Journal of Occupational Rehabilitation
https://doi.org/10.1007/s10926-019-09843-4

Validity of the Work Assessment Triage Tool for Selecting Rehabilitation Interventions for Workers' Compensation Claimants with Musculoskeletal Conditions

Douglas P. Gross[1] · Ivan A. Steenstra[2] · William Shaw[3] · Parnian Yousefi[4] · Colin Bellinger[5] · Osmar Zaïane[4]

Abstract

Purpose The Work Assessment Triage Tool (WATT) is a clinical decision support tool developed using machine learning to help select interventions for patients with musculoskeletal disorders. The WATT categorizes patients based on individual characteristics according to likelihood of successful return to work following rehabilitation. A previous validation showed acceptable classification accuracy, but we re-examined accuracy using a new dataset drawn from the same system 2 years later. *Methods* A population-based cohort design was used, with data extracted from a Canadian compensation database on workers considered for rehabilitation between January 2013 and December 2016. Data were obtained on demographic, clinical, and occupational characteristics, type of rehabilitation undertaken, and return to work outcomes. Analysis included classification accuracy statistics of WATT recommendations. *Results* The sample included 28,919 workers (mean age 43.9 years, median duration 56 days), of whom 23,124 experienced a positive outcome within 30 days following return to work assessment. Sensitivity of the WATT for selecting successful programs was 0.13 while specificity was 0.87. Overall accuracy was 0.60 while human recommendations were higher at 0.72. *Conclusions* Overall accuracy of the WATT for selecting successful rehabilitation programs declined in a more recent cohort and proved less accurate than human clinical recommendations. Algorithm revision and further validation is needed.

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

Introduction

Musculoskeletal disorders are some of the most common and disabling health problems worldwide [1, 2]. Given the substantial human, economic and societal burden, improved health care and rehabilitation strategies are needed [3]. Especially important are strategies aimed at facilitating sustainable return to work (RTW) [4]. This need has led to the development of various treatments and rehabilitation programs aimed at enhancing work ability or job placement.

Ideally, workers at risk of delayed recovery and RTW would be identified early and effective interventions would be targeted towards this high-risk group [3, 5–7]. Many musculoskeletal conditions recover quickly and often do not require treatment beyond reassurance and advice to stay active [8–10]. However, many evidence-based guidelines recommend further assessment and a risk stratification approach if recovery has not occured within the first few weeks to avoid progression to chronic pain and disability [9,

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s10926-019-09843-4) contains supplementary material, which is available to authorized users.

✉ Douglas P. Gross
dgross@ualberta.ca

[1] Department of Physical Therapy, University of Alberta, 2-50 Corbett Hall, Edmonton, AB T6G 2G4, Canada

[2] Morneau Shepell, Toronto, Canada

[3] Department of Medicine, University of Connecticut School of Medicine, Farmington, CT, USA

[4] Department of Computing Science, University of Alberta, Edmonton, Canada

[5] National Research Council of Canada, Ottawa, Canada

Published online: 02 July 2019

11]. If barriers to recovery and return to work are observed, various interventions are recommended including structured exercise and physical conditioning programs [10, 12], workplace-based interventions [13, 14], and psychological therapies or more complex multidisciplinary biopsychosocial rehabilitation [11, 15, 16]. However, individual response to these interventions is highly variable. Despite having access to clinical practice guidelines and other available evidence, clinicians are often unable to identify with complete accuracy which patients will respond best to the various treatment options. This leads clinicians to select treatment options they expect to have a high probability of success, and then to iteratively update the approach if the response is not favourable.

Recently development in computer technology, machine intelligence, and analytical techniques has led to the development of clinical decision support tools that present knowledge to health care decision-makers to inform treatment choices [17]. These tools are often designed as computer-based point-of-care resources that match individual patient characteristics to available interventions. Use of such clinical decision support tools may lead to improved effectiveness of rehabilitation programs as well as more rapid and sustainable RTW because referrals are optimally targeted, personalized to individual patients, and human errors in decision making are minimized.

The Work Assessment Triage Tool (WATT) is a clinical decision support tool for classifying injured workers to appropriate rehabilitation programs [18]. Machine learning classification techniques were used to learn parameters and develop the algorithm, and the WATT uses claimant characteristics that influence RTW including demographic, occupational, injury-related, functional, and psychosocial factors. The WATT was developed for use by front line staff such as rehabilitation professionals, physicians, and compensation case managers, but is also applicable to compensation policy makers. In the initial validation study, classification accuracy was tested using tenfold cross validation procedures and accuracy of the computer-based WATT (ROC Area = 0.94) was better than human clinical decision-making (ROC Area = 0.86) for identifying rehabilitation programs that led to successful RTW [18]. However, preliminary validation was conducted on a validation subset from the original dataset, and validation from a new patient cohort might provide a stronger test of accuracy [19]. The final algorithm was integrated into a computer-based clinical decision support tool that now requires additional validation and impact evaluation. Recently, the WATT algorithm was compared to clinician decisions [20]. Results indicated the WATT is more likely than clinicians to recommend some rehabilitation interventions supported by current evidence, such as workplace-based interventions. Further research was recommended.

Reproducibility of machine learning and other artificial intelligence results on health and medical datasets can be problematic. It has been reported that the majority of studies of machine learning solutions for health applications evaluate their results on a single dataset [21]. These limited scope 'internal validity' evaluations do not provide a complete picture of the conceptual reproducibility of machine learning solutions. In particular, results may not generalize to future datasets. Therefore, the purpose of this study was to examine the external validity and accuracy of the WATT algorithm for making treatment decisions for workers with musculoskeletal conditions using more recent data from the jurisdiction where the WATT was initially developed.

Methods

Design

A population-based cohort design was used, with data extracted from the administrative and clinical databases of the Workers' Compensation Board—Alberta (WCB-Alberta). These provincial databases contain information on thousands of claimants from rehabilitation providers across the province. Contracted providers file reports at time of claimants' RTW assessment, admission for rehabilitation, and discharge. This mandatory reporting and data collection on a large province-wide scale provided a unique opportunity to conduct this study. The University of Alberta's Health Research Ethics Board approved this study.

We hypothesized that successful RTW outcomes would be more likely when actual treatments undertaken were consistent with the WATT decision-making algorithm and that fewer cases would have successful RTW when actual treatments represented a mismatch with WATT model recommendations. Using this approach, 2×2 matrices can be constructed for each individual treatment and tested relative to whether it was matched/unmatched with the WATT recommendation (similar to the development study procedures). A 'true positive' case would exist when the actual treatment undertaken was consistent with the WATT decision-making algorithm, while a 'true negative' would exist when the WATT did not recommended a treatment and in fact some other treatment was undertaken. A 'false positive' case would exist when the treatment under consideration was recommended by the WATT but another treatment was actually undertaken leading to the successful RTW outcome, while a 'false negative' a case exists when the treatment being considered had not been recommended by the WATT but was actually undertaken. See Fig. 1 for a graphical representation of the 2×2 matrix approach used for evaluating WATT accuracy.

	Treatment was Undertaken	Another Treatment was Undertaken
WATT Recommended the Treatment	True Positive	False Positive
WATT did NOT Recommend the Treatment	False Negative	True Negative

Fig. 1 2×2 matrix for evaluating WATT accuracy for selecting individual treatments within cases with successful return to work

Population

Province-wide data were available on all WCB-Alberta claimants with musculoskeletal injuries referred to RTW assessment facilities across the province. Based on a continuum of care model, claimants are referred for RTW assessment when they have met or surpassed expected injury healing times (typically 4–8 weeks) and have plateaued with medical interventions, yet 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. The assessing clinician interprets findings and claimants are triaged to what is deemed the most appropriate rehabilitation program. Since the WATT is still undergoing validation, it is not currently used within the jurisdiction for clinical decision-making.

The original WATT algorithm was developed using data (n = 8611) from WCB-Alberta claimants undergoing RTW assessment between December 1, 2009 and January 1, 2011. The current study used data collected between January 1, 2013 and December 31, 2016. We included all workers' compensation claimants with musculoskeletal disorders assessed during the study period. We excluded claimants with compensable neurological or psychological conditions, such as brain injury, spinal cord lesion, common mental health disorders, and traumatic psychological injury. These inclusion and exclusion criteria were the same as those used in the initial development study. Otherwise, all claimants undergoing RTW assessment during the timeframe were included. We did not restrict the cohort based on injury duration to increase generalizability to all claimants undergoing work assessment, but included duration as a factor within the WATT algorithm.

Study Procedures

This study was limited to archived clinical and administrative data of the WCB-Alberta. No claimants were directly recruited or consented. Within Alberta, workers' compensation reports are electronic and data is automatically entered into WCB-Alberta databases. For this study, demographic and work-related characteristics of individual claimants were linked with data from health care providers on type of rehabilitation program undertaken and other clinical characteristics. Compensation outcomes following rehabilitation were also linked, including wage replacement benefits data. Extraction of this information and formation of the database repeated methods and measures used when developing the WATT [18].

Measures

WATT

We obtained data on the 17 variables used in the WATT algorithm to make rehabilitation program recommendations. This included: employment status (i.e., job attachment status), National Occupational Classification code, current work status, availability of modified work, Occupational item from the Pain Disability Index (PDI) [22], score on the pain Visual Analogue Scale (VAS) [23], primary diagnostic code, injury duration in days, and items 2, 4, 5, 7, 12, 14, 18, 21, and 25 from the 36-Item Short Form Health Survey (SF-36) [24]. Individual claimant scores on these variables were then entered into the WATT algorithm to determine WATT recommendations for each claimant. The WATT recommends one of six rehabilitation programs including: community physical therapy, multidisciplinary functional restoration, workplace-based intervention, hybrid program (functional restoration with integrated workplace-based intervention), complex biopsychosocial chronic pain program, or no further rehabilitation. This was done using computer programming to avoid human error in data entry. This method allowed us to determine: (1) the rehabilitation program recommended by the WATT with highest confidence for successful return to work (using WATT positive rules); and (2) the rehabilitation program recommended by the WATT with highest confidence for unsuccessful return to work (using the WATT negative rules).

Human Recommendations

We obtained information on the recommendation made by the clinician conducting the return to work assessment at the time where information on the WATT variables were obtained. The clinicians, therefore, had access to all information within the WATT along with all other information obtained during this assessment. This consists of detailed medical history information, physical examination, scores on the patient reported outcomes mentioned above, as well as functional testing. The clinicians did not, however, have access to the WATT recommendations or the algorithm to inform their decision making. Possible recommendations included all of the same categories available in the WATT along with a new rehabilitation program for traumatic psychological injury, which is rarely recommended.

Program Undertaken After Assessment

We obtained information on the actual rehabilitation undertaken by the claimant after the return to work assessment. Often this is the same as the program recommended by the clinician conducting the assessment, however, at times the recommendations are overruled by the claimant's case manager who is responsible for deciding what program will be most cost-effective. As with the human clinicians, the case managers did not have access to WATT recommendations or the algorithm to inform their decision making. Potential programs undertaken were the same as those potentially recommended by the WATT.

Descriptive Variables

Additional information was collected for descriptive purposes and included: age in years, sex, nature of injury code, comorbid injury/condition code, urban/rural residence, number of previous compensation claims, overall scores on several self-report clinical measures (PDI as a percentage, SF-36 domains, Quick Disabilities of the Arm, Shoulder, and Hand measure, Lower Extremity Functional Scale, Orebro Musculoskeletal Pain Questionnaire) whether the claimant was receiving time loss benefits at time of assessment, and whether the claimant underwent a repeat program sometime in the follow-up year.

Outcome Variable

We obtained information on the same outcome variable used in the original WATT development paper [18]. This was the claimant's wage replacement status 30 days following the return to work assessment. Reception of wage replacement benefits is a key outcome within workers' compensation jurisdictions, and a surrogate indicator of work status. Receiving time loss benefits indicates that the worker is off work for an entire day. Thirty days after assessment was used in the original paper as this time point provided sufficient variability in the outcome to conduct machine learning analyses.

Statistical Analysis

Initially, all data records were reviewed to determine if any data issues such as missing data, outliers or out of range values existed within the dataset. Such occurrences were infrequent, however, substantial missing data existed for the self-report questionnaires. When data were missing these were replaced with 'unknown' for use with the WATT algorithm. Descriptive statistics were then calculated including means and standard deviations for continuous variables, modes and percentages for categorical variables.

WATT and human recommendations were described using counts and percentages, with this analysis stratified according to type of rehabilitation program undertaken following return to work assessment and according to program success (i.e. were claimants receiving wage replacement benefits 30 days after return to work assessment). To examine classification accuracy of the WATT, we examined claimants no longer receiving wage replacement benefits at 30 days following assessment (i.e. experienced successful outcome). We developed 2×2 tables for individual treatment programs (see Fig. 1) comparing WATT recommendations with actual programs undertaken to calculate sensitivity, specificity, overall accuracy, and area under the receiver operating characteristic curve (AUC) of the WATT for selecting these successful programs. We repeated this analysis for human clinical recommendations following RTW assessment for comparison with WATT recommendations. We also calculated positive and negative likelihood ratios for the WATT in comparison to human decisions. Lastly, we examined WATT negative rule recommendations according to type of rehabilitation program undertaken after return to work assessment in claimants experiencing an unsuccessful outcome. This allowed us to determine the proportion of claimants who may have avoided an unsuccessful outcome had WATT negative rule recommendations been followed. All analyses were completed in IBM SPSS v25 (Armonk, New York). Due to the very large sample size available, statistical testing was not undertaken.

Results

Population Characteristics

The dataset included 28,919 individuals with open workers' compensation claims related to a wide variety of

musculoskeletal disorders. The majority of claimants were employed (86.4%) and working in trades and transport occupations (44.5%) (see Table 1). Average age of the sample was 43.9 years and median duration of injury was 56 days. Characteristics of the individuals according to the type of rehabilitation program undertaken is also shown in Table 1. Of note is the very small number of claimants (n = 13, 0.04%) who underwent workplace-based interventions. This program is now rarely used within the WCB-Alberta jurisdiction. The majority of claimants (80.0%) experienced a successful return to work outcome as indicated by no longer receiving wage replacement benefits 30 days following return to work assessment.

Table 2 shows clinical characteristics of those completing the various patient reported pain intensity, disability, and health-related quality of life measures (n = 7002). Average pain intensity was moderate at 4.8/10 and disability was also moderate (44.8/100 on the PDI). Highest pain and disability levels were observed in claimants who underwent complex multidisciplinary chronic pain rehabilitation, while lowest levels were observed in those not undergoing rehabilitation or else undergoing a hybrid program (i.e., functional restoration with combined workplace intervention).

Comparison of WATT and Human Recommendations

Table 3 shows recommendations made by the WATT and human clinicians stratified according to type of rehabilitation undertaken following return to work assessment. In 4502 claimants (15.6%), the rehabilitation program undertaken matched the WATT recommendation of highest confidence. This is lower than the 15,441 claimants (53.3%) where the rehabilitation program undertaken matched the clinician's recommendation. The WATT was more likely to recommend any form of rehabilitation than human clinicians (97.0% vs. 74.2%). The WATT also more frequently recommended workplace-based interventions (15.0% vs. 0.2%), functional restoration (65.9% vs. 24.3%), and complex chronic pain programs (6.9% vs. 0.5%). Human clinicians were more likely to recommend community physical therapy (33.3% vs. 0.03%).

WATT Versus Human Accuracy

Table 4 shows sensitivity, specificity, overall accuracy levels, and AUC for the WATT and human clinicians in claimants who had experienced a successful outcome within 30 days after return to work assessment (i.e., made successful decisions). Overall, human clinicians outperformed the WATT in nearly every program with an overall accuracy of 0.72 and AUC of 0.69 compared to the WATT accuracy of 0.60 and AUC 0.50. However, the sensitivity of human decisions was still low at 0.54. Specificity was higher for both the WATT (0.87) and human (0.85) recommendations, however, this was due to the large number of programs that were not recommended by either. For the WATT, these values correspond to an overall positive likelihood ratio of 1.08 and negative likelihood ratio of 0.99. For human decisions, these values correspond to an overall positive likelihood ratio of 3.60 and negative likelihood ratio of 0.54.

WATT Negative Rules Recommendations

A comparison of the WATT negative rules (i.e. rehabilitation programs to avoid) with actual rehabilitation programs undertaken is shown in Table 5. In 1430 cases, use of the WATT negative rules may have prevented an unsuccessful outcome. This represents 24.7% of all claimants with negative outcome and 4.9% overall.

Discussion

Despite very high classification accuracy in the development study, we observed low WATT accuracy in this external validation compared to human clinicians. The human clinician recommendations outperformed the WATT overall and within every category of rehabilitation program except community physical therapy, which is rarely recommended by the WATT. High specificity values were observed for the WATT, but this was due to the large number of programs that were not recommended by the algorithm (i.e., many true negative recommendations). However, for this clinical situation we believe sensitivity is more important than specificity since the objective of the WATT is to improve selection of programs leading to rapid and sustainable return to work (i.e., true positive recommendations). The low sensitivity and overall accuracy combined with marginally useful likelihood ratios observed for the WATT algorithm indicates it is not yet ready for widespread clinical use.

A variety of factors may have led to an inability to replicate initial findings. The goal of machine learning classification is to learn the parameters of some function, such that it can accurately classify new cases to the correct class in the future. For finite datasets, estimates of performance are produced using resampling methods, such as bootstrapping or k-fold cross-validation [25]. In order for performance estimates to hold, the training data must be drawn independently and identically from the unknown, underlying population data distribution. Moreover, the underlying distribution should be static overtime. More specifically, the training data and the future application data must be drawn from the same probability distribution. In many real-world applications, however, the data can change either abruptly or slowly over time due to changes in personal interest and decision making, government policy, the environment, or other unknown

Table 1 Characteristics of claimants undergoing initial return-to-work assessment by rehabilitation program undertaken following assessment

	All claimants	No rehab/medical/other	Community physical therapy	Functional restoration	Workplace-based	Hybrid rehab	Complex chronic pain
	n = 28,919	n = 12,759	n = 6680	n = 5731	n = 13	n = 3449	n = 287
	All values represent either mean (standard deviation) or n (percentage)						
Age (years)	43.9 (12.8)	43.6 (13.1)	45.0 (12.9)	43.4 (12.5)	42.6 (8.6)	43.4 (12.2)	46.8 (10.5)
Injury duration (days)	265.9 (1019)	338.3 (1177)	264.9 (1054)	149 (682.2)	529.4 (1008)	105.9 (222.2)	1305.7 (2293)
	Median = 56	Median = 57	Median = 48	Median = 53	Median = 266	Median = 66	Median = 477
Number of previous claims	3.8 (5.0)	3.7 (5.0)	4.1 (5.3)	4.0 (5.1)	2.5(1.9)	3.7 (4.7)	3.9 (4.5)
Sex							
Male	11,579 (40.0%)	5137 (40.3%)	2712 (40.6%)	2377 (41.5%)	4 (30.8%)	1247 (36.2%)	102 (35.5%)
Female	6962 (24.1%)	2954 (23.2%)	1455 (21.8%)	1448 (25.3%)	3 (23.1%)	1054 (30.6%)	48 (16.7%)
Not specified	10,378 (35.9%)	4668 (36.6%)	2513 (37.6%)	1906 (33.3%)	6 (46.2%)	1148 (33.3%)	137 (47.7%)
Occupational category							
Management	951 (3.3%)	407 (3.2%)	259 (3.9%)	131 (2.3%)	0	149 (4.3%)	5 (1.7%)
Business, finance, admin	1753 (6.1%)	754 (5.9%)	365 (5.5%)	316 (5.5%)	2 (15.4%)	307 (8.9%)	9 (3.1%)
Sciences and related	691 (2.4%)	311 (2.4%)	136 (2.0%)	143 (2.5%)	1 (7.7%)	97 (2.8%)	3 (1.0%)
Health	2688 (9.3%)	1032 (8.1%)	464 (6.9%)	728 (12.7%)	0	445 (12.9%)	19 (6.6%)
Education, law and services	1372 (4.7%)	628 (4.9%)	319 (4.8%)	259 (4.5%)	1 (7.7%)	157 (4.6%)	8 (2.8%)
Art, culture, etc.	198 (0.7%)	87 (0.7%)	40 (0.6%)	31 (0.5%)	0	38 (1.1%)	2 (0.7%)
Sales and service	5663 (19.6%)	2549 (20.0%)	1259 (18.8%)	1059 (18.5%)	7 (53.8%)	740 (21.5%)	49 (17.1%)
Trades and transport	12,868 (44.5%)	5741 (45.0%)	3169 (47.4%)	2562 (44.7%)	0	1241 (36.0%)	155 (54.0%)
Production	947 (3.3%)	430 (3.4%)	241 (3.6%)	194 (3.4%)	1 (7.7%)	69 (2.0%)	12 (4.2%)
Manufacturing	1628 (5.6%)	716 (5.6%)	388 (5.8%)	295 (5.1%)	1 (7.7%)	203 (5.9%)	25 (8.7%)
Unknown	160 (0.6%)	104 (0.8%)	40 (0.6%)	13 (0.2%)	0	3 (0.1%)	0
Employed (% yes)	24,996 (86.4%)	10,817 (84.8%)	5798 (86.8%)	4888 (85.3%)	12 (92.3%)	3321 (96.3%)	160 (55.7%)
Currently working (% yes)	14,398 (49.8%)	6281 (49.2%)	2992 (44.8%)	2168 (37.8%)	11 (84.6%)	2891 (83.8%)	55 (19.2%)
Modified work available (% yes)	15,481 (53.5%)	6822 (53.5%	3569 (53.4%)	2576 (44.9%)	11 (84.6%)	2421 (70.2%)	82 (28.5%)
Diagnosis							
Sprain/strain	13,144 (45.5%)	5445 (42.7%)	2629 (39.4%)	2789 (48.7%)	10 (76.9%)	2182 (63.3%)	89 (31.0%)
Joint disorder	6989 (24.2%)	2966 (23.2%)	1575 (23.6%)	1632 (28.5%)	1 (7.7%)	742 (21.5%)	73 (25.4%)
Fracture	3144 (10.9%)	1255 (9.8%)	1099 (16.5%)	549 (9.6%)	1 (7.7%)	193 (5.6%)	47 (16.4%)
Contusion	1399 (4.8%)	678 (5.3%)	355 (5.3%)	225 (3.9%)	0	129 (3.7%)	12 (4.2%)
Laceration	675 (2.3%)	409 (3.2%)	143 (2.1%)	76 (1.3%)	1 (7.7%)	34 (1.0%)	12 (4.2%)
Dislocation	49 (1.7%)	200 (1.6%)	164 (2.5%)	79 (1.4%)	0	47 (1.4%)	5 (1.7%)
Nerve damage	562 (1.9%)	265 (2.1%)	185 (2.8%)	83 (1.4%)	0	19 (0.6%)	10 (3.5%)
Other	2511 (8.6%)	833 (6.5%)	200 (3.0%)	148 (2.6%)	0	48 (1.4%)	26 (9.1%)
Part of body							
Upper extremity	9827 (34.0%)	4171 (32.7%)	2407 (36.0%)	1676 (29.2%)	5 (38.5%)	1487 (43.1%)	81 (28.2%)
Back	5523 (19.1%)	2502 (19.6%)	1135 (17.0%)	1345 (23.5%)	0	483 (14.0%)	58 (20.2%)
Lower extremity	5057 (17.5%)	2092 (16.4%)	1484 (22.2%)	916 (16.0%)	6 (46.2%)	514 (14.9%)	45 (15.7%)

Table 1 (continued)

	All claimants	No rehab/medical/other	Community physical therapy	Functional restoration	Workplace-based	Hybrid rehab	Complex chronic pain
	n = 28,919	n = 12,759	n = 6680	n = 5731	n = 13	n = 3449	n = 287
Neck	4133 (14.3%)	1732 (13.6%)	595 (8.9%)	1085 (18.9%)	0	675 (19.6%)	46 (16.0%)
Other	4379 (15.1%)	2262 (17.7%)	1059 (15.9%)	709 (12.2%)	2 (15.4%)	290 (8.4%)	57 (19.8%)
Comorbid injury (% yes)	4482 (15.5%)	1913 (15.0%)	1154 (17.3%)	971 (16.9%)	2 (15.4%)	376 (10.9%)	66 (23.0%)
Rural residence (% yes)	8780 (30.4%)	3980 (31.2%)	2382 (35.7%)	1464 (25.5%)	9 (69.2%)	864 (25.1%)	81 (28.2%)
Receiving benefits at assessment (% yes)	15,302 (52.9%)	6047 (47.4)	3898 (58.4%)	4198 (73.3%)	2 (15.4%)	951 (27.6%)	206 (71.8%)
Receiving benefits 30 days after assessment (% yes)	5795 (20.0%)	1517 (11.9%)	1997 (29.9%)	2002 (34.9%)	1 (7.7%)	93 (2.7%)	185 (64.5%)
Repeat program? (% yes)	2031 (7.0%)	0	1671 (25.0%)	250 (4.4%)	0	90 (2.6%)	20 (7.0%)

Table 2 Clinical characteristics of claimants completing self-report questionnaires by rehabilitation program undertaken following assessment

	All claimants	No rehab	Community physical therapy	Functional restoration	Workplace-based	Hybrid rehabilitation	Complex chronic pain
	n = 7002	n = 2668	n = 1271	n = 1872	n = 4	n = 1131	n = 56
	All values represent mean (standard deviation)						
Pain visual analogue Scale (out of 10)	4.8 (2.5)	4.4 (2.6)	5.0 (2.5)	5.3 (2.3)	5.8 (2.1)	4.3 (2.4)	6.5 (2.5)
Pain disability index (out of 100)	44.8 (22.2)	41.6 (23.1)	47.6 (22.9)	50.6 (19.7)	51.1 (14.6)	38.7 (19.6)	66.2 (19.0)
SF-36 domain (out of 100)							
Physical function	50.9 (26.3)	53.9 (27.2)	50.0 (26.7)	46.5 (24.4)	29.7 (20.9)	53.4 (25.1)	32.0 (24.3)
Role physical	31.0 (26.7)	34.2 (28.1)	29.7 (27.6)	24.0 (22.5)	18.8 (37.5)	37.5 (25.8)	14.3 (20.6)
Bodily pain	30.6 (22.6)	32.6 (23.6)	27.1 (21.4)	25.5 (20.1)	35.0 (30.0)	38.6 (22.7)	14.3 (16.9)
General health	67.6 (18.6)	68.0 (18.9)	66.6 (19.4)	66.4 (18.1)	75.5 (17.3)	70.6 (16.8)	49.5 (19.2)
Vitality	49.0 (20.3)	50.5 (21.0)	48.9 (21.4)	45.5 (18.9)	60.9 (20.7)	52.2 (18.7)	32.9 (18.9)
Social function	54.0 (26.6)	55.7 (27.1)	52.4 (27.7)	48.6 (25.0)	56.3 (29.8)	62.1 (24.2)	29.9 (24.3)
Role emotional	57.7 (32.4)	60.0 (32.5)	56.4 (33.7)	51.5 (32.1)	56.3 (37.5)	65.6 (28.5)	30.1 (27.2)
Mental health	63.4 (20.9)	64.4 (20.9)	63.6 (21.3)	59.7 (20.8)	62.8 (23.6)	67.8 (18.9)	44.2 (19.2)
QuickDASH (n = 1580)	43.8 (20.0)	39.6 (20.2)	48.8 (20.5)	47.9 (18.2)	Insufficient cases	39.1 (17.4)	58.5 (19.1)
QuickDASH work module cases (n = 1497)	53.6 (27.6)	47.4 (27.7)	59.6 (27.0)	64.4 (24.0)	Insufficient cases	43.8 (25.9)	62.5 (20.3)
Lower extremity functional scale (n = 1840)	49.5 (21.8)	52.6 (21.4)	43.7 (20.2)	45.5 (20.7)	Insufficient cases	59.0 (22.3)	27.8 (20.4)
Orebro musculoskeletal pain cases questionnaire (n = 4197)	98.5 (26.5)	94.1 (29.2)	102.5 (25.2)	105.0 (23.3)	Insufficient cases	89.8 (22.4)	125.9 (21.3)

Table 3 Recommendations made by the Work Assessment Triage Tool (WATT) and human clinicians according to type of rehabilitation undertaken following assessment

	No rehab	Community physical therapy	Functional restoration	Workplace-based	Hybrid rehabilitation	Complex chronic pain	Total
	All values represent n (percentage)						
Entire sample	12,759 (44.1)	n = 6680 (23.1)	5731 (19.8)	13 (0.04)	3449 (11.9)	287 (1.0)	28,919 (100)
WATT recommendation							
No further rehabilitation	533 (4.2)	209 (3.1)	52 (0.9)	1 (7.7)	12 (0.3)	58 (20.2)	865 (3.0)
Community physical therapy	6 (0.05)	2 (0.03)	2 (0.03)	0 (0)	1 (0.03)	0 (0)	11 (0.03)
Functional restoration	8743 (68.5)	4936 (73.9)	3399 (59.3)	7 (53.8)	1831 (53.1)	149 (51.9)	19,065 (65.9)
Workplace-based	1786 (14.0)	687 (10.3)	940 (16.4)	1 (7.7)	921 (26.7)	8 (2.8)	4343 (15.0)
Hybrid	927 (7.3)	458 (6.9)	694 (12.1)	4 (30.8)	523 (15.2)	28 (12.1)	2634 (9.1)
Complex chronic pain	764 (6.0)	388 (5.8)	644 (11.2)	0 (0)	161 (4.7)	44 (15.3)	2001 (6.9)
Human recommendation							
No further rehabilitation	4603 (36.1)	1940 (29.0)	599 (10.5)	1 (7.7)	243 (7.0)	62 (21.6)	7448 (25.8)
Community physical therapy	4471 (35.0)	3964 (59.3)	860 (15.0)	1 (7.7)	310 (9.0)	33 (11.5)	9639 (33.3)
Functional restoration	2155 (16.9)	504 (7.6)	4085 (71.3)	0	269 (7.8)	19 (6.6)	7033 (24.3)
Workplace-based	24 (0.2)	6 (0.1)	13 (0.2)	8 (61.5)	2 (0.1)	0	53 (0.2)
Hybrid	1311 (10.3)	213 (3.2)	149 (2.6)	2 (15.4)	2613 (75.8)	4 (1.4)	4292 (14.8)
Complex chronic pain	68 (0.5)	38 (0.6)	19 (0.3)	0	10 (0.3)	168 (58.5)	303 (1.0)
Other	127 (1.0)	15 (0.2)	6 (0.1)	0	2 (0.1)	1 (0.3)	151 (0.5)
Positive cases (successful RTW outcome)	11,242 (48.6)	4683 (20.3)	3729 (16.1)	12 (0.05)	3356 (14.5)	102 (0.4)	23,124 (100)
WATT positive rule recommendations							
No further rehabilitation	468 (4.2)	143 (3.1)	21 (0.6)	21 (0.6)	12 (0.4)	17 (16.7)	662 (2.9)
Community physical therapy	6 (0.1)	1 (0)	1 (0)	0 (0)	1 (0)	0 (0)	9 (0.0)
Functional restoration	7526 (66.9)	3295 (70.4)	2075 (55.6)	6 (50.0)	1764 (52.6)	11 (10.8)	14,716 (63.6)
Workplace-based	1707 (15.2)	607 (13.0)	707 (19.0)	1 (8.3)	907 (27.0)	4 (3.9)	3933 (17.0)
Hybrid	889 (7.9)	394 (8.4)	569 (15.3)	4 (33.3)	516 (15.4)	20 (19.6)	2392 (10.3)
Complex chronic pain	646 (5.7)	243 (5.2)	356 (9.5)	0 (0)	156 (4.6)	11 (10.8)	1412 (6.1)
Human recommendation							
No further rehabilitation	3995 (35.5)	1078 (23.0)	247 (6.6)	0 (0)	222 (6.6)	28 (27.5)	5570 (24.1)
Community physical therapy	3998 (35.6)	3026 (64.6)	460 (12.3)	1 (8.3)	284 (8.5)	16 (15.7)	7785 (33.7)

Table 3 (continued)

	No rehab	Community physical therapy	Functional restoration	Workplace-based	Hybrid rehabilitation	Complex chronic pain	Total
Functional restoration	1829 (16.3)	350 (7.5)	2866 (76.9)	1 (8.3)	260 (7.7)	11 (10.8)	5317 (23.0)
Workplace-based	23 (0.2)	5 (0.1)	10 (0.3)	8 (66.7)	2 (0.1)	0 (0)	48 (0.2)
Hybrid	1289 (11.5)	202 (4.3)	132 (3.5)	2 (16.7)	2576 (76.8)	3 (2.9)	4204 (18.2)
Complex chronic pain	50 (0.4)	17 (0.4)	11 (0.3)	0 (0)	10 (0.3)	44 (43.1)	132 (0.6)
Other	58 (0.5)	5 (0.1)	3 (0.1)	0 (0)	2 (0.1)	0 (0)	68 (0.3)
Negative cases (unsuccessful RTW outcome)	1517 (26.2)	1997 (34.5)	2002 (34.5)	1 (0.02)	93 (1.6)	185 (3.2)	5795 (100)
WATT positive rule recommendation							
No further rehabilitation	65 (4.3)	66 (3.3)	31 (1.5)	0 (0)	0 (0)	41 (22.2)	203 (3.5)
Community physical therapy	0 (0)	1 (0.1)	1 (0.5)	0 (0)	0(0)	0 (0)	2 (0.03)
Functional restoration	1217 (80.2)	1641 (82.2)	1324 (66.1)	1 (100)	67 (72.0)	99 (53.5)	4349 (75.0)
Workplace-based	79 (5.2)	80 (4.0)	233 (11.6)	0 (0)	14 (15.1)	4 (2.2)	410 (7.1)
Hybrid	38 (2.5)	64 (3.2)	125 (6.2)	0 (0)	7 (7.5)	8 (4.3)	242 (4.2)
Complex chronic pain	118 (7.8)	145 (7.3)	288 (14.4)	0 (0)	5 (5.4)	33 (17.8)	589 (10.2)
Human recommendation							
No further rehabilitation	608 (40.1)	862 (43.2)	352 (17.6)	1 (100)	21 (22.6)	34 (18.4)	1878 (32.4)
Community physical therapy	473 (31.2)	938 (47.0)	400 (20.0)	0 (0)	26 (28.0)	17 (9.2)	1854 (32.0)
Functional restoration	326 (21.5)	154 (7.7)	1219 (60.9)	0 (0)	0 (0)	9 (9.7)	1716 (29.6)
Workplace-based	1 (0.1)	1 (0.1)	3 (0.1)	0 (0)	0 (0)	0 (0)	5 (0.1)
Hybrid	22 (1.5)	11 (0.6)	17 (0.8)	0 (0)	37 (39.8)	1 (0.5)	88 (1.5)
Complex chronic pain	18 (1.2)	21 (1.1)	8 (0.4)	0 (0)	0 (0)	124 (67.0	171 (3.0)

Table 4 Classification accuracy of the Work Assessment Triage Tool and human clinicians for selecting programs that led to successful return to work outcomes

	Sensitivity	Specificity	Overall accuracy	Area under the receiver operating curve
	All values represent human clinician/WATT algorithm performance			
Functional restoration	0.77/0.56	0.87/0.35	0.86/0.38	0.82/0.45
Complex chronic pain management	0.43/0.11	0.996/0.94	0.99/0.94	0.71/0.52
Workplace-based program	0.67/0.08	0.998/0.83	0.998/0.83	0.83/0.46
Hybrid (functional restoration with workplace component)	0.77/0.15	0.92/0.91	0.90/0.80	0.84/0.53
Community physical therapy	0.65/0.0002	0.74/0.999	0.72/0.80	0.69/0.50
No further rehabilitation	0.36/0.04	0.87/0.98	0.62/0.53	0.61/0.51
Weighted averages	0.54/0.13	0.85/0.87	0.72/0.60	0.69/0.50

Table 5 Comparison of Work Assessment Triage Tool negative rule recommendations and type of rehabilitation undertaken in claimants experiencing an unsuccessful return to work outcome (n = 5795)

	No rehab	Community physical therapy	Functional restoration	Workplace-based	Hybrid rehabilitation	Complex chronic pain	Total
	All values represent n (percentage)						
WATT negative rule recommendations							
Not no further rehabilitation	2 (0.1)	6 (0.3)	2 (0.1)	0 (0)	2 (2.2)	1 (0.5)	13 (0.2)
Not community physical therapy	50 (3.3)	11 (0.6)	19 (0.9)	0 (0)	1 (1.1)	7 (3.8)	88 (1.5)
Not functional restoration	1276 (84.1)	1745 (87.4)	1389 (69.4)	1 (100)	69 (74.2)	149 (80.5)	4629 (79.9)
Not complex chronic pain	189 (12.5)	235 (11.8)	592 (29.6)	0 (0)	21 (22.6)	28 (15.1)	1065 (18.4)

factors. Changes in the underlying data can lead to deterioration of classifier accuracy to levels that are no longer appropriate for the application [26]. In machine learning, slow changes in the underlying distribution are known as concept drift and abrupt changes as concept shift [27].

Concept drift is likely to have deteriorated accuracy of the WATT algorithm. Importantly, the rehabilitation decisions and processes within the jurisdiction appear to have changed. Claimants are now more likely to undergo no rehabilitation (44.1% in this study vs. 19.0% in the development dataset) or community physical therapy (23.1% vs. 16.8%), and less likely to undergo for functional restoration (19.8% vs. 50.3%) and workplace-based interventions (0.2% vs. 1.4%). Changes in clinician and case manager behaviours within the jurisdiction combined with other sampling fluctuations could have led to inability to replicate WATT accuracy results. In situations of concept drift, machine learning algorithms need to be incrementally and automatically updated or adjusted as changes within the system occur [27–29]. This requires a means of detecting and reacting to such changes, enabling the learnt function to evolve overtime by periodically retraining or updating itself with newly acquired data [29]. Alternatively, the data and performance of the classifier can automatically be monitored to detect harmful changes. When such changes are detected, the old training data is replaced by new data and the classification model retrained [30]. In the WCB-Alberta and related insurance systems, this would require directly linking the learning algorithm to the administrative claims database and undertaking periodic updates to the algorithm as new data were received. Classification models can also be updated and improved through ongoing input from expert clinicians [31].

Alternatively, it has been long recognized that statistical models perform better in development stages than when undergoing external validation, and this same problem likely holds for machine learning algorithms [19]. Prior to incorporating machine learning classifiers into compensation and insurance databases, it is strongly recommended that such tools are tested in randomized controlled trials and shown to improve desired outcomes. Such trials would compare clinical and return to work outcomes between groups of workers treated by therapists with and without access to clinical decision support tools. This strategy would also reduce potential biases inherent in studies such as ours that rely on historical data from clinicians who do not have access to WATT recommendations (e.g., clinician recommendations likely strongly influenced treatment selections within our study).

Several data-related concerns have been raised by authors about the use of data mining in healthcare for predicting health outcomes or response to treatment [32–35]. Most concerns relate to the completeness and integrity of datasets, the large number of potential dimensions, variables, and variable combinations (i.e., a "high dimensionality" problem when all available medical information is used) [34], problems with collinearity or multivariate collinearity between predictor variables and their combinations, and the potential for model overfitting. In our application, a minimal set of variables and treatments were included in the modeling, and none were correlated to the extent that collinearity would create unstable models. We also compared descriptive statistics on predictor variables between our two samples, and found few differences at the aggregate level and no major changes in demographics within the injured worker population (See Supplemental Table). The only variable with a potentially clinically important difference was injury duration (mean of 74 vs. 56 days), otherwise the other 18 variables demonstrated small (< 5%), clinically insignificant changes. Initial development of the WATT model used a standard approach whereby the entire dataset was randomly divided into ten equal subsets, with nine datasets being used to "train" the model, and the tenth validation dataset being

used to estimate accuracy. This should have avoided any serious overfitting of the model, but it's possible that other factors could have changed between time periods that would have added unexplained sources of variance that were not incorporated in our original model. We have clear evidence that sorting strategies and referral patterns of providers did change over this period, though it's difficult to say what secular changes in regulations, peer practices, or other circumstances may have reduced the predictive accuracy of the model within just a few years. The possibility remains that workers' compensation systems may have fluid properties that limit the applicability of data mining procedures when designed to be finely tuned to current practices and population characteristics.

The reduced use of functional restoration and workplace-based interventions within the jurisdiction is surprising. These interventions are supported by strong evidence indicating their success for helping individuals with subacute and chronic musculoskeletal conditions to return to work [12–16]. It appears that claimants are more likely to undergo no further rehabilitation or community physical therapy, which are less expensive options. And in fact, 11,242 (48.6%) claimants who did not undergo rehabilitation experienced a successful outcome 30 days following assessment. This may represent a higher reliance on return to work assessment within the jurisdiction for determining whether claimants are capable of safely returning to work.

Further validation and refinement of the WATT algorithm is needed. The model was built using Repeated Incremental Pruning to Produce Error Reduction (RIPPER) models. Newer approaches such as neural networks or deep learning strategies may result in better models with higher accuracy. Integrating additional clinical and contextual variables such as patient expectations of recovery or other psychosocial variables may also help improve the validity of computerized clinical decision support tools since these variables have been found important predictors of RTW [1, 5, 6, 36].

Limitations

This study was limited by reliance on archived data from WCB-Alberta. This led to a large amount of missing data on patient reported outcome measures that are incorporated into the WATT. While claimants cannot be compelled to complete questionnaires, completing the study in the context of a controlled trial may have resulted in a higher completion rate. The WATT algorithm is capable of using individual claimant data even with missing data on the self-report questionnaires (i.e., the option 'unknown' is entered into the model when data is missing), however, these clinical variables are important classifiers and the high amount of missing data may have reduced the accuracy of WATT recommendations. However, rates of missing data on the self-report questionnaires in this study were comparable to rates in the original validation study. Additionally, prospective validation in the context of a controlled trial where clinical decision support tools are used in actual decision-making may provide better a test of whether they improve clinical outcomes.

Conclusions

Accuracy of the WATT for selecting successful rehabilitation programs was modest and less accurate than human clinical recommendations in a new cohort of workers' compensation claimants drawn from the same system. The low accuracy observed for the WATT algorithm indicates it is not yet ready for widespread use. Algorithm revision and further validation is needed to create a more robust model.

Acknowledgements The Workers' Compensation Board of Alberta assisted with data collection.

Funding Funding was provided by the Workers' Compensation Board of Alberta.

Compliance with Ethical Standards

Conflict of interest The research team received a research grant from the Workers' Compensation Board of Alberta but otherwise declares that they have no relevant conflicts of interest. Douglas Gross declares that he has no conflict of interest. Ivan A. Steenstra declares that he has no conflict of interest. William Shaw declares that he has no conflict of interest. Parnian Yousefi declares that she has no conflict of interest. Colin Bellinger declares that he has no conflict of interest. Osmar Zaïane declares that he has no conflict of interest.

Ethical Approval All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards.

Informed Consent This study was an analysis of an administrative dataset therefore informed consent for the study was not obtained.

References

1. Hartvigsen J, Hancock MJ, Kongsted A, Louw Q, Ferreira ML, Genevay S, et al. What low back pain is and why we need to pay attention. Lancet. 2018;391(10137):2356–2367. https://doi.org/10.1016/S0140-6736(18)30480-X.
2. DALYs GBD, Collaborators H. Global, regional, and national disability-adjusted life-years (DALYs) for 315 diseases and injuries and healthy life expectancy (HALE), 1990-2015: a systematic analysis for the Global Burden of Disease Study 2015. Lancet. 2016;388(10053):1603–1658. https://doi.org/10.1016/S0140-6736(16)31460-X.
3. Foster NE, Anema JR, Cherkin D, Chou R, Cohen SP, Gross DP, et al. Prevention and treatment of low back pain: evidence, challenges, and promising directions. Lancet.

2018;391(10137):2368–2383. https://doi.org/10.1016/S0140-6736(18)30489-6.
4. Collie A, Di Donato M, Iles R. Work disability in Australia: an overview of prevalence, expenditure, support systems and services. J Occup Rehabil. 2018. https://doi.org/10.1007/s10926-018-9816-4.
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(1):64–80. https://doi.org/10.1007/s10926-008-9159-7.
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(3):394–405. https://doi.org/10.1007/s10926-009-9218-8.
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(4):591–605. https://doi.org/10.1007/s10926-006-9061-0.
8. Cote P, Wong JJ, Sutton D, Shearer HM, Mior S, Randhawa K, et al. Management of neck pain and associated disorders: a clinical practice guideline from the Ontario Protocol for Traffic Injury Management (OPTIMa) Collaboration. Eur Spine J. 2016;25(7):2000–2022. https://doi.org/10.1007/s00586-016-4467-7.
9. Wong JJ, Cote P, Shearer HM, Carroll LJ, Yu H, Varatharajan S, et al. Clinical practice guidelines for the management of conditions related to traffic collisions: a systematic review by the OPTIMa Collaboration. Disabil Rehabil. 2015;37(6):471–489. https://doi.org/10.3109/09638288.2014.932448.
10. Qaseem A, Wilt TJ, McLean RM, Forciea MA, Clinical Guidelines Committee of the American College of Physicians. Noninvasive treatments for acute, subacute, and chronic low back pain: a clinical practice guideline from the American College of Physicians. Ann Intern Med. 2017;166(7):514–530. https://doi.org/10.7326/m16-2367.
11. Almeida M, Saragiotto B, Richards B, Maher CG. Primary care management of non-specific low back pain: key messages from recent clinical guidelines. Med J Aust. 2018;208(6):272–275.
12. 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;(1):CD001822. https://doi.org/10.1002/14651858.cd001822.pub2.
13. van Oostrom SH, Driessen MT, de Vet HC, Franche RL, Schonstein E, Loisel P, et al. Workplace interventions for preventing work disability. Cochrane Database Syst Rev. 2009;(1):CD006955. https://doi.org/10.1002/14651858.cd006955.pub2.
14. Cullen KL, Irvin E, Collie A, Clay F, Gensby U, Jennings PA, et al. Effectiveness of workplace interventions in return-to-work for musculoskeletal, pain-related and mental health conditions: an update of the evidence and messages for practitioners. J Occup Rehabil. 2018;28(1):1–15. https://doi.org/10.1007/s10926-016-9690-x.
15. Saragiotto BT, de Almeida MO, Yamato TP, Maher CG. Multidisciplinary biopsychosocial rehabilitation for nonspecific chronic low back pain. Phys Ther. 2016;96(6):759–763. https://doi.org/10.2522/ptj.20150359.
16. Kamper SJ, Apeldoorn AT, Chiarotto A, Smeets RJ, Ostelo RW, Guzman J, et al. Multidisciplinary biopsychosocial rehabilitation for chronic low back pain: cochrane systematic review and meta-analysis. BMJ. 2015;350:h444. https://doi.org/10.1136/bmj.h444.
17. Gross DP, Armijo-Olivo S, Shaw WS, Williams-Whitt K, Shaw NT, Hartvigsen J, et al. Clinical decision support tools for selecting interventions for patients with disabling musculoskeletal disorders: a scoping review. J Occup Rehabil. 2016;26(3):286–318. https://doi.org/10.1007/s10926-015-9614-1.
18. Gross DP, Zhang J, Steenstra I, Barnsley S, Haws C, Amell T, et al. Development of a computer-based clinical decision support tool for selecting appropriate rehabilitation interventions for injured workers. J Occup Rehabil. 2013;23(4):597–609. https://doi.org/10.1007/s10926-013-9430-4.
19. Collins GS, de Groot JA, Dutton S, Omar O, Shanyinde M, Tajar A, et al. External validation of multivariable prediction models: a systematic review of methodological conduct and reporting. BMC Med Res Methodol. 2014;14:40. https://doi.org/10.1186/1471-2288-14-40.
20. Qin Z, Armijo-Olivo S, Woodhouse LJ, Gross DP. An investigation of the validity of the Work Assessment Triage Tool clinical decision support tool for selecting optimal rehabilitation interventions for workers with musculoskeletal injuries. Clin Rehabil. 2016;30(3):277–287. https://doi.org/10.1177/0269215515578696.
21. Johnson AEW, Pollard TJ, Mark RG. Reproducibility in critical care: a mortality prediction case study. Proc Mach Learn Res. 2017;68:361–376.
22. Pollard CA. Preliminary validity study of the Pain Disability Index. Percept Mot Skills. 1984;59(3):974.
23. 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.
24. 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(3):247–263.
25. Zhao M. K-fold cross-validation for improving medical classification accuracy and model selection in K-nearest neighbors classifiers. Basic Clin Pharmacol. 2016;118(Suppl 1):107.
26. Žliobaitė I, Pechenizkiy M, Gama J. An overview of concept drift applications. In: Japkowicz N, Stefanowski J, editors. Big data analysis: new algorithms for a new society. Cham: Springer; 2016. p. 91–114.
27. Gama J, Žliobaitė I, Bifet A, Pechenizkiy M, Bouchachia A. A survey on concept drift adaptation. ACM Comput Surv. 2014;46(4):1–37. https://doi.org/10.1145/2523813.
28. Webb GI, Lee LK, Goethals B, Petitjean F. Analyzing concept drift and shift from sample data. Data Min Knowl Discov. 2018;32(5):1179–1199. https://doi.org/10.1007/s10618-018-0554-1.
29. Widmer G, Kubat M. Learning in the presence of concept drift and hidden contexts. Mach Learn. 1996;23(1):69–101. https://doi.org/10.1023/a:1018046501280.
30. Gama J, Castillo G. Learning with local drift detection. In: Li X, Zaïane OR, Li Z, editors. Advanced data mining and applications. ADMA 2006. Lecture notes in computer science. Vol. 4093. Springer, Berlin, Heidelberg; 2006.
31. Ambrosino R, Buchanan BG. The use of physician domain knowledge to improve the learning of rule-based models for decision-support. Proc AMIA Symp. 1999;192–196.
32. Lee CH, Yoon HJ. Medical big data: promise and challenges. Kidney Res Clin Pract. 2017;36(1):3–11. https://doi.org/10.23876/j.krcp.2017.36.1.3.
33. Dinov ID. Methodological challenges and analytic opportunities for modeling and interpreting Big Healthcare Data. Gigascience. 2016. https://doi.org/10.1186/s13742-016-0117-6.

34. Sinha A, Hripcsak G, Markatou M. Large datasets in biomedicine: a discussion of salient analytic issues. J Am Med Inform Assoc. 2009;16(6):759–767. https://doi.org/10.1197/jamia.M2780.
35. Taranu I. Data mining in healthcare: decision making and precision. Database Syst J. 2015;6(4):33–40.
36. Iles RA, Davidson M, Taylor NF, O'Halloran P. Systematic review of the ability of recovery expectations to predict outcomes in non-chronic non-specific low back pain. J Occup Rehabil. 2009;19(1):25–40. https://doi.org/10.1007/s10926-008-9161-0.