Journal of Safety Research 43 (2012) 327–332

Contents lists available at SciVerse ScienceDirect

Journal of Safety Research

journal homepage: www.elsevier.com/locate/jsr

Development and evaluation of a Naïve Bayesian model for coding causation of workers' compensation claims ☆

S.J. Bertke [a,*], A.R. Meyers [a], S.J. Wurzelbacher [a], J. Bell [b], M.L. Lampl [c], D. Robins [c]

[a] National Institute for Occupational Safety and Health, Division of Surveillance, Hazard Evaluations, and Field Studies, Industrywide Studies Branch, 4676 Columbia Parkway, Cincinnati, OH 45226, USA
[b] National Institute for Occupational Safety and Health, Division of Safety Research, Analysis and Field Evaluations Branch, 1095 Willowdale Road, Morgantown, WV 26505, USA
[c] Ohio Bureau of Workers' Compensation, Division of Safety & Hygiene, 13430 Yarmouth Drive, Pickerington, OH 43147, USA

ARTICLE INFO

Article history:
Received 18 April 2012
Received in revised form 27 August 2012
Accepted 23 October 2012
Available online 1 November 2012

Keywords:
Data-mining
Text-mining
Bayes
Accident narratives
Text classification

ABSTRACT

Introduction: Tracking and trending rates of injuries and illnesses classified as musculoskeletal disorders caused by ergonomic risk factors such as overexertion and repetitive motion (MSDs) and slips, trips, or falls (STFs) in different industry sectors is of high interest to many researchers. Unfortunately, identifying the cause of injuries and illnesses in large datasets such as workers' compensation systems often requires reading and coding the free form accident text narrative for potentially millions of records. *Method:* To alleviate the need for manual coding, this paper describes and evaluates a computer auto-coding algorithm that demonstrated the ability to code millions of claims quickly and accurately by learning from a set of previously manually coded claims. *Conclusions:* The auto-coding program was able to code claims as a musculoskeletal disorders, STF or other with approximately 90% accuracy. *Impact on industry:* The program developed and discussed in this paper provides an accurate and efficient method for identifying the causation of workers' compensation claims as a STF or MSD in a large database based on the unstructured text narrative and resulting injury diagnoses. The program coded thousands of claims in minutes. The method described in this paper can be used by researchers and practitioners to relieve the manual burden of reading and identifying the causation of claims as a STF or MSD. Furthermore, the method can be easily generalized to code/classify other unstructured text narratives.

1. Introduction

Work-related musculoskeletal disorders caused by ergonomic risk factors (MSDs) such as overexertion and repetitive motion as well as injuries caused by a slip, trip, or fall (STF) are common among workers and result in pain, disability, and substantial cost to workers and employers (Bureau of Labor Statistics [BLS], 2011; Liberty Mutual Research Institute for Safety, 2011). The majority of work-related occupational injuries and illnesses can be categorized into two mutually exclusive categories — MSD or a STF (BLS, 2011). Improved surveillance of occupational illnesses and injuries classified as MSDs and STFs has been a high national priority, as determined by the National Institute for Occupational Safety and Health (NIOSH) National Occupational Research Agenda (NORA). In fact, 90% of the time, surveillance of MSDs and STFs were included as strategic goals among the 10 NORA sectors' (e.g., manufacturing, construction, wholesale/retail trade [WRT]) agendas (NIOSH, CDC, 2012). Tracking the incidence and prevalence of MSDs and STFs among Ohio workers is one aim of the partnership between the NIOSH and the Ohio Bureau of Workers' Compensation (OBWC).

The OBWC collects claims data primarily to manage claims and determine future workers' compensation premiums. Prior to 2007, OBWC had no systematic way of tracking events or exposures (i.e., causation) such as ergonomic risk factors and slips, trips, or falls. Causation was only recorded in a free-text field (unstructured data) used to describe the work-related cause of the claim. Tracking the incidence and prevalence of MSDs and STFs among Ohio workers would therefore require manually coding millions of unstructured fields and was not feasible.

Recently, researchers (Lehto, Marucci-Wellman, & Corns, 2009; Marucci-Wellman, Lehto, & Corns, 2011; Wellman, Lehto, & Sorock, 2004) demonstrated that computer learning algorithms using Bayesian methods could auto-code injury narratives into different causation groups, without any manual intervention, efficiently and accurately. The authors demonstrated that the algorithms could code thousands of claims in a matter of minutes or hours with a high degree of accuracy by "learning" from claims previously coded by experts, referred to as a

☆ Disclaimer: The findings and conclusions in this report are those of the authors and do not necessarily represent the views of the National Institute for Occupational Safety and Health.

* Corresponding author at: Industrywide Studies Branch, Division of Surveillance, Hazard Evaluations and Field Studies, The National Institute for Occupational Safety and Health, 4676 Columbia Parkway, R-15, Cincinnati, OH 45226, USA. Tel.: +1 513 841 4493; fax: +1 513 841 4486.
E-mail addresses: inh4@cdc.gov (S.J. Bertke), itm4@cdc.gov (A.R. Meyers), srw3@cdc.gov (S.J. Wurzelbacher), zvd4@cdc.gov (J. Bell), Michael.L.1@bwc.state.oh.us (M.L. Lampl), David.R.1@bwc.state.oh.us (D. Robins).

0022-4375/$ – see front matter. National Safety Council and Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.jsr.2012.10.012

training set. Furthermore, these algorithms provided a score for each claim that reflected the algorithm's confidence in the prediction and, therefore, claims with low confidence scores could be flagged for manual review.

The main goal of this project was to develop and evaluate an auto-coding method that could be used to aid the manual coding of OBWC claim causations as MSD, STF, or other (OTH). Three additional issues were investigated regarding the future implementation of the auto-coding method: (a) the effect of increasing the number of predicted categories on prediction accuracy, (b) the effect of the size of the training set on the effectiveness of the program, and (c) the sensitivity of using a training set comprised of claims from one sector in predicting claims from a different sector.

2. Methods

2.1. Case definitions

MSDs are a unique class of work-related injuries/illnesses caused by ergonomic hazards. The BLS case definition for MSDs is based on a combination of two Occupational Injury and Illness Classification System (OIICS; BLS, 2012) codes: (a) nature of injury and (b) event or exposure. Our case definition for MSDs was developed to reflect the BLS case definition of an MSD. Specifically, *possible* MSD cases were the subset of claims where the nature of injury included sprains, strains, tears; back pain, hurt back; soreness, pain, hurt, except the back; carpal tunnel syndrome; hernia; or musculoskeletal system and connective tissue diseases and disorders. Claims with any other natures of injury (e.g., fractures, respiratory diseases) were ineligible for classification as an MSD. *Confirmed* MSD cases were identified as possible MSD (based on nature of injury) where the cause of the injury/illness was one of the following OIICS event or exposure categories: bodily reaction (bending, climbing, crawling, reaching, twisting); overexertion; repetition; rubbed or abraded by friction or pressure (contact stress); rubbed or abraded by friction or vibration. All claims that were not classified as an MSD were coded into two other mutually exclusive causation categories, STF or Other (OTH). All claims caused by slips, trips or falls, as defined by OIICS, were classified as STF cases. This would include a slip or trip without a fall as well as jumps to a lower level. The third category, OTH, included all injuries/illnesses not classified as either a MSD or a STF.

The auto-coding program (described below) was used to identify the causation category of various OBWC claims. For the purposes of this study, causation category was explained by an 'accident narrative' and 'injury category' fields. The unstructured accident narrative is a brief description of how the injury or illness occurred. The most influential field for a manual coder is the accident narrative; however, narratives tend to be noisy, with misspellings, abbreviations, and grammatical errors. For example, a STF narrative reads "IN COOLER, CARRING CRATE TRIP OVER CASE OF BEER HIT CEMENT FLOOR." The structured injury category field was created by OBWC for internal purposes and gives a description of the nature of the injury. It is a categorical field with 50 levels assigned based on the claim's most severe *International Classification of Diseases Ninth Revision Clinical Modification* (ICD-9 CM) code. The most severe injury, in the event multiple injuries were listed, was the ICD-9 code considered optimal for return to work based on the Degree of Disability Measurement measures. It is the one allowed ICD-9 that most likely will keep the injured worker off for the longest period of disability.

2.2. Auto-coding Procedure

The auto-coding procedure developed for this project was based on a process referred to as Naïve Bayes analysis, which is a common text classifier technique (Sebastiani, 2002), and attempted to build upon the work of Lehto et al. (2009) in this area. Details of the procedure can be found in Appendix A. In short, the procedure first attempts to calculate the probability a given claim belongs to each possible causation category. The probabilities are estimated by considering the relevant words of a text narrative and investigating their frequency in the text narratives of all the claims in a training set. For example, the word "FELL" frequently occurs in the narratives of STF claims in the training set and as a result any unknown claim with the word "FELL" in its narrative will be assigned a high probability of being a STF. In addition to considering the accident text narrative, the injury category description field was also considered since, for our study, the definition of an MSD is dependent on how the injury occurred as well as the resulting injury. Consideration of this additional structured field is an extension of the work of Lehto et al. (2009), which only considered the unstructured accident text. After probabilities have been estimated for all outcomes, the causation category with the highest probability is assigned to the claim. Finally, a score value reflecting the probability the claim was coded correctly is assigned.

2.3. Method of Evaluation

NIOSH evaluated the algorithm on the set of 10,132 un-coded OBWC-insured, single location employers, WRT Sector claims from 2008. To implement our method, NIOSH randomly sampled 2,400 claims out of the 10,132 to use as a training set for the algorithm. The claims were randomly sampled evenly across each month and between two claim severity types (lost-time, medical only). Three NIOSH safety and ergonomics experts independently coded each of the 2,400 claims as a MSD, a STF, another claim type (OTH), or not otherwise classified (NOC). NOC claims were usually missing an accident narrative or the narrative was too vague to make a determination. Of the 2,400 claims, the three coders disagreed on 148 (6.2%) claims and 12 (0.5%) claims were coded as NOC. These 160 claims were removed from the training set resulting in a set containing 2,240 manually coded claims.

The auto-coding method was then applied to the remaining 7,732 (10,132 minus the 2,400 sampled for the training set) un-coded OBWC WRT Sector claims from 2008. As a quality control (QC) measure to evaluate the effectiveness of the algorithm, an additional 800 claims (over 10% of the 7,732 un-coded claims) were sampled. These claims were then manually coded by 1 of the three NIOSH experts, blinded to the auto-coded results. The results from the manual coding (which were assumed to be accurate) were then compared to the auto-coded results. The effectiveness of the auto-coding program was measured by the sensitivity, specificity, and positive predictive value (PPV).

To evaluate the strength of the Naïve Bayes method in predicting the causation of claims beyond 3 categories, the claims coded as OTH in the 2,240 training set and the 800 QC set were further coded as one of the remaining OIICS classifications: violence or other injuries by persons or animals, transportation incidents, fires and explosions, contact with objects or equipment, exposure to harmful substances or environments, or non-MSD overexertion and bodily reaction. Non-MSD overexertion and bodily reaction claims were those in which the accident narrative met the definition for that OIICS event or exposure but the injury category description was not consistent with the MSD case definition developed for this study. The 800 claims were then auto-coded again with the refined 2,240 training set and compared to the manual code.

To assess the effect of the size of the training set on the accuracy of coding claims as MSD or STF, a convenience sample of 3,040 coded claims was created by combining the 2,240 WRT training set claims with the 800 QC set. Although the methods used to code these two sets of claims were slightly different, the same case definitions were used. From this set of 3,040 coded claims, 1,000 were randomly selected and treated as "un-coded," and a random sample from the remaining 2,040 claims was selected and used as a training set to code the un-coded claims. The sensitivity, specificity, and PPV of

Table 1
Performance statistics of the auto-coding program in classifying claims as STF, MSD or other (OTH).

		Text Only				Text + Injury Code			
	N[a]	N[b]	Sensitivity	Specificity	PPV	N[b]	Sensitivity	Specificity	PPV
All Claims	800		88.4%[c]				89.9%[c]		
NOC	6	0	0.0%	100.0%	-	0	0.0%	100.0%	-
MSD	144	147	85.4%	96.3%	83.7%	146	90.3%	97.6%	89.0%
STF	190	205	90.0%	94.4%	83.4%	215	90.5%	93.0%	80.0%
OTH	460	448	89.8%	89.7%	92.2%	439	90.7%	93.5%	95.0%

[a] – Actual number of claims in each causation category.
[b] – Number of claims predicted by auto-coding program in each category.
[c] – Overall percent of claims coded correctly by the auto-coding program.

coding an MSD and STF were recorded. This entire process of randomly selecting 1,000 "un-coded" claims and then randomly selecting a training set was repeated ten times for each training set size, which varied from 300 to 2,000 claims, in increments of 100 claims.

Finally, as MSDs and STFs are of concern across many sectors due to their high frequency and cost, in the future NIOSH intends to use this Naïve Bayes method to code claims from other industry sectors in addition to the work done in the WRT Sector. To assess the need for obtaining a separate training set for each sector, a convenience sample of coded agricultural claims was obtained. The agriculture claims were selected based on OBWC's internal industry categorization method and may not be truly representative of the Agriculture Sector as defined by NORA. One of the three NIOSH expert raters had previously coded a set of agriculture claims using the MSD, STF, OTH, or NOC coding scheme. A random sample of 2,240 claims were selected from this set, to match the size of the training set size from the WRT Sector, and these coded claims were also used as a training set to code 800 QC WRT Sector claims. Again, the manually coded and the auto-coded causation categories were compared.

3. Results and Discussion

3.1. Prediction of WRT Sector claims' causation

The Naïve Bayes auto-coding program developed in this project took less than 5 minutes to auto-code the 7,732 WRT 2008 claims using the 2,240 previously coded training set. For comparison purposes, it took one manual coder about 10 hours to code the 2,400 claims in the training set, which extrapolates to an estimate of over 30 hours for a manual coder to code 7,732 claims. However, due to fatigue the manual coders did not code the 2,400 claims in one sitting; claims were coded over the course of a couple weeks by each manual coder.

Table 1 lists the performance of the method in categorizing the 800 randomly sampled QC set into the three causation categories. Overall, when using only the unstructured text narrative to code claims, the auto-coding method predicted 88.4% of the claims correctly. When the structured injury category code was also considered, there was modest improvement overall (89.9%) in predicting claims. However, there was a large improvement in identifying MSDs, with the sensitivity increasing from 85.4% to 90.3% and the positive predictive value (PPV) increasing from 83.7% to 89.0%. This improvement in identifying MSDs is not surprising since the definition of a MSD depends not only on the cause of the injury/ illness but also the nature of the injury/ illness.

The fact that the auto-coding program agreed with the manual coder on approximately 90% of the claims in the QC set is particularly notable considering that the three manual coders agreed on only 93.8% of the claims when coding the training set. The disagreements between manual coders were generally a result of human error and/or ambiguities in the claim narratives. Therefore, the inconsistencies seen from the program are not much worse than the general inconsistencies expected between manual coders.

Based on these results, the auto-coding program has demonstrated the ability to identify STF and MSD claims with a high degree of accuracy. However, since there are so few categories, these results may not be overly impressive and may not fully demonstrate the usefulness of this program. For instance, any program will code about 33% of claims correctly by random chance; furthermore, the STF causation category would seem to be fairly easy to identify from the text since one would expect almost any narrative with the words "SLIP," "TRIP," or "FALL" will likely fall into this category. To further demonstrate the strength of this program, Table 2 lists the performance from categorizing the same 800 randomly sampled QC set into eight possible OIICS event or exposure categories. Again, overall the Naïve Bayes method performs remarkably well, correctly coding 84.1% of the claims overall with consideration of the narrative only. When the injury category description was also considered, the program's overall accuracy improved to 86.0%.

3.2. Effect of training set size on prediction accuracy

The program performed well overall when using a training set with 2,240 claims and, theoretically, the program should auto-code

Table 2
Performance statistics of the auto-coding program in classifying claims into 8 causation categories.

		Text Only				Text + Injury Code			
	N[a]	N[b]	Sensitivity	Specificity	PPV	N[b]	Sensitivity	Specificity	PPV
All Claims	800		84.1%[c]				86.0%[c]		
NOC	6	0	0.0%	100.0%	-	0	0.0%	100.0%	-
MSD	144	148	86.8%	96.5%	84.5%	146	90.3%	97.6%	89.0%
STF	190	205	90.0%	94.4%	83.4%	214	91.6%	93.4%	81.3%
Violence and other injuries by persons or animals	13	3	23.1%	100.0%	100.0%	7	30.8%	99.6%	57.1%
Transportation incidents	32	38	75.0%	98.2%	63.2%	39	78.1%	98.2%	64.1%
Fires and explosions	5	4	60.0%	99.9%	75.0%	4	80.0%	100.0%	100.0%
Contact with object or equipment	394	393	86.5%	87.2%	86.8%	375	86.8%	91.9%	91.2%
Exposure to harmful substances or environment	15	9	40.0%	99.6%	66.7%	15	60.0%	99.2%	60.0%
Overexertion and bodily reaction	1	0	0.0%	100.0%	-	0	0.0%	100.0%	-

[a] Actual number of claims in each causation category.
[b] Number of claims predicted by auto-coding program in each category.
[c] Overall percent of claims coded correctly by the auto-coding program.

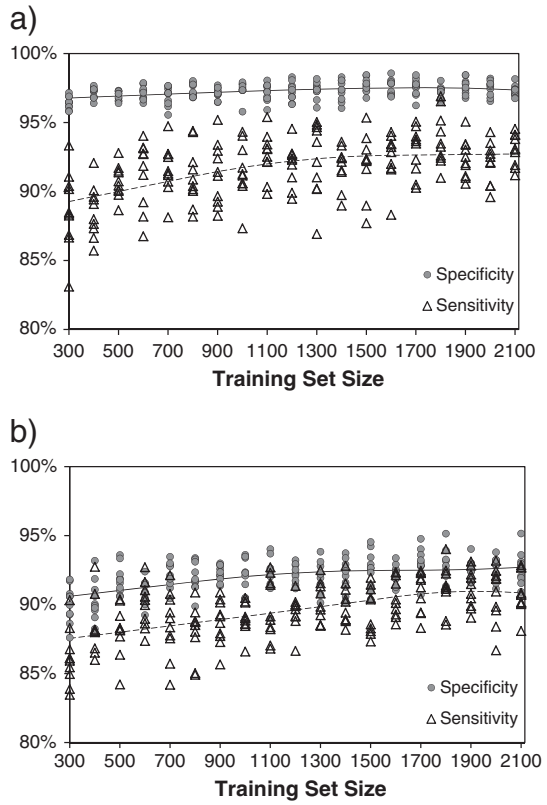

Fig. 1. Sensitivity and Specificity from using different training set sizes to code claims as a) MSD and b) STF.

claims with more precision with larger training sets. To obtain a better understanding of the effect of the training set size in identifying STFs and MSDs, the specificity and sensitivity in coding MSDs and STFs for various training set sizes, ranging from 300 to 2,000 claims is presented in Fig. 1, as described in the methods section. Notably, the auto-coding program performed well with smaller training set sizes. For example, using training sets with only 300 randomly selected claims, the sensitivity ranged from 83%–91% and from 83%–93.5% for coding STFs and MSDs, respectively, and the specificity ranged from 87%–92% and 95.5%–97.5% for coding STFs and MSDs, respectively. The auto-coding program results appear to improve with larger training set sizes, but only slight improvements were observed from using a training set with 1,000 claims compared to a training set with 2,000 claims.

The fact that the program performs well even with relatively small training set sizes is likely due to the fact that the prevalence of STFs and MSDs is very high overall and are therefore well represented in any sample of claims. These favorable results would most likely not generalize to situations where the intent is to categorize narratives into many different categories that rarely occur.

3.3. Effect of coding claims from different industry sectors

Another concern is whether the set of 2,240 WRT Sector claims could be used as a training set for other sectors, or whether a new training set should be created for each sector. As a sensitivity analysis, a convenience sample of agricultural claims were used as a training set to code claims from the WRT Sector. Anecdotally, the narratives from the agricultural claims appeared to differ when compared to that of the WRT Sector. For instance, an example of a claim coded as a STF reads "I SLIPPED ON A MAT THAT HAD EGGS ON IT. I SLIPPED ON EGGS" and an example of an MSD claim reads "MOVING PIGS, PICKED ONE UP THE WRONG WAY AND PULLED THE MUSCLE." Agricultural words such as PIG, HORSE, EGG, etc. occur more frequently in this set of claims and the impact of these different words in predicting causation category was unclear. In addition, the probabilities of each causation category differ between sectors. Most notably, 14.6% of the agriculture claims were coded as an MSD compared to 24.6% of the WRT Sector claims. For this reason, there was a concern that if the agriculture claims were used as a training set to code WRT Sector claims, the auto-coding program may have a tendency to under-represent MSDs. The effect these issues had on identifying claims as MSD and STF were investigated.

Table 3 illustrates the improvements from using claims from a common sector as a training set to code new claims by comparing the results from using the 2,240 WRT Sector claims as a training set and using the 2,240 agricultural claims as a training set to code the 800 WRT QC set. As would be expected, using the WRT Sector claims to code additional WRT Sector claims performed better, but only slightly (overall 89.9% vs. 89.6%). The largest discrepancy was observed in coding claims as MSDs. With WRT Sector claims as a training set, MSDs were coded with 90.3% sensitivity versus 84.7% sensitivity when agricultural claims were used as a training set. There are minor differences in the resultant predictive words and their associated conditional probabilities between each training set, as shown in Table 4. In fact, the top four words are identical for MSDs and STFs between training sets. These results are promising and suggest that obtaining a new training set for each sector is unnecessary.

3.4. Recommendations and quality control

The Naïve Bayes auto-coding program developed for this project and described in this paper performed very effectively in the identification and coding of un-coded claims. There was little evidence for the need to add manually coded claims to the training set for the purposes of identifying STFs and MSDs. The success of using this current training set is largely due to the high prevalence of STFs and MSDs occurring in the training set, as well as their very different case definitions. However, further evaluation of the training set size is needed if one wishes to use this method to code claims into additional causation categories, rarely occurring categories, or categories that are very similar. Additional manually coded claims, beyond the 2,240 used in this study, may be required, or additional quality control techniques may be required to identify potentially miscoded claims (see Marucci-Wellman et al., 2011).

Table 3
Comparison of performance statistics in coding WRT Sector claims using a training set from the WRT Sector vs using a training set of agricultural claims.

		WRT Coding WRT				Ag coding WRT			
Outcome	N[a]	N[b]	Sensitivity	Specificity	PPV	N[b]	Sensitivity	Specificity	PPV
All Claims	800		89.9%[c]				89.6%[c]		
NOC	6	0	0.0%	100.0%	-	0	0.0%	100.0%	-
MSD	144	146	90.3%	97.6%	89.0%	141	84.7%	97.1%	86.5%
STF	190	215	90.5%	93.0%	80.0%	207	90.0%	94.1%	82.6%
OTH	460	439	90.7%	93.5%	95.0%	452	92.2%	91.8%	93.8%

[a] – Actual number of claims in each causation category.
[b] – Number of claims predicted by auto-coding program in each causation category.
[c] – Overall percent of claims coded correctly by the auto-coding program.

Table 4
Words with highest predictive value for each causation category and their associated $P(vi|c)$ value.

WRT Training Set Top Words						Agriculture Training Set Top Words					
MSD		STF		OTH		MSD		STF		OTH	
BACK	0.33	FELL	0.51	FINGER	0.2	BACK	0.33	FELL	0.5	LEFT	0.18
PAIN	0.24	SLIPPED	0.31	HIT	0.19	PAIN	0.22	SLIPPED	0.3	HAND	0.16
FELT	0.24	BACK	0.18	CUT	0.18	LIFTING	0.16	BACK	0.18	HIT	0.15
LIFTING	0.22	OFF	0.17	LEFT	0.18	FELT	0.15	OFF	0.17	RIGHT	0.14
UP	0.2	RIGHT	0.17	HAND	0.16	UP	0.14	LEFT	0.16	EYE	0.12
RIGHT	0.15	ICE	0.17	RIGHT	0.15	LEFT	0.13	DOWN	0.16	CUT	0.12
SHOULDER	0.14	LEFT	0.16	FELL	0.11	RIGHT	0.12	RIGHT	0.14	PROVIDER	0.11
LEFT	0.12	OUT	0.15	WHILE	0.1	WHILE	0.12	OUT	0.12	FINGER	0.11
TRUCK	0.11	DOWN	0.14	OFF	0.09	OUT	0.1	HIT	0.11	INTO	0.1
DOWN	0.1	TRUCK	0.14	UP	0.09	PROVIDER	0.1	ANKLE	0.1	FAX	0.1

For the purposes of this study, any additional manual coding resources would be better spent performing a quality control on the auto-coded claims by flagging claims that appear suspicious. One method of flagging claims could be to identify claims with low score values as calculated by the auto-coding program. The score value is intended to reflect the probability that each claim is coded correctly. To investigate how well this score value represents this probability, Fig. 2 graphs the percent of claims predicted correctly versus the score value assigned by the auto-coding program. There is a definite trend that claims with lower scores were less likely than claims with higher scores to be coded correctly. However, it appears that the score value tended to slightly overestimate the prediction strength. For example, only 70% of claims with a score between .83 and .85 were coded correctly. Even so, this score can be useful in flagging claims for manual review. For example, when only the bottom 15% of the 800 WRT QC claims with the lowest score values were corrected, the sensitivity for all three causation categories increased to above 95% and the PPV for all three improved to above 92%.

3.5. Limitations

There were several limitations in this study. First, not all 50 injury categories were represented in the training set, which reduced the predicted accuracy for claims with less common injury categories. Fortunately, for MSD claims it would be easy to flag additional claims for manual review when the injury category is not consistent with our MSD case definition or seems unlikely for a STF (e.g., diseases of the circulatory, digestive, or genitourinary systems). Less than 0.01% of claims would need to be reviewed to correct these potential erroneous predictions. Second, although a small proportion of claims have multiple injuries, our program used only the most severe ICD-9-CM. Using only the most severe ICD-9-CM may underestimate the number of MSD claims if the accident narrative clearly describes an ergonomic related injury/illness but the most severe ICD-9-CM was not consistent with our case definition (e.g., fractures, contusions, major depressive disorder) so the claim was coded as OTH. Finally, some of the expert coded claims were coded by only one expert rater. This includes the 800 claims coded for QC as well as the eight causation categories. It is unclear how this limitation would influence our results.

As next steps, NIOSH will apply the program described in this paper using the 2,240 training set obtained from the 2008 WRT Sector claims to code all OBWC claims from 2001–2009 as MSD, STF, or OTH. Additional manual coding QC will be used to verify the accuracy of auto coding for other sectors. This QC process will first include manual review of claims flagged when the injury category is not consistent with our MSD case definition or seems unlikely for a STF. A random sampling of coded claims will then be manually coded and checked for accuracy. As a third step, additional claims that have low auto coding probabilities may also be manually coded and checked for accuracy.

3.6. Conclusion and Impact on Industry

We replicated and expanded upon a Bayesian machine learning auto-coding technique that has been shown to be an effective, accurate, and fast technique of identifying the accident causation category for a claim. Our work extended the previous efforts of others in this area by not only considering the accident text narrative, but also the injury category field; these two fields taken together improved the program's overall accuracy. This program will allow us to code many years of OBWC claims data in order to calculate rates of STF and MSD claims by sector and sub-sector. Eventually this benchmarking information will help to target occupational safety and health intervention efforts for Ohio employers. Additionally, it will allow researchers to evaluate the effectiveness of injury reduction efforts at larger scales. Similar techniques as described in this paper could be used by other public health practitioners to analyze large sets of existing unstructured text data that are not currently useful.

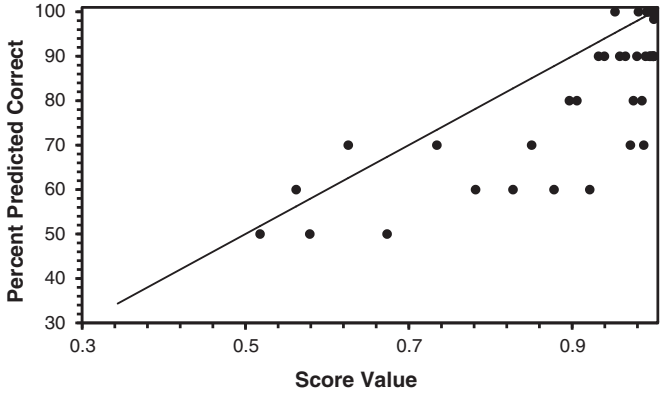

Fig. 2. Graph of percent of claims coded correctly vs. their score value calculated by the auto-coding procedure.

Acknowledgements

The authors would like to thank Dr. Mark Lehto for his valuable input and correspondence during the development of the program for this project.

References

Bureau of Labor Statistics [BLS] (2011). *Nonfatal occupational injuries and illnesses requiring days away from work, 2010.* Bureau of Labor Statistics News Release. Washington, DC: U.S. Department of Labor, Bureau of Labor Statistics.

Bureau of Labor Statistics [BLS] (2012). http://www.bls.gov/iif/oiics_manual_2010.pdf (accessed on February 16, 2012)

Domingos, P., & Pazzani, M. (1997). On the optimality of the simple Bayesian classifier under zero–one loss. *Machine Learning, 29*, 103–130.

Koller, D., & Sahami, M. (1997). Hierarchically classifying documents using very few words. In Proceedings of ICML-97, *14th International Conference on Machine Learning* (pp. 170–178). Nashville, US.

Lehto, M., Marucci-Wellman, H., & Corns, H. (2009). Bayesian method6s: a useful tool for classifying injury narratives into cause groups. *Injury Prevention, 15*, 259–265.

Liberty Mutual Research Institute for Safety (2011). *2011 Liberty Mutual Workplace Safety Index* (pp. 2). Hopkinton, MA: Author.
Marucci-Wellman, H., Lehto, M., & Corns, H. (2011). A combined Fuzzy and Naïve Bayesian strategy can be used to assign event codes to injury narratives. *Injury Prevention*, *17*, 407–414.
National Institute for Occupational Safety and Health (NIOSH) Office of the Director Centers for Disease Control and Prevention (CDC) (2012). *The National Occupational Research Agenda (NORA)*. February 8, 2012. (http://www.cdc.gov/niosh/nora/ accessed on February 16, 2012)
Sebastiani, F. (2002). Machine Learning in Automated Text Categorization. *ACM Computing Surveys*, *34*, 1–47.
van Rijsbergen, C. (1977). A theoretical basis for the use of co-occurrence data in information retrieval. *Journal of Documentation*, *33*, 106–119.
Wellman, H. M., Lehto, M. R., & Sorock, G. S. (2004). Computerized coding of injury narrative data from the National Health Interview Survey. *Accident; Analysis and Prevention*, *36*, 165–171.

Steve Bertke, Ph.D., is a statistician with National Institute for Occupational Safety and Health (NIOSH) and since 2009 has been involved with the design and analysis of a wide range of occupational health studies. Steve recently completed his PhD in Mathematical Sciences with an emphasis in Statistics from the University of Cincinnati.

Alysha Meyers, Ph.D., is an epidemiologist with NIOSH and since 2010 has served as the epidemiologist on a series of projects that have resulted from the overall collaboration between NIOSH and OBWC. Alysha has experience in designing, overseeing, implementing and evaluating epidemiologic studies, developing data collection instruments and questionnaires, and developing new methodology to analyze unstructured text data.

Steve Wurzelbacher, PhD, CPE, is a researcher at the NIOSH where he conducts studies to determine the effectiveness of occupational safety and health programs. Steve worked as a senior loss control consultant with MSIG insurance from 2001–2007 and as a research industrial hygienist with NIOSH from 1998–2001. Steve has specialized in developing exposure metrics, program evaluations, and practical workplace controls, including a patented postural support device. Steve has a PhD in Occupational Safety and Ergonomics from the University of Cincinnati, a BS in Chemical Science from Xavier University, and is a Certified Professional Ergonomist (CPE).

Jennifer L. Bell, PhD earned her doctoral degree at West Virginia University and has been working as a Research Epidemiologist at NIOSH since 1998. Her primary research focus has been evaluating the injury prevention effectiveness and cost-effectiveness of injury prevention methods, training, and equipment in the workplace.

Mike Lampl, MS, CPE, is an ergonomics technical advisor for the Ohio Bureau of Workers' Compensation (BWC). He has worked in the occupational safety and health field both in private industry and at BWC since 1993. Mike has a BS in Industrial & Systems Engineering from the Ohio State University, an MS in Occupational Health from the Medical College of Ohio, and is a Certified Professional Ergonomist (CPE).

David Robins is a Management Analyst Supervisor with the Ohio Bureau of Worker's Compensation and has been with the Bureau since 1990. David has years of experience in developing data reports, SQL queries, and data report applications. David has worked on innumerable projects to coalesce multiple data-sources to analyze and summarize large data sets.

Appendix A.

The auto-coding procedure developed for this project was based on a process referred to as Naïve Bayes, which is a common text classifier technique. To implement this model, first a list of key-words of interest was compiled. This list of keywords was comprised of all words that occurred in the narrative of at least 4 claims in a training set and that did not appear in a list of stop words (i.e. a list of words with little predictive value such as "the", "a", "an"…). Little effort was made to correct misspellings and grammatical errors (of which, there appeared to be many) to evaluate the effectiveness of this model even in the situation of very "noisy" narratives.

The next step involved representing the narrative field of each claim as a vector of "features." The features of a narrative are the occurrence (represented by a 1) and non-occurrence (represented by a 0) of the key-words in the text narrative. As an illustrative example, suppose the key-words of interest are (*fall*, *floor*, *hit*, *lift*, *trip*). In reality, the list of keywords used in our program consisted of about 900 words as opposed the 5 used in this example. With this small list of keywords, the narrative "IN COOLER, CARRING CRATE TRIP OVER CASE OF BEER HIT CEMENT FLOOR" would be then represented as (0 1 1 0 1). All other words in the narrative would be ignored since they are not in the list of key-words. The Naïve Bayes model then attempts to calculate the probability of each causation category given the vector of features using Bayes Rule. That is, given the vector of features $\boldsymbol{v} = (v_1\ v_2\ \ldots\ v_f)$, of 1's and 0's, the probability this claim belongs to causation category c is:

$$P(c|v) = \frac{P(c)P(v|c)}{P(v)} \propto P(c)P(v|c) \tag{1}$$

where $P(c)$ denotes the probability a claim belongs to causation category c, $P(\boldsymbol{v})$ denotes the probability a claim has vector of features $\boldsymbol{v}$ and $P(\boldsymbol{v}|c)$ denotes the probability a claim known to belong to causation category c has vector of features $\boldsymbol{v}$. The term $P(\boldsymbol{v})$ is not calculated directly in practice since it does not depend on each causation category and will thus not affect the resulting decision as to which causation category the claim should be assigned.The term $P(c)$ can be estimated in the obvious way by calculating the proportion of claims in a training set assigned to causation category c. Estimating $P(\boldsymbol{v}|c)$ is less obvious. To make this estimation, each of the features of the claim are naively assumed to be conditionally independent and therefore $P(v|c) = \prod_{i=1}^{f} P(v_i|c)$. The term $P(v_i|c)$ is then estimated in the following manner:

$$P(v_i|c) = \frac{count(v_i|c) + \alpha * count(v_i)}{count(c) + \alpha * N} \tag{2}$$

where $count(v_i|c)$ is the number of claims with feature v_i assigned to causation category c, $count(v_i)$ is the number of claims with feature v_i, $count(c)$ is the total number of claims assigned to causation category c, and N is the total number of claims in the training set. This estimation of $P(v_i|c)$ attempts to reduce the effects of noise in the narrative as described by Lehto et al. (2009). The α term is a smoothing constant and was assigned a value of 0.05 for this study. Therefore, for each claim with vector of features $\boldsymbol{v} = (v_1\ v_2\ \ldots\ v_f)$, the following score is calculated for each causation category, c:

$$\left(\frac{count(c)}{N}\right) \prod_{i=1}^{f} \frac{count(v_i|c) + 0.05 * count(v_i)}{count(c) + 0.05 * N} \tag{3}$$

The causation category with the highest score is assigned to the claim. The scores of each category for a given claim can then be normalized so that the sum across causation categories totals to one. The normalized scores then have the interpretation of being an estimate of the probability that a given claim belongs to a particular category.The assumption of conditional independence of features is not verified, and is most likely not valid. For example, if the word "fell" occurs in a text narrative then the word "off" is more likely to occur in the same text. Attempts have been made to improve the model by relaxing the independence assumption; however this had modest improvements in performance but severe computational cost (Koller & Sahami, 1997; van Rijsbergen, 1977). Furthermore, there has been some theoretical justification regarding the optimization of this model even when the independence assumption is violated (Domingos & Pazzani, 1997).The Naïve Bayes model has the attractive property that additional fields other than the accident narrative text of a claim can be included in the analysis. For this study, the definition of an MSD is dependent on the event or exposure leading to the injury (captured by the accident narrative text) as well as the nature of the resulting injury. Therefore, consideration of the injury category description field would provide information as to how to categorize a claim. The value in the injury category description field can simply be included as a single additional feature of a claim and since this field is structured, the one additional feature (taking on fifty possible values) is added to the vector of features and analysis proceeds identically as described above.The algorithm described above was written and performed in SAS® version 9.2 (SAS Institute Inc., Cary, NC).