PERGAMON

Accident Analysis and Prevention 36 (2004) 165–171

ACCIDENT ANALYSIS & PREVENTION

www.elsevier.com/locate/aap

Computerized coding of injury narrative data from the National Health Interview Survey

Helen M. Wellman [a,*], Mark R. Lehto [b], Gary S. Sorock [a], Gordon S. Smith [a]

[a] *Liberty Mutual Research Institute for Safety, 71 Frankland Road, Hopkinton, MA 01748, USA*
[b] *School of Industrial Engineering, Purdue University, 1287 Grissom Hall, West Lafayette, IN 47907-1287, USA*

Received 15 June 2002; received in revised form 8 November 2002; accepted 18 November 2002

Abstract

Objective: To investigate the accuracy of a computerized method for classifying injury narratives into external-cause-of-injury and poisoning (E-code) categories.

Methods: This study used injury narratives and corresponding E-codes assigned by experts from the 1997 and 1998 US National Health Interview Survey (NHIS). A Fuzzy Bayesian model was used to assign injury descriptions to 13 E-code categories. Sensitivity, specificity and positive predictive value were measured by comparing the computer generated codes with E-code categories assigned by experts.

Results: The computer program correctly classified 4695 (82.7%) of the 5677 injury narratives when multiple words were included as keywords in the model. The use of multiple-word predictors compared with using single words alone improved both the sensitivity and specificity of the computer generated codes. The program is capable of identifying and filtering out cases that would benefit most from manual coding. For example, the program could be used to code the narrative if the maximum probability of a category given the keywords in the narrative was at least 0.9. If the maximum probability was lower than 0.9 (which will be the case for approximately 33% of the narratives) the case would be filtered out for manual review.

Conclusions: A computer program based on Fuzzy Bayes logic is capable of accurately categorizing cause-of-injury codes from injury narratives. The capacity to filter out certain cases for manual coding improves the utility of this process.
© 2003 Elsevier Science Ltd. All rights reserved.

Keywords: Injury; Narrative text; E-code; Fuzzy Bayes

1. Introduction

Analysis of the circumstances surrounding an injury-producing event are essential for determining injury mechanisms and guiding prevention efforts. Central to this effort is the assignment of meaningful cause-of-injury codes for data analysis and comparison. A widely used system for coding causes of injury is the external-cause-of-injury and poisoning (E-codes) of the World Health Organization's (WHO's) International Classification of Diseases (ICD-9, World Health Organization, 1977). Although ICD-9 cause coding has limitations in the specificity of its codes (Sorock et al., 1993), it provides a useful means of standardizing external causes across different data sources (Williamson et al., 2001).

Since 1957, the National Center for Health Statistics (NCHS) has conducted the National Health Interview Survey (NHIS), which collects annual health survey data on the US population. In 1997, the survey was redesigned to include detailed questions about injuries including free text narratives of the circumstances surrounding the injury event (Warner et al., 2000). Trained coders hired by the NCHS (experts) code this information into ICD-9 E-code categories. Branching text questions were also added to obtain more specific information about the circumstances for certain injuries (e.g. motor vehicle crashes, gunshots, falls, burns, and drownings).

The addition of narrative text information in electronic format to injury databases can be a useful adjunct to epidemiological analysis and provide valuable information (Sorock et al., 1997; Smith, 2001). Accident descriptions can be used to identify and prioritize prevention efforts. Grouping the data (or coding) is an essential part of the analytic process. However, manual coding, especially on large datasets can be burdensome and use up valuable resources. Several papers have evaluated the benefits of coding and analyzing narrative text using computer algorithms (Buckely et al., 1993; Lehto and Sorock, 1996; Sorock et al., 1996,

* Corresponding author. Tel.: +1-508-435-9061x206; fax: +1-508-435-8136.
E-mail address: helen.wellman@libertymutual.com (H.M. Wellman).

0001-4575/$ – see front matter © 2003 Elsevier Science Ltd. All rights reserved.
doi:10.1016/S0001-4575(02)00146-X

1997; Davies et al., 1998; Lortie and Rizzo, 1999; Smith, 2001). Keyword searching of narrative data can be used to more completely identify specific injuries of interest (Buckely et al., 1993) or to determine specific, uncoded circumstances of injury (McLoughlin et al., 1986; Jenkins and Hard, 1992; Sorock et al., 1997). However, it is uncertain whether narrative text can be E-coded by computer with sufficient accuracy to be useful for analytic purposes (NCHS, 1999). A recent study which investigated the utility of coding the cause of work-related fatalities using text searches of narratives demonstrated high sensitivity for computer coding of occupational cases involving drowning, explosion, gunshot, rollover and falls but low sensitivity for cases involving chemical exposure and being trapped (Williamson et al., 2001).

The purpose of this study was to develop a computerized method for coding the cause of injury from injury narratives. The hypotheses were: (1) that we would be able to predict injury cause categories by using a computer program based on Fuzzy Bayes logic and (2) including multiple-word combinations would increase the sensitivity and specificity of the predictions over single words alone (McCullough and Smith, 1998).

2. Methods

Data from the 1997 and 1998 NHIS were downloaded from the Center for Disease Control's (CDC's) NCHS website (NCHS, 2001) as four separate ASCII text files: two injury episode files, and two verbatim or narrative text files. These files were then converted into Statistical Analysis Software (SAS) datasets for analysis. NHIS injury case definitions and survey methods are described fully in a NCHS report (Warner et al., 2000).

The files were merged using unique person-episode identifiers and one SAS file was created including demographic and injury episode identifiers, E-codes assigned by NCHS, narrative verbatim text and responses to branching questions. The narrative text was entered in response to the following question: "*How did the (person's) injury happen*? *Please describe fully the injury circumstances or events leading to the injury, and any object, substance or other person involved*."

The SAS dataset was exported to a Microsoft Excel spreadsheet, which was then imported into a HyperCard database developed by one of the authors (ML). This HyperCard program also included an interface and functions for parsing and indexing free text, and a machine learning (Lehto and Sorock, 1996) module which implements the Fuzzy Bayes approach discussed later. The program can be run in either a MacIntosh environment or on a PC Microsoft platform. For more information about the program's availability contact one of the authors (ML).

After the data were read in, the HyperCard program provided an alphabetical list of how often every word was found in the narratives. Each word was read and assigned to either a "keyword" list or a "drop word" list by one of the authors (HW). Words that were felt to contain the same classification value (such as "automobile" and "car") were identified and one of the words from each group of synonyms, and word misspellings was assigned as the "keyword". The "drop word" list contained words with no classification value such as "it," "a," "and" and "the". The program then matched each word in a narrative to a word found on either the "keyword" list or "drop word" list. If the word matched a word on the keyword list it was transformed to its keyword equivalent. If the word was found on the drop word list it was excluded from the narrative. The resultant narratives contained keywords only. Therefore, if the original narrative was "Riding a bicycle he hit a bump and went ovr [over] the top and fell and brke [broke] his hand", the narrative would be transformed (including the correct spelling of misspelled words) to "Ride bike he hit bump went over top fell broke his hand" for use in the model.

The program then used a machine learning approach based on Fuzzy Bayes logic to classify narratives into 1 of 13 mutually exclusive categories (e.g. overexertion, falls). The Fuzzy Bayesian model was developed within the HyperCard program to first calculate word probabilities from narratives that had been classified into E-code categories by expert coders. This was the training part of the program. The program was then used to independently classify accident narratives into one of the categories.

The Fuzzy Bayes approach calculates $P(A_i|n)$ using the expression:

$$P(A_i|n) = \mathrm{MAX}_j \frac{P(n_j|A_i)P(A_i)}{P(n_j)} \quad (1)$$

where $P(A_i|n)$ is the probability of E-code category A_i given the set of n words or word combinations in the narrative. $P(n_j|A_i)$ is the probability of word n_j given category A_i. $P(A_i)$ is the probability of category A_i and $P(n_j)$ is the probability of word n_j in the entire keyword list. $P(n_j|A_i)$, $P(A_i)$ and $P(n_j)$ are all estimated using relative word frequencies from the database.

The model attempts to select the E-code category which has the maximum probability for the set of keywords found in the narrative. This is an iterative process which includes assigning a probability for each category given each word in the narrative calculated from the learned dataset. Then the category most associated with a keyword (or keyword combination) is used to make the prediction.

The Fuzzy Bayesian approach unlike Classic Bayes assumes conditional dependence between words (Zhu and Lehto, 1999). Negative evidence (i.e. words that indicate an index term is not relevant) is considered indirectly. That is, each word or word combination increases the probability of alternative categories rather than reduce the probability of a particular category. For example, the transformed narrative "Ride bike he hit bump went over top fell broke his hand", evaluated using the single keyword model maximizes the probability of the "all falls" category given the word "fell"

($P(A_i|n) = 0.82$). However, the multiple-word model maximizes the "other transport" category given the combination of words "bike & went" ($P(A_i|n) = 1.0$), thus excluding the "all falls" category. The inclusion of multiple-word strings as keywords was introduced in the current study as a means of aggregating evidence in the Fuzzy Bayes approach. The use of word combinations in Fuzzy Bayes provides a means of aggregating evidence without making the controversial independence assumption of Classic Bayes (Zhu and Lehto, 1999). Two other examples are as follows:

1. "WHILE USING CHAIN SAW TO **CUT** BRANCH IT **FELL** ON HER HEAD AND HIT **TOP**"
   - The single-word predictor "FELL" incorrectly classified the narrative into the "all falls" category (strength = 0.82).
   - The multiple-word predictor "CUT & FELL & TOP" correctly classified the narrative into the "struck by caught against" category (strength = 1.0).
2. "**SHARPEN**ING AN AXE AND **PIECE** FLEW **IN EYE**"
   - The single-word predictor "SHARPEN" incorrectly classified the narrative into the "cutting/piercing" category (strength = 0.75).
   - The multiple-word predictor "PIECE & IN & EYE" correctly classified the narrative into the "foreign body eye" category (strength = 0.78).

The program was run using two different keyword lists to explore the potential value of combining keywords. The two separate keyword lists input into the program included: (1) single words alone from the narratives (single-word model) and (2) multiple words from the narratives (multiple-word model). The multiple-word model included single words and combinations of up to four words. To be included a keyword had to occur at least three times in the entire dataset.

To evaluate the program's performance, the computer-generated codes were compared with the expert assigned E-codes "gold standard" found in the NHIS file. The sensitivity, specificity and positive predictive value of model predictions was calculated. Sensitivity (true positives) is the percentage of narratives coded by experts into each category that were also coded to the same category by the computer program and specificity (true negatives) is the percentage of narratives coded into any other category by NCHS that were excluded by the computer program from each specific category. Positive predictive value was defined as the percentage of narratives correctly coded into a specific category out of all narratives coded by the computer program into that category.

3. Results

3.1. Model performance

There were 5677 injury episodes with narratives reported in NHIS in the 2-year period 1997–1998. Four records coded as misadventures during medical care were excluded because they did not meet our injury definition. The multiple-word model was the optimum model, with the highest sensitivity, and specificity overall and with a sensitivity of at least 63% in all but the "other unspecified/specified" categories and a specificity of at least 87% (Table 1). Overall percent agreement between model predictions and E-codes assigned by experts increased over the 13 categories from 71.3% to 82.7% using the single- and multiple-word models, respectively. Sensitivity and specificity increased in all categories when multiple words were included in the model. In some cases, the sensitivity increased to a great extent (e.g. an increase from 26.9% to 75.6% for the "other transport" category).

The single-word model would often place an injury narrative with the word "fall or fell" into the "all falls" category even when another category was correct. The multiple-word model however, could use multiple words with higher probabilities to exclude the "all falls" category from first-choice consideration. Multiple words will usually have higher probabilities than any single word alone because the likelihood that two or more words are found together in more than one category is less than the likelihood that single keywords are found in more than one category as is illustrated in the methods section (see examples 1 and 2 in Section 2).

To further illustrate why prediction performance increases when using the multiple-word model, consider the results of the "other transport" category. Here, only 43 of the 160 (26.9%) narratives were correctly classified into the "other transport" category using the single-word model, whereas 121 of the 160 (75.6%) were correctly classified in the multiple-word model (Fig. 1). Many bicycle accidents with the word "fell" in them should have been classified into the "other transport" category. However, the single-word model placed these episodes into the "all falls" category because of the high probability of the "all falls" category given the word "fell" ($P(A_i|n) = 0.82$) in the narrative. Examples of accident narratives that contained the word "bicycle" but were not "other transport" accidents were: "hand got caught in a bicycle" or "motorized bicycles". These happened enough that the probability of the "other transport" category given the word "bicycle" was reduced to 0.72. So if an injury narrative included "fell from bicycle" the word "fell" had a higher predictive value for the "all falls" category than the word "bicycle" had for the "other transport" category. The word "fell" became a good predictor for the "other transport" category when combined with words such as bicycle. For example, the probability of the "other transport" category given the words "bicycle & fell" was $P(A_i|n) = 0.87$ and was even higher at $P(A_i|n) = 0.93$ for the words "bicycle & fell & ride".

The use of branching text (text from responses to other parts of the survey) was also evaluated since expert coders have the narrative *and* additional responses to other questions in the survey to assign an E-code to the injury episode. The overall agreement changed from 82.7% to 83.2% when

Table 1
Sensitivity, specificity and positive predictive value of three-digit E-code categories assigned by the computer comparison of two different keyword lists (single words and multiple words)

ICD-9 E-codes	E-code description	N_{expert}[a]	Percentage	Single-word prediction model			Multiple-word prediction model		
				Sensitivity[b] (%)	Specificity[c] (%)	Positive predictive value[d] (%)	Sensitivity[b] (%)	Specificity[c] (%)	Positive predictive value[d] (%)
810–825	Motor vehicle	699	12	93.0	95.1	72.6	95.3	97.5	84.4
800–807, 826–848	Other transport	160	3	26.9	99.7	71.7	75.6	99.8	91.7
880–888	All falls	1960	35	92.9	77.2	68.2	95.9	87.0	79.6
890–899	Fire/flames	16	0	–	–	–	–	–	–
905–906	Animal-related injury	126	2	79.4	99.5	76.9	88.1	99.7	88.8
914	Foreign body eye	56	1	28.6	99.9	80.0	67.9	99.9	82.6
916–918	Struck by/falling/against object/person	915	16	46.7	97.3	76.7	70.1	97.9	86.3
919	Machinery	99	2	19.2	99.8	67.9	62.6	99.9	89.9
920	Cutting/piercing	470	8	70.4	98.6	81.7	86.0	99.1	90.0
924	Hot substances, caustic or corrosive material	88	2	60.2	99.7	75.7	76.1	99.8	85.9
927	Overexertion	700	12	65.7	97.1	76.2	79.7	98.4	87.7
928.9	Other unspecified accident	255	4	22.4	99.5	69.5	26.3	99.6	77.9
960–976	Assault/legal intervention	81	1	67.9	99.8	82.1	75.3	99.8	85.9
	Other specified[e]	52	1	25.0	100.0	92.9	30.8	100.0	100.0
Overall agreement[f]		5677		71.3			82.7		

[a] Number of injury episodes assigned by expert coders to each E-code category.
[b] The percentage of narratives coded to each category by National Centers for Health Statistics (NCHS) expert coders that were also coded to the same category by the computer program (percent true positives).
[c] Percentage of narratives coded into any other category by NCHS expert coders that were excluded by the computer program form each specific category (true negatives).
[d] Percentage of narratives correctly coded by the computer into a specific category out of all narratives coded by the computer program into that category.
[e] Other specified includes: E849—place of occurrence; E900, E901, E907, E908—accidents due to natural and environmental factors except animals; E910–913, E915—accidents caused by submersion or suffocation; E921–923—accidents caused by explosion, pressure vessel or firearm missile; E925–accidents caused by electric current; and E986—injury undetermined.
[f] Total number of narratives classified correctly into one of the categories, excludes E876—misadventures in medical care ($N = 4$).

the branching text was included in the narratives for the multiple-word model. The sensitivity of the predictions was better for three categories [motor vehicle (98.4%), falls (99.5%), and hot substances, caustic, or corrosive material (88.6%)] in which the branching text was used to further identify the circumstances surrounding the injury event. For the remaining 10 categories it was unchanged or somewhat less sensitive, but more so for the other transport category (75.6% versus 67.5%).

One additional trial was run using the multiple-word model to code narratives to the more discrete three-digit E-code level. For example, while all E-codes between E810 through E825 are grouped into the motor vehicle category, E-code E814 is more specifically defined as “motor vehicle traffic accident involving collision with pedestrian.” Overall agreement using the three-digit E-code model was lower than for the categorical model (58.4% versus 82.7%, respectively). However, the sensitivity of one or more of the largest three-digit E-code groups within each category was higher than the sensitivity of the category as a whole (except for the “all falls” category). For example, three-digit E-code E826 (pedal cycle accident) had a sensitivity of 97% while the entire “other transport” category had a sensitivity of only 76%.

3.2. The influence of probability thresholds on coding procedures

The Fuzzy Bayes classifier in this study chose the word or word combination with the highest probability of association with an E-code category. This is equivalent to setting a threshold level of “>0” in Table 2. Table 2 shows a range of pre-specified thresholds that could be used to limit the number of narratives coded by the computer to only those that could be coded based on a probability greater than the threshold. This would allow the user to increase the accuracy of the computer-coded narratives, and filter out the more difficult narratives for manual review.

For example, using the multiple-word model to code at a very high accuracy level (97%), you would set an acceptable threshold probability for the program of 0.9. The program would only be used to code the narrative if the maximum probability of a category given the keywords in the narrative was at least 0.9. If the maximum probability was lower than

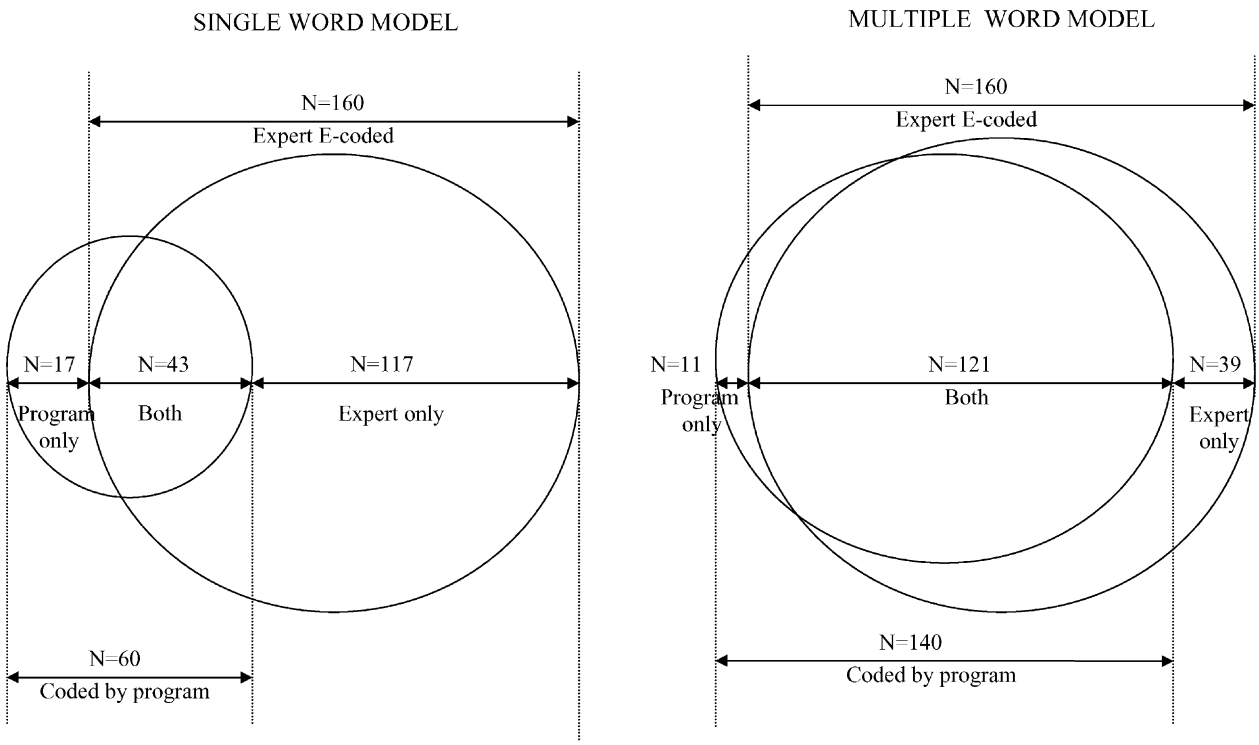

Fig. 1. Comparison of records E-coded by an expert vs. those coded by the computer program for the "other transport" category (E-codes: E800–807, E826–848).

Table 2
Overall accuracy expected[a] at different threshold values ($N = 5677$)

Threshold	Single-word model				Mutliple-word model			
	Total	Accurately coded by the model		Narratives to manually code (%)	Total	Accurately coded by the model		Narratives to manually code (%)
		N	Percentage			*N*	Percentage	
>0	5677	4045	71.3	0	5677	4694	82.7	0
>0.3	5587	4042	72.3	1.6	5591	4691	83.9	1.5
>0.4	5530	4027	72.8	2.6	5545	4675	84.3	2.3
>0.5	5322	3952	74.3	6.3	5471	4649	85.0	3.6
>0.6	4875	3719	76.3	14.1	5307	4564	86.0	6.5
>0.7	4046	3242	80.1	28.7	4891	4323	88.4	13.8
>0.8	3419	2880	84.2	39.8	4499	4100	91.1	20.8
>0.9	996	958	96.2	82.5	3814	3698	97.0	32.8

[a] For example, using the multiple-word model, if you set the maximum probability for the program at 0.9, 97% of the narratives where there was a keyword with a probability greater than 0.9, would be correctly coded ($N = 3698$), leaving 32.8% ($N = 1862$) of 5677 for manual review.

0.9 (which will be the case for approximately 33% of the narratives) the case would be filtered out for manual review (see Table 2, threshold level >0.9). However, if the acceptable accuracy level were lower at 88.4%, and you set the threshold probability at 0.7, the program would classify most of the narratives (88%), leaving only 12% to manually code. Therefore, the number of narratives you have and the accuracy required can be used to set the threshold level. Also, for a threshold level above 0.4, one would need to code over twice as many narratives manually when using the single-word model as compared with the multiple-word model.

4. Discussion

In this study, we used the available language contained in the narrative text to classify each narrative into 1 of 13 categories based on the probability of word occurrence. The results suggest that automated coding of injury narratives to E-code categories is feasible and reasonably accurate; thus hypothesis one was supported. We were able to predict injury cause categories by using a computer program based on Fuzzy Bayes logic. Errors in assigning categories and codes remain, as will the presence of non-specific

E-codes necessitating manual review of some narrative descriptions.

The use of multiple words in the Fuzzy Bayes model introduced a way to consider negative evidence indirectly, resulting in substantial improvement over the single-word model; therefore, hypothesis 2 was also supported by the data "which states that including multiple-word combinations would increase the sensitivity and specificity of the predictions over single words alone". However, the multiple-word model was substantially more complex. The computation involved millions of multiple-word combinations versus thousands for the single-word model. The combinations increase dramatically when additional text is added. Computational time is therefore a factor that needs to be considered. The use of multiple-word combinations may become prohibitive for datasets with very long narratives.

Recent studies where text mining techniques were used to code injury narratives have relied on more traditional data mining techniques where human-defined decision rules or predefined "text strings" were developed by experts in the field to mine the database (Williamson et al., 2001; Kennedy et al., 2001). For example, the Federal Aviation Administration developed a set of decision rules to code aviation safety narratives looking for specific safety issues such as a "go around" or "hard landing". While this approach worked well for this application, the development of the decision rules required a human expert knowledgeable in aviation linguistics, acronyms and abbreviations. The rules used in that study took considerable time to develop and could not be applied to other safety narratives outside of the aviation industry. The program developed for this study used a machine-learning approach, instead of a human-knowledge-acquisition approach, thus transferring the majority of the effort to the computer. This approach could be applied to a variety of different coding situations. For example this method could be used to code the narratives on death certificates or accident investigation narratives in other large databases into E-code categories. Another application includes coding motor vehicle crash narratives into different driving situations such as "driving in the fog" versus "driving in the rain" or "driving at night," etc. The method is versatile because the program is trained with word probabilities from narratives classified by experts depending on the outcome of interest.

4.1. Limitations

In order to optimize the computer program's learning capacity, all of the available NHIS injury narratives were used to both develop and test the model. It will be important to test the performance of this approach on new NHIS narratives as they become available. Moreover, use of training and testing datasets from the same data source may cause an optimistic bias in classification analysis (Clancy, 1997). However, in this study we dropped infrequent words and word combinations from the model (if they were used less than three times) to reduce the optimistic bias. Also, the same model was shown earlier to be insensitive to optimistic bias using the leave-one out bootstrapping method of outcome prediction (Lehto and Sorock, 1996).

It is assumed that the learned program will perform better on datasets that have the same underlying distribution of categories as the dataset that the program learned from. However, different distributions of event categories can be expected from different sources of data. For example, hospital discharge data will be weighted more towards severe injuries than the NHIS data and the distribution of events may be different. An important next step is to determine how the program which learned from NHIS injury data will perform on other datasets with different underlying E-code distributions.

Common words (words that could be found frequently in more than one classification) used as predictors will always be weighted toward the category with the greatest numbers of words in the learned dataset. The learned dataset had more "all falls" injury episodes than any other category, and so some frequently used words would always be used to classify narratives into the "all falls" category. The over-classification into the "all falls" category because of common words was dramatically reduced using multiple-word keywords, since these were more specific to one category.

The low accuracy by the program in predicting the "other unspecified" category is probably due to the fact that there are no unique predictors for this category and many of the narratives contained words common to other categories. This category will require manual review.

4.2. Branching text

Analysis of the results when branching text was used revealed that model sensitivity increased for the categories that the branching text was designed to address, but decreased for the remaining categories. We also found that the responses to the branching text should not contain highly predictive words from more than one category.

4.3. Research needs

The three-digit trial analysis demonstrated that with additional records per individual E-code, this program could be used to group narratives into more specific categories. More detailed codes are useful for informing prevention strategies. However, to differentiate between more detailed codes, narratives themselves may need to be longer and the learning dataset may need to contain a larger amount of records in order to get high enough word frequencies for words specific to each category. Consequently in this trial, the program appeared to have more difficulty "learning" from the less populated three-digit codes. When the less populated E-codes are clumped into a category the model does a better job at classifying them into that category. Therefore, for

the three-digit analysis it may be useful to determine if the program performs better by first "learning" the general category before "learning" the three-digit group.

Generation of the keyword list is a very important step because all predictions rely on the "matching" of the keywords to a specific category. The "keyword" list should be refined to improve sensitivity. Keywords that were used frequently by the program to make incorrect predictions may need to be broken out further. For example, performance may improve if we split out single keywords into two keywords based on the context of the word as it relates to a category such as "fell" versus "falling," "stair" versus "upstairs," or "automobile" versus "SUV".

The narratives used in this study were short (mean length 11 words) and were generated from one question administered in the NHIS. The program should be tested and perhaps modified to determine whether it can be used to E-code longer narratives collected from other surveys, inspections or investigations where several questions are asked with many parts. Longer narratives will contain several events preceding the injury in a sequence of "precipitating" events (NCHS, 1999).

5. Conclusion

These results suggest that accident narratives can be E-coded by machine with reasonable accuracy especially by using a multiple-word classification approach. It appears that manual coding of narratives can be reduced in part, but will not be eliminated. The ability to set threshold levels significantly reduces the amount of manual coding required without sacrificing accuracy and allows the user to focus on the difficult narratives. Public health programs, hospitals and other organizations with limited resources might benefit by implementing a machine learning approach to coding at least a portion of their narratives. Program output should always be tested by manual coders for accuracy. Further research is needed with training and testing the Fuzzy Bayes multiple-word model on other accident narratives from different injury databases.

Acknowledgements

The authors would like to express their gratitude to Theodore Courtney, Barbara Webster, and Dr. Margaret Warner for reviewing the manuscript.

References

Buckely, S.M., Chalmers, D.J., Langley, J.D., 1993. Injuries due to falls from horses. Aust. J. Pub. Health 3, 269–271.

Clancy, E.A., 1997. Factors influencing the resubstitution accuracy in multivariate classification analysis: implications for study design in ergonomics. Ergonomics 40 (4), 417–427.

Davies, J.C., Stevens, G., Manning, D.P., 1998. Understanding accident mechanisms: an analysis of the components of 2516 accidents collected in a MAIM database. Safety Sci. 29, 25–58.

Jenkins, E.L., Hard, D.L., 1992. Implications for the use of E-codes of the international classification of diseases and narrative data in identifying tractor-related deaths in agriculture, United States, 1980–1986. Scand. J. Work Environ. Health 18 (Suppl.), 49–50.

Kennedy B., Ramos-Santacruz M., Sada B., Dodd R., 2001. Making effective use of aviation narratives: safety event coding using text mining. In: Proceedings of the Federal Database Colloquium and Exposition, San Diego, 28–30 August 2001, pp. 357–373.

Lehto, M., Sorock, G., 1996. Machine learning of motor vehicle accident categories from narrative data. Methods Info. Med. 35 (4–5), 309–316.

Lortie, M., Rizzo, P., 1999. The classification of accident data. Safety Sci. 31, 31–57.

McLoughlin, E., Langley, J.D., Laing, R.M., 1986. Prevention of children's burns: legislation and fabric flammability. N. Z. Med. J. 99, 804–807.

McCullough, P.A., Smith, G.S., 1998. Evaluation of narrative text for case finding: the need for accuracy measurement. Am. J. Ind. Med. 34, 133–136.

National Center for Health Statistics, 1999. In: Proceedings of the International Collaborative Effort on Automating Mortality Statistics, vol. 1. Centers for Disease Control and Prevention, US Department of Health and Human Services, Publication No. (PHS) 99-1252.

National Center for Health Statistics, 2001. National Health Interview Survey injuries data and documentation: 1997 and 1998. http://www.cdc.gov/nchs/nhis.htm. Accessed 20 December 2001. Hyattsville, MD, July 1999, pp. 1–11.

Smith, G.S., 2001. Public health approaches to occupational injury prevention: do they work? Inj. Prev. 7 (Suppl. I), i3–i10.

Sorock, G., Smith, E., Hall, N., 1993. An evaluation of New Jersey's hospital discharge database for surveillance of severe occupational injuries. Am. J. Ind. Med. 23, 427–437.

Sorock, G., Ranney, T., Lehto, M., 1996. Motor vehicle crashes in roadway construction work zones: an analysis using narrative text from insurance claims. Accid. Anal. Prev. 28, 131–138.

Sorock, G., Smith, G., Reeve, G., et al., 1997. Three perspectives on work-related injury surveillance systems. Am. J. Ind. Med. 32, 116–128.

Warner M., Barnes P.M., Fingerhut L.A., 2000. Injury and poisoning episodes and conditions: National Health Interview Survey. Vital Health Stat. 10, 202.

Williamson, A., Feyer, A.-M., Stout, N., 2001. Use of narrative analysis for comparisons of the causes of fatal accidents in three countries: New Zealand, Australia, and the United States. Inj. Prev. 7 (Suppl. I), i15–i26.

World Health Organization, 1977. International Classification of Disease—Ninth Revision. World Health Organization, Geneva.

Zhu, W., Lehto, M., 1999. Decision support for indexing and retrieval of information in hypertext systems. Int. J. Hum. Comput. Interact. 11 (4), 349–371.