Development and Evaluation of an Auto-Coding Model for Coding Unstructured Text Data Among Workers' Compensation Claims[1]

Bertke SJ§, Meyers AR§, Wurzelbacher SJ§, Bell J§, Lampl ML*, Robins D*
§National Institute for Occupational Safety and Health, *Ohio Bureau of Workers' Compensation

Introduction

Work-related musculoskeletal disorders caused by ergonomic risk factors (MSDs) such as overexertion and repetitive motion and 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, 2011; Liberty Mutual Research Institute for Safety, 2011). The majority of work-related occupational injuries and illnesses can be categorized as a MSD or a STF (Bureau of Labor Statistics, 2011). Improved surveillance of occupational illnesses and injuries (II) classified as MSDs and STFs has been a high national priority, as determined by the National Occupational Research Agenda (NORA). In fact, ninety percent of the time, surveillance of MSDs and STFs were included as strategic goals among the ten NORA sectors' (e.g. manufacturing, construction, wholesale/retail trade [WRT]) agendas. Tracking the incidence and prevalence of MSDs and STFs among Ohio workers is one aim of the partnership between the National Institute for Occupational Safety and Health (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 coding causation for millions of unstructured fields and to do this manually was not feasible.

Recently, Lehto et al (Lehto et al 2009; Wellman et al, 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 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 which could be used to aid the manual coding of OBWC claim causations as MSD, STF, or other (OTH).

Methods

Case definitions

The case definition for a MSD developed for this study reflected the MSD case definition used by the BLS, which uses the Occupational Injury and Illness Classification System (OIICS) to code nature of injury and event or exposure. The first criteria for MSD cases were those where the nature of injury included sprains, strains, tears; back pain, hurt back; soreness, pain, hurt, except the back; carpal tunnel syndrome;

[1]The findings and conclusions in this paper are those of the authors and do not necessarily represent the views of the Ohio Bureau of Workers' Compensation or the National Institute for Occupational Safety and Health.

hernia; or musculoskeletal system and connective tissue diseases and disorders. The second criteria for MSD cases, with few exceptions, were those where the event or exposure leading to the injury or illness was one of the following OIICS codes: 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. Almost all of STF cases were injuries caused by slips, trips and falls, as defined by OIICS. Claims were also coded as a third category, Other (OTH), which included all other II events not classified above as either an MSD or STF. The OTH category included events such as assaults, motor vehicle crashes, contact with objects and equipment, and exposure to harmful substances.

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 fifty levels assigned based on the claim's most severe International Classification of Diseases Ninth Revision Clinical Modification (ICD-9 CM) code.

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. In short, the procedure attempts to calculate the probability a given claim belongs to each possible causation category and the causation category with the highest probability is assigned to the claim. Also, a score value reflecting the probability the claim was coded correctly is assigned. 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 nature of 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.

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 2400 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 2400 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 2400 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 2240 manually coded claims.

The auto-coding method was then applied to the remaining 7732 (10 132 minus the 2400 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 7732 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).

Results and Discussion

The Naïve Bayes auto-coding program developed in this project took less than 5 minutes to auto-code the 7732 WRT 2008 claims using the 2240 previously coded training set. Table 1 lists the performance of the method in categorizing the 800 randomly sampled QC set into the 3 causation categories. Overall, when using only the text narrative to code claims, the auto-coding method predicted 88.4% of the claims correctly. When the injury narrative 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 II but also the nature of II.

To investigate how well the score value represents the auto-coding program's accuracy, Figure 1 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.

Conclusions

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 is not currently useful.

References

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

Bureau of Labor Statistics. [2010]. Occupational injury and illness classification manual, V. 2.0. US Department of Labor, Bureau of Labor Statistics, September 2010.

Lehto M, Marucci-Wellman H, Corns H. [2009]. Bayesian methods: 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.

Sebastiani F. [2002]. Machine Learning in Automated Text Categorization. ACM Computing Surveys, 34: 1–47.

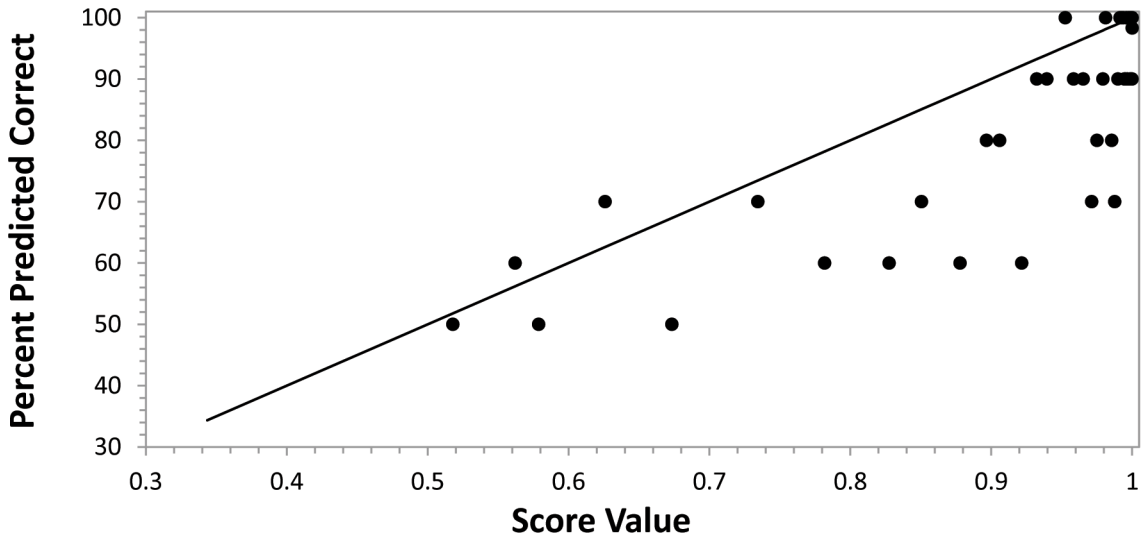
Wellman HM, Lehto MR, and Sorock GS. [2004]. Computerized coding of injury narrative data from the National Health Interview Survey. Accid Anal Prev, 36: 165–71.

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

Figure 1. Graph of percent of claims coded correctly vs. their score value calculated by the auto-coding procedure.



Use of Workers' Compensation Data for Occupational Safety and Health: Proceedings from June 2012 Workshop

Department of Health and Human Services
Centers for Disease Control and Prevention
National Institute for Occupational Safety and Health

Use of Workers' Compensation Data for Occupational Safety and Health: Proceedings from June 2012 Workshop

David F. Utterback and Teresa M. Schnorr, Editors

Department of Health and Human Services
Centers for Disease Control and Prevention
National Institute for Occupational Safety and Health

May 2013

Delivering on the Nation's promise: safety and health at work for all people through research and prevention

To receive documents or other information about occupational safety and health topics, contact NIOSH

Telephone: 1-800-CDC-INFO (1-800-232-4636)
TTY: 1-888-232-6348
email: cdcinfo@cdc.gov

or visit the NIOSH website http://www.cdc.gov/niosh/

For a monthly update on news at NIOSH, subscribe to NIOSH eNews by visiting http://www.cdc.gov/niosh/eNews.

DHHS (NIOSH) Publication No. 2013–147
May 2013

Department of Health and Human Services
Centers for Disease Control and Prevention
National Institute for Occupational Safety and Health

SAFER • HEALTHIER • PEOPLE™