Article

Predicting the outcome of occupational accidents by CART and CHAID methods at a steel factory in Iran

Gholam Abbas Shirali,[1] Moloud Valipour Noroozi,[1] Amal Saki Malehi[2]

[1]Department of Occupational Health Engineering, and [2]Department of Biostatistics and Epidemiology, Faculty of Public Health, Ahvaz Jundishapur University of Medical Sciences, Ahvaz, Iran

Significance for public health

The aim of this paper is to show that classification and regression trees (CART) and chi-square automatic interaction detection (CHAID) techniques can be used for detection of vulnerable individuals exposed to the occupational accidents at steel plants.

Abstract

Background: A large number of occupational accidents happen at steel industries in Iran. The information about these accidents is recorded by safety offices. Data mining methods are one of the suitable ways for using these databases to create useful information. Classification and regression trees (CART) and chi-square automatic interaction detection (CHAID) are two types of a decision tree which are used in data mining for creating predictions. These predictions could show characteristics of susceptible people exposed to occupational accidents. This study was aimed to predict the outcome of occupational accidents by CART and CHAID methods at a steel factory in Iran.

Design and methods: In this study, the data of 12 variables for 2127 cases of occupational injuries (including three categories of minor, severe and fatal) from 2001 to 2014 were collected. CART and CHAID algorithms in IBM SPSS Modeler version 18 were used to create decision trees and predictions.

Results: Five predictions for the outcome of occupational accidents were created for each method. The most important predictor variables for CART method included age, the cause of accident and level of education respectively. For CHAID method, age, place of accident and level of education were the most important predictor variables respectively. Furthermore the accuracy of CART and CHAID methods were 81.78% and 80.73%, respectively for predictions.

Conclusions: CART and CHAID methods can be used to predict the outcome of occupational accidents in the steel industry. Thus the rate of injuries can be reduced by using the predictions for employing preventive measures and training in the steel industry.

Introduction

Occupational accidents threaten the lives of many people annually and are also the cause of a high percentage of disabilities. Occupational accidents can lead to negative consequences such as fatalities, disability and missed work days.[1,2] The international labour organization (ILO) announced every 15 seconds, 153 workers experience occupational accidents and one worker dies owing to the work-related illnesses. In addition, ILO stated that approximately 4% of the gross domestic product (GDPs) of countries is spent to compensate for damages caused by accidents.[3] In Iran approximately 14,000 occupational accidents occur annually.[2] The steel industry is one of the world's most dangerous industries. The frequency of occupational accidents in the steel industry is higher than others. Data on occupational accidents are usually recorded, but the main challenge is the extraction of useful information from suitable sources.[4] One solution to this problem is the use of data mining methods.[5-7] Data mining techniques are one of the most useful methods of extracting meaningful data. This includes using a set of statistical methods.[4,8-12] Studies in many different countries have been conducted using these data mining methods in various scientific fields.[7,13-15] Data mining methods can be used to reduce accidents by uncovering the vulnerable individuals at industrial workplaces.[2,7] Among the different methods of data mining, the decision tree is one of the powerful and common tools for creating predictions.[10] In this study, classification and regression trees (CART) and chi-square automatic interaction detection (CHAID) methods were chosen. The CART method was designed by Breiman *et al.* in 1984. The CART method generates a decision tree by Gini index.[16] The Gini index is a data categorizing scale that creates nodes.[16-20] The CHAID method was introduced by Kass in 1980.[21] In CHAID method, production of nodes carried out by the chi-square test.[21] Previous studies show that it is possible to predict the outcome of occupational accidents by using the CART and CHAID methods.[14,15] Furthermore, the results obtained by the CART and CHAID methods are easier to understand and have a desirable accuracy.[22,23] The current study was carried out to predict the outcome of occupational accidents by the CART and CHAID methods at a steel factory in Ahvaz, a south-western city of Iran.

Materials and methods

Study design

The current retrospective study was conducted at a steel factory in Ahvaz, the centre of Khuzestan province in south-west Iran at the year 2015 (Figure 1). The study site (factory) was a manufacturing complex of 6 units (steel making, pipe rolling, beam rolling, bar rolling, kowsar rolling and machinery). The inclusion criteria for cases were included; 1) accidents with complete data for 12 variable of this study, 2) accidents took place between 2001 and 2014. The criteria for exclusion of cases were included; 1) accidents with no data for any variable of this study, 2) accidents that happen before 2001 and after 2014. Informed consent was taken from all individuals that their information was used in this study. In this study, data of 2127 occupational injuries were col-

lected. All data were related to male workers. Then, 207 cases were ignored owing to missing data. The injuries were categorized as minor (1554 cases), severe (357 cases) and fatal (9 cases) damages based on working day lost.[22] The data were related to 12 variables (Figure 2). The outcome of occupational accidents was the main (target) variable. The predictor variables included age, level of education, work shift, place of accident, time of the accident, the season of the accident, the day of the accident, the cause of the accident, using protective equipment, work experience, and marital status (Figure 2). They were categorized based on previous studies (Figure 2).[7,23]

Statistical analysis

Data were processed by CART and CHAID algorithms in IBM SPSS Modeler version 18. To prepare and test the model, data were categorized into two groups of training and testing, about 70% of data were randomly assigned to model training and the rest 30% was considered for model assessments.[22]
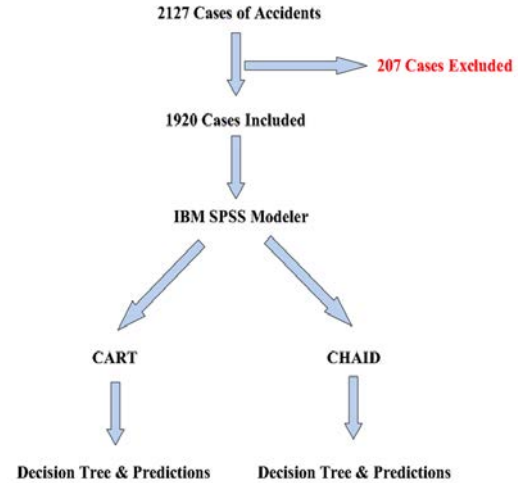

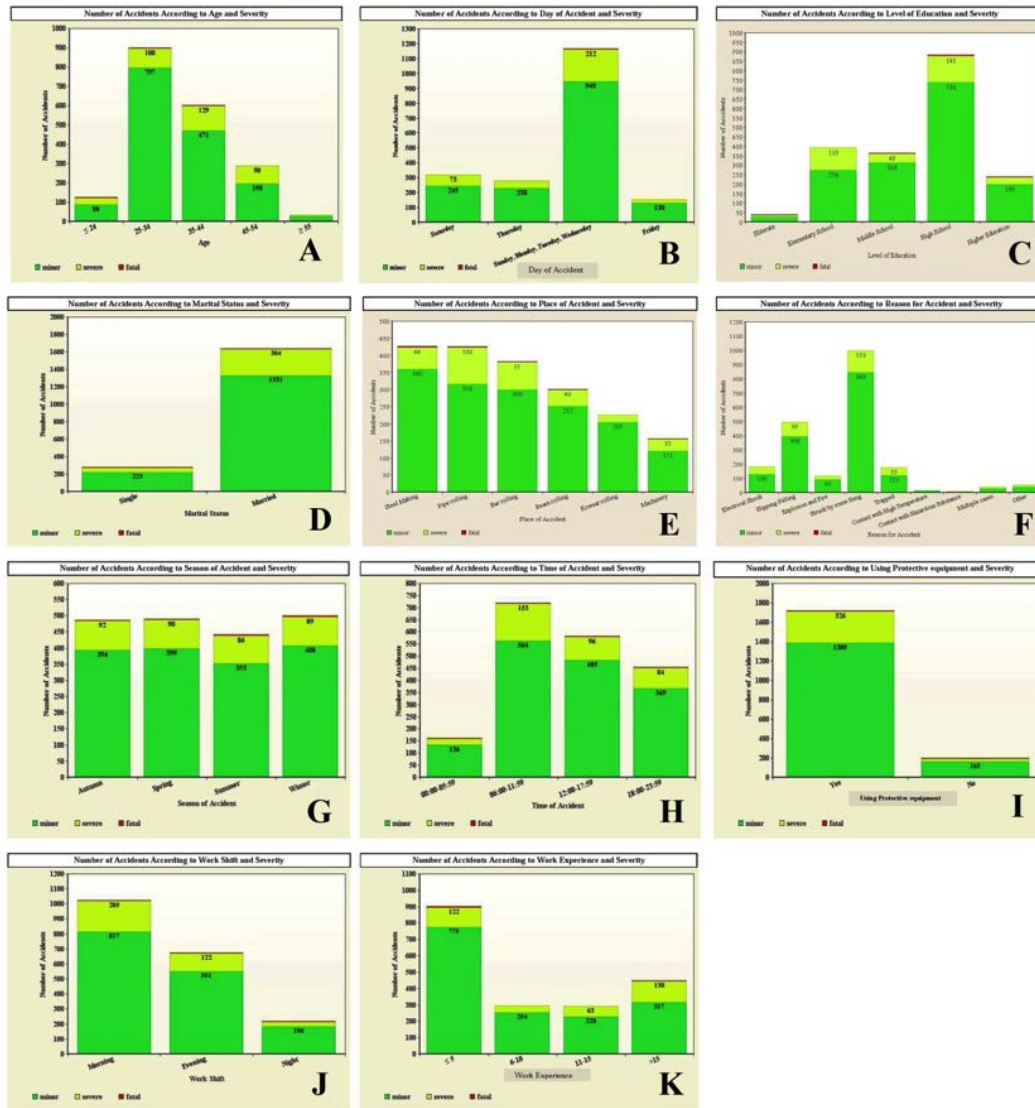**Figure 1. Diagram of the study.**

Figure 2. Classification of predictor variables for prediction of the outcome of occupational accidents and demographic characteristics of participants (The colors in all charts represents the outcome of occupational accidents, green; minor injury, yellow; severe injury and red; fatal injury).

Results

Accuracy of models and important predictor variables

The predictive accuracy of the CART and CHAID methods were almost similar (81.78% and 80.73%, respectively). The most important predictor variables based on the CART method in predicting the outcome of occupational accidents included the age, the cause of the accident, level of education, using protective equipment, the day of the accident, marital status, working experience and place of accident respectively. The most important predictor variables based on the CHAID method included age, place of accident, level of education, using protective equipment, and time of accident respectively. For both methods, age was the most important predictor variable.

Decision tree and predictions

A decision tree and five predictions were created for the CART and CHAID methods (Figures 3 and 4; Tables 1 and 2). All predictions achieved by the CHAID and CART methods were related to minor injuries, except for one prediction by the CART method that was related to severe injuries (Tables 1 and 2).
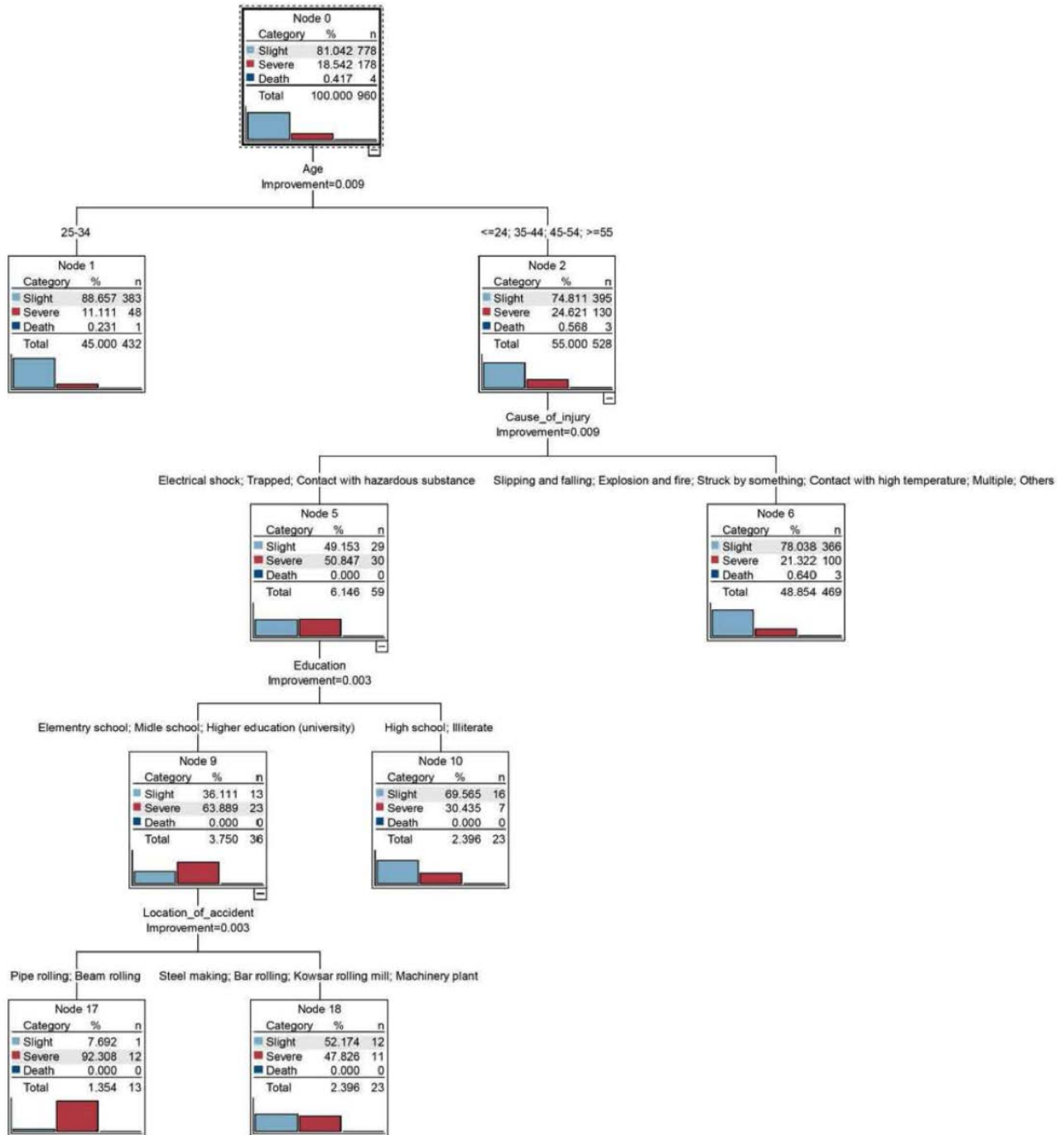

Figure 3. The decision tree for prediction of the outcome of occupational accidents based on the CART method.

Interpretation of the CART model

Node 0 was divided into two branches based on the age. This indicates that the most important predictor variable in the categorization and prediction of the outcome of occupational accidents was age (Figure 3). Node 1 on the left side of the tree was related to injured subjects in an age range of 25-34 years. This indicates that people within this age range are more vulnerable to occupational injuries when compared to others (Figure 3; Table 1). In terms of the causes of accidents, node 2 on the right side of the tree was branched into node 5 and node 6. With respect to node 6, the frequency of minor injuries (78%) among individuals within the age ranges of ≤24, 35-44, 45-54, and ≥ 55 years who were injured owing to slipping, falling, explosion and fire, struck by something, contact with high temperature, or multiple cases was higher than those of workers injured owing to other factors (trapped, contact with hazardous substance, or electrical shock) (Figure 3; Table 1). Node 9 was branched into nodes 17 and 18 based on the accident site. This indicates that the risk of severe injuries was higher in the pipe and beam rolling units, compared with other sections (units for steel making, machinery, bar rolling, etc.) (Figure 3; Table 1).

Interpretation of CHAID model

The similarity to CART model, in CHAID model also age was the most important predictor variable in the categorization and prediction of the outcome of occupational accidents (Figure 4). Node 0 was branched into the node 1, node 2 and node 3 (Figure 4). Node 1 was branched into node 4 and node 5. Node 4 was branched into node 8 and node 9. Node 8 shows that the frequency of severe injuries among workers with an elementary education was higher than that of others (45.63%) (Figure 4; Table 2). Node 9 was branched into node 10 and node 11. This shows that the risk of severe injuries was higher in the pipe and machinery units (37%) compared with other sections (units for steel making, beam rolling, bar rolling, etc.) (Figure 4). Node 5 shows that the frequency of minor and fatal injuries among individuals within the age ranges of ≤24 or 45-54 years (who did not use protective equipment) was 70% and 3%, respectively (Figure 4). Node 2 shows that the risk of minor injuries was higher in the age range of 25-34 years (88%) compared with other age groups (Figure 4; Table 2). Node 3 was branched into final nodes 6 and 7. Node 6 shows that the frequency of fatal injuries in individuals within the age ranges of 35-44 or ≥55 years was 5%, during 00:00-5:59 (Figure 4; Table 2).

Discussion

More than 300,000 workers die annually due to occupational accidents worldwide.[24] In addition, occupational accidents impose huge financial losses on industries and society.[25] The steel industry is among the industries that workers face many dangers due to the nature of their jobs.[26] Therefore, the possibility of occupational accidents is high in steel industry.[27] Predicting the outcome of occupational accidents is useful when implementing preventive strategies.[25] Safety managers can reduce the prevalence of accidents by using the predictions of CART and CHAID for detecting susceptible people. According to this study, age was the main pre-

Table 1. Predictions of CART method for the outcome of occupational accidents.

Node	Predictions	The outcome of occupational accidents	Probability (%)	Population
1	If (age range: 25-34)	Minor injury	88.65	383
17	If (age range: ≤24; 35-44; 45-54; ≥55) and (reason for accident: trapped; contact with hazardous substance; electrical shock) and (level of education: elementary or middle school; higher education) and (place of accident: pipe rolling; beam rolling)	Severe injury	92.3	12
18	If (age range: ≤24; 35-44; 45-54; ≥55) and (reason of accident: trapped; contact with hazardous substance; electrical shock) and (level of education: elementary or middle school; higher education) and (place of accident: steel making; bar rolling or Kowsar rolling; machinery)	Minor injury	52.17	12
10	If (age range: ≤24; 35-44; 45-54; ≥55) and (reason of accident: trapped; contact with hazardous substance; electrical shock) and (level of education: illiterate; high school)	Minor injury	69.56	16
6	If (age range: ≤24; 35-44; 45-54; ≥55) and (reason of accident: slipping-falling; explosion and fire; struck by something; contact with high temperature; multiple cases; other)	Minor injury	78.03	366

Table 2. Predictions of CHAID method for the outcome of occupational accidents

Node	Predictions	The outcome of occupational accidents	Probability (%)	Population
8	If (age range: ≤24; 45-54) and (Using protective equipment: yes) and (level of education: elementary school)	Minor injury	54.36	56
10	If (age range: ≤24; 45-54) and (Using protective equipment : yes) and (level of education: illiterate; middle or high school; higher education) and (place of accident: steel making; bar rolling; Kowsar rolling)	Minor injury	86.13	87
2	If (age range: 25-34)	Minor injury	86.08	525
6	If (age range: 35-44; ≥55) and (time of accident: 0-6)	Minor injury	77.77	28
7	If (age range: 35-44; ≥55) and (time of accident: 6-12; 12-18; 18-24)	Minor injury	78.71	318

dictor variable in predicting the outcome of occupational accidents for both models. Age is a personal factor that can change the physical and cognitive skills of workers, and affect their ability to perform their duties.[28] The highest rate of injuries occurred in the age range of 25-34 years, in both methods (Figures 3 and 4). This result can be justified owing to unfamiliarity with environmental conditions, low experience, and lack of skill in using the equipment. In the older age ranges, the rate of occupational accidents was lower, but the outcome of occupational accidents was worse. The higher

rate of severe injuries in the older age ranges may occur because employees were accustomed to the work and consequently do not comply with safety procedures. These results were consistent with previous studies.[28-30] According to CART and CHAID method, the prevalence of injuries among individuals with high school education and lower ones were 69.56% and 54.36% respectively (Figure 2C). This may be related to their inadequate awareness of safety measures. These findings correlate with previous studies.[28,31] Meanwhile, in the CART model, the maximum rate of injuries
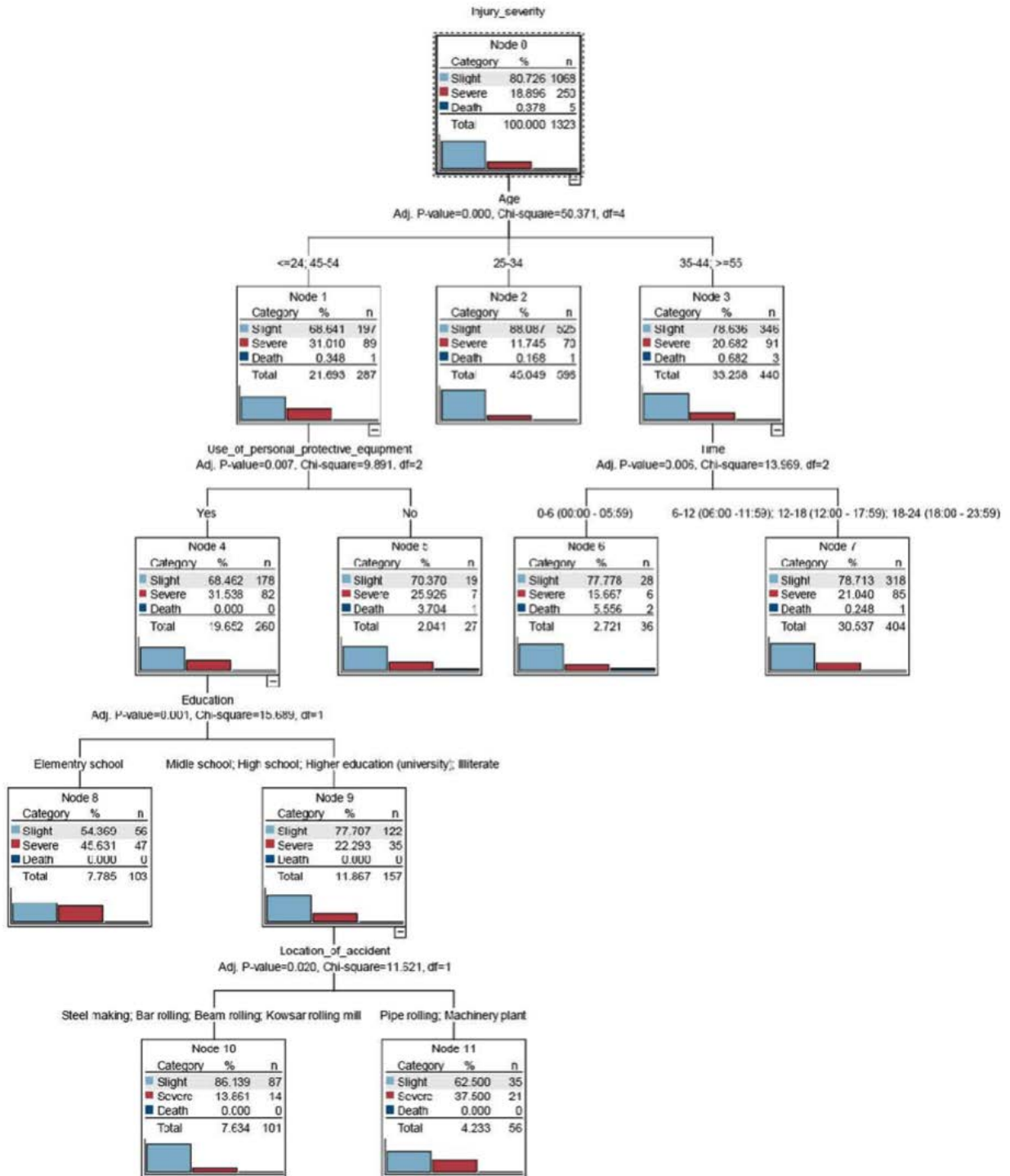

Figure 4. The decision tree for prediction of the outcome of occupational accidents based on the CHAID method.

(61%) occurred from Sundays to Wednesdays (Figure 2B). This may result from a heavy workload during these days. These results were similar to those of Rahmani *et al*.[25] The higher frequency of minor and severe injuries during the morning shift (6:00-11:59) may be related to the high workload of the staff at morning shift. In this regard, the current study was in agreement with the previous studies.[7,31] In addition, the higher frequency of fatal injuries during the night shift (00:00-5:59) may be associated with drowsiness during this time interval. In this study, accidents mostly occur among married subjects, regardless of the outcome of occupational accidents, the reason can be attributed to the high number of married subjects. These results were similar to those of Rahmani *et al*. (Figure 2D).[25] Moreover, this study shows injuries mostly occurred in subjects with work experience <5 years (Figure 2K). This finding was in agreement with the study of Ujwala *et al*. The reason can be attributed to less experience, inadequate skill, and lower concentration over job among younger workers, compared with their older counterparts.[24] According to the result of this study, the CART and CHAID algorithms are two useful methods for safety offices in order to develop a strategy for prevention of accidents at steel plants in Iran.

Conclusions

The CART and CHAID algorithms were suitable for predicting the outcome of occupational accidents in a steel factory. Furthermore, safety officers can reduce the rate of accidents by using the predictions for detecting vulnerable workers in steel plants. The data of accidents that happen before 2001 were not available and it was limitation of the current study. Finally, it is recommended to employ other methods of data mining such as C5, support vector machine (SVM) and Bayesian networks to predict the outcome of injuries in steel industries for future studies.[32-34]

References

1. Omidvari M, Rasooli- Gharmaroudi M. Analysis of human error in occupational accidents in the power plant industries using combining innovative FTA and meta-heuristic algorithms. J Saf Health Work 2015;5:1-12. [in Persian].
2. Atrkarroshan S, Alizadeh SS. Estimate of economic costs of accidents at work in Iran: A case study of occupational accidents in 2012. Iran Occup Health 2015;12:12-9. [in Persian].
3. International Labour Organization [Internet]. Safety and health at work. ILO web site; 2017 [cited 26 Sep, 2017]; Available from: http://www.ilo.org/global/topics/safety-and-health-at-work/lang—en/index.htm.
4. Palamara F, Piglione F, Piccinini N. Self-organizing map and clustering algorithms for the analysis of occupational accident databases. Saf Sci 2011;49:1215-30.
5. Tsay YJ, Chiang JY. CBAR: an efficient method for mining association rules. Knowl Based Syst 2005;18:99-105.
6. Liao CW, Perng YH. Data mining for occupational injuries in the Taiwan construction industry. Saf Sci 2008;46:1091-102.
7. Cheng CW, Yao HQ, Wu TC. Applying data mining techniques to analyze the causes of major occupational accidents in the petrochemical industry. J Loss Prev Process Ind 2013;26:1269-78.
8. Witten IH, Frank E. Data mining: Practical Machine learning tools and techniques. 2nd ed. San Francisco: Morgan Kaufmann; 2005. pp. 423-560.
9. Cios KJ, Pedrycz W, Swiniarski RW, Kurgan L. Data Mining: A knowledge discovery approach. Berlin (Germany): Springer; 2007. pp. 544-606.
10. Koyuncugil AS, Ozgulbas N. Financial early warning system model and data mining application for risk detection. Expert Syst Appl 2012;39:6238-53.
11. Paolo G. Applied data mining: statistical methods for business and industry. New York: Wiley; 2003.pp. 1-50.
12. Anand S, Keren N, Tretter MJ, et al. Harnessing data mining to explore incident databases. J Hazard Mater 2006;130:33-41.
13. Delen D,Sharda R, Bessonov M. Identifying significant predictors of injury severity in traffic accidents using a series of artificial neural networks. Accid Anal Prev 2006;38:434-44.
14. Kim TW, Koh DH, Park CY. Decision tree of occupational lung cancer using classification and regression analysis. Saf Health Work 2010;1:140-8.
15. Mistikoglu G, Gerek IH, Erdis E, et al. Decision tree analysis of construction fall accidents involving roofers. Expert Syst Appl 2015 42:2256-63.
16. Breiman L, Friedman JH, Stone CJ, Olshen RA. Classification and regression trees. New York: Taylor & Francis; 1984.pp. 305-368.
17. Marshall RJ. The use of classification and regression trees in clinical epidemiology. J Clin Epidemiol 2001;54:603-9.
18. Lemon SC, Roy J, Clark MA, et al. Classification and regression tree analysis in public health: methodological review and comparison with logistic regression. Ann Behav Med 2003;26:172-81.
19. Muller R, Mockel M. Logistic regression and CART in the analysis of multimarker studies. Clin Chim Acta 2008;394:1-6.
20. Gupta B, Rawat A, Jain A, et al. Analysis of various decision tree algorithms for classification in data mining. Int J Comput Appl 2017;8:15-9.
21. Kass GV. An exploratory technique for investigating large quantities of categorical data. Appl Stat 1980;29:119-27.
22. Kashani AT, Shariat- Mohaymany A, Ranjbari A. Analysis of factors associated with traffic injury severity on rural roads in

Iran. J Inj Violence Res 2012;4:36-41.

23. De Oña J, Lopez G, Abellan J. Extracting decision rules from police accident reports through decision trees. Accid Anal Prev 2013;50:1151-60.

24. Ukey-Ujwala U, Chitre-Dhruvs S, Padmasree D, Satyanarayan D. Occupational injuries in workers of steel plant at Visakhapatnam. Int J Prev Med 2015;1:79-83.

25. Rahmani A, Khadem M, Madreseh E, et al. Descriptive study of occupational accidents and their causes among electricity distribution company workers at an eight-year period in Iran. Saf Health Work 2013;4:160-5.

26. Saha A, Kumar S, Vasudevan DM. Occupational injury surveillance: A study in a metal smelting industry. Indian J Occup Environ Med 2008;11:103-7.

27 Nordlöf H, Wiitavaara B, Winblad U, et al. Safety culture and reasons for risk-taking at a large steel-manufacturing company: Investigating the worker perspective. Saf Sci 2015;73:126-35.

28. Aghilinejad M, Kouhpayezade J, Noori MK, Golabadi M. Association of age and work experience with work-related injuries in mining and mineral industries in Iran 2003-2011. Razi J Med Sci 2013;19:20-8. [in Persian].

29. Alizadeh SH, Mortazavi B, Sepehri MM. Analysis of occupational accident fatalities and injuries among male group in Iran between 2008 and 2012. Iran J Med Sci 2015;17:1-6. [in Persian].

30. Salminen S. Have young workers more injuries than older ones? An international literature review. J Safety Res 2004;35:513-21.

31. Gulhan B, Iihanb M, Civil EF. Occupational accidents and affecting factors of metal industry in a factory in Ankara. Turk J Public Health 2012;10:76-85.

32. Sarkar S, Verma A, Maiti J. Prediction of Occupational Incidents Using Proactive and Reactive Data: A Data Mining Approach. Industrial safety management: Springer; 2018. pp.65-79.

33. Sanchez AS, Fernandez PR, Lasheras FS, et al. Prediction of work-related accidents according to working conditions using support vector machines. Appl Math Comput 2011;218:3539-52.

34. Rivas T, Paz M, Martin JE, Matias JM, Garcia JF, Taboada J. Explaining and predicting workplace accidents using data-mining techniques. Reliab Eng Syst Safe 2011;96:739-47.