If we include the effects of manual review of the filtered (most weakly predicted) subset of cases, we obtain a combined (or team) level of performance. The overall combined performance when cases are strategically filtered for manual review is shown in table 3 for two semicomputerised coding strategies. Using the first strategy, in which 1928 cases in the *agree* dataset are classified by the

Table 1 Evaluation of strategic selection of computer-assigned codes in which fuzzy and naive agree on classification alone and when the naive prediction strength is greater than 0.89

		Computer-assigned codes in which naive and fuzzy algorithms agree on classification (n=1928)					Computer-assigned codes in which fuzzy and naive agree on classification and the naive prediction strength is >0.89 (n=1500)				
BLS category	Description	Gold standard (n)	Predicted by naive and fuzzy into category (n)	Correct (n)	Sensitivity	PPV	Gold standard (n)	Predicted by naive and fuzzy into category (n)	Correct (n)	Sensitivity	PPV
Contact: Group 0											
01	Struck against	52	25	22	0.42	0.88	27	13	12	0.44	0.92
02	Struck by	127	140	108	0.85	0.77	85	88	75	0.88	0.85
03	Caught/compressed	42	43	34	0.81	0.79	34	35	31	0.91	0.89
Fall: Group 1											
11	Fall to lower level	113	119	94	0.83	0.79	83	86	74	0.89	0.86
13	Fall on same level	217	244	193	0.89	0.79	158	180	151	0.96	0.84
Bodily motion: Group 2											
21	Bodily reaction	212	147	133	0.63	0.90	120	83	80	0.67	0.96
22	Overexertion	452	536	446	0.99	0.83	389	427	387	0.99	0.91
23	Repetitive motion	59	65	57	0.97	0.88	55	59	54	0.98	0.92
Exposure to harmful substances or environment: Group 3											
31	Contact with electric current	14	15	13	0.93	0.87	12	11	11	0.92	1.00
32	Contact with temperature extremes	77	82	74	0.96	0.90	69	75	68	0.99	0.91
34	Exposure to caustic substance	71	65	60	0.85	0.92	57	57	53	0.93	0.93
35	Exposure to noise	32	34	32	1.00	0.94	32	33	32	1.00	0.97
37	Exposure to stress	28	33	25	0.89	0.76	25	26	22	0.88	0.85
Transportation: Group 4											
41	Highway accident	204	217	203	1.00	0.94	194	203	194	1.00	0.96
42	Non-highway accident	17	18	12	0.71	0.67	16	16	12	0.75	0.75
43	Pedestrian struck by vehicle	46	41	39	0.85	0.95	36	33	32	0.89	0.97
Fire or explosion: Group 5											
52	Explosion	5	3	3	0.60	1.00	3	1	1	0.33	1.00
Assaults and violent acts: Group 6											
61	Assaults	52	57	47	0.90	0.82	43	47	41	0.95	0.87
Non-classifiable											
9999	Non-classifiable	77	40	38	0.49	0.95	43	24	23	0.53	0.96
Other											
	General unclassifiable*	19	0	0	0.00	0.00	10	0	0	0.00	0.00
	All categories n<10†	12	4	3	0.25	0.75	9	3	3	0.33	1.00
Overall		1928	1928	1636	0.85	–	1500	1500	1356	0.90	–

*Unspecified and non-classifiable within category, ie, 10, contact unspecified.

†All categories with less than 10 cases including: rubbed or abraded, jump to lower level, bodily conditions, exposure to air pressure, assaults by animals.

BLS, Bureau of Labor Statistic; PPV, positive predictive value.

computer and the remaining 1072 cases in which the naive and fuzzy classifications disagree are manually reviewed results in high overall sensitivity (0.90), and PPV above 0.85 for most categories. This combined performance is quite good, and requires only 36% of the cases to be manually coded.

Using the second strategy, in which another 428 narratives are filtered from the *agree* dataset when naive prediction strength is less than 0.89 and added to the *disagree* cases, increases the proportion of manually coded cases to 50% and further improves the results. This strategy results in an observed sensitivity above 0.95 for the overall dataset (table 3), and PPV above 0.92 for nearly all categories (table 3).

In actual practice, an organisation with limited coding resources that could ideally be deployed can reach a targeted level of accuracy by varying the number of manually reviewed cases. Figure 3 shows how targeted levels of sensitivity on the x-axis can be reached by manually reviewing different proportions of the cases. Each point on the curve shows the resulting sensitivity and proportion of computer-assigned cases for particular prediction strength thresholds used to filter the *agree* and *disagree* datasets. The two strategies discussed earlier correspond to two different points on the curve. Strategy 1 corresponds to using a prediction strength threshold of 1 for the *disagree* set and a threshold of 0 for the *agree* dataset as indicated by the point (1, 0) on the curve. Strategy 2 corresponds to the point (1, 0.89) and results in more manual coding, but a higher overall sensitivity.

Figure 2, therefore, shows the trade-off between accuracy and the number of manually reviewed cases. As stated earlier, all computer classifications of the *agree* dataset were predicted at a sensitivity of 0.85. To improve on that we filtered out more narratives for manual review from the *agree* dataset.

DISCUSSION

The findings from this study suggest that, for classification of large administrative database injury narratives into discrete categories, a human–machine integrated approach may be preferable to either the human or machine approach alone. A strategy of reducing human resources by half using Bayesian models and strategic but simple assignment of targeted narratives revealed a sensitivity of at least 0.90 for the final computer–manual assigned codes. While human coding will probably always be necessary for the classification of complex, ambiguous situations or for emerging issues, the ability of the

Table 2 Evaluation of computer-assigned codes in which naive and fuzzy algorithms disagree on classification (n=1072)

			Naive algorithm				Fuzzy algorithm			
BLS category	Description	Gold standard (n)	Predicted by naive (n)	Correct (n)	Sensitivity	PPV	Predicted by fuzzy (n)	Correct (n)	Sensitivity	PPV
Contact: Group 0										
01	Struck against	94	81	36	0.38	0.44	29	14	0.15	0.48
02	Struck by	166	196	82	0.49	0.42	106	37	0.22	0.35
03	Caught/compressed	32	40	15	0.47	0.38	15	3	0.09	0.20
Fall: Group 1										
11	Fall to lower level	74	99	30	0.41	0.30	65	23	0.31	0.35
13	Fall on same level	104	167	45	0.43	0.27	107	32	0.31	0.30
Bodily motion: Group 2										
21	Bodily reaction	158	127	66	0.42	0.52	135	53	0.34	0.39
22	Overexertion	83	51	14	0.17	0.27	314	53	0.64	0.17
23	Repetitive motion	17	29	12	0.71	0.41	15	2	0.12	0.13
Exposure to harmful substances or environment: Group 3										
31	Contact with electric current	14	10	7	0.50	0.70	10	6	0.43	0.60
32	Contact with temperature extremes	15	18	5	0.33	0.28	32	10	0.67	0.31
34	Exposure to caustic substance	39	19	12	0.31	0.63	33	17	0.44	0.52
35	Exposure to noise	5	4	0	0.00	0.00	5	4	0.80	0.80
37	Exposure to stress	5	4	2	0.40	0.50	11	3	0.60	0.27
Transportation: Group 4										
41	Highway accident	16	13	3	0.19	0.23	97	11	0.69	0.11
42	Non-highway accident	39	72	27	0.69	0.38	13	2	0.05	0.15
43	Pedestrian struck by vehicle	58	57	36	0.62	0.63	28	7	0.12	0.25
Fire or explosion: Group 5										
52	Explosion	5	4	0	0.00	0.00	3	3	0.60	1.00
Assaults and violent acts: Group 6										
61	Assaults	30	25	11	0.37	0.44	22	9	0.30	0.41
Non-classifiable										
9999	Non-classifiable	74	43	26	0.35	0.60	25	15	0.20	0.60
Other										
	General unclassifiable*	22	10	1	0.05	0.10	1	0	0.00	0.00
	All categories n<10†	22	3	2	0.09	0.67	6	6	0.27	1.00
Overall		1072	1072	432	0.40	—	1072		0.29	—

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

†All categories with less than 10 cases including: rubbed or abraded, jump to lower level, bodily conditions, exposure to air pressure, assaults by animals.

BLS, Bureau of Labor Statistic; PPV, positive predictive value.

computer to target those types of narratives for human investigation is what makes the human–machine integration a highly valuable tool.

While both the naive and fuzzy models alone performed fairly well at the two-digit level in our earlier analyses,[10] the high performance achieved when the fuzzy and naive algorithms agreed on a classification and the quantity of narratives that could be classified correctly using this filter is noteworthy. By following this approach, the computer was able to classify 64% of the narratives with an overall sensitivity of 0.85 of the

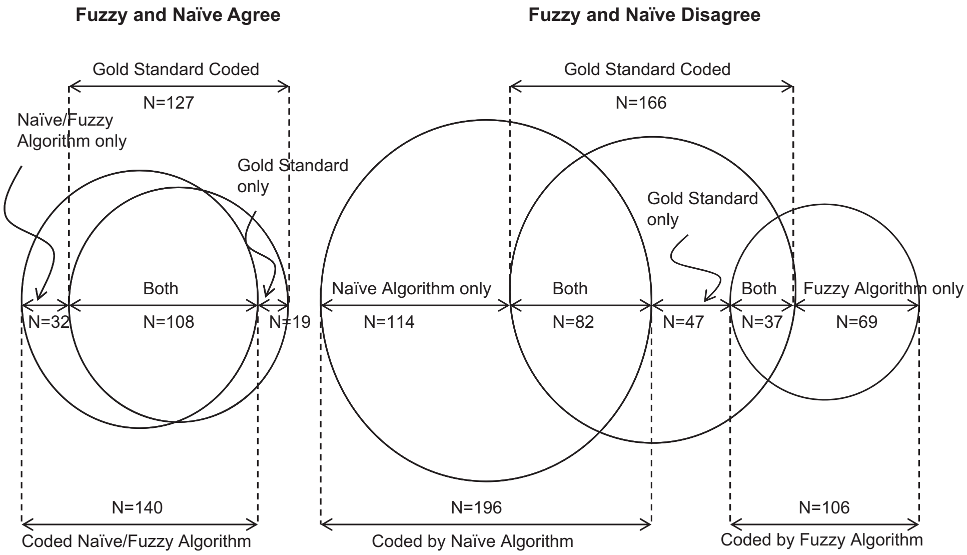


Figure 2 Comparison of gold standard records versus those coded by the computer algorithm (fuzzy and naive algorithms) for the 'struck by' category (Bureau of Labor Statistic event code group 02). This comparison is shown separately for the subset of cases in which fuzzy and naive agreed on the classification and in which fuzzy and naive disagreed on the classification.

Table 3 Performance of two semicomputerised coding strategies

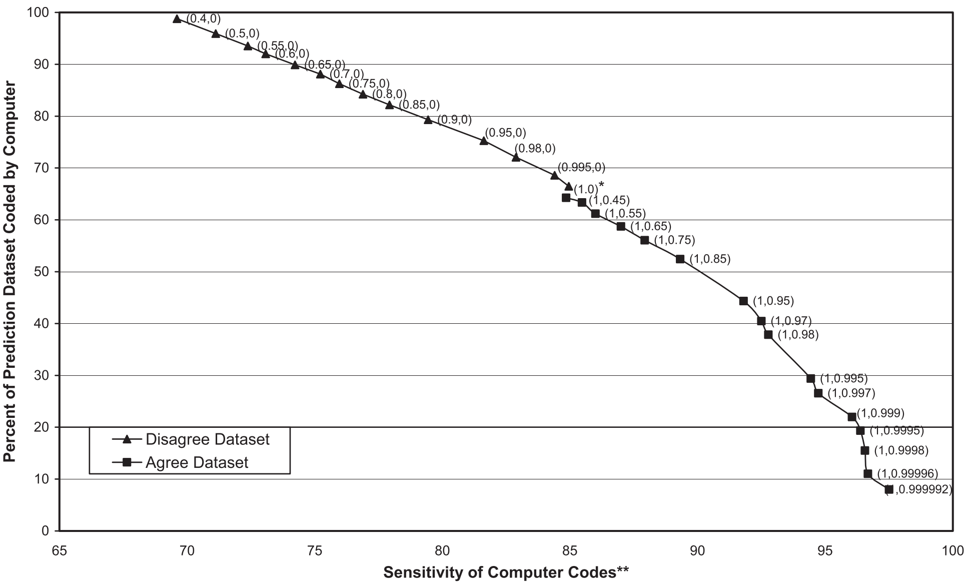
				Filter 1:* Fuzzy and naive agree (n=1928), manually review (n=1072)				Filter 2:† Computer classify (n=1500), manually review (n=1500)			
BLS category	**Description**	**Gold standard (n)**	**%**	**Predicted into category (n)**	**Correct (n)**	**Sensitivity‡**	**PPV***	**Predicted into category (n)**	**Correct (n)**	**Sensitivity‡**	**PPV***
Contact: Group 0											
01	Struck against	146	5	119	116	0.79	0.97	132	131	0.90	0.99
02	Struck by	293	10	306	274	0.94	0.90	296	283	0.97	0.96
03	Caught/compressed	74	2	75	66	0.89	0.88	75	71	0.96	0.95
Fall: Group 1											
11	Fall to lower level	187	6	193	168	0.90	0.87	190	178	0.95	0.94
13	Fall on same level	321	11	348	297	0.93	0.85	343	314	0.98	0.92
Bodily motion: Group 2											
21	Bodily reaction	370	12	305	291	0.79	0.95	333	330	0.89	0.99
22	Overexertion	535	18	619	529	0.99	0.85	573	533	1.00	0.93
23	Repetitive motion	76	3	82	74	0.97	0.90	80	75	0.99	0.94
Exposure to harmful substances or environment: Group 3											
31	Contact with electric current	28	1	29	27	0.96	0.93	27	27	0.96	1.00
32	Contact with temperature extremes	92	3	97	89	0.97	0.92	98	91	0.99	0.93
34	Exposure to caustic substance	110	4	104	99	0.90	0.95	110	106	0.96	0.96
35	Exposure to noise	37	1	39	37	1.00	0.95	38	37	1.00	0.97
37	Exposure to stress	33	1	38	30	0.91	0.79	34	30	0.91	0.88
Transportation: Group 4											
41	Highway accident	220	7	233	219	1.00	0.94	229	220	1.00	0.96
42	Non-highway accident	56	2	57	51	0.91	0.89	56	52	0.93	0.93
43	Pedestrian struck by vehicle	104	3	99	97	0.93	0.98	101	100	0.96	0.99
Fire or explosion: Group 5											
52	Explosion	10	0	8	8	0.80	1.00	8	8	0.80	1.00
Assaults and violent acts: Group 6											
61	Assaults	82	3	87	77	0.94	0.89	86	80	0.98	0.93
Non-classifiable											
9999	Non-classifiable	151	5	114	112	0.74	0.98	132	131	0.87	0.99
Other											
	General non-classifiable§	41	1	22	22	0.54	1.00	31	31	0.76	1.00
	All categories n<10¶	34	1	26	25	0.74	0.96	28	28	0.82	1.00
Overall			10								
		3000	0	3000	2708	0.90	—	3000	2856	0.95	—

*Filter 1: Manually review cases in which fuzzy and naive disagree.
†Filter 2: Manually review cases in which fuzzy and naive disagree and when naive strength is less than 0.89, naive prediction and accuracy used for computer-classified cases.
‡Manually classified cases assumed correctly classified.
§Unspecified and non-classifiable within category, ie, 10, contact unspecified.
¶All categories with less than 10 cases including rubbed or abraded, jump to lower level, bodily conditions, exposure to air pressure, assaults by animals.
BLS, Bureau of Labor Statistic; PPV, positive predictive value.

computer-assigned codes, confirming that using fuzzy and naive agreement as a filter strategy could be highly effective. In addition, over 50% of the original (struck by, bodily reaction, non-highway, pedestrian, explosion and other classifiable) cases were targeted for manual review resulting in a large increase in capture (sensitivity increased ≥0.79 and PPV increased >0.85) in categories with originally low sensitivities in the earlier analyses (eg. struck against, bodily reaction).[10] In situations in which the computer classification is made with high confidence, it is likely that the computer may assign a classification more systematically (with repeated evidence that words in the narrative are associated with a particular category) than a human coder. Also similarly, the classifications made by the computer may be used to identify coding errors of human classifications.

While both algorithms use the same evidence from the training dataset, each may come up with a different classification. Disagreement between the fuzzy and naive algorithms may indicate that these narratives were more ambiguous than usual. This follows because one algorithm (fuzzy) focuses only on the strongest single piece of evidence while the other (naive) combines all the evidence, so both algorithms will agree when the evidence is consistent or unambiguous for a particular category.

Applying a second filter using the prediction strength with the aim of 50% manual coding and 50% computer coding was very simple and resulted in an improvement in the final coded dataset. The overall accuracy at the two-digit level was 0.95 and the PPV ranged from 0.88 for the 'exposure to stress' category to 1.0 for the 'contact with electric current' category. Even more critical was that the greatest improvement in sensitivity and PPV occurred in the categories that had lower sensitivities and PPV after the first filter. This demonstrates the power of using the prediction strengths as a filter and that, from a combined fuzzy—naive Bayesian approach, one not only gets a prediction, but there is additional information about confidence in that prediction, enough to target certain narratives strategically for manual review. This method showed that with half the resources (only half coded by humans) we could obtain a distribution of the final coded dataset (human and computer codes strategically assigned) that was highly representative of the gold standard distribution and the PPV of the classifications were also very high.



* A (1,0) threshold is defined as using all computer coded Agree cases and no computer coded Disagree cases
** Manually classified cases assumed correctly classified, Naïve prediction and accuracy used for computer classified cases

Note: Parenthesis (x,y) indicate prediction strength threshold level of filter, x=threshold for disagree dataset, y=threshold for agree dataset

Figure 3 Trade-off between sensitivity and proportion coded by computer using a semi-autonomous classification strategy.

It should also be mentioned that more sophisticated filtering strategies could yield further improvements. For example, a plot of PPV compared with sensitivity achieved at different threshold levels for each category could be used to pick out optimum threshold levels for each category. Other information in both the fuzzy and naive algorithms could also be used to aid in filtering. In particular, the naive prediction strength could be adjusted simultaneously considering other information from the algorithms (such as the fuzzy strength, the difference between the highest and the second highest fuzzy strengths, if the naive classification agreed with the second fuzzy classification (yes/no), etc).

An important next step is to determine how the algorithms and filtering strategies tested here will perform on other datasets with different underlying causes of injury distributions, different coding protocols, or longer or shorter narratives. Future studies might also focus on issues such as the feasibility of predicting more detailed codes, or improving performance on the few categories observed in this study to be difficult to predict. In particular, the naive and fuzzy Bayes algorithms both tended to under-predict cases manually assigned to the 'unspecified categories'. This result can partly be explained by both the small number of training narratives in the dataset for the unspecified categories and the lack of unique predictors indicating that particular narratives are not classifiable. Also, with difficult narratives, human coders may be more likely to assign the narrative into an 'unspecified' category even though the computer may be able to identify detail in the narrative allowing for a specific code to be assigned. This suggests that computer coding might also offer a strategy for reducing undercounts in applications in which human coders tend to over-assign cases into the 'unspecified' category.[11]

CONCLUSION

This study indicates that utilising an integrated computer—human approach, with strategic assignment of manual

What is already known on the subject

- The Centers for Disease Control and Prevention have recently been actively promoting strategies for improving the quality and completeness of external cause of injury coding in the USA, including the use of automated systems that assist coders in assigning event codes.
- Computerised automated systems have been recognised as a potential solution to improve accuracy and reduce resource requirements needed for the manual classification of narratives in large administrative databases.
- Two different Bayesian models (naive vs fuzzy) have been shown to be able to classify injury narratives from large administrative datasets into broad (one-digit) classifications with high accuracy and with fair accuracy at a more detailed level.

What this study adds

- A semi-automatic approach to assigning classifications based on a combined naive—fuzzy Bayesian approach to strategic filtering to identify narratives most beneficial for manual review can be implemented with minimal text processing and result in high accuracy for assigning two-digit event code classifications.
- When the fuzzy and naive Bayesian models agreed on a classification, the sensitivity was very high (0.85) for classifying injury narratives into two-digit categories.
- It follows that agreement can be used as an easily implementable and effective filtering strategy for a combined partial manual, partial computer approach to classifying injury narratives in large administrative databases. When using agreement alone as a filtering strategy, manual coding was reduced to only one-third of the dataset and resulted in a sensitivity of 0.90 for the final coded dataset.
- Other possible applications of the combined fuzzy—naive approach can include the identification of coding errors.

narratives, results in a very high final accuracy at the two-digit classification level. While some specific categories have lower predicted strengths and are more difficult to code, these are likely to be filtered out for manual review, thus maintaining a high accuracy even in these categories. In order to maintain high accuracy requirements or to identify emerging issues, manual coding may never be totally replaced. However, strategically assigning manual coders to the more ambiguous narratives, while allowing the computer to classify common incident narratives, allows for a more efficient and cost-effective utilisation of resources. The strategies utilised reduced manual coding by half and improved accuracy beyond what would be expected with manual coding alone. Most importantly, there is reason to believe further improvements in performance are quite achievable, due to both model refinements and learning that occurs when new manually coded narratives are fed back into the system to fine tune the predictive models in real-world settings.

Acknowledgements The authors would like to thank Ms. Barbara Webster, Dr Manuel Cifuentes and Mr. Theodore Courtney for their thoughtful reviews and recommendations and Ms Peg Rothwell for editorial input on the final manuscript.

Competing interests None.

Ethics approval This study was conducted with the approval of the Liberty Mutual Research Institute for Safety and Purdue University.

Contributors HM-W: Co-principal investigator, lead author. ML: Co-principal investigator, contributed to development of methods used, study design, data interpretation and writing. HC: Data analyses and interpretation.

Provenance and peer review Not commissioned; externally peer reviewed.

REFERENCES

1. **Sorock G,** Smith G, Reeve G, *et al*. Three perspectives on work-related injury surveillance systems. *Am J Ind Med* 1997;**32**:116—28.
2. **Smith GS.** Public health approaches to occupational injury prevention: do they work? *Inj Prev* 2001;**7**(Suppl I):i3—10.
3. **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.
4. **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.
5. **Annest JL,** Fingerhut LA, Gallagher SS, *et al*. *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 March 28, 2008/57(RR01);1—15)*. Atlanta, GA: Centers for Disease Control and Prevention (CDC). http://www.cdc.gov/mmwr/preview/mmwrhtml/rr5701a1.htm (accessed 15 Jun 2010).
6. **Noorinaeini A,** Lehto MR. Hybrid singular value decomposition: a model of text classification. *International Journal of Human Factors Modelling and Simulation (IJHFMS)* 2006;**1**:95—118.
7. **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.
8. **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.
9. **McKenzie K,** Scott DA, Campbell MA, *et al*. The use of narrative text for injury surveillance research: a systematic review. *Accid Anal Prev* 2010;**42**:354—63.
10. **Lehto M,** Marucci-Wellman H, Corns H. Bayesian methods: a useful tool for classifying injury narratives into cause groups. *Inj Prev* 2009;**15**:259—65.
11. **Kaida AK,** Marko J, Hagel B, *et al*. Unspecified falls among youth: predictors of coding specificity in the emergency department 2006 *Inj Prev* 2006;**12**:302—7.

Amish men jailed for not displaying buggy safety signs

Eight members of a traditional Amish sect were imprisoned in Kentucky after refusing to pay fines for failure to display orange—red safety triangles on their horse-drawn buggies. They refused on religious grounds. The Amish contend that paying the fines would amount to complying with a law that violates their religious restrictions against wearing or displaying bright colours or relying upon man-made symbols for their safety. All the defendants are members of a traditional Amish group; other Amish comply with the requirements.

Preventing spinal cord injuries in children

There are 12 000 new cases of spinal cord injury in the USA each year; approximately 10% affect children under the age of 16 years. Shriners Hospitals for Children, a leader in paediatric spinal cord injury treatment and rehabilitation, provides important educational information for families through their 'Prevention Begins with Awareness' campaign. Some of the most common causes of spinal cord injuries in children and adolescents include motor vehicle accidents, diving accidents, trampolines and falls.