Evaluation of A Clinical Decision Support Tool for Selecting Optimal Rehabilitation Intervention for Injured Workers

by

Ziling Qin

A thesis submitted in partial fulfillment of the requirements for the degree of

Master of Science

in

Rehabilitation Science

Faculty of Rehabilitation Medicine

University of Alberta

Abstract

Objective: To evaluate the concurrent validity of a newly developed clinical decision support tool (Work Assessment Triage Tool, WATT) by comparing the rehabilitation interventions determined using the WATT with the current gold standard–clinician recommendations.

Methods: This is a secondary data analysis study. Data were collected in a clinical trial conducted previously at the Workers' Compensation Board of Alberta rehabilitation facility. A variety of statistical methods were used to compare recommendations for rehabilitation strategies determined using the WATT, clinician recommendations, actual programs claimants undertook and return-to-work outcomes. Analyses included percent agreement, crosstabs, and likelihood ratios.

Results: Percent agreement between clinician recommendations and WATT recommendations were low (r = 0.19) to moderate (r = 0.46). The WATT does not appear to improve upon clinician recommendations as only half of the RTW claimants whose actual rehabilitation programs did not match those of the clinician recommendations, matched recommendations identified using WATT.

Discussions: Contrary to internal validation demonstrating that the WATT outperformed clinician recommendations; results of the external validation of the WATT were not as promising. Findings do not provide evidence of concurrent validity of the WATT against the current gold standard. Four possible reasons could explain the results: (1) important differences were observed in claimant characteristics between the original WATT development data and our validation dataset; (2) insufficient data for claimants who failed RTW and those with successful RTW whose actual rehabilitation program did not match with the clinician recommendations; (3) data processing techniques that were used to overcome rehabilitation class imbalance when building the WATT, which may contribute to errors in the WATT recommendations; (4) clinician recommendations conflicted somewhat with existing evidence as some rehabilitation programs that were highly supported by research evidence (i.e. workplace interventions) were rarely recommended by clinicians in our validation dataset.

Conclusion: WATT recommendations do not concur with clinician recommendations. With respect to concurrent validity, no conclusion can be drawn as to which method, WATT or clinician judgment, provides better recommendations for return-to-work in actual practice. Further research is needed to resolve this uncertainty.

Keywords

Clinical Decision Support Tool, Work Assessment Triage Tool, concurrent validity, return-to-work

Preface

This thesis is an original work by Ziling Qin. The research project, of which this thesis is a part, received research ethics approval from the University of Alberta Research Ethics Board. The project name is “Evaluation of the Work Assessment Triage Tool” and project number is Pro00036317, Date June 14, 2013.

Acknowledgement

First of all, I would like to express my deepest gratitude to my supervisor, Dr. Douglas Gross, who offered me the data for the research. The data was obtained from a previous study done at the Workers' Compensation Board of Alberta at Millard Health that was funded by WorkSafe BC. Not only did Dr. Gross provide access to the data, but he also guided me in doing research and preparing the thesis. His professionalism, patience, kindness and enthusiasm greatly impact me. I could not imagine having a better supervisor and mentor during my master study.

My heartfelt appreciation goes to the rest of my committee: Dr. Susan Armijo Olivo and Dr. Linda Woodhouse. Both of them gave me lots of encouragement, guidance and insightful comments for my thesis.

Besides, I would like to offer my special thanks to my friend Yi Shen for assisting me writing codes in the data analysis part, and for his persistent support and encouragement.

Last but not least, I owe my greatest thanks to my parents, my father Dejin Qin and my mother Yingqing Zhang, for giving birth to me at the first place and supporting me spiritually throughout my life.

TABLE OF CONTENTS

Chapter 1 1

Introduction 1

1.1. Problem Statement 1

1.2. Definition Terms 3

Chapter 2 5

Literature Review 5

2.1. Work-related Musculoskeletal Disorders 5

2.2. WCB-Alberta Soft Tissue Injury Continuum of Care Model 6

2.3. Clinical Decision Support Tools 9

2.4. Validity Testing of Clinical Decision Support Tools and the Gold Standard 13

2.5. Work Assessment Triage Tool 18

2.6. Objective of the Study 23

2.7. Research Hypotheses 23

Chapter 3 25

Methods and Procedures 25

3.1. Study Design 25

3.2. Database Information 25

3.3. Sample 25

3.4. Data Collection 26

3.5. Statistical Analysis 31

Chapter 4 38

Results 38

4.1. Data Reviewing and Claimants' Demographics 38

4.2. Summary of Rehabilitation Programs Frequency Distribution 45

4.3. Percent Agreements of Recommendation 47

4.4. Comparison of WATT and Clinician Recommendations 48

4.5. Likelihood Ratio 50

Chapter 5 56

Discussion 56

5.1. Characteristics of Claimants 56

5.2. Discussion on three objectives 60

5.3. Clinical Evidence in Support of WATT Recommendation 63

5.4. Study Strengths 65

5.5. Study Limitations 66

5.6. Clinical Significance 67

5.7. Suggestions for Future Research 67

5.8. Summary and Conclusion 68

References 70

Appendix 74

LIST OF FIGURES

Figure 1Interface of the Work Assessment Triage Tool (developed by Gross DP, etc.) 2
Figure 2 WCB- Alberta Continuum of Care Model (22) 7
Figure 3 A hypothetical risk factor based intervention strategy for low back pain(28)................. 16
Figure 4 Example of the diagnostic interface (32).. 17
Figure 6 Evaluation framework - focus, method and metrics (32) 18
Figure 6 Final class distribution comparison after all data pre-processing was completed(9)...... 20
Figure 7 Performance of the clinician baseline recommendation compare to the final machine learning algorithm (9) 22
Figure 8 Likelihood of RTW when WATT/ clinician recommendation matched 52
Figure 9 Likelihood of RTW when clinician/ WATT Top recommendation matched 53
Figure 10 Likelihood of RTW when clinician/ real program matched 54

LIST OF TABLES

Table 1 Frequency of WATT/ clinician recommendation for provider-based rehabilitation 33

Table 2 Frequency of WATT/ clinician recommendation for worksite-based rehabilitation 33

Table 3 Frequency of matching program (WATT and real program comparison as column) for claimants RTW 35

Table 4 Frequency of matching program (WATT and clinician comparison as column) for claimants RTW 35

Table 5 Frequency of matching recommendation considering RTW status 36

Table 6 Likelihood ratio interpretation 37

Table 7 Comparison of claimants' characteristics 38

Table 8 Characteristics of claimants in various rehabilitation programs undertaken overall 41

Table 9 Characteristics of BFCE claimants in various rehabilitation programs undertaken 43

Table 10 Characteristics of CFCE claimants in various rehabilitation programs undertaken 44

Table 11 Rehabilitation program frequency distribution overall 46

Table 12 Rehabilitation program frequency distribution for claimants returning to work 46

Table 13 Rehabilitation program frequency distribution for claimants failing to return to work.. 46

Table 14 Rehabilitation program frequency distribution for BFCE claimants 46

Table 15 Rehabilitation programs frequency distribution for CFCE claimants......... 47

Table 16 Percent agreement between WATT and clinician recommendations (overall) 48

Table 17 Percent agreement between WATT and clinician recommendations (BFCE only) 48

Table 18 Frequency of matching program (WATT and real program comparison as column) for claimants RTW 49

Table 19 Frequency of matching program (WATT and clinician comparison as column) for claimants RTW 49

Table 20 Frequency of WATT/ clinician recommendation matching level considering RTW status50

Table 21 Frequency of WATT top/ clinician recommendation matching level considering RTW status 50

Table 22 Frequency of Real program/ clinician recommendation matching level considering RTW status 51

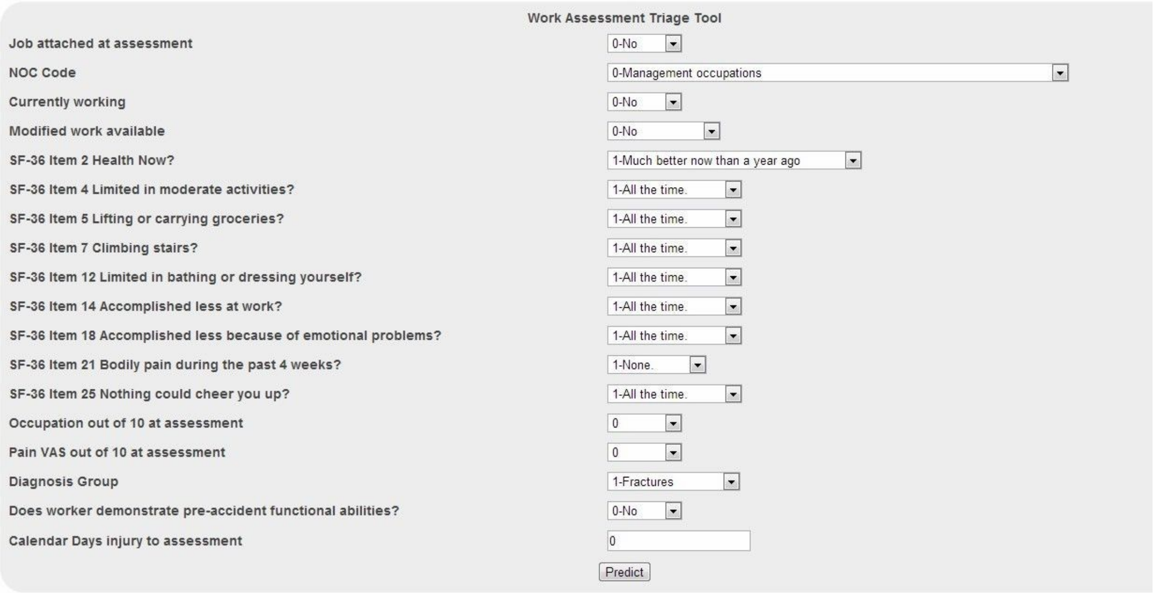
Chapter 1
Introduction

1.1. Problem Statement

Work-related musculoskeletal disorders (WMSDs) are some of the most common and costly problems among injured workers in Canada (1, 2). These disorders are also leading causes of workers' compensation time loss claims across Canada. According to records of the Workers Compensation Board of Alberta (WCB-Alberta), most injured workers recover and return-to-work (RTW) quickly. However, a minority of injured workers remain off work for longer periods of time and are responsible for the majority of associated health care and compensation costs(3).Various rehabilitation programs exist for injured workers with time loss injuries (for example, single service physical therapy or chiropractic treatment, multidisciplinary rehabilitation, chronic pain management programs). However, it is difficult to select the optimal treatment that will lead to successful RTW outcomes because individual response to treatment is highly variable. Currently, clinicians are unable to identify which claimants will respond best to the various treatment options resulting in referrals being made in a trail-and-error fashion(4). Given the substantial human, economic and societal burden of work injuries, improved health care and rehabilitation strategies are needed. Especially important are strategies aimed at reducing work absence and facilitating sustainable RTW.

Clinician decision support tools (CDSTs) are being widely researched in medical treatment, diagnosis and prognosis, and some of them have been applied in clinical practice (5-7). However, CDSTs related to musculoskeletal disorders are relatively rare (8). Given the high probability of workers suffering from musculoskeletal disorders, the uncertainty of appropriate rehabilitation methods recommended by clinicians, and the limitation of current clinical decision models, it would be useful to develop a validated CDST that could be used for many different types of musculoskeletal disorders. Accordingly, a triage algorithm and computer-based CDST named the Work Assessment Triage Tool (WATT) has been developed using data from WCB-Alberta(9). A graphical representation of the tool can be seen in Figure 1 below and the tool is available at the following website:

http://www.rehabresearch.ualberta.ca/doug_gross/watt/decisionApp.html.



Work Assessment Triage Tool

Job attached at assessment	0-No
NOC Code	0-Management occupations
Currently working	0-No
Modified work available	0-No
SF-36 Item 2 Health Now?	1-Much better now than a year ago
SF-36 Item 4 Limited in moderate activities?	1-All the time.
SF-36 Item 5 Lifting or carrying groceries?	1-All the time.
SF-36 Item 7 Climbing stairs?	1-All the time.
SF-36 Item 12 Limited in bathing or dressing yourself?	1-All the time.
SF-36 Item 14 Accomplished less at work?	1-All the time.
SF-36 Item 18 Accomplished less because of emotional problems?	1-All the time.
SF-36 Item 21 Bodily pain during the past 4 weeks?	1-None.
SF-36 Item 25 Nothing could cheer you up?	1-All the time.
Occupation out of 10 at assessment	0
Pain VAS out of 10 at assessment	0
Diagnosis Group	1-Fractures
Does worker demonstrate pre-accident functional abilities?	0-No
Calendar Days injury to assessment	0

Figure 1Interface of the Work Assessment Triage Tool (developed by Gross DP, etc.)

Promising results were reported for the internal validation study (i.e. the computer outperformed the human baseline recommendations). However the WATT has not been externally validated in another dataset, neither in a clinical trial nor using data from another rehabilitation setting. Prior to implementation into clinical practice, validity testing is a necessary and important procedure.

The current study was designed to evaluate the concurrent validity of the WATT against clinician recommendations made on workers' compensation claimants who were not part of the original development database (i.e. external validation).

1.2. Definition Terms

*Work-related Musculoskeletal Disorders(WMSDs):*WMSDs include a wide range of traumatic, inflammatory and/or degenerative conditions affecting the muscles, tendons, ligaments, joints, peripheral nerves, and supporting blood vessels. These include clinical syndromes such as tendonitis and related conditions, nerve compression disorders (carpal tunnel syndrome, sciatica), osteoarthritis, low back pain and other regional pain syndromes not attributable to known pathology (10).

Workers' Compensation Board of Alberta (WCB): The WCB is a statutory corporation created by government under the Workers' Compensation Act to administer a system of workplace insurance for the workers and employers of the province of Alberta. The organization is employer funded and aims to provide cost-effective disability and liability insurance. WCB compensates injured workers for lost income, health care and other costs related to a work-related injury (see WCB website: www.wcb.ab.ca).

Clinical Decision Support Tools (CDSTs): CDSTs are defined as any electronic or non-electronic system designed to aid directly in clinical decision making, and use characteristics of individual patients and their signs and symptoms to generate patient-specific assessments or recommendations that are then presented to clinicians for consideration(11).

*Return-to-work (RTW):*RTW can refer to the process of returning an injured worker to work or it can refer to vocational outcomes, which may involve return to the pre-injury employer or the pre-injury job (12). In this study, RTW is a vocational outcome judged as whether or not the injured worker receives compensation (i.e. wage replacement benefits) from WCB-Alberta.

Work Assessment Triage Tool (WATT): The WATT is a web-based online clinical decision support system designed to assist with identifying the most appropriate rehabilitation program for claimants off work due to a variety of musculoskeletal disorders (e.g.: fractures, dislocations, joint disorders, etc.) (9).

Concurrent Validity: This form of validity is assessed empirically by comparing a new test/ instrument with an established gold standard measuring the same phenomenon. If the new test/ instrument and the gold standard exhibit good agreement, the new test/ instrument can be deemed to hold concurrent validity (also referred to as criterion validity) (13). In this study, the gold standard refers to rehabilitation recommendations made by clinicians while the WATT is the new test/ instrument.

External Validity: Validity of generalized inferences in scientific studies. It explains whether or not an observed causal relationship should be generalized to and across different measures, persons, settings and times(14). In this study, external validation will be conducted to test the WATT, in a separate dataset from the one it was developed on, prior to its use in clinical practice.

Chapter 2

Literature Review

2.1.Work-related Musculoskeletal Disorders

WRMDs are widespread internationally with substantial associated costs along with personal and social burden. According to the Association of Workers' Compensation Boards of Canada, the frequency of work-related musculoskeletal disorders in workplaces in Canada was 2.26 per 100 workers. Of those experiencing such injuries, the number of accepted time loss claims was 0.3 million in Canada in 2007 (15). In addition, several studies demonstrate that in the United States, Canada, Finland, Sweden, and England, musculoskeletal disorders cause more work absenteeism and disability than any other disease group (16-19). Alberta is located in western Canada and is Canada's largest producer of conventional crude oil, synthetic crude, and natural gas. As a major industrial province, the percentage of workers and the frequency of WRMDs occurring among workers are relatively high, which leads to many workers suffering from MSK conditions.

Among the WRMDs, low back pain and associated disorders are some of the most common and have been studied the most(20). Upper extremity musculoskeletal symptoms also rank high among all MSK disorders. Numerous surveys of working populations have reported upper extremity symptom prevalence of 20% to 30% or higher (10). The least often affected area is the lower extremity.

Following the onset of WRMDs, most injured workers recover and RTW quickly. However, a small minority remains off work for longer periods of time and are the group that is responsible for the majority of associated health care and compensation costs. Given that individual response to treatment is highly variable, clinicians are currently unable to identify which claimants will respond best to the various conservative treatment options available, thus referral is often made in a trail-and-error fashion (4). Because of the substantial human, economic and societal burden, improved and creative health care and rehabilitation strategies are urgently needed.

2.2.WCB-Alberta Soft Tissue Injury Continuum of Care Model

From 1996 to 1997, WCB-Alberta implemented a continuum of care healthcare delivery model for soft tissue conditions. A continuum of care has been defined broadly as a coordinated array of settings, services, providers, and care levels in which health, medical, and supportive services are provided in the appropriate care setting. This may include treatment in an acute hospital, outpatient department, or community setting. Ideally, the patient receives health care at the most appropriate time and site according to their stage of recovery and level of need, and strong continuity and linkages exist between services within the system(21) . Continuums of care have been used successfully in various areas of health care, including nutrition and dietetics services, provision of hospital-based nursing and multidisciplinary care, as well as psychiatry and mental health. A study conducted by Stephens and Gross comparing two groups (an intervention group of claimants with soft injuries and a control group of claimants with non-soft injuries) using the WCB-Alberta administrative database demonstrated that the implementation of a soft tissue injury continuum of care involving staged application of various types of rehabilitation services lead to more rapid RTW and sustained recovery (22).

The WCB-Alberta Continuum of Care Model (See Figure 2) involved 3 main components: 1) staged application of different types of rehabilitation services depending on the progress of recovery; 2) case management protocols and checkpoints integrated into case planning; and 3) contracted services with 4 types of rehabilitation service providers (physical therapy, chiropractors, multidisciplinary assessment centers, and multidisciplinary rehabilitation providers) (22).

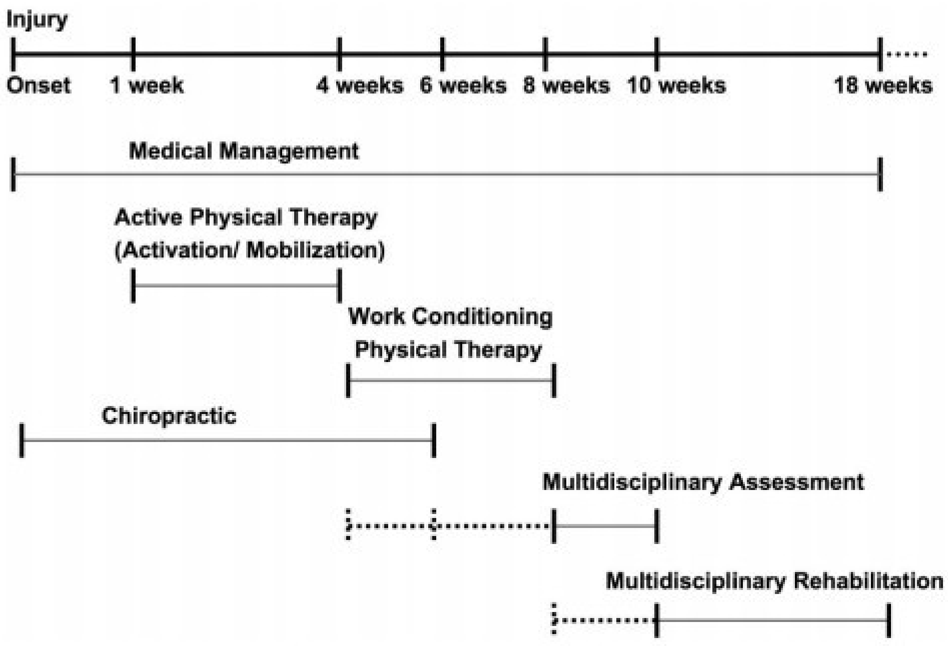


Figure 2 WCB- Alberta Continuum of Care Model (22)

There are five rehabilitation interventions derived from the WCB-Alberta Soft Tissue Injury Continuum of Care Model which will be studied in this research(9).These include:

A. **Provider Site-based Rehabilitation**: 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.

B. **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 (i.e. low intensity transitional work) to help claimants stay at work.

C. **"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.

D. **Complex Interdisciplinary Bio-psychosocial 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-behavioral approach, and RTW planning through stakeholder negotiation.

E. **Other intervention**: This involves either no rehabilitation or referral back to a single service provider (i.e. physical therapy or chiropractic).

While these various rehabilitation programs exist for injured workers with time loss injuries, it is currently difficult to select the optimal treatment that will lead to successful RTW outcomes. Individual response to these treatments is highly variable. Currently, clinicians are unable to identify which claimants will respond best to the various treatment options and treatment is often made in a trail-and-error fashion (4). Given the substantial human, economic and societal burden, improved health care and rehabilitation strategies are needed. One option is the development and use of CDST.

2.3.Clinical Decision Support Tools

Health care policy makers, researchers and front line clinicians are making efforts to find improved and creative health care and rehabilitation strategies that benefit injured workers and facilitate RTW. In the recent decade, researchers have developed a variety of models that consider the multidimensional aspects of RTW. Those models include biomedical models, bio-psychosocial models, and more complex models that identify multiple legal, administrative, social, political, health care and cultural factors that influence RTW(15). It is now recognized that individual, environmental, as well as other psychosocial factors, play an important role in RTW interventions and are thus increasingly the subject of more and more research in this area.

Individual factors seem to play a role in predicting which injured workers will respond best to rehabilitation interventions. Several studies (23-27) have demonstrated that younger age, fewer days of work loss, availability of a job to return to or a strong connection to the work force, and high patient expectations of recovery have consistently been reported to increase the likelihood of RTW following treatment. Moreover, in recent years researchers have identified individuals' characteristics that can be used to target interventions to achieve optimal outcomes. Statistical models, classification algorithms and experts' models have been built to help manage patients with musculoskeletal disorders and to facilitate early RTW (25, 28, 29). These models/ algorithms have the potential to identify which interventions are best for sub-groups of injured workers and identifying the optimal intervention for each individualized based on unique worker characteristics toward optimal rehabilitation. Some have been validated in rehabilitation settings or using data of specific patients, but most still need further evaluation and external validation.

2.3.1. A Brief History of Clinical Decision Support Tools

CDSTs are used to improve medical performance and patients' outcomes, to reduce errors in diagnosis and management, and to reduce health care expenditures. The scientific evidence about CDSTs has grown in recent years, with contributions from researchers in various disciplines and fields. CDSTs have dramatically improved in many aspects, including an increase in the number of health care settings where CDSTs are used, along with increasing types and sophistication of CDSTs.

CDSTs have a history of nearly 40 years. Early in 1975, Shortliffeetet al. described the MYCIN system used for clinical therapeutics. This was an early expert system that used artificial intelligence to identify bacteria causing severe infections, such as bacteraemia and meningitis, and to recommend antibiotics with the dosage adjusted for the patient's body weight. The name derived from the antibiotics themselves, as many antibiotics have the suffix "-mycin". This is one of the earliest famous examples of CDSTs (5). Another early example is the Quick Medical Reference (QMR) (6), which was developed from an expert consultation program for diagnoses in general medicine. This is a three-level program containing an average of 85 findings and 8 associated disorders relevant to the diagnosis of approximately 600 disorders in internal medicine.

The twentieth century witnessed a number of CDSTs being produced in many aspects of general medicine, including drug dosing, preventive care, and other aspects of diagnosis and management. PROTÉGÉ is another example (7). It is a suite of tools and methodologies for building knowledge-based systems and domain-specific knowledge-acquisition tools. PROTÉGÉ was first used in providing protocol-based decision support in the domain of treating diseases such as Human immunodeficiency virus (HIV). Other representatives of CDSTs developed at that time are CADUCEUS, DiagnosisPro, Dxplain, among others (30)

CDSTs have developed rapidly since the start of the twenty-first century because of the development of modern technology and computer science. CDST has also been recently used for the diagnosis and management of musculoskeletal disorders (8, 28, 31-33). The recent categorization model proposed by Shaw et al. for the rehabilitation of claimants with low back pain is one example. This theoretical model indicates that specific forms of rehabilitation maybe selected based on specific barriers to recovery identified during clinical evaluation (28).

2.3.2. The Three Forms of Clinical Decision Support Systems

There are three approaches to developing CDSSs that have been used to date, including the Knowledge Base, Expert System and Predictive Algorithm approach (34). Each of these will be discussed in detail.

(1) Knowledge Base

Knowledge Bases are the oldest form of CDSS, and among which clinical textbooks and journals are some of the oldest forms of Knowledge Base. Knowledge Bases have become more prevalent and more available than ever before with the advent of modern computer technology and pervasive computer networks. However, as vast as these online Knowledge Bases are, they require clinicians to conduct a search, find the information of interest, then read and interpret the findings. Hence, efforts have been made to try and develop more “directed” Knowledge Bases that could assist clinicians in finding the right information and attempt to provide some quality control.

(2) Expert Systems

The most traditional and widespread CDSSs are known as Expert Systems. Expert Systems are largely static, deterministic systems developed as sets of rules by groups of peer-acknowledged experts. The best-known examples are the drug-drug interaction warnings built into many commercial computerized physician order entry systems. These Expert Systems are developed using information from the research literature, but also allow individual institutions to implement additional rules based upon their individual experience. Most clinical guidelines are also Expert Systems, where expert panels digest the literature to develop rules and guidelines that are relatively simple and easy to follow.

(3) Predictive Algorithms

Most CDSSs being developed today have some type of Predictive Algorithm as their foundation. The term “Predictive Algorithm,” is broad and rather vague, and it can represent anything from a simple two-part scoring system through to complex computational models. The process of developing Predictive Algorithms is almost as varied as the algorithms themselves. Some algorithms are as simple as calculating Hazard Ratios, some use Regression Models or a Logistic Regression variant, while others use more complex methods that can represent extremely complex systems, such as Neural Networks, Bayesian Belief Networks, or Decision Trees. Some of these methods can be enhanced by applying Machine Learning algorithms to discover additional information, Bootstrapping to evaluate sample population robustness, or Random Forest analysis to address consistency of modeling.

2.3.3. The Use of Clinical Decision Support Tools in Work-Related Musculoskeletal Disorders

CDSTs have been utilized in many aspects of general medicine, and they are also useful in the rehabilitation of painful musculoskeletal disorders. Over the past decades, research has contributed to the development of CDSTs for WRMDs, especially back pain and regional musculoskeletal pain (31-33). The application of CDSTs has been shown to be beneficial in the diagnosis of back pain (32). Lin and colleagues implemented and evaluated a web-based decision support system that used an intuitive and easy-to-use framework for assessing the patient information. The tool provides a diagnosis consisting of one or multiple parts. When compared with expert opinion, the system performed at a comparable level, which confirmed the effectiveness of such decision support systems (32).

Unlike back pain, studies on the use of CDSTs in neck, shoulder or other parts of body are relatively rare. It is difficult to find related articles about the use of CDSTs in other WRMDs. For example, we were able to find only one article about a CDST for shoulder pain pathology (35). Although back pain makes up the largest diagnostic category of WRMDs, it typically accounts for less than 40% of claimants who are off work due to musculoskeletal conditions. Thus, it is important to expand research on CDSTs to other musculoskeletal pain and related diseases besides back pain, in order to address the needs of all injured workers requiring rehabilitation.

2.4. Validity Testing of Clinical Decision Support Tools and the Gold Standard

It is important and necessary to test the validity of any CDST by comparing its results with the currently accepted gold standard before applying it widely into clinical settings. Omission of this step may lead to dangerous and potentially fatal errors if clinicians rely only on a system with outputs of uncertain quality. Unfortunately, the majority of CDSTs available are not well tested prior to release (36). The main theory of validation testing is to determine the level of agreement and relationship between a certain CDST and a “gold standard”. When a measure or tool is highly correlated with a gold standard, the specific CDST has a greater possibility of being clinically valid. On the contrary, a low correlation would indicate lower validity and less trustworthiness when implemented into practice. In medicine, the term “gold standard” refers to the most reliable and accurate diagnostic method in the clinic, which is also called standard diagnosis method. For example, commonly used gold standards are diagnostic imaging (i.e. CT, B ultrasonic), surgeon ratings, or pathological examination. However, in practice there are sometimes no true "gold standard" tests. Sometimes the gold standard are called "perfect" or "alloyed" gold standards (37). In the diagnosis and management of regional pain disorders where no specific findings can be seen on imaging studies, the gold standard is often the clinical diagnosis provided by experienced clinicians, including doctors and/or therapists. Two clinical decision support tools with validation will be discussed below.

2.4.1. The Shaw Model

In 2006, Shaw et al. (28) proposed a knowledge-based clinical decision support approach developed by reviewing 17 articles obtained from all English publications within past 5 years. The authors’ initial objective was to assess the extent to which effective strategies for reducing work absence after acute low back pain matched empirical risk factors. Two literature searches were conducted. The first search identified current review articles summarizing risk factors for low back disability. The second search was for recent review articles summarizing effective RTW interventions to reduce sickness absence following the onset of acute low back pain. Then, the authors evaluated the correspondence between effective interventions and known risk factors by analyzing the responses to four criterions: answering “yes” to at least three questions indicates a high correspondence; answering “yes” to two questions indicates a moderate correspondence; otherwise the correspondence is low. Through the evaluation of both groups of articles, the authors developed two figures that describe low back disability risk factors and low back interventions, respectively. They then proposed a hypothetical risk factor based intervention strategy for low back pain(28). This model indicates that specific forms of rehabilitation interventions could be selected based on specific barriers identified during clinical evaluation. It also highlights important directions for future research and some new approaches to workplace and clinical intervention. (Figure 3)

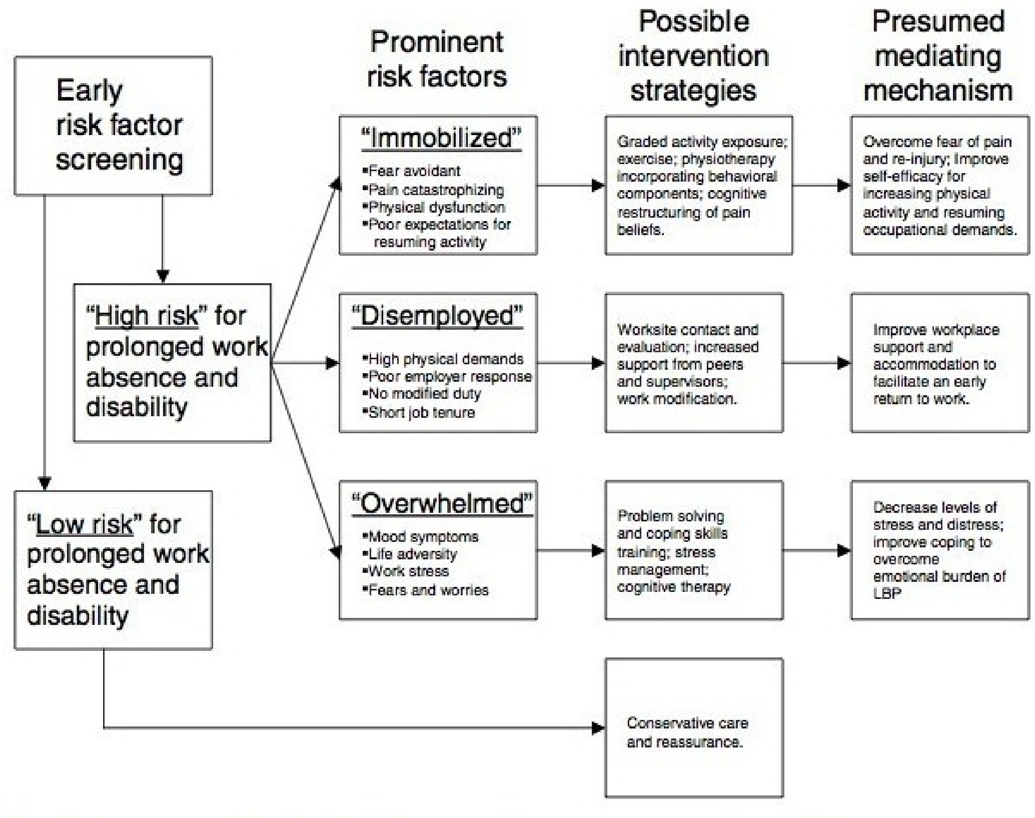

Early risk factor screening
Prominent risk factors
Possible intervention strategies
Presumed mediating mechanism
"High risk" for prolonged work absence and disability
"Low risk" for prolonged work absence and disability
"Immobilized"
•Fear avoidant
•Pain catastrophizing
•Physical dysfunction
•Poor expectations for resuming activity
Graded activity exposure; exercise; physiotherapy incorporating behavioral components; cognitive restructuring of pain beliefs.
Overcome fear of pain and re-injury; Improve self-efficacy for increasing physical activity and resuming occupational demands.
"Disemployed"
•High physical demands
•Poor employer response
•No modified duty
•Short job tenure
Worksite contact and evaluation; increased support from peers and supervisors; work modification.
Improve workplace support and accommodation to facilitate an early return to work.
"Overwhelmed"
•Mood symptoms
•Life adversity
•Work stress
•Fears and worries
Problem solving and coping skills training; stress management; cognitive therapy
Decrease levels of stress and distress; improve coping to overcome emotional burden of LBP
Conservative care and reassurance.

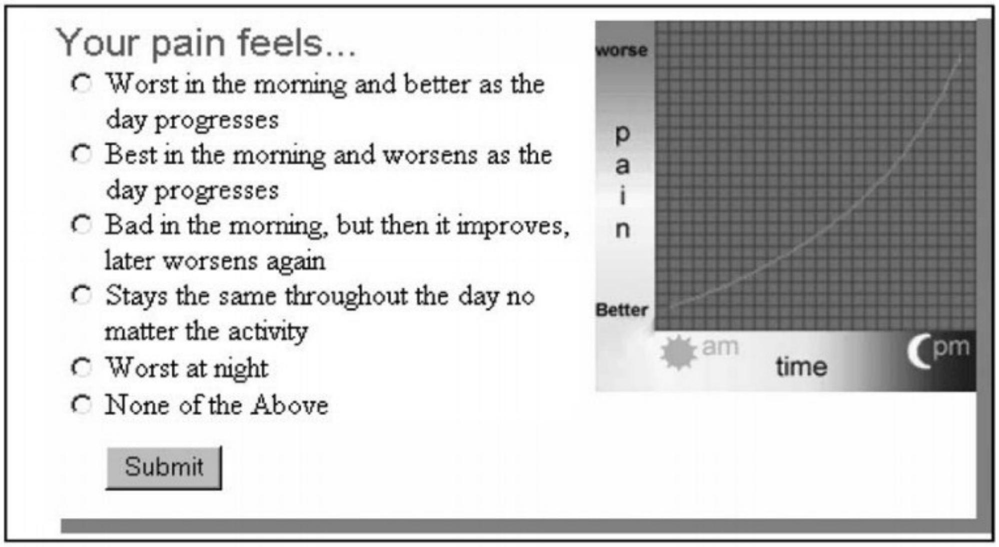
Figure 3 A hypothetical risk factor based intervention strategy for low back pain(28)

Steenstra et al. tested the Shaw Model in 2010 in a cohort of participants off work due to low back pain (8). The validation study focused on 442 claimants in the Readiness for Return-to-Work Cohort Study. Claimants in the cohort who had already returned to work approximately 1 month post-injury (n=259) were categorized as the low-risk group. A latent class analysis was performed on 183 workers still absent from work, categorized as the high-risk group. The results showed that three classes were identified: (1) workers with workplace issues; (2) workers without workplace issues, but with back pain; and (3) workers having multiple issues. Classes 2 and 3 had a similar rate of RTW; both worse than the rate of RTW for class 1.However, RTW status and recurrences at 6 months were similar in all 3 groups. Results largely confirm that several subgroups could be identified based on previously defined risk factors as suggested in the

Shaw model. Thus, different groups of workers might be identified and might benefit from different interventions.

2.4.2. The Lin Decision Support Systems(32)

Lin et al. designed and implemented a Web-based decision support system that employs an intuitive and easy-to-use rule based framework to assess the patient's information and recommend a diagnosis for low back pain (See Figure 4). In the same study, the authors also conducted comprehensive evaluation including knowledge base verification, system validation using a modified Turing test which is a test of machine's ability to exhibit intelligence behaviors equivalent to a human and clinical efficacy assessment involving 5 clinicians and 180 real-world cases collected from geographically dispersed clinics. The authors were confident enough about their system and evaluation to claim that the proposed system was ready for clinical use.

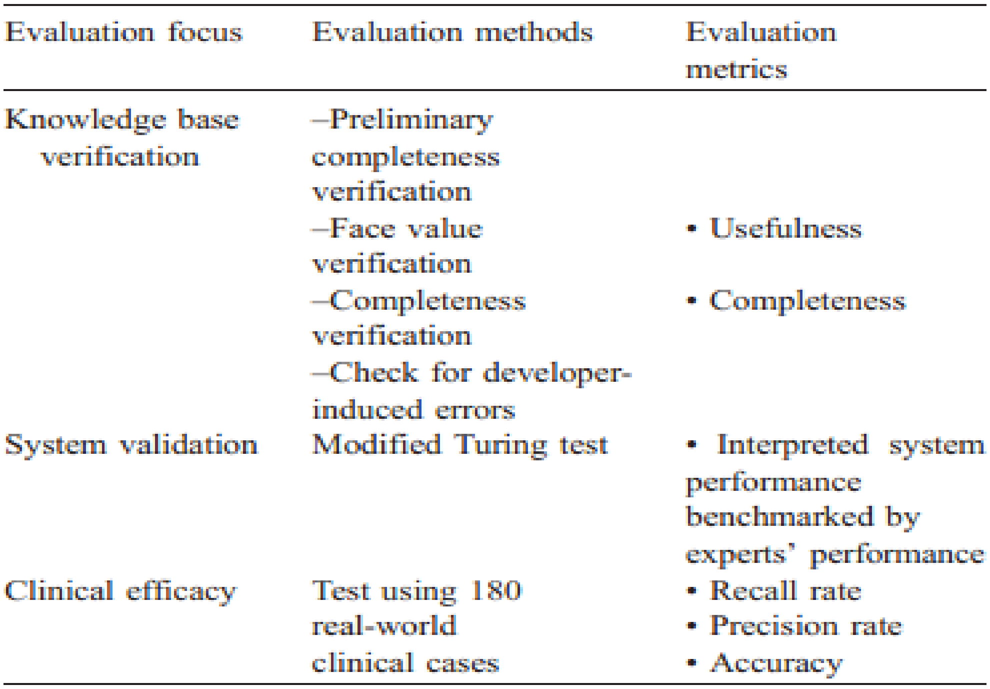


Your pain feels...

- ○ Worst in the morning and better as the day progresses
- ○ Best in the morning and worsens as the day progresses
- ○ Bad in the morning, but then it improves, later worsens again
- ○ Stays the same throughout the day no matter the activity
- ○ Worst at night
- ○ None of the Above

Figure 4 Example of the diagnostic interface (32)

The Lin System was inherently a predictive diagnostic algorithm in nature. The process of the system development was described in detail, and the authors also provided a summary of their evaluation process (See Figure 5). In this study, the real-world clinical cases (clinical recommendations) were viewed as the gold standard. The authors obtained generally positive results during the evaluation of functional capacity and health performance based on multiple performance metrics. Thus, they further confirmed the effectiveness of their CDST. The system seems to embrace important and verifiable diagnostic knowledge, and is capable of performing at a level comparable to domain experts and exhibits encouraging clinical efficacy.



Evaluation focus	Evaluation methods	Evaluation metrics
Knowledge base verification	–Preliminary completeness verification	
	–Face value verification	• Usefulness
	–Completeness verification	• Completeness
	–Check for developer-induced errors	
System validation	Modified Turing test	• Interpreted system performance benchmarked by experts' performance
Clinical efficacy	Test using 180 real-world clinical cases	• Recall rate • Precision rate • Accuracy

Figure 5 Evaluation framework - focus, method and metrics (32)

2.5.Work Assessment Triage Tool

In order to address the limitations and shortcomings identified above (such as limited research on patients with conditions other than low back pain), a triage algorithm and computer-based decision support tool named the Work Assessment Triage Tool (WATT) was developed for injured workers in Alberta using data from WCB-Alberta (9).

2.5.1. The Process of Developing the WATT (9)

The WATT was developed using machine learning technology. Machine learning is a technology in computing science, which builds classification systems using multiple independent and dependent variables (38). The goal of machine learning is to build a concise model of the distribution of categorical labels in terms of predictive features to overcome human errors in making complex classifications (38). All data used for developing the WATT was extracted from the WCB-Alberta provincial database of 8,611claimants undergoing RTW assessment between December 2009 and January 2011. Data were available on more than 200 features of these claimants including numerous personal, clinical, occupational and social variables measured at time of RTW assessment. Additionally, information was available on rehabilitation programs undertaken and RTW outcomes following the interventions. The process of model development and evaluation involved identifying desirable characteristics, machine learning algorithm selection and training, and finally evaluation of the model and comparison with the accuracy of the clinician recommendations that were made at baseline.

The researchers extracted 4,876 successful cases from the database as the initial development dataset in order to create a tool to recommend a rehabilitation program that leads to successful RTW. The rest of the cases were used to train another model of negative rules which tell clinicians not to recommend a certain rehabilitation program. The training dataset consisted of five categories (provider site-based rehabilitation, worksite-based rehabilitation, hybrid functional restoration/ worksite-based rehabilitation, complex and other), each representing a specific rehabilitation intervention. However, the class distribution in the dataset (i.e. the number of claimants in each intervention) was severely imbalanced with very few claimants undergoing complex, workplace-based or hybrid interventions. In order to solve this problem the Synthetic Minority Over-sampling Technique (SMOTE) was used (39). Additionally, the researchers used the Tomek Link method to overcome class overlaps after sampling in order to avoid minority data generated by SMOTE from invading the majority class too deeply and causing classification difficulties (40). Moreover, since there were more variables than desired in the final model, a feature selection process was undertaken based on statistical data processing to select a certain amount of variables. Finally, there were 30 variables that remained in consideration for the final model. The process of sampling and cleaning described are shown in Figure 6.

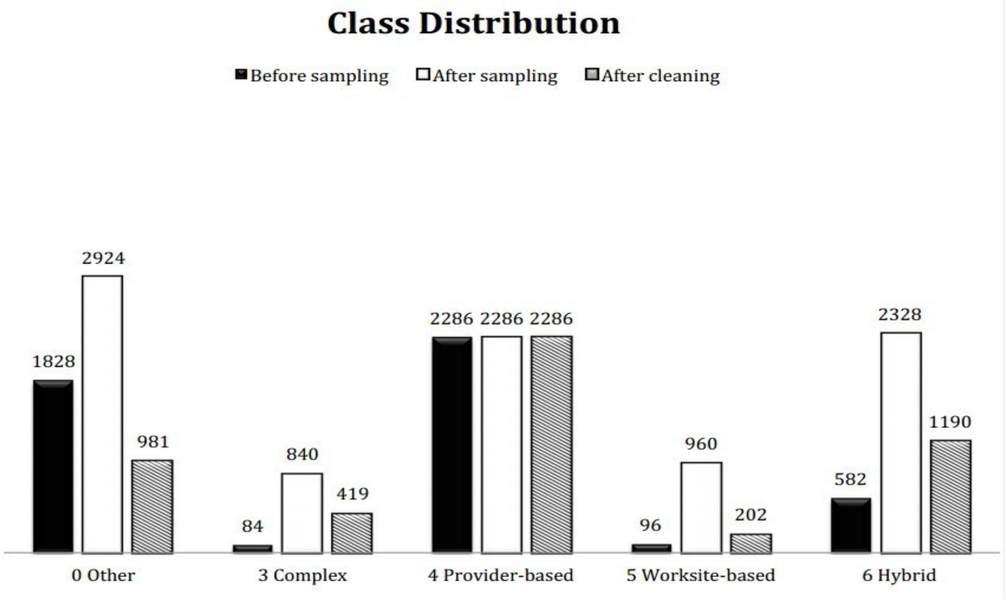


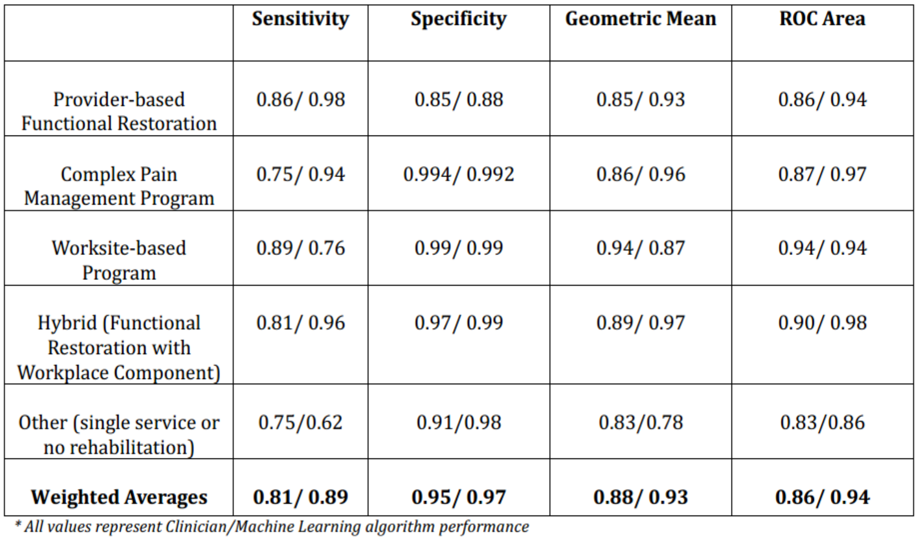
Figure 6 Final class distribution comparison after all data pre-processing was completed(9)

After data pre-processing and feature selection, Repeated Incremental Pruning to Produce Error Reduction (RIPPER) 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 process of RIPPER was beyond the scope of this research and details will not be discussed here (41).

Using the RIPPER algorithm, seventeen features were identified as important in the final rule set. The content of those features and their selectable answers are listed in Appendix 1. Other measures used in the WATT including theSF-36 Health Questionnaire, Pain Visual Analogue Scale (Pain VAS) and website of National Occupational Category code and Diagnosis Group code are also attached as appendixes (Appendix 2-4).

2.5.2. Evaluation of the Machine Learning Model and Clinician Baseline (9)

The clinician baseline was defined as the number of clinician recommendations of the "correct" rehabilitation program. The clinician recommendations were deemed successful when the recommendation matched the actual program undertaken which led to successful RTW and no repeat program was needed. Both the clinician baseline and machine learning models were evaluated using Sensitivity, Specificity, Geometric Mean of Sensitivity and Specificity, and Receiver Operating Characteristic (ROC) Area. The result of the evaluation process is shown in Figure 7.



	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**

** All values represent Clinician/Machine Learning algorithm performance*

Figure 7 Performance of the clinician baseline recommendation compare to the final machine learning algorithm (9)

Figure 7 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. However, the validation results show that machine performance was substantially higher than clinician decisions with sensitivity, specificity and ROC of 0.89, 0.97, and 0.94 in the positive rule set. Thus, the authors concluded that the use of machine learning classification techniques appears to have resulted in classification performance that was higher than clinician decision-making.

In conclusion, the authors suggested that the computer-based clinical decision support tool required additional validation and impact evaluation in clinical samples, ideally through randomized controlled trials. External validation of the WATT is the objective of this thesis. This study will contribute to the investigation of the psychometric properties of the WATT by testing its concurrent validity against a currently accepted gold standard: the recommendations of experienced clinicians.

2.6.Objective of this Study

The general goal of this study was to evaluate the concurrent validity of the WATT. Specific objectives included:

(1) evaluating the level of agreement between the rehabilitation programs recommended by the WATT and a gold standard defined as the rehabilitation intervention actually recommended by clinicians. In both cases, the potential interventions included: A. provider site-based functional restoration; B. worksite-based rehabilitation; C. “hybrid” functional restoration/ worksite-based rehabilitation; D. complex interdisciplinary bio-psychosocial rehabilitation; E. no further rehabilitation.

(2) evaluating whether WATT recommendations could improve upon the gold standard by investigating whether claimants with successful RTW but whose rehabilitation programs did not match with clinician recommendations would match better with the WATT recommendations.

(3) evaluating the ability of WATT to predict RTW when WATT recommendation/ WATT top recommendation/ actual rehabilitation programs that claimants undertook matched with the baseline clinician recommendations.

2.7.Research Hypotheses

The hypotheses for this study included that:

(1) rehabilitation recommendations from the WATT for injured workers will highly agree (i.e. ratio of agreement >0.7) with clinician recommendations in general, but will not be

exactly the same as the goal of the WATT is to improve upon clinician recommendations.

(2) the WATT will perform better than clinician recommendations for recommending rehabilitation programs that lead to successful RTW for injured workers. We hypothesize that claimants who successfully RTW, but whose rehabilitation program did not match with clinician recommendations would match better with the WATT recommendations,

(3) there will be at least a moderate increase in the likelihood of successful RTW for the injured workers when WATT/ WATT top/ actual programs match the clinician's recommendation.

Chapter 3

Methods and Procedures

3.1.Study Design

This is a cross-sectional concurrent validity study. Since concurrent validity is assessed empirically by comparing a new test or instrument with an established gold standard that measures the same phenomenon, this study design was deemed most appropriate. Approval was obtained from the University of Alberta Health Research Ethics Board.

3.2.Database Information

This study is a secondary analysis as data were obtained from a separate study involving injured workers. All data for this study were from a previous clinical trial evaluating RTW assessment techniques conducted at the WCB-Alberta Millard Health Centre(42). This database contains information on 434 workers' compensation claimants seen at Millard Health for RTW assessment, and is held by Dr. Doug Gross in the University of Alberta's Common Spinal Disorders lab. The database contains:1) individual claimant-level information used by the WATT algorithm; 2) demographic information such as age, gender and education level; 3) information on the rehabilitation intervention recommendations made by clinicians and the corresponding actual interventions they truly undertook; and 4) the associated RTW outcome for each claimant.

3.3.Sample

From November 2011 to June 2012, consecutive claimants seen at the rehabilitation facility were enrolled in the clinical trial and entered into the database. Although we included all the clients at this period of time, the data was only from WCB-Alberta Millard Health which is an occupational rehabilitation and disability management service in Edmonton. Thus, the sample for this study is 434 workers' compensation claimants undergoing RTW assessment. In the database, the label "assessment type" divided the claimants into two groups: claimants undergoing basic functional capacity evaluation (BFCE) and claimants receiving comprehensive functional capacity evaluation (CFCE). Data used to develop the WATT were from claimants who primarily received BFCE, thus, it was necessary to divide this dataset into two groups when validating the WATT. BFCEs provide a "snapshot" of a client's current level of function compared to the demands of their job (what they can and cannot do). This assessment helps determine if the claimant requires more rehabilitative treatment or services in order to meet the demands of their job. CFCEs test the claimant's overall work capabilities. The two-day CFCE assessment helps determine the claimant's actual work abilities and possible restrictions through a physical exam and functional testing (definitions of BFCE and CFCE come from WCB-Alberta website). Basically, claimants undergoing CFCE have more chronic conditions and severe barriers to RTW than claimants receiving BFCE. For this study, both claimants undergoing BFCE and CFCE were included. However, stratified analysis was also performed according to type of assessment received. Since the WATT is applicable to all injured workers undergoing RTW assessment, there were no other inclusion or exclusion criteria.

3.4.Data Collection

The independent variable in this study is the WATT recommendation while the dependent variable is the criterion gold standard of clinician recommendation. We also have information on the actual rehabilitation program undertaken after the RTW assessment as approved by the WCB-Alberta case managers, and claimants' RTW outcome 30 days after the clinicians' assessment determined by wage replacement status (RTW or failed RTW).

3.4.1. Gold Standard – Clinician Recommendations

As mentioned, the gold standard in this study was clinician recommendations made after RTW assessment. After RTW assessment, every claimant had only one intervention recommended from the assessing clinician. And all the recommendations were based on the WCB-Alberta soft tissue injury continuum care model that outlines possible rehabilitation programs (22). The rehabilitation options for clinicians are listed below.(9)

A. **Provider Site-based Rehabilitation**: 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.

B. **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, confirming job demands, participatory ergonomics and identification of suitable duties to help claimants stay at work.

C. "**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.

D. **Complex Interdisciplinary Bio-psychosocial Rehabilitation**: This is a comprehensive pain management program for claimants with more 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-behavioral approach, and RTW planning through stakeholder negotiation.

E. **Other intervention**: This involves either no rehabilitation or referral back to a single service provider (i.e. physical therapy or chiropractic).

3.4.2. The WATT Recommendation

Clinician recommendation variable was available in the database; thus the main task of data collection was determining the rehabilitation program recommended by the WATT. Rehabilitation options provided by the WATT were exactly the same as the possible clinician recommendations discussed above. However, unlike the clinicians who can make only one recommendation for each claimant, the WATT often provides multiple recommendations for each claimant. The WATT output includes recommended program names, anticipated duration, level of confidence in the recommendation and specific rules leading to the recommendation. Among all the positive recommendations provided, a top recommendation can be selected based on: 1) shorter anticipated program duration; 2) more specific rules supporting the recommendation; and 3) higher confidence in the recommendation. In addition to providing positive recommendations (suggesting programs to consider), the WATT also provides negative recommendations (suggesting programs to avoid).

As mentioned above, there are eighteen features identified by machine learning techniques in the WATT interface: job attachment and working status at time of RTW assessment, availability of modified work, National Occupational Classification Code, diagnostic group, calendar days injury to assessment, the Pain Disability Index (PDI) 'Occupation' item out of 10 at assessment, Pain Visual Analog Scale (VAS) out of 10, and the following items from the SF-36: 2 (health now?), 4 (limited in moderate activities?), 5 (lifting or carrying groceries?), 7 (climbing stairs), 12 (limited in bathing or dressing yourself), 14 (accomplished less at work), 18 (accomplished less work because of emotional problem), 21(bodily pain during the past 4 weeks), 25 (nothing could cheer you up). Demonstration of pre-accident functional ability was added to the WATT based on expert opinion and is not available in the dataset, but is not entirely relevant for this study because it was not derived from the computing technique that includes the programs that formed the basis of the recommendations. The self-report clinical measures used in the WATT are shown in the Appendix l.

To determine WATT recommendations, a computing program was developed and used to input data on the WATT features from the WCB dataset directly into the WATT by one of its developers. This allowed the researcher to overcome potential mistakes from manual data entry. Through this process, the WATT recommendations were obtained for all subjects. Three examples of case scenarios for obtaining the WATT recommendations are given in Appendix 5 to better clarify the WATT recommendation data collection process.

3.4.3. List of Descriptive Variables

Variables listed below are the claimants' demographics that were included in the dataset and were used for descriptive analyses. They are: age; days from injury to assessment; gender; education level; marital status; job attached at the time of assessment; working status at the time of assessment; availability of modified work; MSK diagnostic group based on ICD9 Coding; the requirement of a interpreter; the assessment recommendation by clinicians; the percentage of wage replacement benefits claimants 30 days post assessment, and the actual intervention programs the claimant participated in.

3.4.4. Real Intervention Program

In WCB-Alberta, case managers evaluate the need for a certain rehabilitation program after receiving recommendations from clinicians conducting the RTW assessment. The case managers will make a final decision based on previous experiences with the claimants, the program cost and other social factors. There is usually high agreement between clinician recommendations and actual programs claimants undertaken because case managers infrequently revise or totally act on clinician recommendations. However there is discrepancy at times in making recommendations as case managers typically have a broader perspective on the case than clinicians and may act on clinician recommendations. (Personal communication with Dr. Doug Gross)

3.4.5. Claimants' RTW Outcome

Successful acute RTW was defined as "when workers were no longer receiving wage replacement benefits at 30 days after RTW assessment". On the contrary, failed RTW was defined as "when workers were still receiving wage replacement benefits at 30 days after RTW assessment"(9). Within the WCB-Alberta jurisdiction, claimants receive daily wage replacement benefits when they are off work for an entire day.

3.5.Statistical Analysis

3.5.1. Statistical analysis using R

In order to overcome potential mistakes in data entry to generate WATT recommendations by hand and to save time, we used R 3.0.2 version for computer programming. The programming included:

(1) Selecting the top WATT recommendation for claimants.

(2) Determining whether two recommendations matched or not, especially when it came to multiple WATT recommendations.

3.5.2. Statistical analysis using Statistic Package for Social Science (SPSS)

The procedures below were conducted using SPSS version 20.0.0 (SPSS Inc., Chicago, IL, USA).

Data reviewing

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 for all claimants included in the study. Data for any claimants with missing data on key variables needed for the WATT were excluded.

Claimants' demographics

Claimants' demographics were calculated including means and standard deviations for continuous variables (age, time from accident to admission), and modes and percentages for categorical variables (e.g. gender, education level, marital status, working status, diagnostic groups, clinicians' recommendations.). Difference on these variables between development database and validation database and also between BFCE group and CFCE group were calculated and compared using student's t-test and chi-square test. Significant differences between the groups for demographic variables(development dataset vs. comparison standard) were determined with one sample t-test by setting development data as the given population for continuous variables and difference on continuous variables between BFCE and CECE group were tested using independent t-testing. All nominal variables were determined using chi-square testing. Alpha level was set as 0.05.

Rehabilitation program frequency distribution

The frequencies of each recommendation made by clinicians and the WATT under all situations (outcome: RTW and failed RTW; assessment type: overall, BFCE and CFCE) were calculated.

Percent agreement of recommendations

In order to accomplish the first objective of the study, we needed to find the level of agreement between the WATT and clinician recommendations. Since both recommendations are categorical data, the most common methods are percent agreement and Cohen's Kappa (43). Percent agreement measures how often test-retest scores agree, which is the simplest index of agreement, and is denoted by the number of exact agreements divided by the number of possible agreements. Though simple and obvious, the method always overestimates true agreement since some proportion of the results could have occurred by chance. Accordingly, the Kappa statistics is more used as a chance-corrected measure to evaluate categorical agreement (43).

Initially, we planned to use Cohen's Kappa; however, it turned out to be inappropriate since the data were not symmetrical. That is to say, the data failed to meet the condition to perform Kappa. For example, Table 1 and Table 2 were Kappa statistics preparation for provider-based rehabilitation and worksite-based rehabilitation.

Table 1 Frequency of WATT/ clinician recommendation for provider-based rehabilitation

	WATT recommend	**WATT not recommend**
Clinician recommend	115	0
Clinician not recommend	317	0

Table 2 Frequency of WATT/ clinician recommendation for worksite-based rehabilitation

	WATT recommend	**WATT not recommend**
Clinician recommend	0	0
Clinician not recommend	209	223

Thus, the WATT recommended provider-based rehabilitation for all the claimants and clinicians never recommended worksite-based rehabilitation to any claimants. Kappa value could not be calculated based on these data. Conclusively, percent agreement, although not perfect, was used as the most appropriate and applicable method for agreement calculation in this case.

The percent agreements were calculated by determining the number of WATT recommendations and clinician recommendations matched for the same claimant divided by the total number of claimants. Given that there was only one recommendation for each claimant made by clinicians but multiple recommendations by the WATT, two percent agreements were calculated: (1) the percent agreement between the gold standard (i.e. clinician recommendation) and the general

WATT recommendations (i.e. whether the rehabilitation program was within the list recommended by the WATT); and (2) the percent agreement between gold standard (i.e. clinician recommendation) and the top WATT recommendation. For the first comparison, once the clinician recommendation matched any one of multiple WATT recommendations, it was considered as “agreement”. For the second comparison, if the top WATT recommendation with shortest treatment duration, the most rules and highest confidence among all the recommendations shown on the WATT matched the clinician recommendation, then was considered as “agreement”. In this case, it was considered as “agreement” when clinician recommendations were exactly the same as the top WATT recommendation. The percent agreements for each of these scenarios were calculated by the number of “agreements” divided by total number of recommendations. We also calculated the ratio of agreement between clinician recommendations and the real program undertaken. We did not compare negative WATT recommendations since clinicians did not provide negative recommendations; thus we only have negative recommendations by WATT. Thus, agreement cannot be calculated in this case.

There are no specific guidelines for interpreting percent of agreement. We based our interpretation of the percentage of agreement following the guidelines used to interpret Cohen Kappa because of the goal of our study. These guidelines are as follows: Excellent agreement: 0.93~1.00; very good agreement: 0.81~0.92; good agreement: 0.61~0.80; fair agreement: 0.41~0.60; slight agreement: 0.21~0.40; poor agreement: 0.01~0.20; no agreement: 0.00 or less.

The comparison between the WATT and clinician recommendations

The initial idea of the second objective was to investigate whether the claimants RTW whose rehabilitation program did not match with the clinicians will match better with the WATT. Table 3 was built to demonstrate the frequency in each cell.

Table 3 Frequency of matching program (WATT and real program comparison as column) for claimants RTW

	Clinician and real program matched	**Clinician and real program not matched**
WATT and real program matched	A	B
WATT and real program not matched	C	D

According to this table, the sum of *b* and *d* is the number of claimants RTW whose rehabilitation program did not match with clinician recommendation. The hypothesis will be confirmed if *b/ (b+d)* could be approaching *1*, indicating those successful rehabilitation programs that did not match with the clinician but that matched well with the WATT recommendation. Thus, the WATT could improve upon the clinician recommendation because the WATT matched more highly with the successful program leading to RTW while clinician recommendations failed in this case. The range of value *b/ (b+d)* is between 0 and 1. In addition, Table 4 was built as a supplement to answer the second objective more comprehensively.

Table 4 Frequency of matching program (WATT and clinician comparison as column) for claimants RTW

	Clinician and real program matched	**Clinician and real program not matched**
WATT and clinician matched	A	b
WATT and clinician not matched	c	d

According to Table 4, the sum of *b* and *d* is the number of claimants RTW whose actual rehabilitation program did not match with clinician recommendation. If *b* could be very small or even zero, it can be concluded that WATT recommendations also did not match with clinician recommendations that turned out to be unsuccessful in terms of RTW. In this case, WATT is different from the failed clinician recommendations that indicated it is not likely to recommend ineffective programs which could not lead to RTW.

Likelihood ratio of matching programs

In order to achieve the third objective, the frequencies of matching programs between actual programs claimants undertook/ WATT/ WATT top recommendation and clinician recommendations considering RTW status were listed in Table 5. Likelihood ratios were calculated based on those frequencies. In total, three pairs of likelihood ratios were obtained.

In this study, likelihood ratio (LR) tells us the predicting ability of RTW when rehabilitation programs matched. There are two kinds of likelihood ratios (LRs): positive LR (LR+) and negative LR (LR-). LRs were calculated based on sensitivity and specificity where LR+ is denoted by sensitivity/ (1-specificity) and LR- is denoted by (1-sensitivity)/ specificity. The LR was calculated using online likelihood ratio calculator (44).

Website: http://www.medcalc.org/calc/diagnostic_test.php

Table 5 Frequency of matching recommendation considering RTW status

	RTW	Failed RTW
WATT/ WATT top/ real program and clinician matched	a	b
WATT/ WATT top/ real program and clinician not matched	c	d

In this case, likelihood ratio was a more important and easy-to-interpret index. Its cut-off scores and interpretation are listed below (45).

Table 6 Likelihood ratio interpretation

LR	**Interpretation**
> 10	Largely and conclusively increase in the likelihood of RTW.
5 – 10	Moderate increase in the likelihood of RTW.
2 – 5	Small increase in the likelihood of RTW.
1 – 2	Minimal increase in the likelihood of RTW.
1	No change in the likelihood of RTW.
0.5 - 1.0	Minimal decrease in the likelihood of RTW.
0.2 - 0.5	Small decrease in the likelihood of RTW.
0.1 - 0.2	Moderate decrease in the likelihood of RTW.
< 0.1	Largely and conclusively decrease in the likelihood of RTW.

Chapter 4

Results

4.1.Data Reviewing and Claimants' Demographics

Among the total 434 claimants in the database, there were two claimants missing key variables needed for the WATT and these claimants were excluded. There were no outliers or out of range values observed within the dataset. Thus, the final number for all claimants included in further analysis was 432.

Table 7 demonstrates the characteristics of the claimants in our study database as well as a comparison between our validation dataset (including BFCE group and CFCE group) and the original WATT development dataset in order to determine if there were any differences between populations.

Table 7 Comparison of claimants' characteristics

	Validation Data (n=432)	**Validation BFCE Data (n=230)**	**Validation CFCE Data (n=202)**	**Development Data (n=7256)**
Mean (SD) or percentage				
Age (years) [a c d]	44.6 (12.6)	43.3 (13.3)	46.0 (11.6)	42.8 (11.9)
Accident to admission (days) [a c d]	517.8 (1066.0)	165.8 (632.3)	1111.1 (1379.6)	215.1 (426.1)
	Median = 187.5	Median = 67.0	Median = 594	Median = 74
Sex (% male) [c d]	68	64	74	64
Education level [a b c d]				
Grade 8 or less	5	4	6	3
Partial high school	14	12	16	11
high school diploma	23	24	22	18
Partial technical school	7	8	6	5
Technical diploma	17	13	22	13

Partial university	3	5	2	3
University degree	7	7	7	5
Not specified	24	28	20	42
Marital status [a b c d]				
Married/common law	53	49	58	39
Single	24	25	23	17
Divorced/separated	10	9	12	7
Widowed	1	1	1	1
Not specified	11	16	6	36
Job attached [a c d]	72	83	58	84
Currently working	49	50	45	46
Modified work available [a b c d]	15	22	6	54
Diagnostic categories [a b c d]				
Sprain/Strain	53	56	49	44
Joint disorder	12	12	11	29
Fracture	15	11	19	12
Contusion	10	10	10	5
Laceration	3	2	3	2
Dislocation	1	0	2	2
Nerve damage	1	1	1	2
Other	5	6	5	5
Interpreter Required (% Yes)	4	4	3	3
Assessment Recommendation [a b c d]				
No intervention required	44	14	78	6
Single service provider	25	35	13	19
Provider-based RTW program	27	45	6	52
worksite-based RTW program	0	0	0	2
Hybrid RTW program	3	4	1	9
Complex RTW program	3	2	3	4
Medical consult	0	0	0	3
Other	0	0	0	5
30-days Post Assessment (Yes %) [a b c d]	11	16	5	26

Actual Program Undertaken [a b c d]				
No rehabilitation	50	15	88	19
Single Service Provider	17	29	3	18
Provider-based RTW Program	27	45	6	50
Worksite-based RTW Program	0	0	1	2
Hybrid RTW Program	4	8	1	9
Complex RTW Program	2	3	2	4

[a] Statistically significant difference between variables in validation data and development data

[b] Statistically significant difference between variables in validation BFCE and development data

[c] Statistically significant difference between variables in validation CFCE and development data

[d] Statistically significant difference between variables in validation BFCE and CFCE data

According to Table 7, the mean age of the claimants in the validation dataset is 44.6 years with a standard deviation 12.6. The mean time from accident to admission is 517.8 days with a standard deviation 1066.0. Among all the claimants, 68% were male, 72% have jobs attached at assessment, 48% were currently working, and 15% have modified job to return to, etc. The percentage distribution of claimants' education level, marriage status and diagnosis categories were highly diverse. For example, 53% of the claimants were diagnosed as sprain/strain which ranked as the top category of types of injury while the percentage of joint disorders (12%), fracture (15%) and contusion (10%) were less frequent and more similar in percentage. But the rest of the categories account for much less proportion with few claimants diagnosed with laceration (3%), dislocation (1%), and nerve damage (1%).

The validation data were highly different from the development data demographically. Numerous continuous variables (age, time from accident to admission) were significantly different across databases, and most of the nominal variables among the four groups were significantly different in their distributions. These variables included education level, marital status, working status, availability of modified work, diagnosis category, assessment recommendation, thirty-day total disability post assessment and actual program undertaken. The CFCE group was even more different from the development dataset compared to the whole validation dataset with older mean age (46.0), longer mean time from accident to admission (1,111.1) and more likely male (74%). However, the BFCE group appeared less different since the mean age (43.3) and mean time from accident to admission (165.8) were not significantly different between those in the BFCE group and development dataset. In addition, the BFCE and CFCE groups were also highly different among almost all variables.

Table 8 demonstrates claimants' characteristics in various rehabilitation programs undertaken. The actual programs undertaken were highly imbalanced such that "no intervention required" and "provider-based RTW" programs account for most of the cases. Additionally, the other programs had fewer claimants, especially worksite-based RTW program which had only one case. Characteristics of claimants in the BFCE and CFCE groups are shown in Table 9 and Table 10. Because of the imbalance, we deleted the rehabilitation programs with less than 10 claimants under their categories from the table in Tables 9 and 10.

Table 8 Characteristics of claimants in various rehabilitation programs undertaken overall

	No intervention required n=213	**Single service provider n=73**	**Provider-based RTW program n=116**	**Worksite-based RTW program n=1**	**Hybrid RTW program n=19**	**Complex RTW program n=10**
Mean (SD) or percentage						
Age (years)	45.9 (12.2)	44.8 (12.8)	42.6 (13.1)	49	41.8 (11.0)	43.4 (12.9)
Accident to admission (days)	847.0 (1256.5)	148.1(340.5)	232.0 (925.6)	67	200.4 (207.3)	358.5 (261.9)
Sex (% male)	71.4	58.9	67.2	0	68.4	90
Education level						
Grade 8 or less	7	3	4	0	0	10
Partial high school	17	5.5	16	0	0	0
high school diploma	22	30	19	0	21	60
Partial technical school	7	10	5	100	11	10
Technical diploma	20	12	14	0	32	0
Partial university	1	6	5	0	11	0
University degree	8	7	5	0	5	0
Not specified	19	27	31	0	21	20
Marital status						
Single	24	27	23	100	21	20
Married	48	43	35	0	47	70
Common law	9	8	12	0	0	10
Divorced	8	4	7	0	5	0
Separated	4	4	3	0	11	0
Widowed	1	1	1	0	0	0
Not specified	8	12	18	0	16	0
Job attached	61	90	79	0	100	50
Currently working	44	71	40	0	74	40
Modified work available	9	14	28	0	11	0
Diagnosis categories						
Fracture	21	8	11	0	5	10
Dislocation	1	1	0	0	0	0
Sprain/St	52	47	60	100	42	60

rain						
Laceration	0	6	1	0	0	0
Contusions	0	19	5	0	11	10
Nerve damage	0	1	1	0	0	0
Joint disorders	20	14	1	0	37	20
Other	10	3	9	0	5	10
Interpreter Required (% Yes)	3	4	4	0	0	0
TTD 30-days Post Assessment (Yes %)	5	8	22	0	5	50

Table 9 Characteristics of BFCE claimants in various rehabilitation programs undertaken

	No intervention required n=35	single service provider n=67	Provider-based RTW program n=104
Mean (SD) or percentage			
Age (years)	45.3 (13.9)	45.2 (13.1)	42.2 (13.5)
Accident to admission (days)	449.2 (1529.2)	107.1 (308.3)	99.7 (121.3)
Sex (% male)	66	55	66
Education level			
Grade 8 or less	9	5	4
Partial high school	20	16	12
high school diploma	20	18	24
Partial technical school	11	6	8
Technical diploma	3	15	13
Partial university	0	6	5
University degree	11	5	7
Not specified	26	29	28
Marital status			
Single	23	28	24
Married	37	42	33
Common law	14	9	14
Divorced	6	3	5
Separated	3	5	4
Widowed	3	2	1

Not specified	14	12	20
Job attached	74	91	83
Currently working	40	72	40
Modified work available	20	15	31
Diagnosis category			
Fracture	20	8	12
Dislocation	0	2	0
Sprain/Strain	57	49	63
Laceration	0	6	1
Contusions	11	20	4
Nerve damage	0	2	1
Joint disorders	6	12	12
Other	6	3	8
Interpreter Required (% Yes)	0	5	5
30-days Post Assessment (Yes %)	9	9	23
Assessment Recommendation			
No rehabilitation	60	10	3
Single Service Provider	37	81	7
Provider-based RTW Program	3	9	90

Table 10 Characteristics of CFCE claimants in various rehabilitation programs undertaken

	No intervention required n=178	**Provider-based RTW program n=12**
Mean (SD) or percentage		
Age (years)	45.8 (11.7)	46.3 (9.6)
Accident to admission (days)	1090.1 (1255.7)	2325.3 (3571.6)
Sex (% male)	73	75
Education level		
Grade 8 or less	6	0
Partial high school	17	17
high school diploma	22	25
Partial technical school	6	0
Technical diploma	23	0
Partial university	1	0

University degree	7	8
Not specified	17	50
Marital status		
Single	24	17
Married	50	58
Common law	8	0
Divorced	8	25
Separated	4	0
Widowed	1	0
Not specified	6	0
Job attached	56	50
Currently working	44	33
Modified work available	5	0
Diagnosis category		
Fracture	24	0
Dislocation	2	0
Sprain/Strain	49	33
Laceration	4	0
Contusions	8	33
Nerve damage	1	0
Joint disorders	9	17
Other	4	17
Interpreter Required (% Yes)	4	0
30-days Post Assessment (Yes %)	3	0
Assessment Recommendation		
No rehabilitation	85	25
Provider-based RTW Program	2	58

According to Table 8, 9, 10, the mean time from accident to admission in the overall validation data and BFCE group is much longer within the “no intervention required” category than in any other program categories. However, it is the opposite situation in the CFCE group with mean time from accident to admission in “no intervention required” shorter than those in the “provider-based RTW program”.

4.2.Summary of Rehabilitation Programs Frequency Distribution

Table11to13 demonstrate the rehabilitation program frequency distribution for all the claimants in the validation dataset. Table 12 and Table 13 show program frequency distributions for claimants returning to work and experiencing failed return to work separately. Table 14 and Table 15 show the rehabilitation program frequency distribution for the claimants in BFCE and CFCE groups separately.

Table 11 Rehabilitation program frequency distribution overall

	Other	**Provider-based**	**Worksite-based**	**Hybrid**	**Complex**	**Summary**
WATT Top Recommendation	59	79	116	118	60	432
Clinicians Recommendation	297	115	0	10	10	432
Real Program	286	116	1	19	10	432

Table 12 Rehabilitation program frequency distribution for claimants returning to work

	Other	**Provider-based**	**Worksite-based**	**Hybrid**	**Complex**	**Summary**
WATT Top Recommendation	59	67	108	102	49	385
Clinicians Recommendation	274	94	0	10	7	385
Real Program	270	91	1	18	5	385

Table 13 Rehabilitation program frequency distribution for claimants failing to return to work

	Other	**Provider-based**	**Worksite-based**	**Hybrid**	**Complex**	**Summary**
WATT Recommendation	0	12	8	16	11	47
Clinicians Recommendation	23	21	0	0	3	47
Real Program	16	25	0	1	5	47

Table 14 Rehabilitation program frequency distribution for BFCE claimants

	Other	Provider-based	Worksite-based	Hybrid	Complex	Summary
WATT Top Recommendation	13	40	79	79	19	230
Clinicians Recommendation	113	103	0	9	5	230
Real Program	102	104	0	18	6	230

Table 15 Rehabilitation programs frequency distribution for CFCE claimants

	Other	Provider-based	Worksite-based	Hybrid	Complex	Summary
WATT Top Recommendation	46	39	37	39	41	202
Clinicians Recommendation	184	12	0	1	5	202
Real Program	184	12	1	1	4	202

According to Tables 11-15, the WATT provides a variety of choices when making recommendations. That is, every rehabilitation program has some claimants in their cells. However, clinicians preferred to recommend the “other” and “provider-based” programs but seldom recommended other programs including “worksite-based”, ”hybrid” and “complex”. This is most obvious when it comes to the “worksite-based” program which was not recommended by any clinician while the WATT selected it as the top recommendation 116 times. However, the actual programs claimants undertook were similar to clinician recommendations in each separate cell. In addition, the general trends remained constant despite RTW outcome (RTW and failed RTW) or assessment type (overall, BFCE, CFCE).

4.3.Percent Agreements of Recommendation

Table 16 and Table 17 show the level of agreement between the programs recommended by the WATT (including multiple WATT recommendations, the top WATT recommendation and the actual program claimants undertaken) and clinician recommendations.

Table 16 Percent agreement between WATT and clinician recommendations (overall)

	Ratio of Agreement
Clinician& WATT Programs	200/432= 0.46
Clinician& WATT Programs Top	82/432 =0.19
Clinician Program & Real Program	390/432=0.90

Table 17 Percent agreement between WATT and clinician recommendations (BFCE only)

	Ratio of Agreement
Clinician& WATT Programs	127/230=0.55
Clinician& WATT Program top	34/ 230 =0.15
Clinician Program & Real Program	203/230=0.88

According to Table 16 and Table 17, the percent agreement between clinician recommendations and multiple WATT recommendations for all the claimants is fair at 0.46 while the value in the BFCE group is slightly higher (0.55). The percent agreements between clinician recommendations and the top WATT recommendation both overall (0.19) and in the BFCE group (0.15) were poor. However, the percent agreement between clinician recommendations and the actual programs undertaken was very good (0.9). The percent agreements in the CFCE group were not calculated because this group was deemed to be very different from both the development dataset and the BFCE group, which would result in less agreement with clinician recommendations. To summarize, the result contradicts with the first hypothesis that rehabilitation recommendations from the WATT for injured workers will highly agree with clinician recommendations in general in both groups.

4.4.Comparison of WATT and Clinician Recommendations

We next investigated whether actual programs matched better with WATT recommendations for claimants who successfully RTW but whose actual rehabilitation programs did not match with clinician recommendations. Table 18 and Table 19 were built to answer this question.

Table 18 Frequency of matching program (WATT and real program comparison as column) for claimants RTW

	Clinician and real program matched	**Clinician and real program not matched**
WATT and real program matched	154	14
WATT and real program not matched	200	17

According to Table 18, the actual rehabilitation program of claimants and clinician recommendation matched for the majority of claimants (92%) among those who successfully RTW. The second column of the table reveals those not matched. To be consistent with the hypothesis in analysis part, the value *b/ (b+d) = 14/ (14+17) * 100% = 45.2%.*This indicates that only half of the claimants whose actual programs did not match with clinician recommendations could match with the WATT recommendation. The other half whose actual program did not match with clinician recommendations also did not match with the WATT.

Table 19 Frequency of matching program (WATT and clinician comparison as column) for claimants RTW

	Clinician and real program matched	**Clinician and real program not matched**
WATT and clinician matched	154	15
WATT and clinician not matched	200	16

Table 19 is a supplement for answering the second objective. The column index was changed into the comparison between WATT and clinician recommendations. The actual rehabilitation program of claimants and clinician recommendation matched for the majority of claimants (92%) among those who RTW successfully. The second column of the table also reveals clinician and real program not matched. In this case, the frequency of b is not very small, from which we could not conclude that the WATT will not recommend ineffective interventions which were unable to lead to RTW.

Table 18 and Table 19 considered the whole validation dataset. However, the BFCE group was not demonstrated because the number of claimants whose actual program and clinician recommendation not matched was very small for this analysis. Thus, the result contradicts the second hypothesis that claimants whose actual rehabilitation program did not match with the clinicians would match better with the WATT.

4.5.Likelihood Ratio

The frequencies of consistency level between actual programs claimants undertook, WATT, WATT top recommendation and clinician recommendations considering RTW status are shown in Tables 20 to 22. Likelihood ratios were calculated based on the frequency tables.

Table 20Frequency of WATT/ clinician recommendation matching level considering RTW status

	RTW	**Failed RTW**
WATT/ clinician matched	169	26
WATT/ clinician not matched	216	21

Table 21 Frequency of WATT top/ clinician recommendation matching level considering RTW status

	RTW	Failed RTW
WATT top/ clinician matched	66	8
WATT top/ clinician not matched	319	39

Table 22Frequency of Real program/ clinician recommendation matching level considering RTW status

	RTW	Failed RTW
Real/ clinician matched	354	36
Real/ clinician not matched	31	11

Figures 8-10 demonstrate the likelihood of successfully RTW. The figures were obtained based on Tables 20-22 correspondingly.

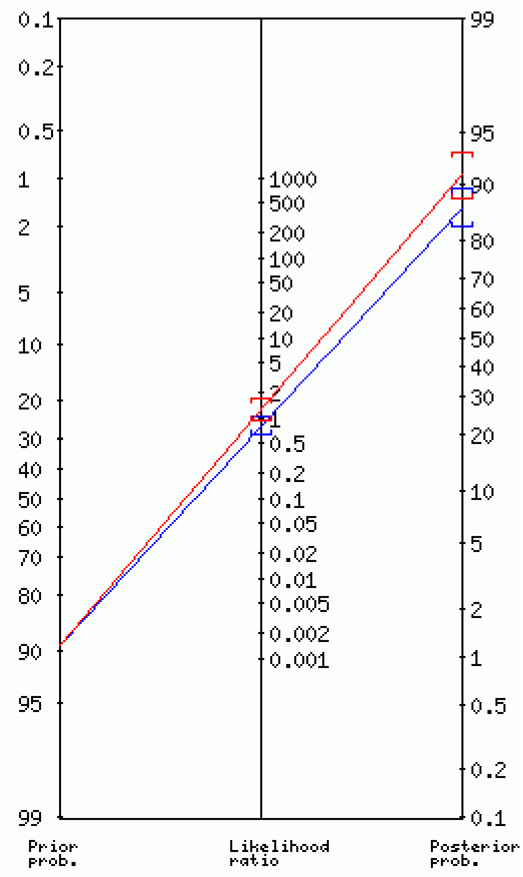


Figure 8 Likelihood of RTW when WATT/ clinician recommendation matched

As shown in Figure 8, the prior probability (prevalence) is 89%. Blue line demonstrates positive likelihood ratio while red line means negative likelihood ratio when WATT and clinician recommendation matched. LR^+ is 0.79 with 95% confidence interval [0.60, 1.05] while LR^- is 1.26 with 95% confidence interval [0.90, 1.75]. According to the likelihood ratio interpretation, the values of LR indicate a minimal increase in the likelihood of RTW when WATT and clinician recommendation are matched. However, the posterior probability in the figure which means the likelihood of RTW shows a very high probability of RTW (from 85% to 95%), which contradicts the interpretation. Thus, the LR here is inconclusive because claimants have a large chance of RTW no matter the WATT and clinician recommendation or despite whether it is matched or not due to the high prevalence of RTW in this dataset.

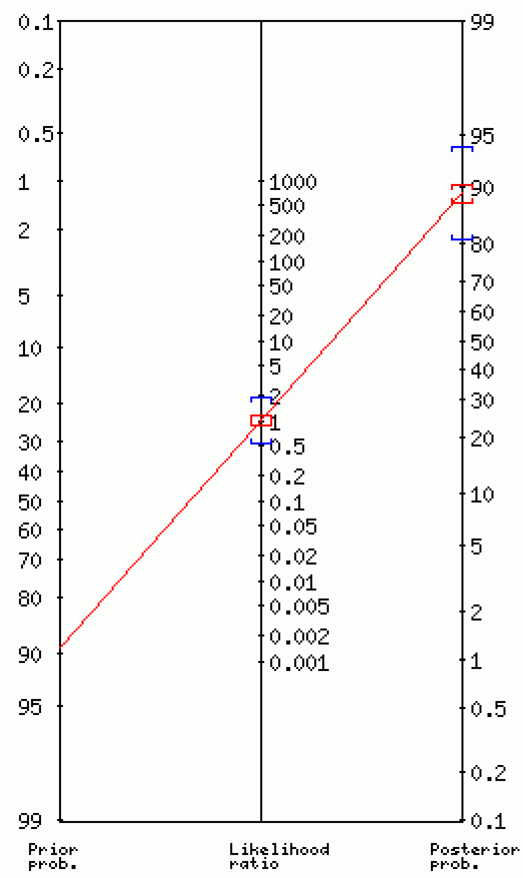


Figure 9 Likelihood of RTW when clinician/ WATT Top recommendation matched

According to Figure 9, the prior probability (prevalence) is 89%. The blue line for positive likelihood ratio overlaps the red line that is negative likelihood ratio. LR^+ is 1.01 with 95% confident interval [0.52, 1.96] while LR^- is 1.00 with 95% confident interval [0.87, 1.15]. According to likelihood ratio interpretation, these values of LR indicate no change in the likelihood of RTW when clinician and WATT top Recommendation matched. However, the posterior probability in the figure shows a very high possibility (around 90%) of RTW for both likelihood ratios, which contradicts the interpretation. Thus, the LR is also inclusive. The same as previously stated, because of the high prevalence of RTW in this dataset, a claimant will have a high probability of RTW regardless of whether the recommendation by WATT top and clinician recommendation is matched or not.

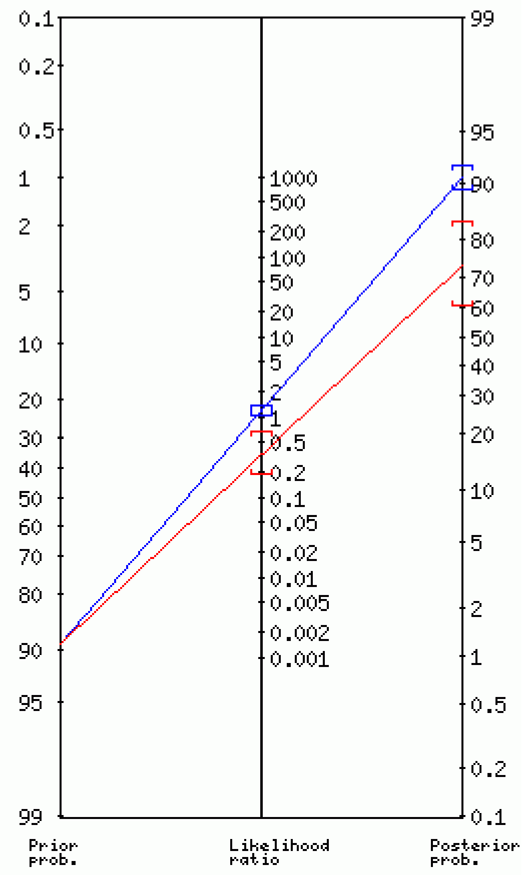


Figure 10 Likelihood of RTW when clinician/ real program matched

As shown in Figure 10, the prior probability (prevalence) is 89%. Blue line indicates positive likelihood ratio while red line means negative likelihood ratio. LR^+ is 1.20 with 95% confident interval [1.02, 1.41] while LR^- is 0.34 with 95% confident interval [0.19, 0.64]. According to likelihood ratio interpretation, these values of LR indicate a minimal likelihood of RTW when clinician and real program are matched. However, the posterior probability in the figure shows high possibility (from 70% to 95%) of RTW, which contradicts the interpretation. Thus, the LR here is not conclusive. But despite this, claimants will have a large chance of RTW regardless of the matching level of clinician and real programs because of the high prevalence of RTW.

To sum up, claimants will generally have a great chance of RTW regardless of the level of matching programs because of the high RTW prevalence in this dataset. Thus, based on the results provided by the likelihoods and the data for RTW of this dataset, subjects will RTW regardless there is good match between WATT, WATT top and clinician recommendations and real program undertaken. Thus, none of these can accurately determine RTW in this validation database.

Chapter 5

Discussion

The purpose of this study was to evaluate concurrent validity of the WATT. There were two specific objectives. The first was to explore agreement between the rehabilitation interventions recommended by the WATT and the gold standard, clinician recommendations. The second objective was to evaluate whether the WATT could improve upon the gold standard by investigating whether claimants with successful RTW whose rehabilitation programs did not match with clinician recommendations would match better with the WATT recommendations.

Findings contradict the hypotheses corresponding to the two objectives, as clinician recommendations did not highly agree with WATT recommendation and the WATT does not appear to improve upon the gold standard for RTW. Potential explanations for the lack of consistency between WATT and clinician recommendations will be discussed, including substantial differences in the development and validation datasets, differences in the characteristics of claimants, clinical recommendations being made, and actual rehabilitation programs undertaken.

5.1.Characteristics of Claimants

Comparison of claimants' characteristics

Numerous claimant demographic characteristics including age, gender, education level, marital status, and diagnostic category were examined. We divided the validation data into two groups (BFCE and CFCE) based on claimants' assessment type, and compared claimants' characteristics across these groups. Therefore, there were 3 datasets including BFCE, CFCE and the whole validation dataset. All of the groups were significantly different from the original development data on most variables. The BFCE group was the most similar since average age and injury duration in this group were not significantly different from the development dataset. The CFCE group was the most different. However, all 3 validation groups were significantly different from the development dataset in the distributions of most nominal variables (diagnostic categories, education level, assessment recommendations, etc.).

Differences between the datasets could be explained by several reasons, but the source of the data is likely to be the most important reason. Although both the validation data and development data came from the WCB-Alberta, they were from slightly different branches/ organizations. The whole development data came from WCB-Alberta Health Care Services and contains 8,611 claimants from across the province who were assessed between December 1, 2009 and January 1, 2011(9). All claimants in this dataset came from various public and private hospitals, clinics, rehabilitation centers and other medical organizations across Alberta. However, the validation data was only from WCB-Alberta Millard Health, an occupational rehabilitation and disability management service in Edmonton. Although clinicians treating injured workers in Alberta use the same soft tissue injury care model when making rehabilitation recommendations, they may likely have different preferences within different rehabilitation settings. This may explain the differences in claimant characteristics and clinician recommendations between the validation and development dataset. For example, the validation dataset had the most claimants (44%) recommended to “no further rehabilitation” while the most preferred program in the development dataset was “provider-based RTW” program (52%). In addition, claimants in different rehabilitation settings likely differ on demographic and clinical factors such as type of injury and occupational category. For example, while the validation group had few claimants with joint disorders (12%), the joint disorder category in the development dataset was much higher (29%).

Second, claimant characteristics in the BFCE and CFCE groups in the validation dataset were significantly different. Given that the two types of assessment were used for different purposes, different claimant characteristics, recommendations and programs undertaken were expected in these two groups. BFCE aims to test claimants' current level of function to determine what they can and cannot do, while CFCE is a more comprehensive assessment that helps determine the claimants' overall work capabilities. Thus, it makes sense that claimants in the CFCE group had more chronic and severe disorders and showed much longer time (1,111 days) from accident to admission than the BFCE group (166 days). CFCE claimants were also significantly older. Another important difference was in clinician recommendations. A very high percentage of CFCE claimants (78%) were recommended to "no intervention required"; however, this particular program only accounted for 14% of BFCE claimants. This can likely be explained by the fact that CFCE claimants were much more chronic and more likely to be considered as unable to succeed within rehabilitation. This indicates the BFCE group was more similar to the development data. Accordingly, we analyzed the BFCE group separately in later analyses.

Third, the year for obtaining the validation and development data sets were different, which may also contribute to the observed differences. As mentioned, the development data contained claimants who were assessed between December 1, 2009 and January 1, 2011, however, claimants in the validation dataset were assessed from November 2011 to January 2012. It is possible that clinicians' preferences when making decisions have changed over the years, thus they may have tended to make different recommendations. Moreover, the claimants themselves may have changed demographically because of the time difference. For example, only 11% of the claimants in the validation dataset claimed disability 30 days after the assessment while 26% did in the development dataset. Additionally, some contextual psychosocial and/or environmental factors such as claimants' attitude towards RTW and economic issues may also have been changed over the years leading to differences.

Characteristics of claimants in various rehabilitation programs undertaken

We also examined characteristics of claimants' in the various rehabilitation programs. In order to be consistent with the WATT development paper, we examined the "no intervention required" and "single service provider" programs separately, thus there were 6 possible rehabilitation programs (shown in Table 8). We found that 3 programs occupied a large proportion of claimants in the validation dataset (provider-based, single service and no intervention required) and there were very few claimants in the workplace-based, hybrid, or complex programs. More extremely, there was only 1claimant who undertook the "worksite-based RTW" program. Because of this, information about the latter 3 programs was limited and relatively meaningless when analyzing the data. Thus, we omitted the programs with extremely few claimants in some analyses on the validation dataset.

When examining the proportion of claimants in the various rehabilitation programs in the development paper (9), the trend of proportion was very similar to the validation dataset: claimants in "provider-based" program, "no intervention required" and "single services" were ranked as the top 3most common programs occupying 86% of all claimants. Similarly, claimants in “worksite based” program were the fewest. Accordingly, we conclude that clinicians prefer to recommend “no further rehabilitation”, “single services” and/or “provider-based” programs, and the preference has little to do with the rehabilitation setting, year or other factors. Additionally, it appears that Millard Health clinicians are particularly averse to recommending workplace-based interventions.

To summarize, claimants’ characteristics were significantly different between the validation and development datasets. We also observed an imbalance among the rehabilitation programs undertaken. Possible reasons were discussed, and these differences and imbalances have laid a foundation for further discussion of the consistency between WATT and clinician recommendations.

5.2.Discussion on the three objectives

Percent agreement

We calculated percent agreements between the WATT and clinician recommendation using the whole validation data and in the BFCE claimants only. As mentioned, we did not calculate percent agreements for the CFCE group because it was substantially different from the development dataset. We compared clinician recommendations against all multiple WATT recommendations, the top WATT recommendation, and the actual programs undertaken by claimants.

Observed percent agreements between clinician recommendations and multiple WATT recommendations were moderate at 0.46, while the value for BFCE claimants was slightly higher at 0.55. The ratio of agreement between clinician recommendations and the top WATT recommendation for both all claimants (0.19) and BFCE claimants (0.15) was low. However, the ratio of agreement observed between clinician recommendations and actual programs undertaken was high (0.90 and 0.88).

This result of low to moderate agreement between the WATT and clinician recommendations was unexpected and contrary to hypotheses. The WATT was built based on data from WCB-Alberta claimants and internal validation demonstrated good performance of the WATT overall (9). However, it is not unusual for a model to perform much more poorly in an external validation study than in the original development study because of machine learning theory. While the computing techniques and statistical methods used to build the WATT were selected to be appropriate and robust, one step which could have been problematic is the over-sampling technique (SMOTE). SMOTE was used to mitigate class imbalance and might be responsible for classification difficulties. More specifically, the WATT developers used SMOTE to over-sample or synthetically enhance the minority classes/ programs ("worksite-based" program, "hybrid" program and "complex" program). Essentially, the imbalance of rehabilitation programs may have been a key limitation of the WATT making it less valid and resulting in low agreement with clinician recommendations in the validation dataset.

We observed that the actual programs undertaken by claimants highly concurred with clinician recommendations. This is likely because clinicians making the recommendation and case managers evaluating and acting on the recommendations came from the same organization, which may lead to very similar preferences and considerations.

The comparison between WATT and clinician recommendations

For those claimants who successfully RTW but whose actual rehabilitation programs did not match with clinician recommendations, we investigated whether the programs matched more highly with WATT recommendations. The result indicated that only half of the claimants whose actual program did not match with clinician recommendations matched with the WATT recommendation. The other half whose actual program did not match with clinician recommendations also did not match with the WATT. This conflicts our hypothesis that WATT recommendations would have a better match with the successful actual programs recommended by case managers in this sub-group. Additionally, a certain number of claimants had a match between WATT and clinician recommendations. Thus, it is inconclusive as to whether the WATT will recommend ineffective interventions that could probably lead to a failure of RTW. Overall, the result contradicts the second hypothesis that claimants whose actual rehabilitation program did not match with clinician recommendations would match better with the WATT. It is also inconclusive which recommendation (clinician or the WATT) is better.

These results might be biased because of the limited number (8%) of claimants whose actual programs were different from the clinician and who also successfully RTW. It has been shown that actual programs recommended by case managers were highly consistent with clinician recommendations mainly because they come from the same rehabilitation setting.

Likelihood of Predicting RTW

Likelihood ratios were calculated based on the frequencies of consistency level between actual programs claimants undertook, WATT recommendations, WATT top recommendations and clinician recommendations while considering RTW status. As shown in Figure 8-10, the prior probability (prevalence) is as high as 89%. But all three pairs of LR^+ / LR^- showed minimal or even no chance of RTW which conflicted with the high posterior probability in the figures. Thus, the LRs were not conclusive in this study. The reason is that because of the high prevalence, a claimant will have a huge probability of RTW regardless of the matching program status.

Though inconclusive, the prevalence is consistent with the literature that most injured workers recover and RTW while a minority remains off work for longer periods of time and are responsible for the majority of associated health care and compensation costs. In this validation dataset, the percentage of claimants RTW was extremely high. Thus, claimants will return to work regardless programs claimants undertook, WATT recommendations, WATT top recommendations, and clinician recommendations matched or not.

5.3.Clinical Evidence in Support of WATT Recommendation

Although overall results demonstrate a low to moderate agreement between WATT recommendations and the gold standard and the WATT did not seem to outperform the clinicians for recommending effective rehabilitation program leading to RTW, no conclusion could be drawn that the WATT is valid or not. Some clinical evidence was found to support WATT recommendations.

According to the available scientific literature, the “provider-based”, “worksite-based” and “complex” programs have supportive evidence of effectiveness for successful RTW in injured workers (46-49). However, clinicians rarely recommend some of the available programs. For example; there were only 10 claimants in each of the “hybrid” program and the “complex” program. More extremely, no “workplace-based” program was recommended among the 434 claimants. It might be reasonable not to recommend “complex” or “hybrid” program since they are more costly than the other programs; however, workplace-based interventions appear to be a cost-effective option (49). In addition, “hybrid”, “worksite-based” and “complex” were also rarely recommended compared to the other programs in the development database. Clinicians and case managers frequently recommend “provider-based” and single service programs partly because of the economic benefits. Though their effectiveness is supported by evidence, “hybrid”, “worksite-based” and “complex” programs were likely do not provide as much benefit to the organization. Thus, those programs were infrequently recommended.

There are several studies supporting “workplace-based” programs. In 2005 Franche et al. (50) published a systematic review examining the effectiveness of workplace-based return-to-work interventions. The authors found strong to moderate evidence to support workplace-based RTW interventions for reducing work disability duration and associated costs. In addition, Franche and other colleagues examined the relationship between early return-to-work strategies and work absence duration using administrative and self-reported data in 632 claimants with work-related musculoskeletal disorders. Their results indicated that a work accommodation offer and acceptance along with advice from health care providers to the workplace regarding re-injury prevention strategies were significant predictors of shorter work absence duration (51). Accordingly, it is surprising that no clinician recommended “workplace-based” programs to claimants with so much supporting evidence of its effectiveness. In addition, literature suggests that clinicians are currently unable to identify which claimants will respond best to various treatment options. Taken everything into consideration, it might be doubted whether clinician recommendations are truly gold standard. This may be an important reason why clinician and WATT recommendations do not highly agree with each other.

5.4.Study Strengths

One strength of this study is that we tested the concurrent validity of the WATT in a dataset that was created fairly close in time to the development of the external dataset. That is to say, the data used for WATT development contains claimants who were assessed between December 1, 2009 and January 1, 2011 and the claimants in the validation dataset were assessed from November 2011 to January 2012. This makes it unlikely that differences in clinician recommendations occurred due to historical changes in policy or practice. This timely validation could help us to discover potential errors of the WATT and make necessary modifications as soon as possible.

The second strength of this study is that we used a computing program to input data from WCB claimants directly into the WATT to overcome potential mistakes from manual entry. This also saved substantial time. Moreover, we wrote codes prior to statistical analysis in order to make the analytic procedure less complicated. For example, we did not need to compare the duration, confidence, or number of rules for multiple recommendations for each claimant manually when selecting the top WATT recommendation. We also did not need to count the frequency of claimants in each cell when trying to find the matching level of recommendations. The application of these computing programs enhanced our statistical methods and appears helpful when dealing with large amounts of data.

5.5.Study Limitations

The first limitation of this study is the source of the data. As mentioned, the development and validation data were collected in different rehabilitation settings affiliated to WCB-Alberta. All the claimants involved in the development dataset came from various public and private hospitals, clinics, and rehabilitation centers (including Millard Health), however, the validation data was only from Millard Health. Although clinicians in different hospitals and clinics may use the same soft tissue injury care model when making rehabilitation recommendations, they probably have diverse preferences within different units. In addition, the validation dataset contained a higher proportion of claimants receiving CFCE since it came from a study of these types of assessments. This required us to conduct separate analyses with and without the CFCE claimants.

This study was a secondary data analysis, thus the researchers have no ability to change or manipulate the existing database. The nature of this secondary analysis may have been a limitation, although all of the information required to form WATT recommendations was available. The number of claimants in the validation dataset also might not have been enough. Although there are 434 claimants in the dataset, it is relatively small compared to the development dataset that contained over 8,000 claimants. The number of claimants who failed to RTW and the number of claimants whose actual program did not match with clinician recommendations were especially limited, which may have introduced a bias when analyzing.

5.6.Clinical Significance

Although no firm conclusions were drawn regarding the concurrent validity of the WATT, the study is still clinically relevant from several aspects. Firstly, the summary of rehabilitation program frequency distributions casted doubts on the gold standard in this study. Clinician recommendations were used as the gold standard; however, clinicians at the WCB-Alberta facility rarely recommended some rehabilitation programs ("workplace-based", "complex" and "hybrid") that have considerable supportive evidence. Accordingly, this study provides some suggestions and evidence for future clinician recommendations. That is, it may be worthwhile considering "workplace-based", "hybrid" and/ or "complex" program to injured workers based on both the WATT recommendation and evidence guidelines. Lastly, we mentioned the WATT may be somewhat problematic due to its original data processing techniques. This study discovered a possible flaw of the WATT that developers and researchers could review and try modify.

To summarize, results indicate WATT recommendations for injured workers should be viewed with caution, especially the positive recommendations which can be quite different from clinician recommendations until more refinement of the WATT is performed.

5.7.Suggestions for Future Research

There are several possibilities for further research. Firstly, more claimants are needed to further test WATT's validity. Ideally the WATT would be tested using data more in concordance with the development dataset. That is to say, claimants should come from across Alberta and not be limited to Millard Health; and all of them should have received only BFCE. Secondly, a larger dataset may be needed to solve the insufficient data of claimants with failed RTW and whose actual program did not match with clinician recommendations. The analysis would be more trustworthy with larger sample size, and some additional statistical models such as logistic regression predicting RTW could be implemented. A randomized clinical trial could also be done to test the effectiveness of the WATT since the most appropriate recommendation is still undefined. The effectiveness of two groups (injured workers follow WATT recommendations and follow clinician recommendations) could be measured through the clinical trial. It would be more powerful than a secondary data analysis and more effectiveness testing is necessary prior to further implementation. Lastly, better computing techniques or additional data to solve the rehabilitation class imbalance maybe needed.

5.8.Summary and Conclusion

Clinical decision support tools have been researched for their potential to facilitate return-to-work and reduce work absence duration for injured workers. An increasing number of clinical decision support tools have been developed recently but most have not been formally tested or rigorously evaluated. Prior to clinical usage, validity testing is necessary and important. This study tested the concurrent validity of the Work Assessment Triage Tool (WATT), a newly developed clinical decision support tool that could potentially help clinicians identify the most appropriate rehabilitation program for claimants off work due to a variety of musculoskeletal disorders. We conducted a secondary analysis using data from a clinical trial conducted previously at the Workers' Compensation Board of Alberta rehabilitation facility. A variety of statistical methods were used to compare recommendations from the WATT, clinician recommendations, actual programs claimants undertaken and return-to-work outcomes.

Results indicate that percent agreements between WATT and clinician recommendations were low (r = 0.19) to moderate (r = 0.46). Additionally, the WATT did not appear to improve upon clinician recommendations for selecting programs that led to RTW. Three possible reasons could explain four results: (1) important differences were observed in claimants characteristics between the original WATT development data and our validation dataset; (2) the insufficient data of claimants who failed RTW and with successful RTW whose actual rehabilitation program did not match with the clinician recommendations (3) data processing techniques were used to overcome rehabilitation class imbalance when building the WATT, which may be a cause of error in WATT recommendations; (4) clinician recommendations somewhat conflicted with existing evidence as clinicians in our validation dataset rarely recommended some rehabilitation programs which were highly supported by research evidence (i.e. workplace interventions). Thus, no firm conclusions could be drawn regards the concurrent validity, and we cannot determine which method (WATT or clinician judgment) would provide better recommendations for return-to-work in actual practice. However, results indicate WATT recommendations for injured workers should be viewed with caution since they can be quite different from clinician recommendations. Further research is needed to resolve this uncertainty.

References

1.Coyte PC, Asche CV, Croxford R, Chan B. The economic cost of musculoskeletal disorders in Canada. Arthritis care and research : the official journal of the Arthritis Health Professions Association. 1998;11(5):315-25. Epub 1998/11/27.

2.Martin BI, Deyo RA, Mirza SK, Turner JA, Comstock BA, Hollingworth W, et al. Expenditures and health status among adults with back and neck problems. JAMA : the journal of the American Medical Association. 2008;299(6):656-64. Epub 2008/02/14.

3.Frank JW, Kerr MS, Brooker AS, DeMaio SE, Maetzel A, Shannon HS, 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(24):2908-17. Epub 1996/12/15.

4.Haldorsen EM. The right treatment to the right patient at the right time. Occupational and environmental medicine. 2003;60(4):235-6. Epub 2003/03/28.

5.Shortliffe EH, Davis R, Axline SG, Buchanan BG, Green CC, Cohen SN. Computer-based consultations in clinical therapeutics: explanation and rule acquisition capabilities of the MYCIN system. Computers and biomedical research, an international journal. 1975;8(4):303-20. Epub 1975/08/01.

6.Miller RA, Masarie FE, Jr. Use of the Quick Medical Reference (QMR) program as a tool for medical education. Methods of information in medicine. 1989;28(4):340-5. Epub 1989/11/01.

7.Tu SW, Eriksson H, Gennari JH, Shahar Y, Musen MA. Ontology-based configuration of problem-solving methods and generation of knowledge-acquisition tools: application of PROTEGE-II to protocol-based decision support. Artificial intelligence in medicine. 1995;7(3):257-89. Epub 1995/06/01.

8.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. 2010;20(3):394-405. Epub 2009/11/11.

9.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. Epub 2013/03/08.

10.Punnett L, Wegman DH. Work-related musculoskeletal disorders: the epidemiologic evidence and the debate. Journal of Electromyography and Kinesiology. 2004;14(1):13-23.

11.Hunt DL, Haynes RB, Hanna SE, Smith K. Effects of computer-based clinical decision support systems on physician performance and patient outcomes: a systematic review. JAMA : the journal of the American Medical Association. 1998;280(15):1339-46. Epub 1998/10/30.

12.Schultz IZ, Stowell AW, Feuerstein M, Gatchel RJ. Models of return to work for musculoskeletal disorders. J Occup Rehabil. 2007;17(2):327-52. Epub 2007/02/09.

13.Jerosch-Herold C. An evidence-based approach to choosing outcome measures: a checklist for the critical appraisal of validity, reliability and responsiveness studies. The British Journal of Occupational Therapy. 2005;68(8):347-53.

14.Calder BJ, Phillips LW, Tybout AM. The concept of external validity. Journal of Consumer Research. 1982:240-4.

15.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(3):367-77. Epub 2010/02/09.

16.Badley E, Rasooly I, Webster G. Relative importance of musculoskeletal disorders as a cause of chronic health problems, disability, and health care utilization: findings from the 1990 Ontario Health Survey. The Journal of rheumatology. 1994;21(3):505.

17.Feeney A, North F, Head J, Canner R, Marmot M. Socioeconomic and sex differentials in reason for sickness absence from the Whitehall II Study. Occupational and environmental medicine. 1998;55(2):91-8.

18.Alexanderson KA, Borg KE, Hensing GK. Sickness absence with low-back, shoulder, or neck diagnoses: an 11-year follow-up regarding gender differences in sickness absence and disability pension. Work: A Journal of Prevention, Assessment and Rehabilitation. 2005;25(2):115-24.

19.Riihimäki H. Hands up or back to work—Mure challenges in epidemiologie research on musculoskeletal diseases. Scandinavian journal of work, environment & health. 1995:401-3.

20.Verhaak PF, Kerssens JJ, Dekker J, Sorbi MJ, Bensing JM. Prevalence of chronic benign pain disorder among adults: a review of the literature. Pain. 1998;77(3):231-9.

21.Position of the American Dietetic Association: nutrition, aging, and the continuum of care. Journal of the American Dietetic Association. 2000;100(5):580-95. Epub 2000/05/17.

22.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(25):2898-904. Epub 2008/02/05.

23.Mayer T, Gatchel RJ, Evans T. Effect of age on outcomes of tertiary rehabilitation for chronic disabling spinal disorders. Spine. 2001;26(12):1378-84. Epub 2001/06/27.

24.Hunter SJ, Shaha S, Flint D, Tracy DM. Predicting return to work. A long-term follow-up study of railroad workers after low back injuries. Spine. 1998;23(21):2319-28. Epub 1998/11/20.

25.Hildebrandt J, Pfingsten M, Saur P, Jansen J. Prediction of success from a multidisciplinary treatment program for chronic low back pain. Spine. 1997;22(9):990-1001. Epub 1997/05/01.

26.Sandstrom J, Esbjornsson E. Return to work after rehabilitation. The significance of the patient's own prediction. Scandinavian journal of rehabilitation medicine. 1986;18(1):29-33. Epub 1986/01/01.

27.van Hooff ML, Spruit M, O'Dowd JK, van Lankveld W, Fairbank JC, van Limbeek J. Predictive factors for successful clinical outcome 1 year after an intensive combined physical and psychological programme for chronic low back pain. European Spine Journal. 2013:1-11.

28.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. Epub 2006/11/07.

29.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. 2009;34(12):1243-9. Epub 2009/05/05.

30.Myers J, Pople H, Miller R, editors. CADUCEUS: a computerized diagnostic consultation system in internal medicine. Proceedings of the Annual Symposium on Computer Application in Medical Care; 1982: American Medical Informatics Association.

31.Forseen SE, Corey AS. Clinical decision support and acute low back pain: Evidence-based order sets. JACR Journal of the American College of Radiology. 2012;9(10):704-12.

32.Lin L, Hu PJH, Sheng ORL. A decision support system for lower back pain diagnosis: Uncertainty management and clinical evaluations. Decis Support Syst. 2006;42(2):1152-69.

33.Patel S, Brown S, Friede T, Griffiths F, Lord J, Ngunjiri A, et al. Study protocol: Improving patient choice in treating low back pain (IMPACT - LBP): A randomised controlled trial of a decision support package for use in physical therapy. BMC Musculoskeletal Disorders. 2011;12.

34.Eberhardt J, Bilchik A, Stojadinovic A. Clinical decision support systems: Potential with pitfalls. Journal of Surgical Oncology. 2012;105(5):502-10.

35.de Ipina KL, Hernandez MC, Grana M, Martinez E, Vaquero C. A Computer-Aided Decision Support System for Shoulder Pain Pathology. In: Demazeau Y, Dignum F, Corchado JM, Bajo J, Corchuelo R, Corchado E, et al., editors. Trends in Practical Applications of Agents and Multiagent Systems2010. p. 705-12.

36.Sailors RM, East TD, Wallace CJ, Carlson DA, Franklin MA, Heermann LK, et al. Testing and validation of computerized decision support systems. Proceedings : a conference of the American Medical Informatics Association / AMIA Annual Fall Symposium AMIA Fall Symposium. 1996:234-8. Epub 1996/01/01.

37.Spiegelman D, Schneeweiss S, McDermott A. Measurement error correction for logistic regression models with an "alloyed gold standard". American journal of epidemiology. 1997;145(2):184-96. Epub 1997/01/15.

38.Kotsiantis SB, Zaharakis I, Pintelas P. Supervised machine learning: A review of classification techniques. 2007.

39.Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: synthetic minority over-sampling technique. arXiv preprint arXiv:11061813. 2011.

40.Batista GE, Prati RC, Monard MC. A study of the behavior of several methods for balancing machine learning training data. ACM SIGKDD Explorations Newsletter. 2004;6(1):20-9.

41.Freund Y, Schapire R, Abe N. A short introduction to boosting. Journal-Japanese Society For Artificial Intelligence. 1999;14(771-780):1612.

42.Gross DP, Asante AK, Miciak M, Battié MC, Carroll LJ, Sun A, et al. A Cluster Randomized Clinical Trial Comparing Functional Capacity Evaluation and Functional Interviewing as Components of Occupational Rehabilitation Programs. Journal of Occupational Rehabilitation. 2013:1-14.

43.Series SC. Measurement of observer agreement. Radiology. 2003;228:303-8.

44.Parshall MB. Unpacking the 2× 2 table. Heart & Lung: The Journal of Acute and Critical Care. 2013;42(3):221-6.

45.Reiman MP, Goode AP, Hegedus EJ, Cook CE, Wright AA. Diagnostic accuracy of clinical tests of the hip: a systematic review with meta-analysis. British journal of sports medicine. 2013;47(14):893-902.
46.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(2):CD006955. Epub 2009/04/17.
47.Karjalainen K, Malmivaara A, van Tulder M, Roine R, Jauhiainen M, Hurri H, et al. Multidisciplinary biopsychosocial rehabilitation for subacute low back pain among working age adults. Cochrane Database Syst Rev. 2003(2):CD002193. Epub 2003/06/14.
48.Schaafsma F, Schonstein E, Ojajarvi A, Verbeek J. Physical conditioning programs for improving work outcomes among workers with back pain. Scandinavian journal of work, environment & health. 2011;37(1):1-5. Epub 2010/08/12.
49.Arnetz BB, Sjögren B, Rydéhn B, Meisel R. Early workplace intervention for employees with musculoskeletal-related absenteeism: a prospective controlled intervention study. Journal of Occupational and Environmental Medicine. 2003;45(5):499-506.
50.Franche R-L, Cullen K, Clarke J, Irvin E, Sinclair S, Frank J. Workplace-based return-to-work interventions: a systematic review of the quantitative literature. Journal of Occupational Rehabilitation. 2005;15(4):607-31.
51.Franche R-L, Severin CN, Hogg-Johnson S, Côté 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. Journal of Occupational and Environmental Medicine. 2007;49(9):960-74.

Appendix

Appendix 1: The WATT features and corresponding answers

(1) Job attached status at time of assessment (job to return to or not).

(2) National Occupational Category code (See Appendix4)

(3) Currently working status (working or not working)

(4) Modified work available (yes or no)

(5) SF-36 Item 2 Healthy Now? (See Appendix 2)

(6) SF-36 Item 4 Limited in moderate activities? (See Appendix 2)

(7) SF-36 Item 5 Lifting or carrying groceries? (See Appendix 2)

(8) SF-36 Item 7 Climbing stairs? (See Appendix 2)

(9) SF-36 Item 12 Limited in bating or dressing yourself? (See Appendix 2)

(10) SF-36 Item 14 Accomplished less at work? (See Appendix 2)

(11) SF-36 Item 18 Accomplished less because of emotional problems? (See Appendix 2)

(12) SF-36 Item 21 Bodily pain during the past 4 weeks? (See Appendix 2)

(13) SF-36 Item 25 Nothing could cheer you up? (See Appendix 2)

(14) Self-rated level of occupational disability at assessment (on a 10-point scale)

(15) Self-rated pain intensity at assessment using a 10-point Pain VAS (See Appendix 3)

(16) Diagnosis Group (from ICD9 category) (See Appendix 4)

(17) Calendar Days injury to assessment

Appendix 2: SF-36 Health Survey

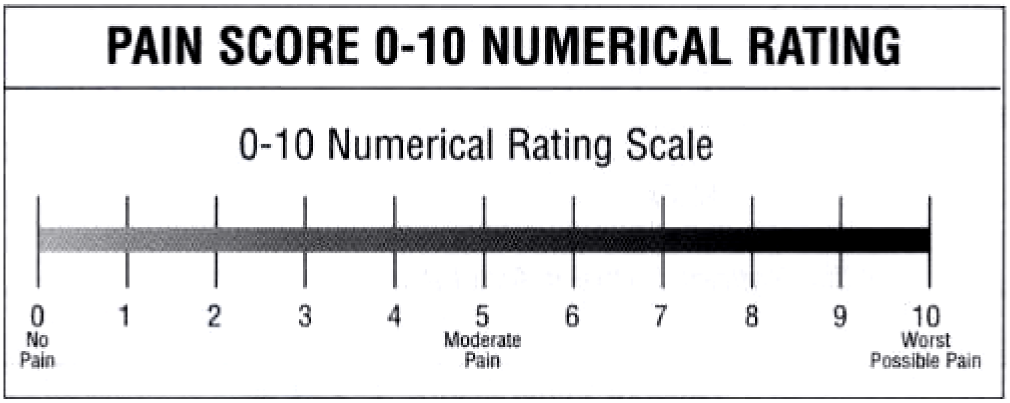
Please refer to the website:

http://www.rand.org/health/surveys_tools/mos/mos_core_36item_survey.html

Appendix 3: Pain Visual Analogue Scale

VISUAL ANALOGUE SCALE

On a scale of 0-10 (where 0 is no pain and 10 is unbearable pain, the worst pain you can imagine), mark where your pain is most of the time.



Appendix 4: Link of National Occupational Category code and Diagnosis Group code

Please refer to the websites.

National Occupational Category code:

www5.hrsdc.gc.ca/NOC/English/NOC/2011/OccupationIndex.aspx

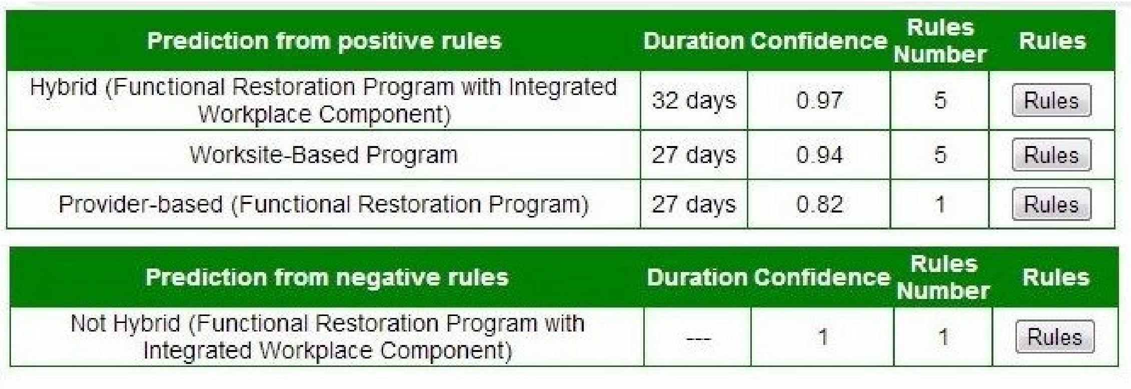
Diagnosis Group code:

http://icd9cm.chrisendres.com/

Appendix 5 Case Scenarios of WATT Recommendations

Case # 1

Mr. A, was a carpenter and was still working at assessment. He also has part-time modified work to return to. The database showed his assessment results for the SF-36. The results are: item 2 (somewhat worse now than a year ago), item 4 (some of the time), item 5 (most of the time), item 7 (little of the time), item 12 (little of the time), item 14 (some of the time), item 18 (none of the time), item 21 (moderate), item 25 (little of the time); occupation 5 out 10; Pain VAS 4 out of 10. He was diagnosed with fracture and there were 97 days from injury to assessment. Rehabilitation interventions recommended by the WATT are shown as below.



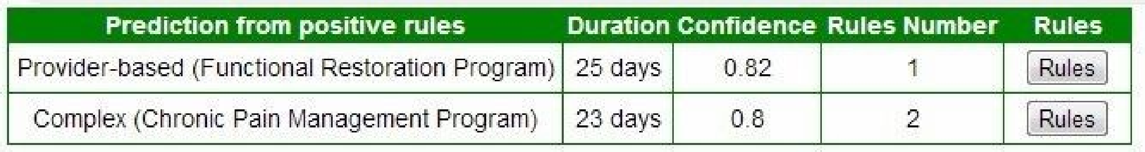
Prediction from positive rules	Duration	Confidence	Rules Number	Rules
Hybrid (Functional Restoration Program with Integrated Workplace Component)	32 days	0.97	5	Rules
Worksite-Based Program	27 days	0.94	5	Rules
Provider-based (Functional Restoration Program)	27 days	0.82	1	Rules

Prediction from negative rules	Duration	Confidence	Rules Number	Rules
Not Hybrid (Functional Restoration Program with Integrated Workplace Component)	---	1	1	Rules

As shown the figure, the WATT provided three positive rehabilitation recommendations and one negative recommendation. Among the positive recommendations, Worksite-Based Program should be selected as the top recommendation based on its shorter duration, higher confidence and more supporting rules. Alternatively, clinicians could consider the three positive recommendations based on their own expertise and not consider the negative one when making actual recommendations.

Case # 2

Mr. B, used to work as an oilman before his arm fracture. He is not job attached and does not have modified work to return to. The database showed his assessment results for the SF-36. The results are: item 2 (much worse now than a year ago), item 4 (some of the time), item 5 (most of the time), item 7 (most of the time), item 12 (some of the time), item 14 (all the time), item 18 (some of the time), item 21 (severe), item 25 (some of the time); PDI occupation 5 out 10; Pain VAS 8 out of 10. And there were 97 days from injury to assessment. Rehabilitation interventions recommended by the WATT are shown as below.

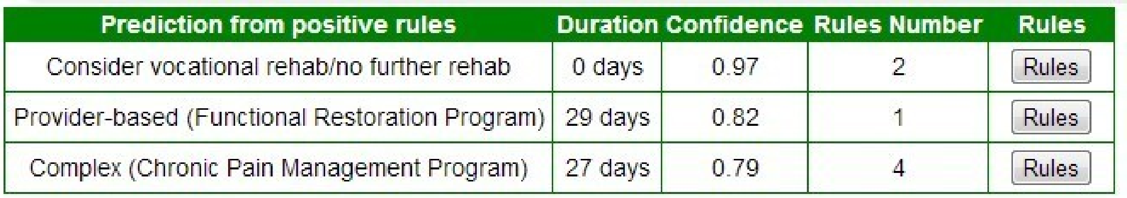


Prediction from positive rules	Duration	Confidence	Rules Number	Rules
Provider-based (Functional Restoration Program)	25 days	0.82	1	Rules
Complex (Chronic Pain Management Program)	23 days	0.8	2	Rules

As shown in the figure the WATT only provided two positive recommendations for Mr. B. Complex program should be selected as the top recommendation for its shorter duration and more supporting rules, even though confidence is slightly higher for the provider-based program.

Case # 3

Ms. C was a sale associate and has not work for years. The database showed her assessment results for the SF-36 are: item 2 (much worse now than a year ago), item 4 (all the time), item 5 (all the time), item 7 (most of the time), item 12 (all the time), item 14 (all the time), item 18 (most of the time), item 21 (severe), item 25 (some of the time). She also scored the following: PDI occupation 8 out 10; Pain VAS 9 out of 10. She was diagnosed with dislocation of the upper arm. And there were 4,564 days since she got injured. Rehabilitation interventions recommended by the WATT are shown as below.



Prediction from positive rules	Duration	Confidence	Rules Number	Rules
Consider vocational rehab/no further rehab	0 days	0.97	2	Rules
Provider-based (Functional Restoration Program)	29 days	0.82	1	Rules
Complex (Chronic Pain Management Program)	27 days	0.79	4	Rules

Similar as Mr. B, the WATT only provided three positive recommendations for Ms. C. Vocational rehab/ no further rehab could be considered as the top recommendation with shortest duration and highest confidence, indicating Ms. C was too chronic to be cured. Clinicians would consider all the three recommendations to help them make the final decision.