J Occup Rehabil (2013) 23:597–609
DOI 10.1007/s10926-013-9430-4

Development of a Computer-Based Clinical Decision Support Tool for Selecting Appropriate Rehabilitation Interventions for Injured Workers

Douglas P. Gross · Jing Zhang · Ivan Steenstra · Susan Barnsley · Calvin Haws · Tyler Amell · Greg McIntosh · Juliette Cooper · Osmar Zaiane

Published online: 7 March 2013

Abstract *Purpose* To develop a classification algorithm and accompanying computer-based clinical decision support tool to help categorize injured workers toward optimal rehabilitation interventions based on unique worker characteristics. *Methods* Population-based historical cohort design. Data were extracted from a Canadian provincial workers' compensation database on all claimants undergoing work assessment between December 2009 and January 2011. Data were available on: (1) numerous personal, clinical, occupational, and social variables; (2) type of rehabilitation undertaken; and (3) outcomes following rehabilitation (receiving time loss benefits or undergoing repeat programs). Machine learning, concerned with the design of algorithms to discriminate between classes based on empirical data, was the foundation of our approach to build a classification system with multiple independent and dependent variables. *Results* The population included 8,611 unique claimants. Subjects were predominantly employed (85 %) males (64 %) with diagnoses of sprain/strain (44 %). Baseline clinician classification accuracy was high (ROC = 0.86) for selecting programs that lead to successful return-to-work. Classification performance for machine learning techniques outperformed the clinician baseline classification (ROC = 0.94). The final classifiers were multifactorial and included the variables: injury duration, occupation, job attachment status, work status, modified work availability, pain intensity rating, self-rated occupational disability, and 9 items from the SF-36 Health Survey. *Conclusions* The use of machine learning classification techniques appears to have resulted in classification performance better than clinician decision-making. The final algorithm has been integrated into a computer-

Electronic supplementary material The online version of this article (doi:10.1007/s10926-013-9430-4) contains supplementary material, which is available to authorized users.

D. P. Gross (✉)
Department of Physical Therapy, University of Alberta, 2-50 Corbett Hall, Edmonton, AB T6G 2G4, Canada
e-mail: dgross@ualberta.ca

D. P. Gross · C. Haws
Workers' Compensation Board-Alberta Millard Health, Edmonton, AB, Canada

J. Zhang · O. Zaiane
Department of Computing Science, University of Alberta, Edmonton, AB, Canada

I. Steenstra
Institute for Work and Health, Toronto, ON, Canada

I. Steenstra
Dalla Lana School of Public Health, University of Toronto, Toronto, ON, Canada

S. Barnsley
Workers' Compensation Board-Alberta Health Care Services, Edmonton, AB, Canada

T. Amell
Centric Health, Calgary, AB, Canada

G. McIntosh
CBI Health, Toronto, ON, Canada

J. Cooper
University of Manitoba, Winnipeg, MB, Canada

based clinical decision support tool that requires additional validation in a clinical sample.

Keywords Rehabilitation · Musculoskeletal diseases · Compensation and redress · Machine learning · Classification · Prediction

Background

Work-related musculoskeletal conditions are some of the most burdensome health conditions internationally in terms of personal, societal and economic costs [1–3]. While the majority of individuals with such conditions return to work quickly, a small minority remain off work for longer periods of time and are responsible for the majority of associated health care and compensation costs [4]. Ideally, those most at risk of delayed recovery and return-to-work (RTW) should be identified and effective interventions targeted towards this high-risk group implemented [5–7]. This is the process of risk-factor classification, which involves assigning patients to appropriate interventions or rehabilitation programs based on individual characteristics and clinical findings [8]. Selecting the optimal intervention that will lead to successful RTW remains a challenging task for clinicians and funders of health care.

As mentioned, many work-related musculoskeletal conditions recover quickly and spontaneously, and often do not require treatment beyond reassurance and advice to stay active [9–11]. It appears that intensive exercise and rehabilitation in the first few weeks may be counter-productive and actually delay recovery in some cases [12, 13]. However, too large of a delay between injury and appropriate rehabilitation has also been found associated with delayed recovery and RTW [14]. Most evidence-based guidelines recommend further assessment and intervention if recovery has not occured within 4–6 weeks to avoid progression to chronic pain and disability [11, 15], In this sub-acute stage various types of rehabilitation interventions have been recommended including physical conditioning or functional restoration programs [16], worksite-based interventions [17], and interdisciplinary biopsychosocial rehabilitation (i.e. chronic pain management programs) [18].

Systematic reviews of these interventions have reported some supportive evidence [16–18]. However, individual response to these interventions is highly variable with some patients benefiting greatly while others do not improve or experience worsening of their problems [19]. Currently, clinicians are unable to identify with complete accuracy what patients will respond best to the various rehabilitation options and referral is often made in a trial-and-error fashion [8]. Individual factors such as expectations of recovery and previous health care seem to predict in general who will respond to rehabilitation [20, 21], but only recently have researchers explored whether specific characteristics can be used to target interventions to achieve optimal outcomes [6, 7, 22–24]. One major limitation of these studies is that most are limited to patients with low back pain. While back pain makes up the largest diagnostic category of injured workers, it typically accounts for less than 40 % of all claimants off work with musculoskeletal conditions [25, 26]. Development of broader classification algorithms are needed to address the needs of all injured workers requiring rehabilitation.

Considerable efforts have been made to increase awareness and appreciation among clinicians of the psychological, social and workplace factors contributing to musculoskeletal disability [5, 27–29]. Unfortunately, uptake in clinical practice has been slow, partly because easy-to-use tools that integrate these factors into routine clinical decision making are lacking [30–32]. Advances in computing technology have allowed the development of expert computer-based clinical decision support (CDS) tools, such as apps for handheld computing devices [33]. These are attractive options given the widespread utilization of handheld computing devices such as tablet computers and smartphone devices. Computers have the ability to rapidly execute advanced statistical algorithms that integrate many factors into decision-making, matching individual patient characteristics with a knowledge base to provide patient-specific recommendations [34]. The use of computerized CDS has the potential to greatly augment clinician decisions [34–37] and was found beneficial in the diagnosis of back pain and use of lumbar spine imaging modalities [38, 39]. To our knowledge computer-based CDS tools have not been evaluated for their ability to augment treatment selection in the realm of work-related musculoskeletal disorders.

The purpose of our project was to develop a classification algorithm and accompanying computer-based CDS tool to help categorize individuals who were not working due to a wide variety of musculoskeletal disorders. The overall aim was to categorize workers based on individual, work-related and clinical characteristics according to their likelihood of successful RTW following various types of rehabilitation. We examined a variety of characteristics likely to influence RTW including self-report measures of physical, functional and psychosocial factors.

Methods

Design

A population-based historical cohort design was used, with data extracted from a Canadian provincial compensation database (WCB-Alberta). Ethical approval was obtained from the University of Alberta's Health Research Ethics Board.

Population

Province-wide data was available on all 8,611 WCB-Alberta claimants with musculoskeletal injuries referred to RTW assessment facilities throughout the province. WCB-Alberta's administrative database was augmented by clinical data from rehabilitation providers who are contracted to file reports at time of claimants' admission and discharge from rehabilitation programs. For this study, data were extracted on all claimants undergoing RTW assessment within the jurisdiction between December 1, 2009 and January 1, 2011.

Rehabilitation of Injured Workers Within the Alberta Context

For the past decade, rehabilitation of injured workers in Alberta has been undertaken within the context of a continuum of care model (shown in Appendix I (ESM)), which has been defined as a coordinated array of settings, services, providers, and care levels in which health, medical, and supportive services are provided in the most appropriate care setting [40]. Details of the Alberta model, its province-wide implementation, and a description of its positive impact on return to work outcomes have been discussed elsewhere [41]. In brief, the model includes a staged approach to rehabilitation, from single service providers through interdisciplinary rehabilitation and complex pain management programs. Progress through the various stages is based on a duration model of occupational disability [4, 42]. Minimal intervention is advocated in early stages followed by interdisciplinary RTW assessment if required.

Claimants are referred for RTW assessment when they have met or surpassed expected injury healing times (i.e. 4–8 weeks for soft tissue injury, longer for specific injuries), have not yet returned to work (not functionally improving or returning to work following a course of single-service physical therapy, chiropractic or medical care), and/or report ongoing difficulties related to their compensable condition. This predominantly represents claimants in the sub-acute phase of recovery, but at times claimants with injuries of longer duration are also assessed. These claimants are evaluated for purposes of: (1) determining readiness to RTW; and (2) need for additional rehabilitation. Based on clinician interpretation of assessment findings, claimants are referred to the most appropriate rehabilitation program. This RTW assessment process and the various tools used for determining readiness for RTW have been previously studied [43–47]. However, little research has been done examining decisions regarding what type of rehabilitation, if any, is required to facilitate recovery and return-to-work.

Various rehabilitation options are currently available that include:

1. Provider Site-Based Program: Interdisciplinary functional restoration at a designated rehabilitation facility. Treatment focuses largely on graded activity, functional restoration, and specific exercise programs, but also includes communication/negotiation with relevant stakeholders such as employers.
2. Worksite-based Rehabilitation: In this program all intervention takes place at the worksite instead of at a rehabilitation facility. Treatment focuses more on maintaining linkages with the workplace, participatory ergonomics and identification of suitable duties to help claimants stay at work.
3. "Hybrid" Functional Restoration/Worksite-based Rehabilitation: This is a combination of provider and worksite based programs. Claimants spend time at both the workplace and rehabilitation facility for treatment. This option is commonly used for claimants with cumulative activity related disorders as opposed to traumatic injuries.
4. Complex Interdisciplinary Biopsychosocial Rehabilitation: This is a comprehensive pain management program for claimants with chronic pain and multiple complex barriers to RTW. Treatment includes counseling psychology sessions to improve coping, decrease stress and overcome emotional burdens, functional restoration with a cognitive-behavioural approach, and RTW planning through stakeholder negotiation.
5. 'Other' intervention: This involves either no rehabilitation or referral back to a single service provider (i.e. physical therapy or chiropractic).

In all of the rehabilitation programs education is provided related to the nature of the injury, normal healing phases and strategies for coping with pain. For this study, information was extracted on the specific rehabilitation option(s) undertaken by each claimant in the database. We also extracted outcome status (RTW or other) following the program and duration of the program.

Measures

All data for this study was extracted from archived WCB-Alberta provincial databases. Data were available on: (1) numerous personal, clinical, occupational, and social variables measured at time of RTW Assessment; (2) type of rehabilitation undertaken (as just described) and whether a repeat program was needed; and (3) RTW outcomes following the intervention. Specific measures from the RTW assessment included:

1. Duration of injury: calendar days between accident and admission for rehabilitation
2. Age in years
3. Sex
4. Marital status
5. If an interpreter was required
6. Education level
7. Job attached status (whether the worker had a job to return to)
8. Currently working? Yes or no.
9. Modified duties available? Yes or no.
10. Number of previous compensation claims
11. Primary diagnosis: ICD9 code
12. If comorbidity were present?
13. Mechanism of injury code [48]
14. Nature of injury code [48]
15. Part of Body code [48]
16. National Occupational Classification code [49]
17. Pain VAS: Pain severity was rated on a visual analogue scale in which the client places a mark on a line, anchored by 0 and 10, to measure their level of pain most of the time. Zero corresponds to no pain and 10 to the worst pain imaginable. Analogue pain scales have been previously shown to have adequate reliability and validity and to be consistently associated with RTW [50].
18. Pain Disability Index (PDI)—The PDI is a commonly used self-report questionnaire measuring perceived disability due to pain [51]. Pollard initially developed this tool and it has since been tested extensively in a variety of populations [52–56]. The PDI asks claimants to rate their level of disability on a 0–10 scale on seven areas of activity: family/home responsibility, recreation, social activity, occupation, sexual behaviour, self-care, and life-support activity. Scores on these seven items are summed and a total out of 70 is obtained. The higher the individual's score, the higher the individual's level of perceived disability. The PDI has been reported to have good psychometric properties, such as test–retest reliability, concurrent validity and internal consistency [52–56]. Individual PDI items were used in analysis of the current study.
19. SF-36v2: This is a commonly used patient assessed generic health outcome measure [57]. It consists of eight scaled scores, which are weighted sums of the items in their section [58]. The eight scales include vitality, physical functioning, bodily pain, general health perceptions, physical role functioning, emotional role functioning, social role functioning, and mental health. These scales are used to calculate physical and mental health summary measures. The SF-36 has previously been used in research with individuals with musculoskeletal conditions and work disability [59–62]. Individual SF-36 items were used in analysis of the current study.

Data on the rehabilitation program recommended by the assessing clinician was also available for each claimant, as well as the anticipated duration of each program. Duration was used in calculating a 'clinician baseline' for accuracy in making treatment recommendations for comparison to the machine learning algorithm.

Outcomes

Since our primary goal was to develop a classification model of successful RTW, we extracted RTW outcome measures commonly used in program evaluation within the WCB-Alberta jurisdiction. This included number of days receiving wage replacement benefits after assessment. For sufficient variability in the measure and to ensure that outcomes were linked to the rehabilitation program, we examined whether the worker was receiving wage replacement benefits at 30 days after assessment. Reception of benefits is a surrogate indicator of RTW, but is commonly used as an outcome within studies of compensation claimants. We also extracted information on whether the worker underwent a subsequent rehabilitation program. Successful RTW was judged when workers were no longer receiving wage replacement benefits at 30 days after assessment and did not undertake a subsequent program during the one-year follow-up.

Analysis

Descriptive statistics were calculated including means and standard deviations for continuous variables, modes and percentages for categorical variables. Significant differences across the rehabilitation programs were determined using Chi square and t-tests. Machine learning, concerned with the design of algorithms to discriminate between classes based on empirical data, was the foundation of our approach to build a classification system with multiple independent and dependent variables [63]. Machine Learning can help overcome human errors inherent in making complex classifications using multiple features or variables. The goal of machine learning is to build a concise model of the distribution of class labels in terms of predictive features [63]. In our case, the 'label' is the appropriate rehabilitation program for each claimant and the predictive features are the various measures available on each claimant at time of assessment. We will describe

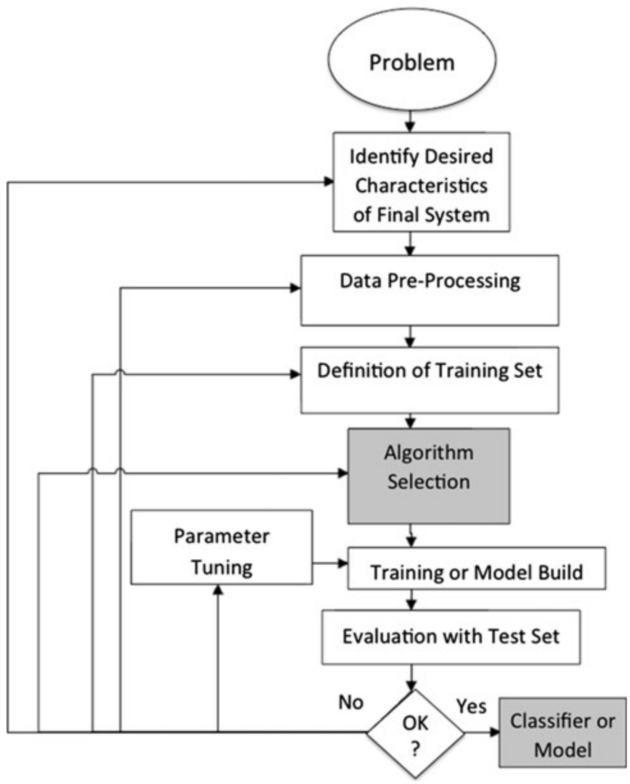

Fig. 1 Process of machine learning (modified from Kotsiantis [63])

the process of model development (i.e. build) and evaluation (i.e. test) that involved identifying desirable characteristics of the final decision support system, data pre-processing, machine learning algorithm selection and 'training' (i.e. model build) and finally evaluation of the model and comparison with the clinician baseline (See Fig. 1). Readers are referred to the review paper by Kotsiantis for more detail about machine learning and the process of model development and testing [63].

Desirable Characteristics of the Final Decision Support System

We wanted the final clinical decision support system to have some fundamental characteristics that would increase the likelihood of usage. We wanted the model to be interpretable and editable. Users should be able to insert their own expertise to modify the model and see the evidence supporting recommendations made by the system, therefore, rule-based algorithms were deemed more desirable. We also wanted the system to consider multiple recommendations so users could choose the most appropriate one under different considerations. Lastly, the system should use web-based technology to provide easy and ubiquitous access to users with all kinds of computer devices. These desirable characteristics guided our selection of analysis strategies.

Data Pre-processing

To recommend a rehabilitation program leading to successful RTW, the machine learning algorithm needed to 'learn' from data consisting of previous cases with a successful RTW result (off benefits at 30 days and no additional rehabilitation programs within 1 year). Thus, we extracted 4,876 successful cases from the dataset as the initial training/development dataset. The rest of the cases were used to train/develop another model of negative rules (i.e. rules indicating what program to avoid).

The training/development dataset consisted of five categories or 'classes', each representing a specific rehabilitation program; however, the class distribution in the dataset was severely imbalanced (very few claimants underwent complex, workplace-based or hybrid interventions). With the presence of class imbalance, a classification model normally will be biased towards the majority classes while the minority classes may be shadowed or be insufficient for learning. To resolve this problem we used an over-sampling technique on the minority classes to mitigate the severity of imbalance. To avoid overfitting and generalizing the decision boundary, we used the Synthetic Minority Over-sampling Technique (SMOTE) [64]. SMOTE is a procedure for dealing with imbalance across classes, and blends under-sampling of majority classes with over-sampling of minority classes. These methods are commonly used in machine learning computation to address imbalanced data [64]. Additionally, we used the Tomek Link method to overcome class overlaps after sampling in order to avoid minority data generated from SMOTE from invading the majority class too deeply and causing classification difficulties [65]. The Tomek Link method is an analytical procedure that removes borderline cases between classes and helps to reduce noise in the dataset, thereby resulting in optimal classification [65].

Since we had more variables or features than desired in our final model, Correlation-based Feature Subset Evaluation [66] with Linear Forward Search [67] feature selection process was undertaken. In general, this algorithm favors individual features that are highly correlated with the class but much less correlated with other features. After feature selection, all the numeric features are further discretized since many machine learning algorithms work better with discrete features. After feature selection, 30 variables remained in consideration in the final model.

Machine Learning Algorithm Development

Decision Trees (C4.5), Repeated Incremental Pruning to Produce Error Reduction (RIPPER), and Naïve Bayes Classifiers were all investigated; however, the best results were obtained with the RIPPER algorithm and this method

Table 1 Characteristics of injured workers in the database

	Entire sample (n = 8,611)	Full data on questionnaires (n = 7256)	Missing data (n = 1355)
Mean (SD) or percent			
Age (years)	42.8 (11.9)	42.8 (11.9)	43.0 (11.9)
Accident to admission[a] (days)	210.5 (419.9)	215.1 (426.1)	185.9 (384.1)
	Median = 71	Median = 74	Median = 61
Number of previous claims[a]	4.2 (5.3)	4.2 (5.3)	4.0 (5.1)
Sex (% male)	64	64	65
Education level			
Grade 8 or less	3	3	3
Partial high school	11	11	9
High school diploma	17	18	14
Partial technical school	5	5	3
Technical diploma	13	13	9
Partial university	3	3	2
University degree	5	5	5
Not specified[a]	44	42	55
Marital status			
Married/common law	37	39	29
Single	17	17	15
Divorced/separated	7	7	6
Widowed	1	1	1
Not specified[a]	38	36	50
Job Attached[a]	85	84	88
Currently working	46	46	46
Modified work available[a] (% yes)	55	54	58
Diagnosis			
Sprain/strain[a]	44	44	48
Joint disorder	28	29	24
Fracture	12	12	11
Contusion	5	5	5
Laceration	3	2	4
Dislocation	2	2	1
Nerve damage	1	2	1
Other	5	5	6
Part of body neck or back[a]	21	20	24
Comorbidity[a] (% with sec. diagnosis)	29	31	20
Interpreter required[a] (% yes)	3	3	5
Assessment recommendation[a]			
No intervention required	6	6	6
Single service provider	18	19	11
Provider-based RTW program	54	52	62
Worksite-based RTW program	1	2	2
Hybrid RTW program	9	9	11
Complex RTW program	4	4	2
Medical consult	3	3	3
Other	4	5	2
Actual program undertaken[a]			
No rehabilitation	19	19	18
Single service provider	17	18	13
Provider-based RTW program	50	50	55
Worksite-based RTW program	1	2	1
Hybrid RTW program	9	9	11
Complex RTW program	3	4	2
Receiving TTD benefits at assess (% yes)	45	46	44
TTD 30-days post assessment[a] (% yes)	25	26	23
Repeat program	8	8	8

[a] Statistically significant difference between those with and without missing data

was used to train the final classification model. RIPPER is an inductive rule-based learner that builds a set of rules to identify the classes while minimizing the amount of error. The error is defined by the number of training examples misclassified by the rules. RIPPER has generally been demonstrated to be more effective than decision-tree learners and more efficient on large samples [68]. RIPPER generates classification rules that follow a specific logical formula that are easily interpreted [69]. The output is a series of first-order logical statements following an 'IF—THEN' framework for recommendations (see Appendix II (ESM) for more information). RIPPER forms rules through a process of repeated *growing* and *pruning*. During the growing phase, the rules are made more restrictive in order to fit the training data as closely as possible. During the pruning phase, the rules are made less restrictive in order to avoid overfitting, which can cause poor classification performance [69]. Concisely, the algorithm proceeds iteratively starting with an empty rule set, and in each iteration, the training data is split into a growing set and a pruning set, then a rule is grown from the growing set and immediately pruned or simplified based on the pruning set. If the error rate of the new rule on the pruning set does not exceed some threshold, the rule is added to the rule set representing the learned model and all examples in the training data covered by this rule are removed before being split again for the next repetition. Otherwise, the iteration is stopped and the rule set is returned. Growing a rule consists of starting with an empty conjunction of conditions and considers adding any condition of the form $A_n = v$, $A_c \leq \theta$, or $A_c \geq \theta$, where A_n is a nominal attribute and v is a legal value of A_n, or A_c is a continuous variable and θ is some value for A_c that occurs in the growing set. The condition that maximizes any information gain criteria is

Table 2 Characteristics of claimants in various rehabilitation programs undertaken

	No rehab n = 1,640	Community PT n = 1,448	Functional restoration n = 4,331	Worksite-based rehab n = 124	Hybrid n = 774	Complex pain n = 294
Mean (SD) or percent						
Age (years)	43.6 (12.2)	42.9 (12.1)	42.5 (11.7)	41.3 (12.7)	42.4 (11.6)	44.6 (10.5)
Accident to admission (days)	394.2 (636.8)	144.4 (289.7)	150.9 (291.5)	103.1 (254.2)	119.1 (193.6)	675.6 (780.0)
Number of previous claims	4.4 (5.6)	4.3 (5.2)	4.2 (5.2)	3.0 (3.8)	3.6 (5.0)	4.5 (5.2)
Sex (% male)	70	62	64	53	57	67
Education level						
Grade 8 or less	2	2	3	2	2	8
Partial high school	10	11	12	5	7	17
High school diploma	18	18	17	7	14	17
Partial technical school	5	6	5	0	3	3
Technical diploma	12	15	12	8	12	12
Partial university	3	3	2	0	3	3
University degree	4	5	5	4	7	6
Not specified	44	39	44	74	53	34
Marital status						
Married/common law	38	42	36	17	33	44
Single	17	18	17	9	12	20
Divorced/separated	7	8	6	5	6	8
Widowed	1	1	1	1	1	1
Not specified	37	32	39	69	48	27
Job attached	74	90	86	99	98	59
Currently working	48	50	38	75	84	16
Modified work available (% yes)	48	60	52	73	81	33
Diagnosis						
Sprain/strain	38	46	45	51	54	29
Joint disorder	25	27	28	32	30	41
Fracture	14	10	12	7	6	17
Contusion	6	6	5	6	4	2
Laceration	5	2	2	2	2	2
Dislocation	2	3	2	0	1	3
Nerve damage	3	2	1	1	1	3
Other	8	4	5	3	3	4
Part of body neck or back	17	11	23	19	23	27
Comorbidity/(% with secondary diagnosis)	32	32	28	12	19	43
Interpreter required (% yes)	3	3	3	3	2	5
Assessment recommendation						
No intervention required	25	7	1	1	1	1
Single service provider	28	59	4	3	3	5
Provider-based RTW program	24	14	91	7	13	11
Worksite-based RTW program	3	1	1	86	1	0
Hybrid RTW program	5	3	1	3	81	0
Complex RTW program	2	1	0	0	0	82
Medical Consult	4	10	1	0	0	0
Other	10	7	2	0	2	2
Receiving TTD benefits at assessment (% yes)	29	44	56	20	10	77
TTD benefits 30-days post assessment (% yes)	13	25	32	0	2	74
Repeat program	7	19	6	1	2	12

continually added until the rule covers no negative examples from the growing dataset. The algorithm uses an information gain function that measures the expected reduction in entropy caused by adding the attribute. Information gain represents the expected amount of information that would be needed to specify whether a new instance should be classified into one class or the other, which is based on the notion of entropy measuring the impurity of a variable. The information gain is equal to the total entropy for an attribute if for each of the attribute values a unique classification can be made for the result attribute. High entropy means there is uniform distribution of the variable values in the sample leading to lack of predictability of the values, while low entropy means the distribution varies and has peaks and valleys leading to predictability.

To prune a rule, RIPPER considers deleting any final sequence of conditions from the rule and chooses the deletion that maximizes the function $(p - n)/(p + n)$ where p is the number of positive examples and n is the number of negative examples in the pruning set. Finally, RIPPER post-processes the rule set by re-pruning or simplifying each rule in the order they were learned. A rule is simplified by deleting any single condition chosen in such a way that the resulting new set of rules yields the greatest reduction of error on the training set [68]. RIPPER handles multiple classes by ordering them from least to most prevalent and then treating them each sequentially as distinct two-class problems (i.e. individual rehabilitation program vs. others).

In Machine Learning, the algorithm 'learns' from successfully classified cases within a 'build' dataset. The contrast is formed from the classes (i.e. rehabilitation programs) within the data. Multiple classes are ordered from least to most prevalent and then treated sequentially as individual two-class problems (i.e. individual rehabilitation program vs. others). In our multi-class situation (i.e. multiple rehabilitation programs could be selected), the rules generated from the RIPPER algorithm are ranked in ascending order based on the number of examples in the class. Unknown instances are tested against the rules in that order. The first rule that covers the test instance "fires" and the testing phase ends. Since we were expecting the model to make multiple recommendations, we modified the algorithm to track all of the rules that cover a given test instance and grouped together the rules recommending the same program. In this way, the model generated a 'rehabilitation intervention' recommendation pool from which the user could choose based on different criteria that will be discussed.

Using data from claimants with unsuccessful outcomes, we also developed a set of negative rules by applying the same procedures as described above. The negative rules apply for all programs except for Worksite-based rehabilitation since there was only one instance of failure in the training data, which resulted in insufficient power for analysis. The duration of a rehabilitation program was also considered an important factor to consider when making recommendations; therefore, we built a linear regression model with 20 features that included the 19 features extracted in the earlier section along with the 'Program' label included as a covariate. This linear regression model predicts the duration of the rehabilitation program based on the clinical characteristics of the claimant and the program undertaken.

Evaluation of Machine Learning Model and Clinician Baseline

The clinician baseline was defined as the rate of clinician recommendations of the 'correct' rehabilitation program. The clinician recommendation was deemed successful when the recommendation matched the actual program undertaken, led to successful RTW and no repeat program was undertaken. Both clinician baseline and the machine learning model were evaluated using Sensitivity, Specificity, the Geometric Mean of Sensitivity and Specificity, and Receiver Operating Characteristic (ROC) Area. We examined each of these measurements with tenfold cross validation [70], which is a technique to estimate the predictive ability of a model in practice. In tenfold cross validation, the original dataset is randomly partitioned into 10 subsets. A single subset is used as a validation set while the other 9 subsets are used as training set. The process repeats 10 times as each subset is used as the validation set exactly once. Since this was a multi-label classification problem, with multiple rehabilitation program options, each program was considered as a binary classification problem (i.e. Sensitivity and Specificity were calculated for each program individually). All the results are then averaged to provide a single estimation of Sensitivity, Specificity and the ROC calculated from the 'validation' datasets. The true and false positive scores for these calculations were obtained by determining the proportion of cases classified by the model compared to actual labels (i.e. actual rehabilitation programs) for those cases within the validation dataset.

Sensitivity = True Positive/True Positives + False Negatives
True Positives = Situations where the classifier correctly labeled the claimant as requiring a particular rehab program and in fact that is the program they underwent
False Negative = Situations where the classifier incorrectly labeled the claimant as NOT requiring a particular rehab program when in fact that is the program they underwent
Specificity = True Negative/True Negative + False Positive
True Negative = Situations where the classifier correctly labels the claimant as NOT requiring a particular rehab program and in fact they did not undertake that program
False Positive = Situations where the classifier incorrectly labeled the claimant as requiring a particular rehab program when in fact they did not undertake that program

Descriptive analyses were conducted with SPSS 19 while machine learning analyses were conducted using the Waikato Environment for Knowledge Analysis (WEKA) [71, 72].

Results

Population Characteristics

The dataset included 8,611 unique injured workers with compensation claims for a wide variety of musculoskeletal disorders. Subjects were predominantly employed (85 %) males (64 %) with diagnoses of sprain/strain (44 %). Table 1 provides a full description of characteristics and Table 2 displays full characteristics of those subjects undergoing the various rehabilitation programs. The largest group of subjects (50 %) underwent functional restoration programs, followed by no rehabilitation (19 %), community physical therapy, and hybrid programs (9 %). Small minorities underwent chronic pain (3 %) or workplace based (1 %) rehabilitation. A portion of subjects (16 %) did not have complete data on the self-report questionnaires. Due to the large sample sizes, these subjects had a large number of statistically significant differences from those without missing data. Differences deemed clinically meaningful included a shorter duration of injury (186 vs. 215 days), fewer comorbidities (20 % vs. 31 %) and involvement of an interpreter (5 % vs. 3 % in those with missing data).

Clinician Baseline

Table 3 shows the accuracy of clinician baseline recommendations, which were accurate with an average sensitivity, specificity and ROC of 0.81, 0.95 and 0.86 respectively. Baseline clinician classification performance was high (sensitivity = 0.81, specificity = 0.95, ROC = 0.86) for selecting programs that led to successful RTW.

Data Pre-processing, Feature Selection and Discretization

During the SMOTE procedure, the final sampling percentage obtained from each minority class was: Complex: 900 %, Worksite-Based: 900 % and Hybrid: 300 %. Figure 2 depicts the final class distribution comparison after all data processing was completed. During feature selection, nineteen features were identified and discretized.

Machine Learning Model Development and Evaluation

Using the RIPPER algorithm, seventeen features were identified as important in the final rule set. The final classifiers were multifactorial and included the variables: job attachment and working status at time of RTW assessment, availability of modified work, National Occupational Classification Code, ICD9 diagnostic group, calendar days injury to assessment, the 'Occupation' item from the PDI, Pain VAS out of 10, and the following SF36 items: 2, 4, 5, 7, 12, 14, 18, 21, 25). Appendix II (ESM) lists all of the unique rules.

Table 3 Performance of the 'clinician baseline' recommendation compared to the final machine learning algorithm

	Sensitivity	Specificity	Geometric mean	ROC area
Provider-based functional restoration	0.86/0.98	0.85/0.88	0.85/0.93	0.86/0.94
Complex pain management program	0.75/0.94	0.994/0.992	0.86/0.96	0.87/0.97
Worksite-based program	0.89/0.76	0.99/0.99	0.94/0.87	0.94/0.94
Hybrid (functional restoration with workplace component)	0.81/0.96	0.97/0.99	0.89/0.97	0.90/0.98
Other (single service or no rehabilitation)	0.75/0.62	0.91/0.98	0.83/0.78	0.83/0.86
Weighted averages	0.81/0.89	0.95/0.97	0.88/0.93	0.86/0.94

[a] All values represent clinician/machine learning algorithm performance

Table 3 compares the final tenfold cross validation analysis on this final processed dataset shown in comparison to the clinician baseline recommendations. Machine performance was substantially higher than clinician decisions (sensitivity = 0.89, specificity = 0.97 and ROC = 0.94). The rule model consists of 61 rules: 24 rules for worksite-based programs, 11 rules for complex pain management programs, 10 rules for hybrid programs, 13 rules for 'other', and 1 default rule for provider-based functional restoration programs. Classification performance for machine learning techniques in the final processed dataset outperformed the clinician baseline (sensitivity = 0.89, specificity = 0.97, ROC = 0.94). Table 4 shows the tenfold cross validation of the negative rule set. The negative rule set consists of 21 rules. Overall performance was high (sensitivity = 0.95, specificity = 0.98 and ROC = 0.95).

Discussion

The use of machine learning classification techniques resulted in classification performance greater than clinician

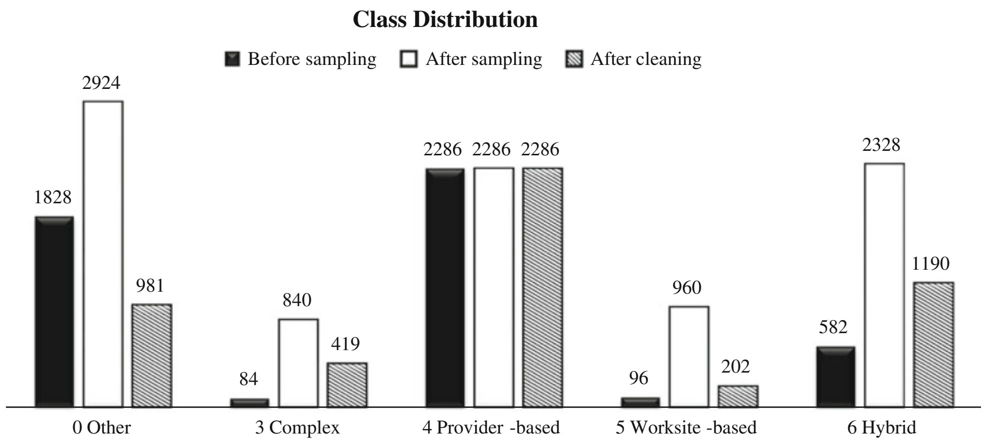

Fig. 2 Final class distribution comparison after all data pre-processing was completed

Table 4 Performance of the negative rule training data

	Sensitivity	Specificity	Geometric mean	ROC area
Provider-based functional restoration	0.98	0.91	0.94	0.95
Complex pain management program	0.97	0.996	0.98	0.98
Hybrid (functional restoration with workplace component)	0.96	1	0.98	0.98
Other (single service or no rehabilitation)	0.85	0.99	0.91	0.93
Weighted averages	0.95	0.98	0.96	0.95

decision-making for selecting rehabilitation programs that lead to successful RTW. Both positive and negative rules were identified and internal validation indicated high levels of performance (ROC > 0.9). The factors included in the rule sets are also theoretically supported because they include several of the variables identified in a previous hypothetical risk-factor model for classification of patients with work-related back pain [7]. Factors included in the final rule set include whether the worker has a job to return to, whether modified duties were available, duration of the condition/injury, pain intensity levels, and reported disturbance in physical functioning.

Computer Interface and Development of Clinical Decision Support Tool

The final algorithm was integrated into a computer-based clinical decision support tool that requires additional validation in a clinical sample. This Clinical Decision Support (CDS) tool is a Web-based application that is accessible on all platforms with JavaScript enabled. Appendix III (ESM) provides a screenshot picture of the tool. The tool integrates both the positive and negative models generated by the RIPPER algorithm. To use the tool to make recommendations for a certain claimant, the user simply selects a value in each of the drop-down menus or enters a value in the textbox according to the characteristics of that claimant, and then presses the “Predict” button. Recommendations made by the system are presented under the data entry form. Each recommendation has the following attributes:

- The name of the rehabilitation program recommended.
- The prediction of the program duration if the claimant were to utilize that program.
- The average rule confidence of the underlying rules supporting a recommendation (interpreted as a probability from 0 (zero likelihood of success) to 1 (complete likelihood of success).
- The number of rules supporting a recommendation.

The user can also view the details of the underlying rules by pressing the “Rules” button. These features provide a way to measure the quality of the recommendations in order for users to make clinical decisions with more confidence. To determine whether the clinical decision support tool augments and improves human decision-making, randomized controlled trials will be needed to compare clinical outcomes when clinicians make decisions that are informed by the tool versus routine clinical decisions not informed by the CDS tool.

Resolving Rule Conflict

Since both the positive and negative models are integrated into the system, it is possible that a given instance is covered by both positive and negative rules, thereby making contradictory recommendations. To resolve this conflict, users can compare different metrics of the recommendations or view the specific rules to decide to accept or reject a recommendation.

Limitations of this research include the reliance on data already available in the WCB-Alberta datasets. Other variables may have allowed for more accurate rule sets; however, a wide variety of variables were available and this resulted in high levels of classification performance. Another limitation is the severe imbalance in rehabilitation classes. An imbalance in the proportions of subjects undergoing the various rehabilitation programs posed some difficulty for this research, and necessitated the use of SMOTE techniques. While not ideal, these techniques are commonly used in the development of rule sets in machine learning analyses [64]. Given the increasing evidence for workplace involvement in the RTW process [16, 73, 74], it was surprising that so few workers underwent workplace-based interventions. But employer involvement and worksite evaluations are typically integrated into all of the rehabilitation programs from the outset and often occur as part of the other rehabilitation programs in the Alberta jurisdiction [41].

Conclusions

The use of machine learning classification techniques appears to have resulted in classification performance higher than clinician decision-making. The final algorithm has been integrated into a computer-based clinical decision support tool that requires additional validation and impact evaluation in clinical samples, ideally through randomized controlled trials.

Acknowledgments The Workers' Compensation Board of Alberta facilitated data collection and provided funding in support of this research.

References

1. Lane R, Desjardins S. Canada, population and public health branch. Strategic policy directorate. Policy research division. Economic burden of illness in Canada, 1998. Ottawa: Health Canada; 2002.
2. Hadler NM. Occupational musculoskeletal disorders. 3rd ed. Philadelphia: Lippincott Williams & Wilkins; 2005.
3. Martin BI, Deyo RA, Mirza SK, et al. Expenditures and health status among adults with back and neck problems. JAMA. 2008;299:656–64.
4. Frank JW, Kerr MS, Brooker AS, et al. Disability resulting from occupational low back pain. Part I: what do we know about primary prevention? A review of the scientific evidence on prevention before disability begins. SPINE. 1996;21:2908–17.
5. Shaw WS, van der Windt DA, Main CJ, Loisel P, Linton SJ. Early patient screening and intervention to address individual-level occupational factors ("blue flags") in back disability. J Occup Rehabil. 2009;19:64–80.
6. Steenstra IA, Ibrahim SA, Franche RL, Hogg-Johnson S, Shaw WS, Pransky GS. Validation of a risk factor-based intervention strategy model using data from the readiness for return to Work Cohort Study. J Occup Rehabil. 2009;20:394–405.
7. Shaw WS, Linton SJ, Pransky G. Reducing sickness absence from work due to low back pain: how well do intervention strategies match modifiable risk factors? J Occup Rehabil. 2006;16:591–605.
8. Haldorsen EM. The right treatment to the right patient at the right time. Occup Environ Med. 2003;60:235–6.
9. Dahm KT, Brurberg KG, Jamtvedt G, Hagen KB. Advice to rest in bed versus advice to stay active for acute low-back pain and sciatica. Cochrane Database Syst Rev. 2010;6:CD007612.
10. van Tulder MW, Tuut M, Pennick V, Bombardier C, Assendelft WJ. Quality of primary care guidelines for acute low back pain. SPINE. 2004;29:E357–62.
11. AAMPG Group. Evidence-based management of acute musculoskeletal pain: a guide for clinicians. Bowen Hills: Australian Government National Health and Medical Research Council; 2004.
12. Sinclair SJ, Hogg-Johnson SH, Mondloch MV, Shields SA. The effectiveness of an early active intervention program for workers with soft-tissue injuries. The Early Claimant Cohort Study. SPINE. 1997;22:2919–31.
13. Cassidy JD, Carroll LJ, Cote P, Frank J. Does multidisciplinary rehabilitation benefit whiplash recovery? Results of a population-based incidence cohort study. SPINE (Phila Pa 1976). 2007;32:126–31.
14. McIntosh G, Frank J, Hogg-Johnson S, Bombardier C, Hall H. Prognostic factors for time receiving workers' compensation benefits in a cohort of patients with low back pain. SPINE. 2000;25:147–57.
15. van Tulder M, Becker A, Bekkering T, et al. Chapter 3. European guidelines for the management of acute nonspecific low back pain in primary care. Eur Spine J. 2006;15(Suppl 2):S169–91.
16. Schaafsma F, Schonstein E, Whelan KM, Ulvestad E, Kenny DT, Verbeek JH. Physical conditioning programs for improving work outcomes in workers with back pain. Cochrane Database Syst Rev. 2010;20:CD001822.
17. van Oostrom SH, Driessen MT, de Vet HC, et al. Workplace interventions for preventing work disability. Cochrane Database Syst Rev. 2009;15:CD006955
18. Karjalainen K, Malmivaara A, van Tulder M, et al. Multidisciplinary biopsychosocial rehabilitation for subacute low back pain among working age adults. Cochrane Database Syst Rev. 2003;2:CD002193.
19. Gross DP, Haws C, Niemelainen R. What is the rate of functional improvement during occupational rehabilitation in workers' compensation claimants? J Occup Rehabil. 2012;22:292–300.
20. Gross DP, Battie MC. Predicting timely recovery and recurrence following multidisciplinary rehabilitation in patients with compensated low back pain. SPINE. 2005;30:235–40.
21. Sandstrom J, Esbjornsson E. Return to work after rehabilitation. The significance of the patient's own prediction. Scand J Rehabil Med. 1986;18:29–33.
22. Steenstra IA, Knol DL, Bongers PM, Anema JR, van Mechelen W, de Vet HC. What works best for whom? An exploratory, subgroup analysis in a randomized, controlled trial on the effectiveness of a workplace intervention in low back pain patients on return to work. SPINE (Phila Pa 1976). 2009;34:1243–9.
23. Gatchel RJ, Polatin PB, Noe C, Gardea M, Pulliam C, Thompson J. Treatment- and cost-effectiveness of early intervention for acute low -back pain patients: a one-year prospective study. J Occup Rehab. 2003;13:1–9.
24. Hill JC, Whitehurst DG, Lewis M, et al. Comparison of stratified primary care management for low back pain with current best

practice (STarT Back): a randomised controlled trial. Lancet. 2011;378:1560–71.
25. Occupational injuries and diseases in Alberta: 2004–2008: Government of Alberta Employment and Immigration; 2009.
26. WorkSafeBC website Accessed January 15, 2013, at http://www2.worksafebc.com/Topics/Ergonomics/BackPainBasics.asp.
27. Vranceanu AM, Barsky A, Ring D. Psychosocial aspects of disabling musculoskeletal pain. J Bone Joint Surg Am. 2009;91:2014–8.
28. Noonan J, Wagner SL. A biopsychosocial perspective on the management of work-related musculoskeletal disorders. AAOHN J. 2010;58:105–14.
29. Nicholas MK, George SZ. Psychologically informed interventions for low back pain: an update for physical therapists. Phys Ther. 2011;91:765–76.
30. Miciak M, Gross DP, Joyce A. A review of the psychotherapeutic 'common factors' model and its application in physical therapy: the need to consider general effects in physical therapy practice. Scand J Caring Sci. 2012;26:394–403
31. Nicholas MK, Linton SJ, Watson PJ, Main CJ. Early identification and management of psychological risk factors ("yellow flags") in patients with low back pain: a reappraisal. Phys Ther. 2011;91:737–53.
32. Fritz JM, Beneciuk JM, George SZ. Relationship between categorization with the STarT back screening tool and prognosis for people receiving physical therapy for low back pain. Phys Ther. 2011;91:722–32.
33. Clinical Prediction Rules: A Physical Therapy Reference. USBMIS, Inc. Accessed January 15, 2013, at http://itunes.apple.com/us/app/clinical-prediction-rules/id439974538?mt=8.
34. Souza NM, Sebaldt RJ, Mackay JA, et al. Computerized clinical decision support systems for primary preventive care: a decision-maker-researcher partnership systematic review of effects on process of care and patient outcomes. Implement Sci. 2011;6:87.
35. Sahota N, Lloyd R, Ramakrishna A, et al. Computerized clinical decision support systems for acute care management: a decision-maker-researcher partnership systematic review of effects on process of care and patient outcomes. Implement Sci. 2011;6:91.
36. Roshanov PS, Misra S, Gerstein HC, et al. Computerized clinical decision support systems for chronic disease management: a decision-maker-researcher partnership systematic review. Implement Sci. 2011;6:92.
37. Hemens BJ, Holbrook A, Tonkin M, et al. Computerized clinical decision support systems for drug prescribing and management: a decision-maker-researcher partnership systematic review. Implement Sci. 2011;6:89.
38. Blackmore CC, Mecklenburg RS, Kaplan GS. Effectiveness of clinical decision support in controlling inappropriate imaging. Journal of the American College of Radiology : JACR. 2011;8:19–25.
39. Lin L, Jen-Hwa H, Sheng ORL. A decision support system for lower back pain diagnosis: uncertainty management and clinical evaluations. Decis Support Syst. 2006;42:1152–69.
40. Position of the American Dietetic Association. Nutrition, aging, and the continuum of care. J Am Diet Assoc. 2000;100:580–95.
41. Stephens B, Gross DP. The influence of a continuum of care model on the rehabilitation of compensation claimants with soft tissue disorders. SPINE. 2007;32:2898–904.
42. Frank JW, Brooker AS, DeMaio SE, et al. Disability resulting from occupational low back pain. Part II: what do we know about secondary prevention? A review of the scientific evidence on prevention after disability begins. SPINE. 1996;21:2918–29.
43. Ikezawa Y, Battie MC, Beach J, Gross D. Do clinicians working within the same context make consistent return-to-work recommendations? J Occup Rehabil. 2010;20:367–77.
44. Gross DP, Battie MC, Cassidy JD. The prognostic value of functional capacity evaluation in patients with chronic low back pain: part 1: timely return to work. SPINE. 2004;29:914–9.
45. Gross DP, Battie MC, Asante AK. The Patient-Specific Functional Scale: validity in workers' compensation claimants. Arch Phys Med Rehabil. 2008;89:1294–9.
46. Gross DP, Battie MC. Recovery expectations predict recovery in workers with back pain but not other musculoskeletal conditions. J Spinal Disord Tech. 2010;23:451–6.
47. Gross DP, Batti MC. Does functional capacity evaluation predict recovery in workers compensation claimants with upper extremity disorders? Occup Environ Med. 2006;63:404–10.
48. Canadian Standards Association. Coding of work injury or disease information. ON, Canada: Mississauga; 2003.
49. National Occupational Classification: Occupational descriptions: Human Resources Development Canada; 2001.
50. Finch E, Brooks D, Stratford P, Mayo N. Physical rehabilitation outcome measures: a guide to enhanced clinical decision making. 2nd ed. Toronto: Canadian Physiotherapy Association; 2002.
51. Pollard CA. Preliminary validity study of the pain disability index. Percept Mot Skills. 1984;59:974.
52. Gronblad M, Jarvinen E, Hurri H, Hupli M, Karaharju EO. Relationship of the Pain Disability Index (PDI) and the Oswestry Disability Questionnaire (ODQ) with three dynamic physical tests in a group of patients with chronic low-back and leg pain. Clin J Pain. 1994;10:197–203.
53. Tait RC, Pollard CA, Margolis RB, Duckro PN, Krause SJ. The Pain Disability Index: psychometric and validity data. Arch Phys Med Rehabil. 1987;68:438–41.
54. Kang SW, Lee WN, Moon JH, Chun SI. Correlation of spinal mobility with the severity of chronic lower back pain. Yonsei Med J. 1995;36:37–44.
55. Gronblad M, Hurri H, Kouri JP. Relationships between spinal mobility, physical performance tests, pain intensity and disability assessments in chronic low back pain patients. Scand J Rehabil Med. 1997;29:17–24.
56. Chibnall JT, Tait RC. The Pain Disability Index: factor structure and normative data. Arch Phys Med Rehabil. 1994;75:1082–6.
57. McHorney CA, Ware JE Jr, Raczek AE. The MOS 36-Item Short-Form Health Survey (SF-36): II. Psychometric and clinical tests of validity in measuring physical and mental health constructs. Med Care. 1993;31:247–63.
58. Ware JE, Gandek B. The SF-36 Health Survey: development and use in mental health research at the IQLOA project. Int J Mental Health. 1994;23:73.
59. Grevitt M, Khazim R, Webb J, Mulholland R, Shepperd J. The short form-36 health survey questionnaire in spine surgery. J Bone Joint Surg Br. 1997;79:48–52.
60. Guilfoyle MR, Seeley H, Laing RJ. The Short Form 36 health survey in spine disease–validation against condition-specific measures. Br J Neurosurg. 2009;23:401–5.
61. Krousel-Wood MA, McCune TW, Abdoh A, Re RN. Predicting work status for patients in an occupational medicine setting who report back pain. Arch Fam Med. 1994;3:349–55.
62. Chen C, Hogg-Johnson S, Smith P. The recovery patterns of back pain among workers with compensated occupational back injuries. Occup Environ Med. 2007;64:534–40.
63. Kotsiantis SB. Supervised machine learning: a review of classification techniques. Informatica. 2007;37:249–68.
64. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: synthetic minority over-sampling technique. J Artif Intell Res. 2002;16:321–57.
65. Batista GEAPA, Prati RC, Monard MC. A study of the behavior of several methods for balancing machine learning training data. In: ACM SIGKDD explorations newsletter—special issue on learning from imbalanced datasets; 2004. p. 20–9.

66. Hall MA. Correlation-based feature selection for discrete and numeric class machine learning. In: 17th international conference on machine learning. Stanford, CA; 2000. p. 359–66.
67. Gutlein M. Large scale attribute selection using wrappers. [Masters' Thesis]. Freiburg: Albert-Ludwigs-Universitat; 2006.
68. Cohen WW. Fast effective rule induction. In: Twelve international conference of machine learning (ICML-95); 1995. p. 115–23.
69. Duma M, Twala B, Marwala T, Newamondo FV. Improving the performance of the ripper in insurance risk classification—A Comparitive Study using feature selection. In: 8th international conference on informatics in control, automation and robotics. Noordwikerhout, The Netherlands; 2011. p. 203–10.
70. Kohavi R. A study of cross-validation and bootstrap for accuracy estimation and model selection. In: Proceedings of the fourteenth international joint conference on artificial intelligence; 1995; San Mateo, California: Morgan Kaufmann; 1995. p. 1137–43.
71. Holmes G, Donkin A, Witten IH. Weka: a machine learning workbench. In: Second Australia and New Zealand conference on intelligent information systems; Brisbane, Australia; 1994. p. 357–61.
72. Witten IH, Frank E, Hall MA. Data mining: practical machine learning tools and techniques. Part III: The weka data mining workbench, 3rd edn. San Francisco, USA: Morgan Kaufmann; 2011.
73. Franche RL, Cullen K, Clarke J, Irvin E, Sinclair S, Frank J. Workplace-based return-to-work interventions: a systematic review of the quantitative literature. J Occup Rehabil. 2005;15: 607–31.
74. Franche RL, Severin CN, Hogg-Johnson S, Cote P, Vidmar M, Lee H. The impact of early workplace-based return-to-work strategies on work absence duration: a 6-month longitudinal study following an occupational musculoskeletal injury. J Occup Environ Med. 2007;49:960–74.