Original Article

Institution of MECHANICAL ENGINEERS

Journal of SPORTS ENGINEERING AND TECHNOLOGY

Proc IMechE Part P:
J Sports Engineering and Technology
1–10

DOI: 10.1177/1754337117728600
journals.sagepub.com/home/pip
SAGE

Ski injury predictive analytics from massive ski lift transportation data

Boris Delibašić[1], Sandro Radovanović[1], Miloš Jovanović[1], Zoran Obradović[2] and Milija Suknović[1]

Abstract

Ski injury research is traditionally studied on small-scale observational studies where risk factors from univariate and multivariate statistical models are extracted. In this article, a large-scale ski injury observational study was conducted by analyzing skier transportation data from six consecutive seasons. Logistic regression and chi-square automatic interaction detection decision tree models for ski injury predictions are proposed. While logistic regression assumes a linearly weighted dependency between the predictors and the response variable, chi-square automatic interaction detection assumes a non-linear and hierarchical dependency. Logistic regression also assumes a monotonic relationship between each predictor variable and the response variable, while chi-square automatic interaction detection does not require such an assumption. In this research, the chi-square automatic interaction detection decision tree model achieved a higher odds ratio and area under the receiver operating characteristic curve in predicting ski injury. Both logistic regression and chi-square automatic interaction detection identified the daily time spent in the ski lift transportation system as the most important feature for ski injury prediction which provides solid evidence that ski injuries are early-failure events. Skiers who are at the highest risk of injury also exhibit higher lift switching behavior while performing faster runs and preferring ski slopes with higher vertical descents. The lowest injury risk is observed for skiers who spend more time in the ski lift transportation system and ski faster than the average population.

Keywords

Ski injury prediction, feature extraction, risk factors, chi-square automatic interaction detection decision tree analysis, logistic regression

Date received: 25 March 2017; accepted: 31 July 2017

Introduction

This article addresses the prediction of ski injuries based on ski lift transportation data. Ski lift transportation data are heavily underused in ski injury research, although these data are massively recorded and stored at ski resorts. Skiing is a multibillion business in the United States with 7.3 billion USD spent at US ski resorts in 2014/2015 and 57.1 million average skier visits (skier-days) yearly since the 2002/2003 season.[1] America (mostly North America) has a total share of 21% of ski visits worldwide with 15% of the 21% from the United States. The other worldwide ski visits comprised the Alps at 43%, Western Europe at 11%, Eastern Europe and Central Asia combined at 9%, and Asia and the Pacific combined at 15% with a total of 400 million skier visits yearly worldwide.[2]

With an average injury rate of two injuries per thousand skier days[3] (IPTSD), each year more than 120,000 skiers are expected to get injured in the United States alone with 800,000 ski injuries expected worldwide each year.

Building predictive models for ski injury from ski lift transportation data is difficult, as the data are highly imbalanced (i.e. only 0.2% of the population will eventually get injured while 99.8% will not get injured). Therefore, ski injury research is usually done on small-scale, case–control studies that analyze the injured population or a small sample of the non-injured population. Whether the sample population is a good

[1]Faculty of Organizational Sciences, University of Belgrade, Belgrade, Serbia
[2]Center for Data Analytics and Biomedical Informatics, Temple University, Philadelphia, PA, USA

Corresponding author:
Sandro Radovanović, Faculty of Organizational Sciences, University of Belgrade, Jove Ilića 154, 11000 Belgrade, Serbia.
Email: sandro.radovanovic@fon.bg.ac.rs

representation is always questionable due to the huge imbalance of injured to non-injured population and heterogeneity of the skiing population.

Ski injury research frequently uses logistic regression for reporting adjusted odds ratios (ORs).[4,5] However, this linear model fails to account for the attribute interactions, where some attributes are good indicators only in the presence of other attributes. Logistic regression also has the limitation that it does not work well when predictors and the response variable have a non-monotonic relationship. Finally, logistic regression assumes that the relationship between the predictors and the outcome is linear with constant log odds increments. In addition to the logistic regression, this study utilized the chi-square automatic interaction detection (CHAID) decision tree algorithm.[6,7] This algorithm produces a decision tree model that captures the hierarchical and conjuncted relationships that exist in real-world data and handles the non-monotonic and non-linear relationship between predictors and the response variable. In addition, CHAID uses chi-square statistics for making decisions on the tree model growth which branches trees on attributes that produce the highest ORs. This is a preferable requirement for ski injury research because of the high class imbalance. The authors showed that the CHAID decision tree model was a good predictor, achieving a higher OR and area under the receiver operating characteristic curve (AUC), as compared to logistic regression. In addition, the decision tree model appeared to be useful, revealing interpretable injury rules on sub-samples of the whole population.

The main contributions of this article include the following:

1. Proposing features from ski lift transportation data that can be used for ski injury research (section "Materials and methods").
2. Proposing the application of CHAID decision tree for ski injury prediction modeling (section "CHAID decision tree analysis").
3. Identification of risk factors that affect the occurrence of ski injuries (sections "Descriptive statistics," "Logistic regression analysis," and "CHAID decision tree analysis").
4. Recognizing ski injuries as early-failure events (sections "Descriptive statistics," "CHAID univariate analysis," and "Conclusion").

A major limitation of this article is the lack of demographic data (i.e. self-reported skiing proficiency, gender, age, etc.), so reported risk factors are demographic-agnostic.

The remainder of the article is structured as follows. In section "Literature review," the background of research similar to the research proposed in this article is provided. In section "Materials and methods," the ski lift transportation data are presented, as well as the features extracted from it that will be used in section "Results" for building univariate and multivariate models for ski injury predictions. The conclusions are summarized in section "Conclusion."

Literature review

Ski injury risk factors are well-studied. Gender,[5] age,[2] skiing errors, fatigue, perception of low difficulty,[8] perceived speed of skiing,[5] skillfulness and experience,[9–11] first-day participation,[12] quality of equipment, type of equipment,[9] lessons taken,[12] quality of ski slopes and quality of their preparation, collision against objects, preexisting medical conditions,[13] snow conditions,[2] poor visibility due to inclement weather,[9] design and maintenance of ski slopes,[14] as well as other factors all influence the occurrence of ski injury. These risk factors have been gathered prevalently through injury reports and field studies. Hume et al.[9] provided a complete meta-analysis on ski injuries with reported ORs for various risk factors.

To the best of the author's knowledge, ski lift transportation data are rarely used for ski injury analysis,[15] although mining sensor-collected data are an emerging research area.[16,17] D'Urso and Massari[16] analyzed one skier-day data and reported two types of skier behavior, variety seekers who tend to switch often between ski lifts and loyal skiers who typically stick to one ski lift. On the other hand, several studies uncovered patterns from transportation data, such as bicycle rentals in Barcelona[18] and the London metro system.[17] Exploiting sensor-based big data to analyze the safety risks in real-time traffic operations for congestion and crash risk[19] is an emerging research field. Still, discovering patterns from ski lift transportation data is not a well-studied area, as the ski resort business has not yet fully exploited the potential of the data being collected.

Predictive analytics methods have already been proposed for ski resort decision-making. King et al.[20] proposed a model for the number of skier-days (ski visits) prediction. Models for predicting the number of ski injuries have been proposed as well. Dalipi et al.[21] proposed using artificial neural networks, while Bohanec and Delibašić[15] proposed using a combined expert-modeling and data-driven approach for the global daily risk of injury on a ski resort. Still, to the best of the author's knowledge, articles that use ski lift transportation data for individual ski injury risk assessment are not reported in the literature.

Materials and methods

Data

This research was conducted on ski lift transportation data obtained from the largest Serbian ski resort, Mt Kopaonik. These data include all ski lift gate entrances from six consecutive seasons between 2006 and 2011. The basic characteristics of the ski lifts in the Mt Kopaonik ski resort are included in Table 1. Lifts with

Table 1. Mt Kopaonik ski lifts and their basic characteristics.

Ski lift	Type	Entrance altitude (m)	Exit altitude (m)	Vertical ascent (m)	Hourly ski lift transportation (skiers/h)	Difficulty of accessible slopes
Krčmar	Two-seat chair lift	1520	1991	471	815	I
Duboka 2	Four-seat chair lift	1520	1904	384	1800	B
Duboka 1	Four-seat chair lift	1616	1976	360	1800	A
Gvozdac	T-bar	1626	1915	289	1191	A
Centar	Two-seat chair lift	1704	1976	272	2070	I
Pančićev vrh	Four-seat chair lift	1728	1976	248	2400	I
Gobelja greben	Four-seat chair lift	1730	1929	199	2400	I
Mali karaman	Four-seat chair lift	1731	1927	196	2400	B
Mali karaman	Platter lift	1731	1927	196	900	B
Suvo rudište	T-bar	1778	1970	192	1191	I
Marine vode	Platter lift	1740	1927	187	900	B
Gobelja rele	Platter lift	1754	1934	180	960	I
Karaman greben	Six-seat chair lift	1725	1904	179	3000	B
Kneževske bare	Platter lift	1768	1917	149	900	B
Karaman	Platter lift	1790	1934	144	900	B
Sunčana dolina	Four-seat chair lift	1640	1778	138	2070	B
Krst	Four-seat chair lift	1720	1823	103	2018	B
Jaram	Platter lift	1791	1859	68	900	B
Malo jezero	Platter lift	1715	1778	63	880	B
Mašinac	Platter lift	1734	1767	33	450	B

Difficulty levels are denoted as B: beginner (blue), I: intermediate (red), and A: advanced (black). Lifts are sorted by vertical ascent in descending order.

higher vertical ascents usually correspond to slopes with a higher degree of difficulty. Each ski lift on Mt Kopaonik has ski lift gates that read ski ticket data at the moment of entrance. Therefore, ski lift gate data for each skier are recorded. In this article, the term "skier" will be used as a generic term for all winter sport participants using ski lifts including skiers, snowboarders, and so on. The ski injury data have been obtained from the mountain rescue service and integrated with the ski lift transportation data.

The ski lift transportation data consist of the following attributes:

- Skier id
- Ski lift gate id
- Date and time of ski lift gate usage

Injury records include the skier id, date, and time of injury.

Skiers with a weekly ski ticket who were allowed to use the ski resort for the entire week with unlimited daily access were included in this study. They represent the largest population of ski lift ticket holders. Exclusion criteria were established for skiers who got injured on their first run because features shown on Table 2 could not be calculated for them. The dataset contained over 6 million ski lift transportations with 503,368 corresponding skier-days. There were 768 all-type injuries present in the data. One skier-day (skier-visit) holds statistics for 1-day skier activity. The average IPTSD over the six seasons was 1.53 which is consistent with injury rates reported in literature.[2] Extracted features from ski lift transportation data with corresponding formulas and descriptions are summarized in Table 2.

Methods

In this article, the authors utilized univariate and multivariate predictive models for ski injury prediction. To be comparable to other researchers, data analysis was performed in the same manner as others. Ruedl et al.[5] analyzed injuries of recreational skiers during four seasons in Austria using multivariate logistic regression with forward stepwise selection to derive adjusted ORs for attributes. The Mann–Whitney U-test was applied for attribute selection by selecting those attributes with a p-value less than or equal to 0.1 to control for confounding of features. This experimental setting was the baseline approach that was benchmarked against the CHAID decision tree analysis.

Results

In section "Descriptive statistics," descriptive statistics of features are shown, as well as a univariate data analysis of the features. In section "CHAID univariate analysis," a Mann–Whitney U-test and a CHAID will be used to test for statistically significant differences on non-injured and injured population distributions of features. In section "Multivariate analysis," logistic regression and a CHAID decision tree will be utilized for the multivariate data analysis.

Table 2. Ski lift transportation features.

Feature	Formula	Description
Normalized variability seeking index (NVSI)	$\frac{1}{N-1}\sum_{i=2}^{N} V_i$ $V_i = 0$ if $s(i) = s(i-1)$ $V_i = 1$, otherwise $s(i)$ presents the ski lift for the ith ski lift transportation; N is the ski lift transportation count	Describes ski lift switching behavior. It takes values from the [0, 1] interval. A higher value indicates that a skier changes ski lifts often.[16]
Normalized distinct lift count (NDLC)	$\frac{l+\lambda\bar{l}}{N+\lambda\bar{N}}$ l is the count of distinct ski lift visits $\bar{l}$ is the average population count of distinct ski lift visits $\bar{N}$ is the average population count of ski lift visits λ is the regularization parameter $[0, \infty]$	Describes ski lift switching behavior. Compared to NVSI, it takes into account distinct ski lifts and uses a regularization term to control for noisy data that might occur when there are a small number of ski lift transportations.
Normalized hourly familiarity index (NHFI)	$\frac{1}{N-1}\sum_{i=2}^{N} \frac{\bar{t}_p - t_i}{\bar{t}_p}$ $\bar{t}_p$ is the average hourly population descent time for a $s(i)$ and $s(i-1)$ ski lift pair t_i is the time between the ski lift gate check-in timestamp i and $i-1$	Measures whether a skier is faster or slower than the skier population on an hourly basis for a specific lift $(i-1)$ to lift (i) transition. Positive values indicate faster skiing while negative values indicate slower skiing. This measure is an adapted measure from Lathia et al.[17]
Elevation (AVG)	$\frac{1}{N}\sum_{i=1}^{N} e_i$ e_i presents the elevation for descent i	Average vertical meters per descent.
Time (AVG)	$\frac{1}{N}\sum_{h=1}^{N} t_h$ t_h presents the length of descent h measured in minutes	Average time per descent.
Time (SUM)	$time_N - time_1$ $time_N$ presents the timestamp of the last ski gate check-in time on a skier-day $time_1$ is the timestamp of the first ski gate entrance on a skier-day	Total time spent in the ski lift transportation system.

Table 3. Descriptive statistics of the extracted features.

Feature	Mean ± standard deviation
NVSI	0.4569 ± 0.2329
NDLC	0.3640 ± 0.0772
NHFI	−0.0741 ± 0.4282
Elevation (AVG), m	208.32 ± 59.17
Time (AVG), hh:mm	00:27 ± 00:13
Time (SUM), hh:mm	04:22 ± 01:35

NVSI: normalized variability seeking index; NDLC: normalized distinct lift count; NHFI: normalized hourly familiarity index.

Descriptive statistics

Table 3 provides basic statistics (mean and standard deviation) for features extracted from Table 2. The average time between two ski lift transportations is 27 min, while skiers on average spend 4 h 22 min in the ski lift transportation system. The average vertical descent between consecutive ski lift transportations is 208 m. Skiers express a 45.7% switching behavior in terms of normalized variability seeking index (NVSI) and 36.4% with respect to normalized distinct lift count (NDLC). The average speed of skiers, represented by normalized hourly familiarity index (NHFI), is 0, which is expected as this feature takes values in the [−1, 1] range and compares individual skiing times with the corresponding population.

In Figure 1, boxplots for each extracted feature are shown over the six analyzed seasons. Feature medians, as well as the first and the third quartiles, were quite stable during the six seasons, indicating the features were not sensitive to a specific season. However, a slightly decreasing trend in average skiing time was evident which is likely a result of ski lifts having been replaced by ones with higher capacity during the observed study period. This change reduced the average time between ski lift transportations.

Injured and non-injured population distributions for each feature are shown in Figure 2. Differences between the non-injured and injured population distributions are most evident on the total amount of time spent in the ski lift transportation system shown on the time (SUM) graph. Skiers tend to get injured early in the ski lift transportation system. On the NDLC, injured skiers demonstrate a higher switching behavior than the injured population. Clear separations between injured and non-injured population distributions on other features are

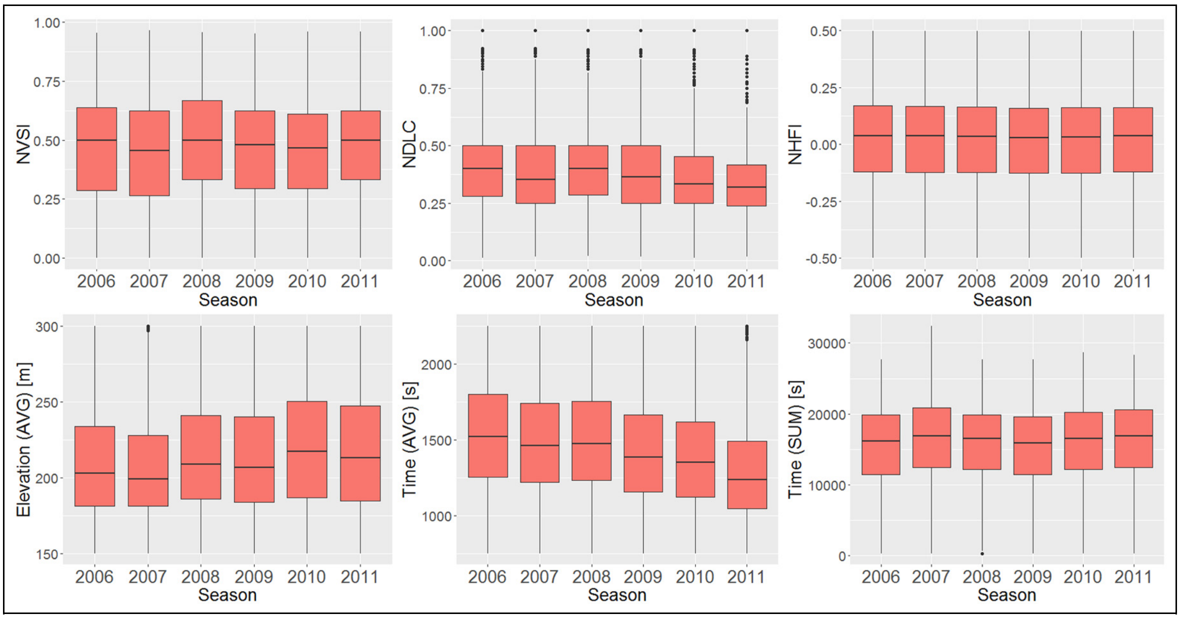

Figure 1. Boxplots of ski lift transportation features over six seasons.

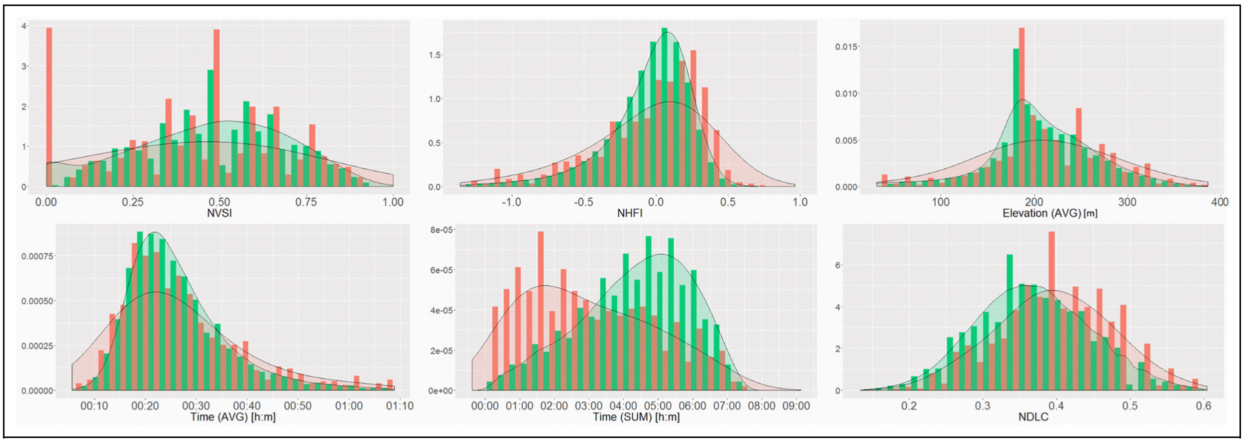

Figure 2. Histogram and density plots (red color represents injured skiers, green non-injured skiers).

not so obvious. Their role will become more clear in the multivariate models proposed in sections "Logistic regression analysis" and "CHAID decision tree analysis" where it will be shown that several features can also discriminate soundly on the injured and non-injured population, yet in conjunction with other features.

The Mann–Whitney *U*-test was used to test the hypothesis whether there were statistically significant differences between the injured and non-injured skier populations for various ski lift transportation features. The test does not assume a normal distribution of features. As shown in Table 4, the most significant differences are observed for NVSI, NDLC, and time (SUM). Time (AVG), elevation (AVG), and NHFI are not significantly different according to this test, yet it will be shown later that these features might express significance when in conjunction with other features. This is why pre-filtering of features based on such test could be deceiving.

CHAID univariate analysis

The Mann–Whitney *U*-test provides a global assessment on whether statistically significant differences exist for features in the injured and non-injured population. The CHAID algorithm,[6,22] besides being a method that produces decision tree models, can be used as a tool for searching for significant differences on segments of a continuous feature utilizing the chi-square test. In Figure 3, the detected significantly different feature segments known as bins are presented, where the height of a bin represents the IPTSD. The green bars present the population with IPTSD below the average and red bars show the injured population equal to or greater than the IPTSD. For the average IPTSD of 1.53 which is assumed to be the decision threshold on whether a skier is in the higher injury risk group or lower injury risk group, the ORs with 95% confidence intervals, true-positive rate (TPR), and false-positive

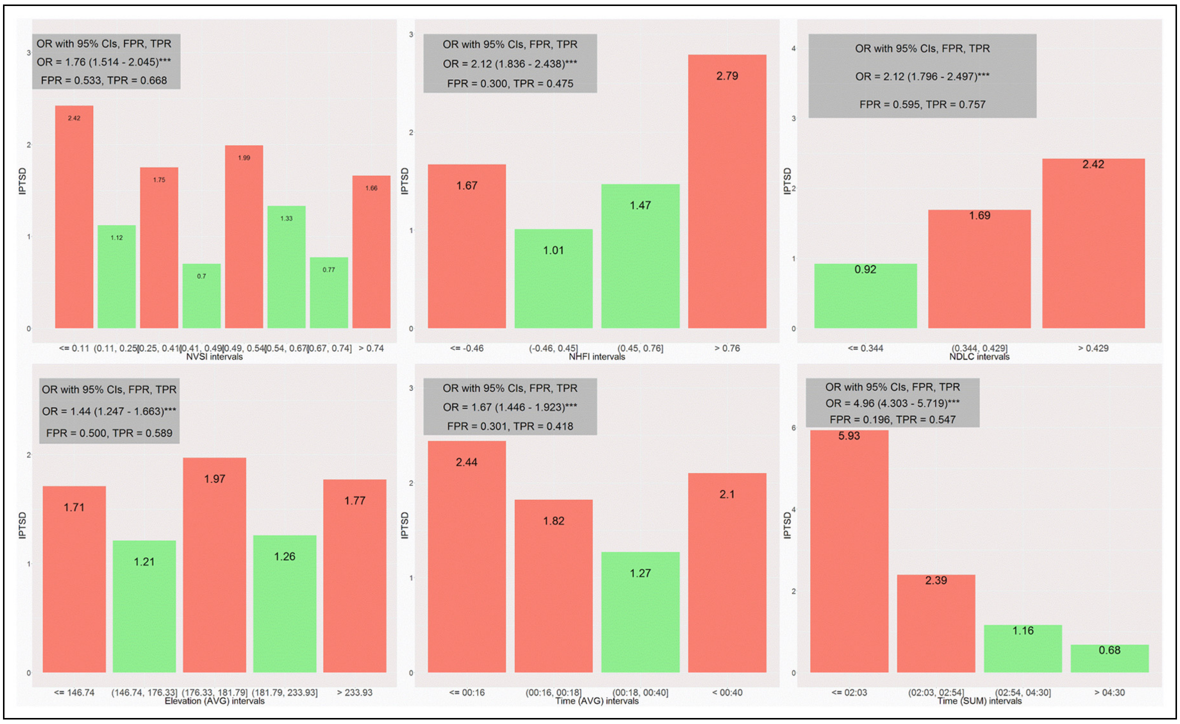

Figure 3. Univariate decision models identified by CHAID.

Table 4. Mann–Whitney *U*-test (characteristics of factors associated with injury).

Feature	Asymptotic significance (two-tailed significance)
NVSI	0.001***
NDLC	0.000***
NHFI	0.492
Elevation (AVG)	0.333
Time (AVG)	0.080
Time (SUM)	0.000***

NVSI: normalized variability seeking index; NDLC: normalized distinct lift count; NHFI: normalized hourly familiarity index.
*$p < 0.05$; **$p < 0.01$; ***$p < 0.001$.

rate (FPR) are given in gray boxes in the upper left corner of each histogram.

CHAID univariate predictive models (decision trees with one level called decision stumps) produce ORs ranging from 1.44 (elevation (AVG)) to 4.96 (time (SUM)) which is a satisfactory segmentation for univariate models. TPR ranged from 41.8% to 75.7% while FPR varied from 19.6% to 59.5%. All identified univariate models achieved a chi-square significance of $p < 0.001$.

One benefit of the CHAID decision stumps is the ability to identify non-linear and non-monotonic dependence in data. Figure 3 shows univariate decision models identified by CHAID, where red represents higher risk of injury and green represents lower risk of injury. On Figure 3, the NVSI, NHFI, elevation (AVG), and time (AVG) express non-linear and non-monotonic behavior. NHFI and time (AVG) have a quadratic form relationship with the response variable which is a common behavior in many advents. For example, the time (AVG) feature reveals that skiers with a shorter time (faster skiers) and longer time (slower skiers) have a higher risk of injury, while medium-speed skiers are at lowest risk of injury. Time (SUM), although having a monotonic relationship with the response variable, expresses a non-linear, exponential relationship. Logistic regression models assume a log linear and monotonic relationship between the predictors and the response variable, so they do not deal properly with features that do not have such properties. On the other hand, decision trees do not require such assumptions and are more appropriate to use when features express non-linear and non-monotonic behavior.

NVSI and elevation (AVG) express the most complex relationship toward the response variable. Such behavior might be an indicator that there are clusters in data. For example, skiers who use different ski lifts would show different behavior. Ski lifts may provide access to several or just a few other ski lifts and the lifts also have intrinsic characteristics of vertical elevations. Therefore, skiers who only use several ski lifts in a specific part of the mountain during a day could express different injury patterns than skiers who use several ski lifts on a different part of the mountain. This is in accordance with the findings of Greve et al.[23] who report that ski injuries are ski resort–dependent.

As shown in Figure 3, time (SUM) expresses an exponential dependence toward injury with the distribution resembling a Weibull early-failure distribution, which models the "time-to-failure" data. As ski injury often occurs early in the ski lift transportation system and the distribution function shows exponential decay, ski injury can be assumed to be an early-failure event. Still, studying ski injuries as early-failure events is out of the scope of this article, so it will be left for verification by future studies.

The Mann–Whitney *U*-test has recognized significant differences in distributions for NVSI, NDLC, and time (SUM), yet it failed to identify that there are significant differences for NHFI, elevation (AVG), and time (AVG) on specific intervals of features.

Multivariate analysis

Multivariate modeling was also applied during this study. Logistic regression and the CHAID decision tree algorithm were applied as they both can produce classifiers, yet with different models. Logistic regression produces models which assume a log linear and monotonic relationship between the predictors and the response variable, while CHAID assumes hierarchical splits of dependencies. Another reason why the CHAID decision tree was used is because it can build models for injury detection on highly imbalanced datasets. Decision trees, such as CART and C4.5, do not have such capabilities because their split growing measures cannot identify significant differences on highly imbalanced populations. To obtain fair performance, evaluation training and testing sets were created. The training set was a random stratified sample on 80% of the whole population, while the remaining 20% were selected for testing. Models were evaluated using OR, AUC, TPR, and FPR. The decision threshold was set at 0.153% which was the average IPTSD, meaning that an injury event was predicted if the probability was greater than or equal to 0.153%. A decision threshold identification procedure proposed by Shi and Abdel-Aty[19] was also tested and the proposed decision thresholds were 0.132% for CHAID and 0.142% for logistic regression. The proposed threshold for CHAID did not influence the classification results and the classification results for logistic regression were slightly worse. Thus, the average IPTSD threshold was adopted for both models.

Logistic regression analysis. Multivariate logistic regression with forward stepwise selection derived adjusted ORs for features as in Ruedl et al.[5] The Mann–Whitney *U*-test was applied for attribute selection by selecting those attributes with a *p*-value less than or equal to 0.1 to control for confounding.[5] In Table 5, logistic regression coefficients with 95% confidence intervals, as well as level of significance, are given. Based on the chosen classification threshold, the AUC, OR, TPR, and FPR were calculated.

Logistic regression achieved AUC 73.27%, OR 4.68, while TPR was 68.35% with a corresponding FPR of 31.57%.

Table 5. Logistic regression coefficients (95% CI).

Feature	Adjusted odds ratios
NVSI	−0.13666* (−0.26173, −0.01280)
NDLC	0.28898*** (0.14870, 0.42990)
Time (SUM)	−0.72327*** (−0.81945, −0.62793)
Intercept	−6.85506*** (−6.96256, −6.75173)

NVSI: normalized variability seeking index; NDLC: normalized distinct lift count.
$p < 0.1$; *$p < 0.05$; **$p < 0.01$; ***$p < 0.001$.

Forward selection eliminated time (AVG), while NHFI and elevation (AVG) were not considered since they were not significant. Logistic regression coefficients suggest that the increase in time (SUM) decreases the injury risk, which means that skiers spending more time in the ski resort are at smaller risk of getting injured. This could also mean that more physically prepared skiers are less prone to injuries. These findings are consistent with results obtained in similar types of research[24–26] which identify that better physical preparation reduces the likelihood of injury.

However, the increase in NDLC which indicates switching behavior increases the chances of injury as well. On the other hand, higher NVSI values signify injury risk decrease. These conclusions are, at first glance, conflicting, as both features indicate switching behavior. Nevertheless, NVSI considers switching behavior between two consecutive ski lift transportations (i.e. it only considers whether the same lift was used for the previous and current ski lift transportation). On the contrary, NDLC takes into consideration the relative number of distinct lifts a skier has used while staying in the ski lift transportation system. Obviously, these two features do not explain the same kind of switching behavior. NDLC signifies that a higher relative use of distinct ski lifts increases the chance of injury. This could be due to more exposure of skiers to different ski slopes and snow conditions. This type of behavior cannot be captured with NVSI which could identify skiers who select only a few, nearby ski lifts, switching often between them, indicating a tendency to stick to a few ski lifts and trails during a skier-day. This conclusion is strengthened by the fact that the features do not have a variance inflation factor greater than 4 (time (SUM) = 1.5752, NDLC = 2.8169, and NVSI = 3.0504), indicating that interpretation of coefficients is not distorted by multicollinearity.[27]

CHAID decision tree analysis. A CHAID decision tree can build models for injury detection on highly imbalanced datasets, as can logistic regression. Decision trees, such as CART and C4.5, do not have such capabilities because their split growing measures cannot identify significant differences on highly imbalanced populations. CHAID grows a decision tree model on features

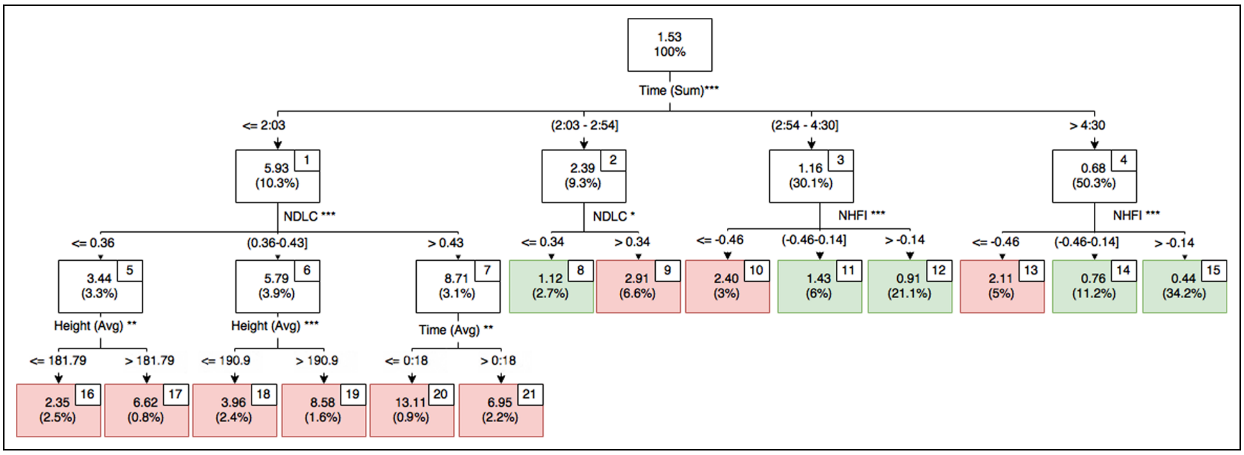

Figure 4. CHAID decision tree for ski injury prediction. Red boxes are decision leaves with IPTSD greater than or equal to the average IPTSD of 1.53. Green boxes denote decision leaves with IPTSD less than the average IPTSD.

which maximize the chi-square statistics, which has the same basis that is used for OR significance calculation. For all these reasons, CHAID was found as a sound choice for ski injury prediction.

A common pitfall of decision trees is that they produce overfitted models that have a high generalization error.[28] To avoid overfitting, the parameters were selected for the CHAID modeling. Minimal skiers in the parent node (preceding) and minimal skiers in the child node (subsequent) were set to 6000 and 4000, respectively, while the maximal tree depth was fixed to 3. The CHAID decision tree model achieved an AUC of 73.78%, OR 6.51, while TPR was 65.82% and FPR 22.83%. The AUC gives an overall assessment of the prediction quality of models based on TPR and FPR, but independent on the decision threshold. Both logistic regression and CHAID achieved a similar AUC, CHAID being better by 0.5%. Still, when comparing OR, TPR, and FPR, which are all dependent on the decision threshold, CHAID outperformed logistic regression (i.e. on the decision threshold of 0.153%, CHAID better discriminated between the injured and non-injured population). Although logistic regression achieved a 3% higher TPR, this resulted in almost a 9% higher FPR.

The decision tree model shown in Figure 4 splits the entire population, shown by the root node with a 1.53 IPTSD, into hierarchically nested subgroups with corresponding population shares and injury rates. A decision model consists of nodes, subgroups of the entire population, and branches and rules for discriminating between nodes, where the final nodes on each branch trace are denoted as leaves. CHAID identified in total 21 nodes, excluding the root node, with 14 decision leaves. Each tree leaf is also characterized by the risk of injury, given by the IPTSD, as well as the percentage of the population for which a decision rule applies. For example, node 6 represents skiers who spend less than 2 h 3 min in the ski resort and whose NDLC is in the (0.36–0.43] range. They have an injury rate of 5.79 and include 3.9% of the population. Red boxes show leaves whose injury rates are greater than or equal to the 1.53 average injury rate (risky skiing patterns) while the green boxes present the leaves whose injury rates are less than the average (safe skiing patterns).

CHAID recognized time (SUM) as the most important attribute for ski injury segmentation. As the time spent in the ski lift transportation system increased, the injury rate dropped. The finding that the total daily time spent in the ski lift transportation system discriminated best between the injured and non-injured population was also suggested by the logistic regression model in section "Logistic regression analysis." This assumption could wrongly imply that skiers who spend less daily time in the ski resort are at higher risk of injury. As noted in Figure 4, the decision tree model identified that 10.3% of the population that spent less than or equal to 2 h 3 min in the ski lift transportation system each day had an injury rate of 5.93. This is a conclusion that cannot be made based on this research as the setup could not evaluate whether skiers who ski less are at higher risk of getting injured. What was observed with a high level of confidence was that ski injuries occurred early in the ski lift transportation system.

For skiers who spend more than 2:54 h in the ski lift transportation system (nodes 3 and 4), it is important to take into consideration NHFI when assessing the injury risk. The results from nodes 10–12 and 13–15 indicate that the injury rate decreased as NHFI (speed of skiing) increased. The injury rates of faster skiers in nodes 12 and 15 were three to five times lower than injury rates of slower skiers in nodes 10 and 13, respectively. This might be explained by advanced and beginner skiers being included in the population and similar rates were reported for beginner and advanced skiers in literature. Ekeland et al.[11] report that beginner skiers have an injury rate three times higher than experts, while Laporte et al.[10] reported beginner skiers having injury rates five times higher than experts.

For skiers who spend less than or equal to 2:54 h in the ski lift transportation system (nodes 1 and 2), the decision tree model chooses the switching behavior

(NDLC) as the next most important feature for risk classification. Higher switching behavior is consistently related to higher injury risk. Nodes 9 and 7 have two and three times higher than the average IPTSD, respectively.

For the group of skiers who spend the least amount of time in the system, less than or equal to 2:03 h (node 1) with the highest switching behavior population (node 7), the average time needed between two ski lift transportations needs to be considered. Skiers who take less than or equal to 18 min between two ski lift transports (node 20) are at highest risk of injury, more than eight times higher than the average ski injury rate. This subgroup, which has 0.9% of the population, has the highest observed injury rate. This is the population with the highest switching behavior and smallest amount of time between two ski lift transports. Skiers who ski less than or equal to 2:03 h with a NDLC less than or equal to 0.43 (nodes 6 and 5) discriminate risk best on the average height of ski lift descents. Skiers with larger average descents in nodes 17 and 19 have injury risks two to three times greater than skiers with smaller average descents in nodes 18 and 16.

Although NVSI was proposed by the Mann–Whitney *U*-test as a significant feature which was also included in the logistic regression model, it was not included in the CHAID model. In addition, features that were determined not to be significant according the Mann–Whitney *U*-test, including NHFI, elevation (AVG), and time (SUM), were included in the CHAID model. This is because decision tree models allow detection of interactions between the predictors and the outcome variable on sub-samples of the population. Identifying conjuncted interactions between predictors and the response variables cannot be recognized with standard logistic regression models. NHFI would not be included in a logistic regression model; still, in conjunction with time (SUM), this predictor demonstrates an important mechanism of ski injury occurrence.

Conclusion

One might expect that injury occurs at a constant rate throughout the entire skiing day. However, most injuries occur in the first couple of hours of skiing, independent of whether a skier-day has started at the opening time of the ski lift transportation system or later during the day. As ski injuries occur early in the ski lift transportation system, they can be characterized as early-failure events. Skiers are mostly injured in the first couple of hours of skiing, and the more risky mechanisms of injury can be explained through higher switching behavior, higher average vertical descents, and shorter average time between ski lift transportations. The results of this research imply that skiers need advice on how to stay injury free for a longer period of time in the system. This need could be addressed by development of safety guidelines on how to behave in the first couple of hours of skiing when the risk of injury is the highest. Higher switching behavior (in terms of NDLC), higher average descents, and smaller amounts of time in the transportation system have shown to induce the highest injury rates. Also, some of the interpreted factors correspond to previous studies. However, this study derived the behavior features from ski transportation data, rather than from questionnaires, which is more readily available. Additionally, ski transportation data include a large sample of control data points, unlike previous case–control studies. The models could be applied to a much broader population.

Several features were proposed which can be used when analyzing ski injury occurrence based on ski lift transportation data. This article emphasized the potential of using the CHAID algorithm for analysis of ski injury which provides an interpretable-rich supplement to logistic regression models, which are usually regarded as baseline models in ski injury research. CHAID also allowed for the discovery of factors which are independently not significant, but jointly show significance regarding the response variable.

This article proposed several ski injury prediction models based on ski lift transportation features. Still, demographic-aware predictors, meteorological predictors, ski slope condition predictors, and so on would certainly enrich and improve these models. This research opens several future research directions. It would be interesting to analyze the potential of ski lift transportation data for early ski injury prediction. Injury days prior to injury could be used to analyze the injury risk as well. Near-real-time ski injury prediction would also be a valuable contribution. That would mean collecting ski lift transportation data in real-time and providing advice to skiers for safer skiing. To conclude, ski lift transportation data provide opportunities for the creation of injury prediction models which can be used for prevention of ski injury and reduction in injury rate.

Acknowledgements

The authors are grateful to the Ski Resorts of Serbia for providing data for this research and for their support throughout the research.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This research was partially funded by the US Department of State CIES Fulbright Visiting

Program grant, conducted at the Center for Data Analysis and Biomedical Informatics (DABI) at Temple University.

References

1. The National Ski Areas Association. National skier/snowboarder visits, 1979–2016, http://www.nsaa.org/media/275017/1516_visits.pdf (accessed 20 March 2017).
2. Vanat L. International report on snow & mountain tourism: overview of the key industry figures for ski resorts, 2017, http://www.vanat.ch/RM-world-report-2017-vanat.pdf
3. Ruedl G, Kopp M, Sommersacher R, et al. Factors associated with injuries occurred on slope intersections and in snow parks compared to on-slope injuries. *Accident Anal Prev* 2013; 50: 1221–1225.
4. Ruedl G, Abart M, Ledochowski L, et al. Self reported risk taking and risk compensation in skiers and snowboarders are associated with sensation seeking. *Accident Anal Prev* 2012; 48: 292–296.
5. Ruedl G, Helle K, Tecklenburg K, et al. Factors associated with self-reported failure of binding release among ACL injured male and female recreational skiers: a catalyst to change ISO binding standards? *Brit J Sport Med* 2016; 50(1): 37–40.
6. Kass GV. An exploratory technique for investigating large quantities of categorical data. *Appl Stat* 1980; 29: 119–127.
7. Biggs D, De Ville B and Suen E. A method of choosing multiway partitions for classification and decision trees. *J Appl Stat* 1991; 18(1): 49–62.
8. Chamarro A and Fernández-Castro J. The perception of causes of accidents in mountain sports: a study based on the experiences of victims. *Accident Anal Prev* 2009; 41(1): 197–201.
9. Hume PA, Lorimer AV, Griffiths PC, et al. Recreational snow-sports injury risk factors and countermeasures: a meta-analysis review and Haddon matrix evaluation. *Sports Med* 2015; 45(8): 1175–1190.
10. Laporte JD, Bajolle L, Lamy D, et al. Winter sports injuries in France over two decades. In: Johnson RJ, Shealy JE, Greenwald RM, et al. (eds) *Skiing trauma and safety: 19th volume*. West Conshohocken, PA: ASTM International, 2012, pp.201–215.
11. Ekeland A, Sulheim S and Rodven A. Injury rates and injury types in alpine skiing, telemarking, and snowboarding. *J ASTM Int* 2005; 2(5): 1–9.
12. Langran M and Selvaraj S. Increased injury risk among first-day skiers, snowboarders, and skiboarders. *Am J Sport Med* 2004; 32(1): 96–103.
13. Hopkins CL, Youngquist ST, Johnson E, et al. Outcome of elderly patients injured at winter resorts. *Am J Emerg Med* 2011; 29(5): 528–533.
14. Siu TL, Chandran KN, Newcombe RL, et al. Snow sports related head and spinal injuries: an eight-year survey from the neurotrauma centre for the Snowy Mountains, Australia. *J Clin Neurosci* 2004; 11(3): 236–242.
15. Bohanec M and Delibašić B. Data-mining and expert models for predicting injury risk in ski resorts. In: *International conference on decision support system technology*, Belgrade, 27 May 2015, pp.46–60. Berlin: Springer International Publishing.
16. D'Urso P and Massari R. Fuzzy clustering of human activity patterns. *Fuzzy Set Syst* 2013; 215: 29–54.
17. Lathia N, Smith C, Froehlich J, et al. Individuals among commuters: building personalised transport information services from fare collection systems. *Pervasive Mob Comput* 2013; 9(5): 643–664.
18. Froehlich J, Neumann J and Oliver N. Sensing and predicting the pulse of the city through shared bicycling. In: *Proceedings of the 21st international joint conference on artificial intelligence (IJCAI'09)*, Pasadena, CA, 11–17 July 2009, vol. 9, pp.1420–1426. New York: ACM.
19. Shi Q and Abdel-Aty M. Big data applications in real-time traffic operation and safety monitoring and improvement on urban expressways. *Transport Res C: Emer* 2015; 58: 380–394.
20. King MA, Abrahams AS and Ragsdale CT. Ensemble methods for advanced skier days prediction. *Expert Syst Appl* 2014; 41(4): 1176–1188.
21. Dalipi F, Mendoza DM, Imran AS, et al. An intelligent model for predicting the occurrence of skiing injuries. In: *2015 5th national symposium on information technology: towards new smart world (NSITNSW)*, Riyadh, Saudi Arabia, 17–19 February 2015, pp.1–7. New York: IEEE.
22. Loh WY. Classification and regression trees. *Wires Data Min Knowl* 2011; 1(1): 14–23.
23. Greve MW, Young DJ, Goss AL, et al. Skiing and snowboarding head injuries in 2 areas of the United States. *Wild Environ Med* 2009; 20(3): 234–238.
24. Raschner C, Platzer HP, Patterson C, et al. The relationship between ACL injuries and physical fitness in young competitive ski racers: a 10-year longitudinal study. *Brit J Sport Med* 2012; 46: 1065–1071.
25. Hébert-Losier K and Holmberg HC. What are the exercise-based injury prevention recommendations for recreational alpine skiing and snowboarding? *Sports Med* 2013; 43(5): 355–366.
26. Schmitt KU, Hörterer N, Vogt M, et al. Investigating physical fitness and race performance as determinants for the ACL injury risk in Alpine ski racing. *BMC Sports Sci Med Rehabil* 2016; 8(1): 23.
27. O'Brien RM. A caution regarding rules of thumb for variance inflation factors. *Qual Quant* 2007; 41(5): 673–690.
28. Dietterich T. Overfitting and undercomputing in machine learning. *ACM Comput Surv* 1995; 27(3): 326–327.