MINIMUM CLINICAL IMPORTANT DIFFERENCES OF HEALTH OUTCOMES IN A CHRONIC PAIN POPULATION: ARE THEY PREDICTIVE OF POOR OUTCOMES?

by

HILARY D. WILSON

Presented to the Faculty of the Graduate School of

The University of Texas at Arlington in Partial Fulfillment

of the Requirements

for the Degree of

DOCTOR OF PHILOSOPHY

THE UNIVERSITY OF TEXAS AT ARLINGTON

December 2007

To Deven

You are my “one big idea”

ACKNOWLEDGEMENTS

This work is the result of a community of efforts, and I would like to acknowledge all the faculty, friends, and family that have played a pivotal role in culmination of this project. First and foremost, I would like to thank all the members of my committee for agreeing to participate in the project. You have all provided a unique perspective, and contributed to the strength of the final work.

I would like to thank my husband for his unwavering support over the past four years. A special thanks for all the evenings and weekends spent staining histology, studying in the library, and enduring endless conversations about what matters to you the least. I do my best work with you at my side.

Thanks mom and dad for providing me with a love for learning, a passion for science, and a respect for the unknown. Mom thanks for the encouragement, gentle nudging, and especially, for the friendship you have provided along the way. Dad, thanks for helping me learn and appreciate the language of mathematics, and establish a confidence in what I know. You have been the inspiration for me to push through the surface, and get at the real meat of the matter. Thanks to my brothers and sisters who have assisted me in numerous ways, the least of which by providing me with emotional support and friendship.

Kathlynn, you have been a wonderful friend, auntie, nanny, grandma, and colleague all rapped up into one. Thanks for being there in every aspect of my life over the past several years. Brian, it has truly been a privilege to have you as a peer in graduate school. I have the utmost respect for your work, and have enjoyed both learning and kicking back with you. Jessica thanks for being there to share the joys and tribulations of being both a graduate student and a mother. Your dedication to both your family and work are an inspiration.

Dr. Fuchs thanks for passing along your passion for your work, and helping me to learn the guts of a complex, multifaceted, academic world. I am most appreciative of your dedication to helping me realize my own interests and passions. Dr. Gatchel and Dr. Mayer, thank you for offering me the opportunity to work on a project of such magnitude. I have enjoyed working with you both and hope to cross paths with you in the future.

September 11, 2007

ABSTRACT

MINIMUM CLINICAL IMPORTANT DIFFERENCES OF HEALTH OUTCOMES IN A CHRONIC PAIN POPULATION: ARE THEY PREDICTIVE OF POOR OUTCOMES?

Publication No. __

Hilary D. Wilson, PhD.

The University of Texas at Arlington, 2007

Supervising Professor: Dr. Bob Gatchel

Psychometric validation of health outcomes measures ensures that the methods utilized to evaluate treatment effects, and aid in individual patient diagnosis are reliable, valid, and meaningful. A relatively new concept within the psychometric process of validation is the assessment of responsiveness, or the ability of an instrument to detect clinically meaningful change. Clinically meaningful change may be defined through subjective, self-reports of change, physician-based assessment, or through objective outcome criteria. The purpose of the current study was to evaluate clinically meaningful changes in chronic pain health outcome measures, as defined by objective outcome

criteria. The average percent change in the Oswestry Disability Index, Million Visual Analog Scale, Short-form 36, Pain Disability Questionnaire, Pain Intensity Scale, and Beck Depression Inventory, were calculated for patients categorized as having Poor, Fair, and Good 1-year socioeconomic outcomes. The predictive ability of the percent change scores were evaluated through logistic regression analysis. Percent difference in BDI and MVAS were predictive of outcome status when combined with pre-treatment scores and age. No other percent difference variables were predictive of outcomes at the individual patient level, negating the application of an MCID for use in a clinical setting for the ODI, PDQ, PI, and BDI. However, a variety of additional pre and post measures were predictive of outcome status. The PDQ was able to predict poor outcomes better than any other scale, and the Pain Intensity, MVAS, and BDI were superior at detecting good outcomes. By combining scales such as MVAS and PI (better sensitivity), which are better at classifying good outcomes, and scales such as the PDQ and ODI, which are better at discriminating among patients that have poor outcomes (better specificity), superior identification of patients at greatest risk for poor outcomes may be realized. The prevalence of pain, critical role of health outcome measures in the field of medicine, and evaluation of MCIDs as a critical aspect in the validation process of measurement scales, highlights the importance of the current project.

TABLE OF CONTENTS

LIST OF ABBREVIATIONS

ANOVA	Analysis of Variance
APA	American Psychological Association
BDI	Beck Depression Inventory
CMSD	Chronic Musculoskeletal Disorder
CMP	Chronic Musculoskeletal Pain
COMPL	Completers
DSM-IV	Diagnostic and Statistical Manual of Mental Disorders – 4th Ed.
GFS	Global Functioning Scale
HRQL	Health Related Quality of Life
LBP	Low Back Pain
LOD	Length of Disability
MCID	Minimum Clinical Important Difference
MHS	Mental Health Component Score (SF-36)
MOS	Medical Outcome Study
MVAS	Million Visual Analog Scale
NA	Not Applicable
ODI	Oswestry Disability Index

PDQ	Pain Disability Questionnaire
PHS	Physical Health Component Score (SF-36)
PI	Pain Intensity
PRIDE	Productive Rehabilitation Institute of Dallas for Ergonomics
SF-36	Short-Form 36
TRT	Treatment
VAS	Visual Analog Scale
ZDS	Zung Depression Scale

CHAPTER 1

INTRODUCTION

Over the past 50 years, the field of Medicine has experienced dynamic change, from theoretical, clinical, and empirical, aspects. Theoretically, the field has been transitioning from a biomedical reductionist view of disease that focused solely on the biological aspects of an illness, to a biopsychosocial model, that recognizes the importance of biological, psychological, social, and cognitive factors in both the etiology and maintenance of a person's illness. Fueled by this multidimensional view of health, clinical practice has begun to incorporate psychological, social, and cognitive aspects into treatment protocols, albeit slowly and not without resistance. Over the past 15 years, a number of multidimensional specialty centers have emerged that address the patient as a whole, rather than simply focusing on the biological aspects of a patient's illness. This model is particularly useful for patients that experience pain, as pain processing is a complex interaction of cognitive, physiological, and psychosocial variables, all of which must be identified and addressed for successful recovery.

In part, the dramatic change that has been experienced in the field of Medicine was a direct result of a heightened focus on health and health reform. This focus was fueled by a number of factors, including a growing aging population, and a concern for

the general quality of life that patients with chronic illness maintained following treatments that prolonged life (Blancett and Flarey, 1998). As patients lived longer, with a higher degree of disability, there was a growing sense that the medical field was missing what matters most to patients when measuring success or failure. Empirically, there was a need to shift away from the traditional biomedical tools of assessment. In response to this need, Health-Related Quality-of-Life (HRQL) instruments were designed as self-report measures of overall health and well-being. Although there is no clear consensus on what concepts should be included in a HRQL measure, all are subjective in nature, and tend to include some combination of physical, emotional, and social well-being concepts (Fayers and Machen, 2001).

A critical phase in the development of HRQL instruments concerns validity, reliability, and responsiveness in the intended populations. Reliability and validity have been assessed for a number of HRQL instruments over the past 30 years; however, responsiveness, or "minimum clinical important difference", is a relatively new concept within the health field. The minimum clinical important difference has been defined as the "the smallest change or difference in an outcome measure that is perceived as beneficial and would lead to a change in the patient's medical management, assuming an absence of excessive side effects and costs" (Wells, Beaton et al. 2001). The concept is distinct from validity, in that it assesses an "important" difference as opposed to a "statistical" difference.

In contrast to the standard methodology that exists for assessing validity and reliability, a number of different techniques abound for assessing the Minimum Clinical

Important Difference (MCID) of a measure. Distribution-based methods are built on the variability in a given measure, and include calculations of effect size, which provide a measure of the magnitude of change from pre to post. A downside to these types of measures is that they are influenced by the variability in the sample, and are not always predictive of individual changes (Testa et al, 1987). Anchor-based methods compare pre- to post-change scores to some "gold standard" outcome measure. Typically, patient self-report measures of change are assessed, and the average change in the HRQL for patients that consider themselves "somewhat better" is defined as the MCID. However, a number of issues exist with the use of a patient self-report as a gold standard, including correlation of error in the self-report HRQL measure and the self-report gold standard. Objectively defined gold-standards, such as rate of work return and health care utilization, circumvent these issues, and provide a clinical, economic, and patient-relevant meaning to quantitative changes in HRQL measures.

Health-Related Quality of Life Instruments are a central component of the larger field of health outcomes research, which is a field that continues to gain momentum in the 21st century. Clinicians are being trained in "Evidence-Based Medicine," and are encouraged to make clinical decisions based on the best empirical information available. Valid and reliable health outcomes are a vital component of this process. In addition, an understanding of what a specific magnitude of change means with regards to clinical, economic, and patient-relevant outcomes is critical to the evaluation of a HRQL instrument. The purpose of the current paper is to make the quantitative change scores in relevant HRQL measures more clinically meaningful, by relating percent

change in HRQL to objective work outcome criteria. MCIDs derived from the analysis may be used to predict patients that will have good or poor outcomes, so that patients at risk for poor outcomes may be targeted for further intervention.

The following work is presented in four sections: I. A literature review of relevant topics, II. Methods, III. Results, and IV. Discussion. Section I includes a literature search presented in three different chapters regarding the 1) theoretical, 2) clinical, and 3) empirical changes in the health field over the past half century (chapters 2-4). Chapter 5 presents the scope of the current project, including the purpose and relevant hypotheses. Methods are presented in Chapter 6, followed by the results, which are presented separately for the demographic variables (Chapter 7), and for each of the six psychosocial measures analyzed in Chapters 8-13. Chapter 14 presents the discussion and general conclusion.

CHAPTER 2

LITERATURE REVIEW: THEORETICAL CHANGES IN PAIN MEDICINE

As defined by the International Association for the Study of Pain, pain is "an unpleasant sensory and emotional experience associated with actual or potential tissue damage or described in terms of such damage." (*International Association for the Study of Pain*, 1994). Theories proposed to understand the development, maintenance, and experience of pain have evolved over the past century to include psychological, cognitive, and social factors that were considered to be independent of pain during the realm of the biomedical reductionist philosophy that dominated the field of medicine until the middle of the twentieth century. Section 2.1 reviews the history and evolution of pain theories.

The modern biopsychosocial model has proven especially useful in chronic pain patients, as psychological, cognitive, and social factors play a key role in the progression from acute to chronic pain. Section 2.2 discusses the model as it relates to the development of chronicity, and Section 2.3 addresses the critical role of psychological factors in this process.

2.1 Review of Pain Theories

Prior to the Renaissance, the idea that one's emotional state could affect pain perception was widely accepted. During the time of Ancient Greece, Hippocrates

theorized that an individual's personality was determined by bodily humors, and these bodily humors could affect one's physical state, and thus an individual's perception of pain (Gatchel, 1993). The idea that pain may be influenced by emotion fell out of favor during the 17th century, when Descartes popularized the idea that the mind and body function independently. Pain was considered a consequence of sensory transmission from the periphery to the brain, and hence independent of the mind. During the time this Cartesian view predominated, many physiological and anatomical discoveries were made, providing a better understanding of the sensory aspects of pain processing. The predominate pain theories during this time were one-dimensional, focusing on the physiological aspect of pain, and are known as the "classical" theories of pain.

Although these theories contributed greatly to the basic underlying physiological mechanisms of pain, the theories did not adequately account for pain experience, and treatments developed based on these classical theories fell short. Due to these shortcomings, theorists began exploring the role that psychosocial factors played in the development and maintenance of pain. Based on the outcome of these studies, theorists postulated that pain was a multidimensional experience that involved biological, psychological, and social aspects. Theories during this time may be considered multidimensional theories of pain, and predominate the current view of pain processing.

2.1.1 Classical Theories of Pain

The classical theories of pain that followed the traditional biomedical model of disease were developed during a time of immense growth in physiological and

anatomical methodology. These new techniques allowed researchers to focus on the physiological process of pain perception.

German physiologist Johannes Müller was the first to put forth the idea that the nerves of sense, motion, and vitality were determined based on the various organs that perceive information (Finger, 1994). Prior to the introduction of this "law of specific sense energies," in the early 1800s it was commonly believed that sensory information was transmitted to some common area in the brain. Müller proposed that there were distinct centers in the brain that were responsible for different sensations, and that nerves transmitting these sensations had some degree of specificity. Von Frey extended this theory (1894) to specific sensations, including pain. He conducted a variety of experiments to explore sensory transmission of mechanical and thermal stimuli, and proposed that there were specialized nerve endings for distinct modalities such as pressure, heat, cold, and pain. He suggested these nerve endings were connected to specific pathways that transmitted specialized information to distinct brain areas. Although the model proved too simplistic, the idea of pain being transmitted from the periphery to the brain via a specific type of receptor and pathway has persisted (Fields, 1999).

Following the discovery of morphologically similar nerve endings in human hairy skin, Weddell suggested that free nerve endings may respond non-selectively to the different modalities of touch. It was this finding that led a group of dissenters to propose that there were not distinct nerve endings responsible for transmitting different modalities, but that distinct patterns of nerve responses led to the perception of different

modalities. This set of theories became known as the "pattern theory of sensation" (Nafe, 1934; Sinclair, 1955; Weddell, 1955). According to this perspective, nociceptive perception was due to the stimulus intensity, combined with a specific pattern of responses (Gatchel, Peng et al. 2007). Although current evidence does not support that a complex pattern of responses is involved in pain perception, it has been demonstrated that impulse frequency and intensity plays a critical role in the perception of pain (Fields, 1999).

2.1.2 Multidimensional Theories of Pain

By the middle of the 20th century, it was recognized that the classical theories of pain did not adequately account for a variety of clinical and experimental findings. For example, the location of pain may be different from tissue damage; pain does not always resolve when tissue heals. The nature of pain, as well as the location, may change over time (Gatchel, Peng et al. 2007). In response to the inadequacy of the classical theories, more integrative and multidimensional theories of pain were developed that involved a combination of social, cognitive, and psychological factors, in addition to the traditional physiological mechanisms of pain.

2.1.2.1 Gate Control Theory

The Gate Control Theory proposed by Melzack and Wall (1965) provided a framework for sensory processing that answered many questions the specificity (Von Frey, 1894) and the pattern theory (Goldschneider, 1894) were unable to do. This was the first theory to consider cognitive, attentional, and psychological mechanisms in pain

processing. This framework was later expanded by Melzack and Casey (1968) to formulate the cognitive, affective/motivational, and sensory theory of pain processing.

Classical theories identified the importance of "specific" high threshold nociceptors. However, the Gate Control Theory recognized that non-nociceptive input may also play a role in pain processing and perception. This non-nociceptive input included descending cognitive input, as well as non-nociceptive peripheral afferent information. The Gate Control Theory proposed five stages of pain transmission: high threshold nociceptors at the periphery; facilitator interneurons in the spinal cord; low-threshold input in the spinal cord; inhibitory interneurons in the spinal cord; and descending modulatory input in the spinal cord (Gatchel, Peng et al. 2007). This was the first theory that included a specific mechanism to account for psychological and cognitive influence on pain perception.

The Gate Control Theory advanced the understanding of pain processing immensely. Overall, the theory provides a strong general framework for how pain information is processed, although specific mechanisms have proven more complex than originally proposed.

2.1.2.2 Neuromatrix Theory

Over thirty years after the initial proposal of the Gate Control Theory, Melzack expanded on the original Melzack and Casey (1968) theory of pain with the Neuromatrix Theory (Melzack, 2001). According to this theory, there is a "neurosignature" pattern of nerve impulses that is generated by a complex neural network in the brain. This widely distributed neural network is known as the "body-self neuromatrix," and incorporates the

cognitive, sensory, and motivational/affective systems described by Melzack and Casey (1968). Central to this theory is the idea that pain is generated by the neural network, rather than directly by nociceptive stimuli. This theory helps to account for the experience of phantom limb pain, where patients experience the sensation of a limb that has been amputated. The Neuromatrix Theory is still developing, and additional research is needed to elucidate specific mechanisms involved in the "neurosignature." However, the theory highlights the importance of viewing pain as a complex, multidimensional experience.

2.2 Acute vs. Chronic Pain

Due to the complex nature of pain processing, pain is difficult to treat. Pain treated in the acute stage has more successful outcomes than chronic pain. The success rates for chronic pain problems rarely exceed 60%, and long-term success rates are less than 30% (Gatchel 2004). Unfortunately, estimates of the prevalence of adults suffering from chronic pain range between 10 and 20% (Guereje, Von Korff et al. 1998; Verhaak, Kerssens et al. 1998; Blyth, March et al. 2001).

Although definitions vary, generally it is accepted that acute injury is approximately 2-4 weeks duration (Fardon, 1997). Acute pain is typically caused by trauma, surgery, or some other "physiological" process (Atlas and Deyo, 2001). Signs of automatic activity such as hypertension, sweating, etc. may be present in acute pain conditions (Cousins and Power, 1999). Most importantly, the pain typically ceases upon healing of the wound or medical condition. When acute pain does not resolve, it may progressively become more chronic. The distinction between acute and chronic pain is

not clear, however, and a variety of factors must be considered when making a chronic pain diagnosis.

Generally, it is accepted that injuries lasting up to twelve weeks fall in the "sub-acute" category, and chronic injury extends beyond twelve weeks (Fardon, 1997). Chronic pain definitions also vary and range from as short a duration of seven weeks to as long as six months (DSM-IV; APA, 1994; Anonymous, 1987). Although time is an easy method of classifying chronic and acute pain, additional factors, such as psychosocial variables, must be taken into consideration when making a diagnosis. A patient with significant psychosocial variables presenting five weeks post-injury may be identified as suffering from chronic pain, whereas a patient four months post-injury may be experiencing a longer sub-acute stage of injury. In recent years, there has been increased attention at identifying variables that predict patients at risk of developing chronic pain for early intervention.

The diathesis-stress model is a 3-stage perspective of pain designed to account for psychological, behavioral, and social factors that play a role in the development of chronic pain (Gatchel, 1991; Gatchel, 1996). In the acute stage (Stage 1), emotional reactions, including anxiety, worry, and fear, develop in response to the experience of pain. If pain has not subsided within two to four months, the patient enters into a more chronic stage of pain (Stage 2). In Stage 2, individuals experience more psychological and behavioral problems, including learned helplessness, anger, and avoidance. An individual's persona, including coping styles and psychological traits, in addition to socioeconomic and environmental conditions, all play a role in the progression of pain

from Stage 1 to Stage 2. In Stage 3, the negative mood, social limitations, and physical disability culminate in adoption of a "sick role," as the patients' life begins to revolve around their pain condition. Anxious patients fear engaging in activities, depressed patients feel helpless, and angry patients have little motivation to comply with recommendations from the healthcare systems (Gatchel et al, 2007). Physical deconditioning results from the lack of activity, attention is focused on the pain condition that has altered the patient's life, and depression is exacerbated by loss of social activities and work. Without the responsibility of work and social obligations, patients grow accustomed to avoidance of responsibility, which maintains the maladaptive behavior. Secondary gains associated with workers' compensation help to maintain the "sick role," as evidenced by poorer treatment outcomes for patients experiencing work-related injuries, as compared to those experiencing injuries that are not work-related (Greenough, Taylor et al. 1994; Roth, Richards et al. 1994).

Critical to the transition from Stage 2 to Stage 3 is the notion of physical and mental deconditioning (Mayer and Gatchel, 1988). Physical deconditioning involves loss of strength, flexibility, and endurance to the injured area resulting from lack of use. Mental deconditioning is the result of decreased motivation to participate in mentally stimulating and challenging activities, and results in decreased confidence, reinforcement of avoidant behaviors, and increased reliance on others for decision making. Both physical and mental re-conditioning are critical aspects of a chronic pain patient's rehabilitation (McMahon, Gatchel et al. 1997).

2.3 Psychopathology and Chronic Pain

The role of psychopathology in the etiology and maintenance of chronic pain has been well documented (Gatchel, Polatin et al. 1994; Gatchel 1996; Fishbain, Cutler et al. 1997). Specifically, the anxiety and depressive disorders have a high comorbidity with chronic pain diagnoses. Evidence supports the idea that people with depression and anxiety are more likely to develop chronic pain (Polatin, Kinney et al. 1993; Gatchel 1996).

Chronic pain also increases the risk of developing psychopathology (Fishbain, Cutler et al. 1997; Dersh 2000). Similar neurochemical systems are involved in pain and psychopathology, and evidence suggests that there is a complex interaction of these systems in chronic pain patients (Ward, Bloom et al., 1982; Roy, Thomas et al., 1984; Ong and Keng, 2003). Attention to psychopathology in the treatment of chronic pain is critical to successful rehabilitation (Gatchel, 1996; Gatchel, Garofalo et al., 1996).

2.3.1 Anxiety

Anxiety is a complex construct, and involves a blend of fear, worry, self-doubt, and hyperactivity of the sympathethic nervous system. Anxiety is commonly reported by pain patients (Wolfe, Smythe et al., 1990; Polatin, Kinney et al., 1993; Vowles, Zvolensky et al., 2004; Dersh, Gatchel et al., 2006), and is a natural reaction to the uncertainty and threat of painful stimuli. High anxiety levels are associated with increased pain perception, whereas low anxiety levels are associated with decreased pain perception (Cornwall and Donderi, 1988; Jones and Zachariae, 2004).

Anxiety is associated with hypervigilance, which increases the attention to pain. It has been suggested that this increased focus on pain actually increases the pain intensity (Rhudy & Meagher, 2000). When the underlying mechanisms of an individuals' pain are poorly understood, which is often the case in chronic pain conditions, this exacerbates a patient's anxiety (Gatchel, Peng et al., 2007). Patients are anxious about the status of their pain condition, whether pain will increase or improve, and whether they will be able to maintain their current physical capacity. Fear and anxiety of increased pain contributes to immobility, which in turn results in the physical deconditioning classically present in chronic pain patients (Boersma and Linton, 2006).

Evidence suggests that the fear of pain is a unique cognitive contributor to the experience of pain, outside of the sensory experience of pain itself. Researchers have demonstrated that fear of re-injury and fear of movement are significant predictors of functional limitations seen in chronic pain patients (Vlaeyen, Kole-Snijders et al., 1995; Crombez, Vlaeyen et al., 1999; Turk, Robinson et al., 2004). It has also been demonstrated that patients experience a reduction in pain and disability following treatment for anxiety (McCracken, Gross et al., 1996).

2.3.2 Depression

Research suggests that as many as 40-50% of all chronic pain patients suffer from some form of depression (Romano and Turner, 1985; Banks and Kerns, 1996; Dersh, Gatchel et al., 2006). Since the identification of the strong association between the two diagnoses, there has been debate as to whether chronic pain precedes depression or vice versa. Epidemiological evidence suggests a large percentage of chronic patients have

suffered from depressive symptoms prior to the development of chronic pain (Katon, Egan et al., 1985; Polatin, Kinney et al., 1993; Gatchel, 1996), although a certain percentage of patients are free of depressive symptoms prior to the onset of their chronic pain condition (Dersh, Gatchel et al., 2001). Identification of similar physiological mechanisms involved in depression and pain suggests that there is no singular antecedent, but there is a complex interaction of both cognitive and biological systems that result in the high comorbidity (Von Korff and Simon, 1996; Ong and Keng, 2003).

2.3.3 Personality Disorders

Personality disorders, as defined by the DSM-IV (APA, 2000), are life-long maladaptive patterns of perceiving, relating to, and thinking about the environment and oneself that are exhibited in a wide range of social and personal contexts. Research suggests that people with personality disorders are at a greater risk of developing chronic pain (Fishbain, Goldberg et al., 1986; Bigos, Battie et al., 1991; Polatin, Kinney et al., 1993; Gatchel, Polatin et al., 1994; Gatchel, 1996; Gatchel, Garofalo et al., 1996). It has been suggested that the maladaptive coping styles prevalent among all personality disorders are central to the development of chronic pain (Bigos, Battie et al., 1991; Gatchel, 1996).

CHAPTER 3

LITERATURE REVIEW: CLINICAL CHANGES IN PAIN MEDICINE

Following the emergence of the biopsychosocial approach in the middle of the 20th century, it was recognized that more interdisciplinary treatment options were needed to care for patients that suffer from chronic pain. The current chapter presents definitions of traditional biomedical reductionist treatments (Section 3.1), multidimensional pain centers (Section 3.2), and the functional restoration approach (Section 3.3).

3.1 Biomedical Reductionist Treatments

Traditionally, quantification of health care relied on biomedical methods including physician-based assessment of x-rays, biopsies, and overall physiological functioning. This method of assessment was based on the "disease model" which indicates there must be some amount of pathological abnormality which is indicated by a set of signs and symptoms.

Typical biomedical treatment options include pharmacological preparations (i.e. opioids, nonsteroidals, anticonvulsants, antidepressants, NMDA antagonists, and topical preparations), operative procedures, physical modalities, regional anesthesia, and neuroaugmentation modalities (i.e. spinal column stimulators, implantable drug delivery systems, etc. (for a review of each of the above listed treatments, and a discussion

relating cost-effectiveness of these therapies, please see Turk and Swanson, 2007). Overall, the above listed methods do not provide adequate pain relief for the majority of patients suffering from chronic pain. In fact, Hansson and Hansson (2001) conducted a multi-national study that found almost none of these traditional methods of pain treatment are effective at assisting patients return to work, or improve health related quality of life. The inadequacy of these biomedical treatment options led to the development of multidimensional pain centers.

3.2 Multidimensional Pain Centers

With the evolution of medical theory to the biopsychosocial approach, there was acknowledgment that a patient may be absent of observable pathological abnormalities and still have signs and symptoms of a disease. Multidimensional pain centers were founded that were geared towards treating psychological, emotional, and cognitive aspects of pain, in addition to the physiological components. Typically, these treatment centers involve some type of cognitive-behavioral therapy, and include a number of psychological tests to identify maladaptive coping patterns, and psychopathology that might be playing a role in maintaining pain. In addition, they attempt to identify secondary gains, such as workers compensation and disability, which may be playing a mediating role in pain maintenance.

A recent analysis provides evidence that multidisciplinary pain centers are substantially more cost-effective than biomedical treatment options (Turk and Swanson, 2007). Despite this fact, it is estimated that only 6% of all chronic pain sufferers are treated at multidisciplinary pain centers (Marketdata Enterprises, 1995). For a full review

of the evolution of multidisciplinary treatment centers and the cost-effectiveness of such centers, please see Turk and Swanson (2007).

3.3 Functional Restoration

Functional restoration was developed in 1983 as a variant of chronic pain management intended for workers' compensation injuries (Mayer, Gatchel et al., 1985). Concordant with the biopsychosocial paradigm, the program assumes that disability related to chronic pain involves both physical and psychosocial components, and treatment of both is necessary for successful intervention. Patients undergo quantitatively directed exercise progression, combined with a multimodal disability management program, which involves case management and psychosocial interventions (Mayer, Gatchel et al., 2006). Key components of the functional restoration approach include objective quantification, an interdisciplinary team, and a sports medicine approach.

Functional restoration is built on the tenant that objective quantification, through evaluation of physical and psychological functioning, is required for the diagnosis and treatment of chronic pain conditions (Mayer and Gatchel, 1988). The physical functional capacity includes evaluation of neurological deficits, strength, cardiovascular endurance, lifting capacity, and overall effort. A patient's current level of function is assessed prior to treatment, and individual goals are designed for the patient to regain function. Psychological assessment provides insight into patients coping styles and psychologic functioning to help tailor the treatment process to the individual. A variety of psychological assessments are administered, including pain, disability, psychopathology, and depression scales. Information provided by the physical and psychological functional

evaluation provides key insights into how a patient views his pain condition, and shapes how the team should tailor treatment for optimal success.

The treatment team includes physical therapists, occupational therapists, psychologists, nurses, and physicians. The role of the physician is to evaluate structural diagnostic testing, determine the need for additional surgical treatment, and participate in medicolegal proceedings (Mayer and Gatchel, 1988). The role of the nurse is to provide counseling on medical matters, educate the patient, and communicate with outside agencies. The physical therapist focuses on functional diagnostic evaluation, and reconditioning the injured part of the body. Occupational therapists provide training in integrative physical tasks, such as lifting, bending, twisting, sitting, and standing. In addition, they play a role in evaluation and counseling regarding the socioeconomic aspects of disability that often play a role in a patient's desire and belief in the ability to return to work. Psychologists identify individual barriers to successful treatment, and help to treat these barriers through a cognitive-behavioral treatment approach. Case managers serve to coordinate treatment and monitor individual patient progress. The team of professionals meets weekly to evaluate patient progress and identify factors that may impede successful return to work.

The sports medicine treatment approach involves reconditioning of the affected area in an attempt to restore functional capacity. Patients undergo a variety of exercises to address mobility, strength, endurance, and cardiovascular deficits (Mayer and Gatchel, 1988). Following rehabilitation of the injured area, a continued maintenance program is recommended, and patients are encouraged to return to work and normal levels of

activity. The return of normal productivity levels addresses the psychological deconditioning that results from loss of physical activity and adoption of the "sick role."

Numerous studies indicate the functional restoration approach is linked to improvements in a variety of outcome measures (Mayer, Gatchel et al., 1985; Mayer, Gatchel et al., 1987; Hazard, Fenwick et al., 1989; Bendix, Bendix et al., 1998; Mayer, McMahon et al., 1998). Consistently across studies, over 80% of patients treated with functional restoration return to work, as compared to only 29-41% of no-treatment controls (Anagnostis, Gatchel et al., 2004). Non-functional restoration treatment comparison groups also have twice the rate of additional surgeries and unsettled compensation litigation, five times the rate of increased health care utilization, and higher rates of re-injury (Anagnostis, Gatchel et al., 2004).

CHAPTER 4

LITERATURE REVIEW: EMPIRICAL CHANGES IN PAIN HEALTH

With the increasing amount of dollars being spent on health and health promotion in the latter half of the 20th century, there was a growing demand by consumers, politicians, and policy-makers for accountability within the health sector. Health outcomes research evolved as a method of quantifying health care information so that it may be evaluated for treatment efficacy, cost-utility, and diagnosis purposes. Section 4.1 addresses the evolution of health outcomes measures, and Section 4.2 discusses the psychometric process of validating health outcome tools. Finally, Section 4.3 discusses the development of a relatively new concept within the psychometric validation process, the Minimum Clinical Important Difference.

4.1 Outcome Measures

Outcomes measurement is defined as "a means of verifying the success of a provider's care in terms of predetermined outcomes (Huber and Oermann, 1998)." Initially, the focus of outcome measures was to provide information on the cost-effectiveness of treatment, although in recent years it has evolved as a means to examine clinical, functional, and patient satisfaction as well (Huber and Oermann, 1998).

Although outcomes such as mortality and morbidity data have been considered for over a century, it was not until the 1980s that the measurement of outcomes became a

central focus to health research. Some of the factors that encouraged the emergence of more patient focused health outcomes were the aging population, increased health care utilization, and variability in treatment application (Wennberg, 1990). Medical advances in chronic diseases led to people living longer, and frequently, more impaired lives. This led to increased healthcare utilization, putting a strain on the healthcare system. In addition, there were concerns that the medical field was missing what matters most to patients when considering treatment options and measuring "success" or "failure." The "quality of life" of patients living with chronic disease became a chief concern, as opposed to whether a patient had a successful surgery, or had an improvement in physiological symptoms. In addition, it was recognized that there was a great deal of variability in treatment recommendations among physicians for patients with similar symptoms (Wennberg, 1990). One doctor might rely heavily on surgery for carpal tunnel diagnoses, while another doctor might prescribe physical therapy more often. All of these factors combined led to international focus on healthcare reform, and a demand of quantifiable information regarding the value of health care dollars.

Functional assessment questionnaires (FAQs) were developed as patient-centered reports of disability, pain, and overall functioning. Functional status has been defined as the "degree to which an individual is able to perform socially allocated roles free of physically or mentally related limitations (Bowling, 1991). Assessment of functional status involves asking patients questions concerning their ability to perform tasks of daily living. The measures focus on what is most important to patients. These measures are

also known as Health Related Quality of Life Questionnaires (HRQLs) and disability scales (these terms will be used interchangeably throughout the current paper).

The World Health Organization (WHO) views these functional outcomes as central to the assessment of disease. In 2001 they published a revised version of the International Classification of Functioning, Disability, and Health (ICF). The purpose of the ICF is to develop a unified language of health and health-related states, and permit communication about these related issues around the world. The ICF is actually a subset of a family of classifications that were developed for this purpose that provides information specifically about functioning. It is an assessment of degree of disability that places emphasis on function and not on condition or disease. It includes personal and environmental determinants of health and disablement, and is interactive as opposed to linear.

Evidence-Based Medicine, which places an emphasis on applying evidence gained from scientific methods to clinical decision, relies on the use of well-validated and reliable health outcomes. Although a variety of health outcomes are utilized in the medical field, currently HRQLs are considered the "gold-standard" in evidence-based medicine.

4.1.1 Types of HRQL Measures

HRQL measures are used in a variety of ways within health care. For example, they may be used in cost-utility analysis, evaluation of clinical trials, and as an evaluative tool within research (Fitzpatirck, Fletcher et al. 1992). Numerous HRQL measures have been designed over the past 3 decades, and they all share one commonality: they are built

on a *subjective* report of health status. However, they vary on two critical dimensions: the *construct design* and *target population*.

4.1.1.1 HRQL Construct Design: One-dimensional vs. Multidimensional Measures

One of the first attempts to capture subjective reports of health status was the Karnofsky Performance Scale (Karnofsky and Burchenal, 1947). The scale ranges from 0, meaning "dead", to 100 indicating "normal, no complaints, no evidence of disease." This global assessment of change addresses the status of a patients overall well-being. A number of additional scales were developed based on this one-dimensional construct of functional ability, or physical functioning (Fayers and Machen, 2001). These types of measures are attractive due to their ease of application and interpretation.

With the evolution of the biopsychosocial approach to medicine, it was recognized that an evaluation solely focused on physical functioning ignored the important aspects of mental, social, and cognitive functioning. Factor analysis of a number of HRQL measures indicate that HRQL may be reduced to a number of lower-level factors, such as physical, emotional, and cognitive functioning (Fayers and Machen, 2001). Thus, a number of HRQLs consider physical, mental, and social functioning, in addition to pain and disease specific symptoms.

4.1.1.2 HRQL Target Population

Some HRQLs are designed for use across disease populations (general HRQLs), while others were designed for use in patient-specific populations (disease-specific HRQLs). One benefit of the broad nature of generic instruments is that patients with

different diseases may be compared to one another, or against the general population. Unfortunately, due to the general nature of these instruments, they fail to focus on issues of particular concern to patients within a specific disease cohort. As a result, it is generally accepted that for a complete clinical picture, both a general and disease specific measure should be examined.

4.2 Psychometric Theory

Just as the medical assessment tools have evolved over the 20th century, the analytic process utilized to evaluate the reliability and validity of these tools has evolved in turn. In fact, traditional biomedical researchers developed functional outcomes measures with a focus on practicality and comprehensiveness, and lacking knowledge of psychometric theory, ignored the empirical issue of validity and reliability altogether (Deyo, Cherkin et al., 1991; Kopec, 2000). However, current standards demand a self-report instrument be subjected to intense psychometric scrutiny to ensure its clinical value (American Psychological Association, 1985).

Psychometric evaluation of an instrument addresses the following: validity, reliability, and responsiveness, or the Minimum Clinical Important Difference. Validation is commonly defined as "the process of determining whether there are grounds for believing that the instrument measures what it is intended to measure, and that it is useful for its intended purpose, (Fayers and Machen, 2001)". The two most common types of validity are *criterion-related validity* and *predictive validity*. *Criterion-related validity* is the ability of a measure to produce results similar to those provided by other established measures of the same variable. The relationship between the measure in question, and

other measures purported to measure the same construct, is typically presented as the Pearson r correlation, or degree of shared variability in the two instruments. High r value indicates two measures are indeed measuring the same construct. For scales that have multiple items that are intended to measure the same construct, analysis of the inter-item correlation structure is relevant, and is presented as Cronbach's α. This type of internal validation may not be reported for single-item scales. This issue may be avoided by including redundant items in the development of the scale as anchors of the intended construct, which are later dropped from the final version. *Predictive validity* refers to the ability of a measure to predict some future behavior. An outcome measure that is designed to assess patient function, that in fact does not accurately predict the patient's ability to function, is clinically useless. Assessment of predictive validity is presented as Wilcoxon within subjects t-test (or F test) of pre-to post change, or between subjects t-test (or F test) of treatment/no-treatment groups. In addition, this may include reports of sensitivity and specificity.

Reliability, also known as repeatability, assesses whether a measure has consistency in obtaining the same results across time. A certain amount of random variability is associated with every measure. When a patient whose condition has not changed repeats a test, the scores should be relatively similar. If reliability results are poor, and there is a high degree of variability associated with repeated measures, the measure is providing no useful information. Reliability statistics are generally reported as Pearson r correlations, and are known as test-retest reliability statistics.

A measure can be reliable, and valid, but provide no clinically relevant information. Responsiveness is known as the ability of a particular instrument to detect clinically *meaningful* change over time (Kirshner and Guyatt, 1985; Deyo and Diehl, 1988; Deyo, Cherkin et al., 1991). The concept of responsiveness, also referred to as "the minimum clinically important difference" (MCID) of a measure, was originally introduced to the health field by Kirshner and Guyatt in 1985; however, the concept has been addressed in the psychology literature since the early 1970s (Cronbach and Furby, 1970; Nunnally, 1975). Although methods for evaluating validity and reliability are standard and well delineated, debate exists regarding the optimal method for determining MCIDs, and very few HRQL measures have been evaluated for responsiveness (Terwee, Dekker et al., 2003). The focus of this paper concerns evaluation of the "minimum clinical important difference" (MCID) of relevant functional outcome measures within the CMSD population, thus Section 4.3 provides a detailed description of the concept of MCID as it relates to the field of HRQLs.

4.3 The Minimum Clinical Important Difference

MCID has been defined as "the smallest change or difference in an outcome measure that is perceived as beneficial and would lead to a change in the patient's medical management, assuming an absence of excessive side effects and costs" (Wells, Beaton et al., 2001). The three main goals of assessing MCIDs are to provide objective differences that may be used to; (1) justify treatment for an individual; (2) assess group differences in treatment efficacy; and (3) as a diagnostic tool (Beaton, Bombadier et al.,

2001)**.** Evaluative techniques for assessing responsiveness will vary based on the goal of application (Testa, 1987).

4.3.1 Minimum Clinical Important Difference Calculation Methods

A variety of methods have been proposed to assess MCIDs, including distribution-based measures, and anchor-based measures. Distribution-based measures are built on the variability of the measure of interest, whereas anchor-based measures compare pre- to post-change scores to some "gold standard" outcome measure.

4.3.1.1 Distribution-based Measures

Distribution measures are based upon statistical distributions rather than direct observation. These measures take into consideration that some of the observed change from pre- to post-treatment is due to random measurement error, and not significant improvement. Several different measures have been used, the majority of which are variants on the basic effect size calculation. The basic effect size calculation is an estimate of the magnitude of between-group differences on a standard scale (Kazis, Anderson et al., 1989). The difference in two means is divided by a standard deviation. When the standard deviation of the measure of interest at *baseline* is used, the calculation is considered a standardized effect size (Cohen, 1988). The standardized response mean (Samsa, Edelman et al. 1999) uses the standard deviation of *difference scores* in the denominator. The responsiveness statistic (Marx, Hudak et al., 1997) uses the standard deviation of the measure of interest in *stable individuals*. An alternative distribution-based measure that has been reported is the minimum detectable change (MDC) statistic (Beaton, Bombadier et al., 2001; Wells, Beaton et al., 2001; Hagg, Fritzell et al., 2002).

This involves adding a specified confidence interval to the SEM, and is based on the premise that if a change score exceeds the value of minimal detectable change, it is true change and not just error.

One of the benefits of distribution-based measures is their simplicity and ease of calculation. In addition, tradition already exists for presenting effect sizes, and they are widely accepted in a variety of fields (Samsa, Edelman et al., 1999). Despite their ease, the application of changes based on distributions must be used with caution when interpreting the level of change for individual patients (Testa, 1987).

4.3.1.2 Anchor-based Measures

Anchor-based measures may be used to analyze differences at the individual and/or the group level. An external criterion is used as the "gold standard" to define improvement. Patients are classified according to this gold standard, and the mean change in patients that are classified as obtaining a minimum outcome is considered the MCID. Alternatively, a classification analysis is conducted on change-scores of the measure of interest, and includes reports of sensitivity and specificity for discriminating between those that did, and did not, achieve a significant level of improvement. The anchor criterion should be independently interpretable (Samsa, Edelman et al., 1999), and related to the field of application. Measures that may be selected for the anchor include physician-based, patient-based (i.e., self-report of pain and disability), or objective measures. Analyses utilizing these various perspectives may lead to very different determinations for what defines a MCID (Beaton, Bombadier et al., 2001).

Physician-based anchors are typically agreed upon by a group of treating physicians about what, through experience, most consistently leads to significant improvement. Outcomes are typically categorical, and fall within an "excellent", "good', "fair", or "poor" category. In this case, the minimum clinically significant difference would be considered the average change on the measure of interest that categorized patients in the "fair" or "good" category, based on the authors' definition. Stratford (1998) and colleagues employed this methodology in an evaluation of the MCID of the Roland-Morris Back Pain Questionnaire for patients undergoing treatment for low back pain. Prior to treatment, therapists met with the patients and established treatment goals related to function. Following treatment, physicians met with their patients and determined if they had met the goals set out prior to treatment. Although physician-based assessment has been frequently used in the past, it has fallen out of favor in recent years due to the increased attention being given to the importance of patient satisfaction.

The most frequently used gold standard is a patient-based anchor, or a patient self-report of what constitutes an important change. In fact, a commonly cited definition of MCID is "the smallest difference in score in the domain of interest which *patients* perceive as beneficial" (Jaeschke, Singer et al., 1989). Often, some subjective global assessment of change is utilized as the gold standard, which includes categories of "much better", "somewhat better", "no change", or "worse". Hagg and colleagues (2003) employed these methods to evaluate the MCID of the ODI, Global Functioning Scale (GFS), Zung Depression Scale (ZDS), and VAS for patients receiving fusion surgeries for low back pain. The MCID of improvement was calculated by taking the difference in

the change score of patients that assessed themselves as "better" and those that assessed themselves as "unchanged." Similar methodology has been employed by a handful of researchers to evaluate the MCID of a variety of psychosocial instruments (Jaeschke, Singer et al., 1989; Kulkarni, 2006).

Patient accounts of important differences are undeniably an important aspect of treatment outcomes. A number of issues are associated with patient self-report level of change, however. Norman and colleagues (1997) point out that when utilizing retrospective *self-report* judgments of change as a gold standard to evaluate the responsiveness of a *self-report* HRQL, there is a violation in the assumption of independence of error, as both measures are being reported by the subject. Physician report of assessment of change does not circumvent the issue, as physician assessment is heavily influenced by the patients' self-report of their condition. A number of researchers have pointed out the need to relate change in HRQL instruments to more objective assessments of change (Testa and Simonson, 1996; Beaton, Boers et al., 2002; Terwee, Dekker et al., 2003). Objective anchors that may be used to assess significant improvement include return-to-work, work retention at one-two years following treatment, healthcare utilization, and case settlement. Despite the obvious significance of an objective perspective in determining MCID, very few studies have employed objective outcome criteria. The few studies that have were analyzed at the group level, and not the individual patient level (Samsa, Edelman et al., 1999).

4.3.2 Relevant Issues with Minimum Clinical Important Difference Measures

Recently, there has been some debate as to what calculation method of MCID is ideal. Some authors suggest that, due to the ease of calculation and interpretability, distribution-based methods are the optimal choice (Samsa, Edelman et al., 1999; Wyrwich, Nienaber et al., 1999). Others suggest that anchor-based approaches are necessary when attempting to apply MCID to the individual level of change (Testa, 1987; Hays and Woolley, 2000; Kulkarni, 2006). Depending on the goal of the research and the target audience, optimal methodology will vary. In addition to the type of methodology selected, a variety of additional factors, such as pre-treatment level of severity, and disease of application, may affect the magnitude of the MCID.

When the main goal of a study is to examine between-group differences, selection of a distribution-based measure of MCID has advantages. Specifically, the calculation is simplistic and is easily interpretable by a wide-audience (Samsa, Edelman et al., 1999). A drug company that is conducting a clinical trial on a new analgesic could perform an effect size calculation very quickly, and physicians, pharmacists, and researchers alike would find no difficulty in interpreting the results. Ultimately, however, findings from between-group comparisons will be applied to a within-group comparison. If a patient tries the new drug, and experiences the same magnitude of change as the MCID estimated from the distribution-based approach, will this be successful enough to justify the cost and potential side effects to the individual patient? Comparisons of methodologies indicate that MCIDs derived from different techniques are variable

(Kulkarni, 2006), with effect size calculations leading to much smaller MCID determinations as compared to anchor-based approaches (Kulkarni, 2006).

It has also been suggested that within the anchor-based methodology, estimates of MCID may vary based on selection of the "gold standard," or anchor criterion (Beaton, Bombadier et al., 2001). As stated previously, the most common anchor selected is a patient self-report of global health. Frequently, authors ask patients "compared to pre-treatment your condition is (a) much better, (b) better, (c) unchanged, (d) worse." (Hagg, Fritzell et al., 2002). Another alternative that may be used is the health transition item on the SF-36, which asks subjects to compare their current health to their health one year ago. One of the pitfalls of utilizing a global health index is that subjects' responses may be influenced by factors unrelated to the treatment in question. In addition, research suggests that peoples' retrospective judgments of change are more heavily influenced by their current states, which makes recall of the pre-treatment severity difficult (Norman, Stratford et al., 1997). Objective criteria, such as return-to-work, work-retention, and health care utilization, are not influenced by such bias, as they are based on concrete criteria relevant to a patient's improvement. Unfortunately, the availability of objective outcome criteria is very limited, preventing objectively driven MCID standards on HRQL measures from being evaluated. Research estimating MCID based on such objective criteria is sorely needed, and efforts to compare estimations based on subjective and objective criteria will provide insight into the degree of relationship among patient self-report and objectively defined success.

An additional factor that has been implicated in MCID variation is pre-treatment level of severity. There is a growing body of literature that indicates people with more functional limitation and disability may have different MCID estimations as compared to people with less moderate disability (Riddle, Stratford et al., 1998; Stratford, Binkley et al., 1998). Based on this evidence, it has been suggested that pre-treatment differences in disability level should be taken into consideration when determining clinically significant change.

The initial purpose of the MCID field was to define a set level of change that is necessary to achieve success of a particular treatment. Researchers would hope to use this degree of change to identify group differences, and doctors would hope to utilize this "magic number" to diagnose patients with a certain condition, or determine if a treatment has been successful for an individual. Hays and Woolley (2000) suggest that, due to the complexity of the issues surrounding MCID, there is no "one" clinically meaningful difference for a particular scale. Although a global "magic number" may not be feasible, through careful comparison of the different methods and evaluation of variables that affect the magnitude of MCID, improvement in diagnostic techniques and superior methodology for evaluating treatment efficacy may be realized.

CHAPTER 5

SCOPE OF THE CURRENT PROJECT

The central focus of the current project is to make quantitative changes in HRQL measures in *individual* patients more clinically meaningful, by relating the percent change in HRQL to objective outcome criteria. The prevalence of pain, critical role of HRQL measures in the field of medicine, and evaluation of MCIDs as a critical aspect in the validation process of measurement scales, highlights the importance of the current project. Section 5.1 discusses the relative incidence of chronic musculoskeletal work related disorders, and is followed by a discussion of relevant outcomes utilized in the pain field (Section 5.2). The chapter concludes with a formal presentation of the purpose (Section 5.3) and relevant hypotheses (Section 5.4).

5.1 Incidence of Chronic Musculoskeletal Work Related Disorders

Work-related injuries continue to be a serious economic and health concern in our country. Approximately four million work-related injuries were reported in 2005 alone, and of these, approximately 2.2 million resulted in lost work days, job transfers, or restriction of duties (U.S. Department of Labor, 2005). Annual costs associated with work-related injuries range from $800 billion to over $1 trillion dollars (U.S. Census Bureau, 1996; Brady, Bass et al., 1997; National Safety Council, 2000; U.S. Department of Labor, 2000; Melhorn and Gardner, 2004).

Of all work-related injuries, musculoskeletal disorders are among the most costly and disabling, and contribute to a significant portion of this annual cost (Schultz, Stowell et al., 2007). Musculoskeletal disorders are injuries or conditions involving the tendons, nerves, and muscles that provide support and structure to the body. Some of the most common musculoskeletal disorders involve the spine. Low back pain is the leading cause of disability in people under the age of forty-five, and approximately 3-4% of the population of industrialized countries is affected by a low back pain episode at some time in their life (Mayer and Gatchel, 1988).

The majority of patients that experience musculoskeletal pain improve rapidly and return to work shortly after their injury. A small subset of this population, however, develops chronic pain resulting in occupational disability (Reid, Haugh et al., 1997). It has been estimated that approximately 60% of patients have returned to work within one week (Seferlis, Nemeth et al., 2000), whereas 10-15% eventually develop a chronic musculoskeletal disorder, resulting in extended absence from work (Spitzer, LeBlanc et al., 1987; Skovron, 1992; Reid, Haugh et al., 1997). It is this 10-15% that account for the majority of lost productivity and wages, and the bulk of health costs for work-related injuries. Specifically, with regards to low back injuries, approximately 50% recover within one month, 90% within six weeks, and 10% result in chronic injury (Beurskens, de Vet et al., 1995).

Of all musculoskeletal disorders, low back pain is one of the most prevalent, and accounts for a significant portion of all disability costs. The costs associated with the diagnosis and treatment of chronic low back pain accounted for $25 billion in 1991

(Kuritzky and Carpenter, 1995), and approximately 33% of healthcare and indemnity costs associated with workers compensation claims have been attributed to occupational low back pain injuries (Anderson, Pope et al., 1991).

Although back injuries are prevalent among all occupations, currently the service industry accounts for the highest percentage of all back injuries (Subramanian, Desai et al., 2006). Approximately 28% of all work-related back injuries from 2000-2002 occurred in the service industry, 19% in manufacturing, and 16% in retail industries. There has been a decrease in back injuries in the transportation, mining, and construction industry over the past decade, however this decrease has not occurred in the service and manufacturing industries. Overexertion injuries account for approximately 70% of all back injuries.

A myriad of factors have been identified that predict individuals that may incur an occupational low back pain disorder, and how long the injury will last. Demographically, work-related back disorders are more prevalent among women than men, and among Caucasians than African-Americans (Praemer, Furnes et al., 1992). Following injury, research suggests that reimbursement is a mediating factor. Sander and Myers (1986) report that individuals who were injured while not at work had an average disability length of four months, compared to a fifteen month length of disability of patients that were injured while at work. Perceived work dissatisfaction and inadequate income have also been linked to recurrent low back pain incidence (Papageorgiou, Macfarlane et al., 1997). In addition, a clear link has been established between a variety of psychological

variables, including mood, anxiety, and cognitive functions, and the development of acute to chronic low back pain (Linton, 2000).

5.2 Outcome Measures and CMSD Disorders

Evaluation of CMSDs must rely on functional status, as physiological factors are not always reliable and do not consistently correlate with pain and disability levels (Deyo, 1988; Beurskens, de Vet et al., 1995; Mayer, Prescott et al., 2000). With the evolution of the health outcomes field, there has been a trend in the past decade to include subjective health questionnaires, rather than relying solely on objective physical measures when evaluating treatment outcomes in various health settings. Functional status includes consideration of muscle strength, spinal mobility, employment status, and psychosocial variables (Mayer, Gatchel et al., 1985; Deyo, Andersson et al. 1994; Flores, Gatchel et al., 1997). A number of FAQs are frequently utilized in CMSD patients, including both general and disease-specific measures.

5.2.1 Oswestry Low Back Pain Disability Questionnaire (ODI)

The Oswestry was developed after a group of people working at an orthopedic clinic observed functional limitations in daily living tasks for people suffering from low back pain (Fairbanks, Couper et al., 1980). The intent of the scale was to capture functional limitations in daily living tasks that occur in response to a patients' injury. The scale has 10 items that involve a variety of daily activities, such as self-care, lifting, walking, standing, and sitting. Each of the items is scored from 0 to 5, and scores are calculated as simple percentages, with high scores indicating high functional loss.

The Oswestry was one of the first measures designed to assess functional loss and disability, and has been thoroughly researched and validated (Ohnmeiss, 2000). The test has high test-retest reliability (Fairbank, Couper et al., 1980; Gronblad, Hupli et al. 1993; Triano, McGregor et al., 1993), and fair internal consistency (Fairbank, Couper et al., 1980; Kopec, Esdaile et al., 1996; Fisher and Johnson, 1997).

Very few criticisms have been directed at the Oswestry scale. It has been suggested, however, that the Oswestry may have a potential floor effect, such that people with low functional disability are less accurately classified as patients that are mildly or moderately disabled (Kopec, Esdaile et al., 1995; Kopec, 2000; Roland and Fairbank, 2000). Another criticism is that the Oswestry only focuses on physical aspects of functioning, and ignores psychosocial concerns that are known to play a key role in chronic pain maintenance (Turk, 1999).

A number of studies indicate the Oswestry is successful at detecting clinically meaningful change (Beurskens, de Vet et al., 1999; Taylor, Taylor et al., 1999; Roland and Fairbank, 2000). The minimum clinical important difference has been reported as 5.2 (Suarez-Almazor, Kendall et al., 2000), 7.0 (Lurie, Hanscom et al., 2001), 10 (Hagg, Fritzell et al., 2002), and 16.3 (Taylor, Taylor et al., 1999).

Clinically, the comparison of the Oswestry and the BDI helps to identify patients who are symptom magnifiers (Mayer and Gatchel, 1988). If a patient scores high on the Oswestry, which indicates high functional loss, and low on the BDI (low depression), they often will reject the emotional component of their pain. Within the functional restoration program, psychologists use the comparison among these scales to identify

patients who are seeking "a medical cure" and need additional education on the emotional aspect of chronic pain maintenance.

5.2.2 Million Visual Analog Scale (MVAS)

The MVAS was developed in an effort to describe both pain and disability (Million, Haavik-Nilsen et al., 1981). The use of the analog scale is beneficial, as this type of measure tends to be highly reproducible and correlate well with "objective" findings by clinicians (Mayer and Gatchel, 1988). The scale is 15 items, takes 5-10 minutes to administer and score, and it is recommended that the instrument be administered verbally by the clinician to the patient (Million, Hall et al., 1982). MVAS scores range from 0-150, with higher scores indicating a more moderate level of disability and pain.

Very few validation studies have been performed on the MVAS (Ohnmeiss, 2000). The initial validation study that was conducted by the creators of the measure indicates that the test-retest reliability is .97 (Million, Hall et al., 1982). Intra-rater reliability was estimated as approximately .97 for the whole scale, and from .85-.94 for individual items, whereas inter-rater reliability was approximately .92 for the total scale, and between .66 and .92 for the individual items (Million, Hall et al., 1982).

Improvements have been demonstrated on the MVAS for patients undergoing functional restoration for chronic pain; however, none have assessed a Minimum Clinical Significant Difference (Mayer, Gatchel et al., 1985; Gatchel, Mayer et al., 1986; Hazard, Fenwick et al., 1989). Anagnostis and colleagues (2003) demonstrated the clinical utility of the measure in predicting treatment outcomes in patients with chronic musculoskeletal

disorders. The authors divided patients into six groups ranging from "no reported disability" to "extreme disability" based on their MVAS scores. Patients in the moderate pretreatment group were less likely to complete the program, and more likely to visit a new health care provider. In addition, patients scoring in the most moderate group at post-treatment assessment were 30% less likely to have had returned to work, 42% less likely to have retained work 1 year following treatment, and 15% less likely to have settled their workers' compensation case, as compared to patients in the no reported disability group. This study illustrates the strength of the scale in accurately distinguishing patients' level of disability.

Additional studies are needed to validate the reliability, validity, and responsiveness of the MVAS. However, initial studies indicate that the scale shows promise as a quick, easily interpretable assessment of functional disability and pain status.

5.2.3 Short-Form 36 (SF-36)

The SF-36 is a general health questionnaire that was developed as an outcome of the longitudinal Medical Outcome Study (MOS) that began in 1986 (Ware and Sherbourne, 1992; Gatchel, Polatin et al., 1998). One of the goals of the study was to develop more efficient tools for evaluation of patient outcomes (Gatchel, Polatin et al., 1998). The SF-36 was designed to assess overall health status in patients in a wide variety of conditions, as opposed to assessment of a specific disease or condition. The form consists of 36 questions, and results in an 8-scale profile, in addition to a physical

component summary (PCS) and mental component summary score (MCS). Lower scores are associated with a higher degree of disability.

Due to the generality of the SF-36, a variety of reliability and validity studies have been conducted in diverse applications. Reliability estimates are all above .80 for each of the 8 scales (Gatchel, Polatin et al., 1998). In addition, the content, criterion, and construct validity have all been demonstrated to be strong in numerous studies (Brazier, Harper et al., 1992; Katz, Larson et al., 1992; Ware, Snow et al., 1993; Brazier, Roberts et al., 2002).

Within the chronic musculoskeletal population, the SF-36 has demonstrated the ability to detect treatment outcome changes in patients (Gatchel, Polatin et al., 1998; Gatchel, Mayer et al., 1999; Taylor, Taylor et al., 1999). Despite the ability of the scale to detect overall group differences, however, one study indicated that it was not predictive of individual success in treatment (Gatchel, Polatin, Mayer, Robinson, & Dersh, 1998). It has been suggested that, due to the brevity of the eight scales, there is low within-subjects reliability, leading to higher confidence intervals around the individual scores (McHorney and Tarlov, 1995). Due to these psychometric limitations, the scale is unable to successfully predict individual outcomes in treatment (Gatchel, Polatin, Mayer, Robinson, & Dersh, 1998).

Although the scale has yet to demonstrate clinical utility in detecting individual treatment outcomes, self-reported pre-program mental health, pain level, and social disability have been identified as important risk factors for non-completion in a

functional restoration program (Gatchel, Mayer et al., 1999), making it an ideal scale for clinical research trials.

5.2.4 Pain Disability Questionnaire (PDQ)

The PDQ was designed as a functional measure of disability for use in the CMSD population (Anagnostis, Gatchel et al., 2004). The scale yields a functional status component, psychosocial component, and a total component score. Each item is scored from 0-10, for a total cumulative possible score of 150. Higher scores on the scale indicate more moderate levels of pain and disability.

Very few studies have explored the reliability and validity of the PDQ. The initial analysis conducted by the authors who designed the scale indicate a test-retest reliability of .97 and internal consistency of .96 (Anagnostis, Gatchel et al., 2004). In addition, multiple studies indicate that patients in a functional restoration program demonstrated significant decreases from pre- to post-treatment (Anagnostis, Gatchel et al., 2004; Gatchel, Mayer et al., 2006). Compared to the MVAS, Oswestry, and SF-36, the PDQ had the largest effect size (Anagnostis, Gatchel et al., 2004), indicating the scale has good responsive properties.

One strength of the scale is the inclusion of a psychosocial component, as the majority of functional status measures ignore this important aspect of functioning that is central to the maintenance of chronic pain conditions (Anagnostis, Gatchel et al., 2004). However, additional studies are needed to explore the validity, reliability, and responsiveness of the PDQ as a measure of functional pain and disability.

5.2.5 Pain Intensity (PI)

Pain analog scales are the most widely used assessment tools to evaluate pain levels in healthcare settings (McGeary, Mayer et al., 2006). Patients are asked to indicate the severity of their pain along a measured line. In the Pain Intensity (PI) Scale, the version of the scale utilized at PRIDE, the scale is scored from 0-10, with 0 being no pain and 10 being the highest level of pain.

Traditionally, pain was measured using categorical scales, such as none, mild, or moderate levels of pain (Wallenstein and Houde, 1975). In the 1970s, the use of visual analog scales grew in popularity. Evidence suggested that VAS measures are more sensitive than categorical measures of pain (Joyce, Zutski et al., 1975; Scott and Huskisson, 1976).

McGeary and colleagues (2006) evaluated the association between PI scores and socioeconomic outcomes in patients with CMSD disorders in a functional restoration program. Higher PI rates prior to treatment were associated with lower completion rates, higher incidence of depression, and increased self-report disability (McGeary, Mayer et al., 2006). In addition, higher post-treatment PI scores were linked to lower likelihood of returning and/or retaining work, higher incidence of visiting a new healthcare provider, and decreased chance of settling workers compensation cases (McGeary, Mayer et al., 2006). Results from this study indicate that the PI scale may be helpful in identifying patients at increased risk for poor treatment outcomes.

5.2.6 Beck Depression Inventory(BDI)

The BDI was developed by Beck (1967) to assess the cognitive components of depression. The scale has 21 items, and includes questions regarding sleep disturbance, sexual dysfunction, weight change, and anaerobia. The form takes approximately five minutes to complete, and scoring takes less than one minute. Low scores on the BDI represent low levels of depression. Cutoff scores have been suggested of <10 for absence of depression, 10-18 for mild depression, 19-29 for moderate depression, and >29 for severe depression (Beck, Steer et al., 1988).

Numerous studies have evaluated the psychometric properties of the BDI, and validity and reliability are consistently strong (Anagnostis, Gatchel et al., 2004). A meta-analysis indicated that test-retest reliability ranged from .60-.83 for nonpsychiatric patients. In addition, the analysis indicated an average internal consistency of .81 across studies, and an average concurrent validity with the Hamilton Depression scale of .73 and MMPI Depression Scale of .76 (Beck, Steer et al., 1988).

5.3 Purpose

The purpose of the current paper is to make the quantitative change scores in relevant HRQL measures more clinically meaningful, by relating change in HRQL to objective work outcome criteria. The proposed use of the MCIDs is to predict patients that will have good or poor outcomes, so that patients at risk for poor outcomes may be targeted for further intervention. Health outcome variables being evaluated are the ODI, SF-36, PDQ, MVAS, PI, and BDI. The following specific aims correspond to the above stated purpose:

1) In order for a MCID to be clinically applicable, the variable in question must be able to discriminate among patients that achieve poor and good outcomes. Thus, the first specific aim is to determine if the Difference in the variable of interest is predictive of the objective outcome criteria.
2) Determine if the difference in the variable of interest varies based on pre-treatment level of severity. Do patients with more severe levels of disability need to experience a greater magnitude of change to obtain good outcomes?
3) Evaluate effect sizes for each instrument associated with poor, fair, and good outcome categories. Effect sizes are frequently used to assess the success of treatment. By understanding what effect size is associated with good outcomes, clinicians may turn a quantitative measure into a clinically meaningful one.
4) Calculate the average percent difference of the variable in question for patients that fall in the poor, fair, and good outcome groups. The average percent difference for patients in the fair category will be considered the MCID. Based on the results from specific aim 2, this number may be different for subjects with varying levels of pre-treatment severity.

5.4 Hypotheses

The following hypotheses have been formulated for the current project:

1) The difference for all variables in question will be predictive of outcomes (specific aim 1).
2) The MCID will vary based on pre-treatment level of severity (specific aim 2).

3) Small effect sizes (Cohen's standard of .2) will be associated with poor outcomes, and large effect sizes (>.8) associated with good outcomes (specific aim 3).

CHAPTER 6

METHODS

6.1 Subjects

There were six separate samples of subjects selected for use in the current experiment. All samples included patients that completed a tertiary functional restoration rehabilitation program for their work-incurred injury claim. Patients were admitted to, and completed, treatment at the Productive Rehabilitation Institute of Dallas for Ergonomics (PRIDE). Various psycho-social instruments were administered at pre-treatment (PRE) and post-treatment (POST). Patients were included in the study if they had a period of more than four months partial/total disability since a work-related injury, failure of non-operative care to achieve functional recovery, surgery that had not produced resolution, and ability to speak English or Spanish. The individual samples varied based on what years individual scales were administered, the type of injury the scales apply to, and the availability of data. All samples were selected from a total of 4,191 cases spanning the years 1992-2004.

6.1.1 ODI Sample

The ODI was administered at PRIDE beginning in 1999, and was designed for use in patients with low back pain. A total of 1,042 subjects with low back injuries were

identified from 1999-2004 (see Table 2 for sample size). Of these, 829 (79.6%) completed the program. ODI and outcome data were missing on 353 of these completers, leaving 476 (56.7%) cases available for analyses. Prior to analysis, 135 cases were randomly selected and reserved for use in a cross-validation study. The total training set was 341, with 41 (12.0%) classified as having poor outcome, 86 (25.2%) as having fair outcome, and 214 (62.8%) as obtaining a good outcome (Table 3).

6.1.2 MVAS Sample

The MVAS was administered from 1993-2002, and is designed for use in patients with spinal disorders. A total of 2,527 subjects with a spinal disorder participated in PRIDE during this ten year span (see Table 2 for sample size). Of these, 2,163 (85.6%) completed the program, and MVAS and outcome data was available on 1,715 (79.3%) of these completers. A total of 528 cases were randomly selected and reserved for use in the cross validation study, leaving a total of 1,187 (7.7% poor, 28.1% fair, and 64.2% good outcome) for use in the training set (Table 3).

6.1.3 SF-36 Sample

The SF-36, which was designed for use in any health population, was administered at PRIDE starting in 1999. A total of 1,904 subjects with any type of musculoskeletal injury were identified in the PRIDE database spanning 1999-2004 (Table 2). Of these, 1,502 completed the program (78.9%). Outcome and PRE/POST SF-36 data was available on 905 of these subjects, and 275 were randomly selected and reserved for use in the cross-validation study. Of the 630 remaining in the training set, 73 (11.6%) fell

in the poor outcome group, 150 (23.8%) were categorized as fair, and 407 (64.6%) classified as good (Table 3).

6.1.4 PDQ Sample

The PDQ was developed at PRIDE for use in patients with musculoskeletal disorders, and was initiated in 2002. A total of 870 patients with any musculoskeletal injury were in the PRIDE database from 2002-2004 (Table 2). Of these, 682 completed the program (78.4%). Data was available on 395 of these cases, and 132 were reserved for the cross validation study. A total of 263 were utilized in the training set, and of these 41 (15.6%) fell in the poor outcome group, 47 (17.9%) in the fair category, and 175 (66.5%) were classified as having good outcomes (Table 3).

6.1.5 PI Sample

The PI Scale was administered from the beginning of the PRIDE program, and applies to patients with any type of injury. A total of 4,134 patients were available from the PRIDE database from 1992-2004 (Table 2). Of these, 3,488 (84.2%) completed the program. Outcome and PI data were available on 2,823 of these cases (80.9%). A total of 874 cases were randomly selected and reserved for use in cross-validation analyses, leaving a total of 1,949 for use in the training set. Approximately 9.6% (188) of these were categorized as having a poor outcome, 26.2% (510) as fair, and 64.2% (1,251) as good (Table 3).

6.1.6 BDI Sample

The BDI has been administered at PRIDE for over ten years. A total of 4,134 patients are available in the dataset from 1992-2004 (Table 2). Of these, approximately

3,488 completed the program, and outcome and BDI data are available on 2,804 (80.4%) of these cases (Table 3). Approximately 30% (828) of these cases were reserved for use in cross validation study, leaving a total of 1,976 for use in the training set. Table 3 presents the outcome classification of the training and test set samples. In the training set, (9.5%) fall in the poor outcome group, 728 (26.0%) in the fair group, 1,811 (64.6%) in the good group.

6.2 Procedure

All patients participated in an intake interview that consisted of an initial evaluation of medical history, physical examination, psychological assessment, medical case management, disability assessment, and a quantitative physical/functional capacity evaluation (Mayer and Gatchel, 1988; Brady, Mayer et al., 1994; Curtis, Mayer et al., 1994; Mayer, Gatchel et al., 1994; Mayer, Gatchel et al., 1994; Mayer, Gatchel et al., 1994; Mayer, Pope et al., 1994; Dersh, Gatchel et al., 2002). At the intake interview, patients were provided with some combination of the following psychosocial instruments, depending on the year of admission: (1) quantified pain drawing with a Visual Analog Scale (VAS) of perceived pain intensity (Capra, Mayer et al., 1985; McGeary, Mayer et al., 2006); (2) the Million Visual Analog Scale (Million, Hall et al., 1982); (3) the Oswestry (Fairbank, Couper et al., 1980), (4) the Short-Form 36 (Ware, Kosinski et al., 1994), (5) the Beck-Depression Inventory (Beck, 1967), and (6) the Pain Disability Questionnaire (Anagnostis, Gatchel et al., 2004). Patients were assessed with all psychosocial instruments at PRE and POST. In addition to psychosocial instruments, demographic information was gathered at intake interviews.

The interdisciplinary treatment program consisted of quantitatively directed physical exercise progression and a multimodal disability management program. Patients were assigned case managers and provided with some combination of individual counseling, group therapy, stress management, biofeedback, coping skills training, and education focusing on disability management, vocational reintegration and future fitness maintenance (Garcy, Mayer et al., 1996; Jordan, Mayer et al., 1998; Mayer, McMahon et al., 1998; Mayer, Gatchel et al., 1999; Wright, Mayer et al., 1999; Mayer, Gatchel et al., 2001; Mayer, Anagnostis et al., 2002).

One year following completion of treatment, case managers attempted to contact all patients for a structured telephone interview. During this interview, patients were evaluated for work and health-related outcome variables (Mayer, Gatchel et al., 2000), including return-to-work, work retention at one-year, healthcare seeking from new providers, number of visits to healthcare providers, additional surgical treatment to the injury sites, recurrent injury claims, medication usage, and case settlement. Historically, contact rates range from 90-95%, and no differences are typically detected between cases with and without missing data.

6.3 Instruments, Difference Scores, and Outcome Measures

6.3.1 Psychosocial Instruments

Four disability/quality of life measures were analyzed (the ODI, MVAS, SF-36, PDQ), in addition to one pain measure (Pain Intensity), and a measure of depression (BDI). All measures have demonstrated predictability and reliability when used with patients with musculoskeletal injuries (Gatchel, 2001; Gatchel, Mayer et al., 2006).

6.3.2 Difference Scores

Percent difference scores were calculated for the variable of interest using the following formula: $[(\mathrm{Pre} - Post)/(\mathrm{Pre} + Post)] \times 100$. In addition, for comparison to previous research studies evaluating the MCID, a raw difference score was calculated by subtracting Post-treatment scores from Pre-treatment scores. The percent difference variable was utilized in all statistical analyses.

6.3.3 Calculation of MCID

The average percent difference (pre-post) was calculated for patients that fell in the poor, fair, and good outcome category. The mean difference associated with patients that fell in the fair category was considered the MCID (specific aim 4).

6.3.4 Outcome Measure

A composite outcome variable was designed utilizing the following 1-year objective outcome variables: work outcome, post-treatment surgery to same compensable body part, and post-treatment healthcare utilization. Patients were categorized as having "poor," "fair," or "good" work outcome, "poor," "fair," or "good" surgery outcome, and "poor," or "good" health utilization. Following these categorizations, an overall 3-level (poor, fair, good) composite outcome variable was calculated.

The following criteria were utilized for categorizing patients on the work, surgery, and healthcare utilization variables:

1. Work outcome

    a. Poor: patients that had never returned to work following treatment at PRIDE, and were not involved in activities that might lead to work (such

as rental property, crafts, etc.) or participating in non-income producing activity (retirement, volunteer work, etc.).

b. Fair: patients that

-denied work but engaged in activities that were potentially income producing

-denied work but participated in a non-income producing activity, or had a comorbid condition

-returned to work, but were not working at 1 year follow-up because of new injury

-returned to work, but were not working at 1 year follow-up because of original injury.

-work return documented during year with patient working at last contact, but no information at one year

-work return documented but off work again at last contact

-patients that were not able to be contacted on work return

c. Good

-patient returned to work and continued to work 1-year following treatment

2. Surgery

a. Poor: had at least one surgery following treatment to original injury

b. Good: had no surgery following treatment to original injury

3. Healthcare utilization

a. Poor: greater than 10 additional treatments following completion of program

b. Fair: between 6-10 additional treatments following completion of program

c. Good: 0-5 additional treatments following completion of program

Due to the overwhelming importance for return to work, decision criteria for the composite outcome variable were determined based on work outcome as the primary variable, and post-treatment surgery and healthcare utilization as secondary variables (see table 1 for decision criteria). If a case was missing either of two secondary outcome criteria (surgery or healthcare utilization), work status categorization determined poor, fair, or good composite outcome. If a case was missing work status, it was dropped from the analyses (see table 2 for presentation of total lost cases within each of the 6 samples).

6.4 Design and Statistical Analyses

Statistical analyses were carried out with SPSS (version 14.0, SPSS Inc., Chicago, IL). All analyses were conducted independently on each variable in question, with the appropriate sample. Prior to analyses, individual samples were selected and identified from the total dataset of PRIDE from 1992-2004, based on availability of data (see section 3.1 for detailed description of individual samples). Approximately 30% of each sample was randomly selected and reserved as a test set for use in cross-validation, to determine how well the results would generalize to a new sample of cases. The remaining 70% was utilized as the training set.

6.4.1 Demographic Analyses

Descriptive statistics were calculated for demographic variables for the total, test, and training sample for each individual sample. Statistical comparisons were made among the test and training set to ensure the test set was representative of the training set. Within the training set, comparisons were made among the patients with poor, fair, and good outcomes on all demographic variables (age, gender, race, and length of disability). In order to identify potential demographic variables important to classification, one-way ANOVAS were run on continuous variables, followed by Fisher LSD post hoc tests. Chi-square analyses were run on categorical variables, followed by individual chi-square analyses for planned comparisons. Variables identified as significantly different among groups were utilized in subsequent regression analyses.

6.4.2 Descriptive Statistics and Analysis of Variance

The first analytical step taken to explore the relationship among outcome, pre-treatment level of severity, and difference, was an Analysis of Variance (ANOVA). A two-way ANOVA, with outcome (poor, fair, and good) and pre-treatment level of severity (mild, moderate, severe) as independent variables, and percent difference as the dependent variable, was utilized to explore differences among outcome groups and patients with varying disability levels. A significant effect for outcome indicated that the amount of pre to post change in the variable of interest varied based on outcome group. This provided the first line of evidence of whether the change in the measure was related to outcome (see specific aim 1). A significant effect for pre-treatment level of severity indicated that the magnitude of change varied based on the level of severity prior to

treatment. A significant interaction between pre-treatment level of severity and outcome was the first line of evidence that the MCID would vary based on pre-treatment level of severity (see specific aim 2). Mean percent changes associated with poor, fair, and good outcome groups provided the gold standard calculations of the MCID. The mean percent change associated with the fair outcome group was considered the MCID (specific aim 4). If a significant interaction between pre-treatment level of severity and outcome was detected, then the MCID for different severity groups was the mean percent change associated with the fair outcome in each level of severity.

All ANOVAS were conducted in SPSS, using PROC GLM and Type III sums of squares to adjust for unequal sample size. Fisher LSD post hoc tests were utilized to determine what groups differed from one another. A Bonferroni correction was utilized to adjust for the number of post hoc comparisons being made.

6.4.3 Effect Size Calculations

Effect-size calculations were made utilizing Cohen's effect size formula: $(\mathrm{X}\,\mathrm{Pr}\,e - \mathrm{X}Post)/SD\,\mathrm{Pr}\,e$. Effect size was calculated for each total sample, as well as individually for poor, fair, and good categories (specific aim 3). These calculations were an estimate of the magnitude of the pre-post change.

6.4.4 Sequential Logistic Regression

A sequential logistic regression analysis was conducted, first on a set of demographic variables, then upon the addition a set of measure-specific variables (i.e. ODI variables or MVAS variables, etc.). The demographic variables utilized were identified as variables that differed among outcome groups in the initial demographic

analyses, and included age, gender, and length of disability. The initial set of difference measures included in the regression model varied for each set of analyses, but was some combination of pre, post, percent difference, and pre x percent difference.

Correlation matrices were evaluated to assess the level of association among the predictors and Outcome, as well as among the predictors themselves. Significant association was anticipated among the measure variables, as two of the variables (difference, and difference x pre) were created from the pre and post variables. Tolerance estimates were evaluated, and a minimum of .20 was set as an acceptable tolerance level for all variables. In all cases, inclusion of all 4 variables in the regression equation led to severe multicollinearity issues (tolerance values of <.20). Thus, at least one of the four variables was dropped from the initial set of measure variables. Decision of what variable to not include was guided by results from the correlation and ANOVA analysis. For example, if ANOVA results indicated no significant interaction between pre-treatment and outcome, than the interaction term was omitted. Chi-square difference between models test was utilized to determine if addition of the set of measure variables provided unique information when combined with the demographic model. If a significant difference in the chi-square test was detected, further deletion of variables, followed by additional chi-square difference tests, was utilized to evaluate the role of the individual predictors.

Prior to analyses, the pre, post and difference variables and the outcome variable were examined through SPSS Explore for accuracy of data entry, missing values, and fit between their distributions. The interaction term was centered to minimize issues with

multicollinearity. The assumption of linearity of the logit was evaluated by computing the interaction of each predictor variable and its natural logarithm, and utilizing these computed logarithm variables as predictors in a regression model with the outcome as the dependent variable. No violations of the assumption of linearity of the logit were detected.

6.4.5 Cross-validation Calculations

If regression analysis indicated that the percent difference score was a significant predictor of outcome, the test set was utilized to assess the accuracy of the MCID derived from the training set in predicting outcomes. Classification results are reported as sensitivity and specificity.

In addition, the mean percent change for patients in the poor, fair, and good outcome groups, and effect sizes for poor, fair, and good outcome groups were calculated for the test set. Statistically, differences in the mean percent change for each outcome group were compared utilizing the Mann-Whitney test.

CHAPTER 7

DEMOGRAPHIC RESULTS

Tables detailing the results of all statistical analyses performed are included in Appendix A, and will be discussed in the RESULTS chapters (7-13). The immediate Chapter presents demographic and descriptive information for individual samples.

7.1 Population Demographics

The population demographics section is subdivided into six sections: ODI, MVAS, SF-36, PDQ, PI, and BDI samples (see Table 2 for respective sample size numbers). Each individual sample was pulled from a total of 4,191 cases of patients spanning the years 1992-2004. Prior to conducting analyses, the composite outcome variable was calculated for the total sample (see Table 3 for total number of patients categorized as poor, fair, and good success).

7.1.1 ODI Sample

Table 4 details the basic demographic variables for the training, test and total ODI sample and statistical analyses for these variables. Table 5 presents the basic demographic variables for the poor, fair, and good outcome groups within the training set, and table 6 contains the statistical analyses for each of these variables. Table 7 contains post hoc and planned comparison analyses that isolate the differences found in

the statistical analyses. The total sample size consisted of 476 subjects, 341 in the training set and 135 in the test set. No significant differences were found for any of the demographic variables between the training and test set.

Within the training set, there were a total of 341 cases: the poor outcome group a total of 41 cases; the fair outcome group 86 cases; and the good outcome group 214 cases (Table 5). A significant difference was found for age, with the poor group averaging 51.7 years, the fair group averaging 47.6 years, and the good group averaging 45.0 years, $F_{2, 338} = 9.61$, $p < .001$ (Table 5 and 6). Post hoc analyses revealed that the poor and good groups significantly differed, $p < .001$ (CI 2.9, 10.5) (Table 7). Length of disability (LOD) was also found to differ significantly among groups, $F_{2, 337} = 3.31$, $p < .05$ (Table 6). Post hoc analyses, however, were not significant for any of the groups (Table 7).

7.1.2 MVAS Sample

Basic demographic variables for the test and training set in the MVAS sample are presented in Table 8. The demographic variables for the poor, fair and good outcome groups are presented in Table 9, with statistical analyses for these variables presented in Table 10. Post hoc analyses for these statistical analyses are presented in Table 11. The total sample size consisted of 1,715 subjects, 1,187 in the training set and 528 in the test set (Table 8). No significant differences were found for any of the demographic variables between the training and test set (Table 8).

Within the training set, the poor outcome group included a total of 91 cases; the fair outcome group 334 cases; and the good outcome group 762 cases (Table 9). A significant difference was found for age, with the poor group averaging 49.3 years, the

fair group averaging 45.2 years, and the good group averaging 42.9 years, $F_{2, 1184}$= 20.13, p < .001 (Table 10). Post hoc analyses revealed that the poor was different from fair, p < .001 (CI 1.4, 6.8), as well as different from good, p < .001 (CI , 3.8, 8.9). In addition, good was different from fair p < .001 (CI -3.8, -0.8, Table 11). Gender was found to differ significantly among groups, χ^2 (2) = 8.89, p <.001, with the poor outcome group averaging 57.1% male, the fair group averaging 54.2 % male, and the good group averaging 63.5% male (Table 9 and 10). Planned comparison indicated that the good group differed significantly from the poor group, χ^2 (1) = 8.46, p <.01 (Table 11). Length of Disability (LOD) was also found to differ significantly among groups $F_{2, 1184}$= 5.04, p < .01 (Table 10). Post hoc analyses indicated the poor group differed from the good group, p <.01 (CI 1.5, 2.7, Table 11).

7.1.3 SF-36 Sample

Table 12 presents the basic demographic variables for the total, training and test set in the SF-36 sample. The demographic variables for the poor, fair and good outcome groups are presented in Table 13, followed by the statistical analyses for these variables in Table 14. Post hoc analyses for these statistical analyses are presented in Table 15. The total sample size consisted of 905 subjects, 630 in the training set and 275 in the test set (Table 12). No significant differences were found for any of the demographic variables between the training and test set.

Of the 630 patients in the training set, 73 were classified as having a poor outcome, 150 as having a fair outcome, and 407 as having a good outcome (Table 13). A significant difference was found for Age $F_{2, 627}$= 14.78, p < .001 (Table 14). Post hoc

analyses revealed that the poor was different from the fair ($p < .05$), and good, ($p < .01$), and the good was different from the fair ($p < .05$, Table 15). Gender was found to differ significantly among groups, χ^2 (2) = 7.5, $p < .05$, with the poor outcome group averaging 42.5% male, the fair group averaging 46 % male, and the good group averaging 56% male (Table 13 and 14). Planned comparison indicated that the good group differed significantly from the poor group, χ^2 (1) = 4.58, $p < .05$ and the good group differed from the fair group , χ^2 (1) = 4.42, $p < .05$ (Table 15). Length of disability (LOD) was also found to differ significantly among groups $F_{2, 626} = 5.32$, $p < .01$ (Table 14). However, post hoc analyses indicated no differences of interest.

7.1.4 PDQ Sample

Table 16 presents demographic variables for the training, test, and total PDQ sample. The total sample size consisted of 395 subjects, 263 in the training set and 132 in the test set. No significant differences were found for any of the demographic variables between the training and test set. The demographic variables for the poor, fair and good outcome groups are presented in Table 17, statistical analyses for these variables in Table 18, and appropriate post hocs for these statistical analyses are presented in Table 19.

Of the 263 patients in the training set, 41 were classified as having a poor outcome, 47 as having a fair outcome, and 175 as having a good outcome (Table 17). A significant difference was found for age $F_{2, 262} = 10.93$, $p < .001$, with the poor group averaging 52.4 years, the fair group 48.6, and the good group 44.9 (Table 18). Post hoc analyses revealed that the poor group was significantly different from the good group, $p < .01$ (Table 18). In addition, LOD was found to differ significantly among groups $F_{2, 262} =$

5.24, $p < .01$ (Table 18). Post hoc analyses indicated the poor group, averaging 27.4 months was significantly different from the good group, which averaged 17.6 months, $p < .01$ (CI 2.0, 17.6. Table 19).

7.1.5 PI Sample

The total PI sample consisted of 2,823 cases, with the training set totaling 1,949 and the test set 874. Demographic variables are presented in Table 20 for the training, test, and total sample. No significant differences were found for any of the demographic variables between the training and test set. The demographic variables for the Poor, Fair and Good Outcome groups are presented in Table 21. The statistical analyses for these variables are presented in Table 22, and are followed by the post hoc results in Table 23.

As presented in Table 21, 188 of the training set were categorized as having a poor outcome, 510 as fair, and 1,251 as good. Consistent with the other samples, a significant difference was found for age $F_{2, 1948}= 32.98$, $p < .001$, with the poor group averaging 49.2 years, the fair group averaging 45.6 years, and the good group averaging 43.4 years (Table 22). Post hoc analyses revealed that the poor and fair ($p < .001$), poor and good ($p < .001$), and fair and good ($p < .001$) categories all differed from one another (Table 23). Gender was also found to significantly differ among groups $\chi^2 (2) = 7.22$, $p <.05$, with the poor outcome group averaging 49.5% male, the fair group averaging 53.1 % male, and the good group averaging 58.1% male (Table 22). Individual comparisons indicated that the poor and good groups differed significantly $\chi^2 (1) = 4.98$, $p <.05$ (Table 23). In addition, LOD was found to differ significantly among groups $F_{2, 1948}= 32.98$, $p < .001$, with the poor group averaging 21.6 months of disability, the fair group averaging

17.1 months, and the good group 15 months. Post hoc analyses indicated the poor and fair ($p < .05$), and poor and good groups ($p < .001$) significantly differed from one another (Table 21 and 22).

7.1.6 BDI Sample

Table 24 presents demographic information for patients in the training, test, and total BDI sample. 1,976 cases make up the training set, 828 the test set, for a total of 2,804 cases. No significant differences were found for any of the demographic variables between the training and test set. The demographic variables for the poor, fair and good outcome groups are presented in Table 25. The statistical analyses for these variables are presented in Table 26, and are followed by the post hoc results in Table 27.

Out of a total of 1,976 in the training set, 181 were classified as having poor outcome, 517 as having fair outcome, and 1278 as having a good outcome (Table 25). Again, a significant difference was found for age $F_{2, 1975} = 40.45$, $p < .001$, with the poor group averaging 49.6 years, the fair group averaging 45.1 years, and the good group averaging 43.0 years (Table 25 and 26). Post hoc analyses revealed that the poor and fair ($p < .001$), poor and good ($p < .001$), and fair and good ($p < .001$) categories all differed from one another (Table 27). Gender was also found to significantly differ among groups $\chi^2 (2) = 12.79$, $p < .01$, with the poor outcome group averaging 48.1% male, the fair group averaging 51.6 % male, and the good group averaging 58.8% male (Table 26 and 27). Individual comparisons indicated that the poor and good groups differed significantly $\chi^2 (1) = 7.53$, $p < .01$, as well as the good and fair groups $\chi^2 (1) = 7.77$, $p < .01$ (Table 27). In addition, LOD was found to differ significantly among groups $F_{2, 1975} = 13.5$, $p < .001$,

with the poor group averaging 22.8 months of disability, the fair group averaging 16.1 months, and the good group 14.8 months (Table 26 and 27). Post hoc analyses indicated the poor and fair ($p < .01$), and poor and good groups ($p < .001$) significantly differed from one another (Table 27).

7.2 Summary of Demographic Information for Individual Samples.

Overall, there were no statistical differences in demographic variables among the test and training sets for any of the samples, ensuring that the cross-validation analyses will be conducted on a representative sub-set of each total sample. Analyses of the individual training samples indicate that age and LOD were consistently different among outcome groups for all samples, with older mean age and lengths of disability seen in the poor outcome group as compared to the good outcome group. In addition, gender was significantly different in the MVAS, SF-36, PI, and BDI samples, with more males in good outcome groups as compared to the poor outcome groups. Based on these results, gender, age, and LOD have been identified as possible demographic variables that are associated with outcome. All three demographic variables will be incorporated into subsequent regression analyses.

CHAPTER 8

ODI RESULTS

The results will be presented in 6 chapters, one for each of the primary variables of interest, in the following order: ODI, MVAS, SF-36, PDQ, PI, and BDI. The current chapter presents the ODI results.

Each set of analyses will be presented in 4 sections:

1) Descriptive Statistics and ANOVA (Specific Aim 1, 2, and 4)

2) Effect-size calculations (Specific Aim 3)

3) Sequential logistic regression analyses to assess predictability of Percent Difference on outcome (Specific Aim 1, 2).

4) Summary of measure-specific results.

8.1 Descriptive Statistics and ANOVA Oswestry Results

To assess group differences in ODI Change Scores, a two-way ANOVA with Outcome group (poor, fair, and good) and pre-treatment level of severity (mild, moderate, and severe) as independent variables, and ODI difference score as the dependent variable was run.

Table 28 presents the mean percent change for patients falling in the poor, fair, and good outcome groups. Patients classified as having poor outcome averaged a percent

difference of 18.92±18.5, those in the fair outcome group 22.41±24.42, and the good outcome group averaged a percent difference of 26.12±27.93. The main effect for outcome was not significant, however ($F_{2, 332} = 1.595$, $p = .204$), indicating that the observed difference in percent change in ODI did not statistically vary among outcome groups. This was the first line of evidence that the pre to post difference in ODI would not be a good predictor of outcome (Specific Aim 1).

Patients that fell in the lowest degree of pre-treatment severity category (mild) experienced the smallest percent change 19.99±27.21 as compared to those in the moderate category 24.0±25.43, and the severe category (29.47±25.05). A main effect for pre-treatment level of severity, $F_{2, 332} = 24.135$, $p < .001$, indicated that magnitude of change for patients with mild, moderate, and severe levels of disability varied statistically. As expected, patients with more severe levels of disability had worse outcomes. The interaction between outcome and pre-treatment level of severity was not significant, however, suggesting that the amount of change in ODI for patients with different outcomes did not vary based on pre-treatment level of severity $F_{4, 332} = 1.620$, $p = .169$. Based on these results, it was concluded that the MCID would be the same for patients of varying degrees of pre-treatment severity (Specific Aim 2). Utilizing the gold standard MCID approach, with fair outcome considered the minimum acceptable outcome, the MCID of the ODI would be 22.41.

8.2 Oswestry Effect Size Calculations

Effect size calculations for each of the 3 outcome groups are presented in Table 29, and all were considered large as defined by Cohen's .8 criteria (poor = .86, fair = .94,

and good = 1.0). Ideally, a questionnaire would have small or negative effect size for patients that did not change, or had poor outcomes. Consistent with initial impressions from descriptive data, these results indicate that the ODI questionnaire was not responsive for patients that had poor outcomes. If the change in ODI was predictive of outcomes, than in a similar population, a treatment study would need to obtain an effect size of at least 1.0 to coincide with good objective 1-year outcomes.

8.3 Oswestry Regression Analysis

To explore the relationship among both the predictors and criterion, and among the predictors themselves, correlational analyses were conducted prior to regression analysis. Table 30 presents the correlations (Spearman Rho) between the ODI variables (pre, post, difference, and pre x difference) and the outcome, and 31 presents the correlations among the ODI predictor variables. The ODI difference was not significantly correlated with outcome ($r = .033$, $p = .541$). However, the pre-treatment ODI scores ($r = -.115$, $p = .009$), and post-treatment scores ($r = -.181$, $p < .001$) were both negatively related to the outcome, indicating that more severe levels of disability were related to worse outcomes. The lack of relationship between ODI percent difference and outcome negates the use of a MCID percent change score as a stand-alone predictive criterion for patients that are likely to have poor outcomes. Despite this fact, logistic regression analyses were conducted to determine whether pre or post ODI scores, which were significantly correlated with outcome, were predictive of outcomes. Initial tolerance estimates indicated severe multicollinearity with inclusion of all four ODI variables in the

model, thus based on ANOVA and correlational analyses, only pre and post ODI variables were included in the model.

There was a good model fit using the demographic variables alone, χ^2 (6, N=341) = 26.33, $p < .001$ (Table 32). Following addition of pre and post ODI variables, χ^2 (10, N=341) = 37.407, $p < .001$, Nagelkerke R^2 = 12.5. Addition of the ODI variables significantly improved the fit of the demographic only model, χ^2 (4, N=341) = 11.077, p = .0257. Tolerance estimates indicated multiciollinearity was not an issue (Table 31a), thus the significance of individual models was utilized to select three insignificant variables for removal from the model (ODI pre, gender, and LOD). The reduced model resulted in χ^2 (4, N=341) = 29.215, $p < .001$, and the difference between the models was not significant, χ^2 (6, N=341) = 8.192, p = .2244. Thus, age and post ODI were retained for the final model.

Regression coefficients, chi-square test, and odds ratios for significant coefficients are presented in Tables 33-34. Age varied significantly between the poor and good outcome groups, with people in the poor outcome group 1.08 times more likely to be older than patients in the good group, $p < .001$ (Table 33). In addition, patients in the poor group were 1.03 times more likely to have a more severe level of disability following treatment (p = .003). In addition, patients in the fair group were 1.03 times more likely to be older than patients in the good group (Table 34).

Despite a significant effect in the prediction of the model, overall classification was not impressive, with 7.3 % of the poor, 0% of the fair, and 99.5 % of the good cases

being correctly classified, for a total of 63.3% of cases being accurately predicted. Cases were grossly over-classified as good.

Following consideration of the lack of correlation between ODI difference and outcome, poor classification, and large confidence intervals surrounding the average change in patients with poor, fair, and good outcomes, change in ODI does not appear to provide adequate information to utilize a clinically important amount of change for individual classification purposes. Post ODI scores, however, are significant predictors of outcomes when combined with age.

8.4 Oswestry Cross-Validation

In order to evaluate the replication of our results, mean percent change in ODI and relevant effect size statistics were calculated in a cross-validation study. Mean percent change for the ODI for both the training and test set are presented in Table 35. As indicated by Mann-Whitney tests, no significant differences were found in mean ODI percent difference between samples. Table 36 presents the effect size calculations for the training and test set. Classification results were not calculated for the MCID in the cross-validation, due to the conclusion that the ODI percent difference would not provide adequate information on an independent basis to predict outcome.

8.5 Summary of Oswestry Results

Table 37 presents summary statistics for the Oswestry analysis. The average raw change for patients in the fair category was 12.87±15.21, and the average percent difference was 22.41±24.42. A student's paired t-test indicated a significant amount of pre-post change on the ODI (t_{340} =17.099, $p < .001$), suggesting that the measure is

sensitive to change in the current population. In addition, an effect size for this change of .96 indicates that the amount of change is of a large magnitude. These two statistics provide evidence that the measure provides adequate power to detect group differences. Despite this fact, the amount of change is not *predictive* of outcome for individual patients. The lack of relationship among the ODI difference score, and the outcome variable, negates the use of the ODI MCID as a method of identifying patients that are likely to have poor outcomes in the current population. The significance of the post ODI score, combined with age, in predicting outcome indicates that classification of patients into high risk categories based on age and post ODI may provide useful in identifying patients at risk for poor outcomes.

CHAPTER 9

MVAS RESULTS

9.1 Descriptive and ANOVA Million Results

Descriptive statistics for gold standard derived percent differences are presented in Table 38. A 2-way between subjects ANOVA, with outcome and pre-treatment severity as independent variables, and percent difference as dependent variable, indicated that the MVAS percent difference varied among outcome groups ($F_{2, 1178} = 14.80$, $p < .001$). Patients classified as having poor outcome averaged a percent difference of 13.76±20.76, whereas the fair outcome group averaged a difference of 18.34±24.46, and the good outcome group a difference of 24.67±26.81. This provided an indication that the percent difference on MVAS may be an adequate individual predictor of outcome results.

In addition, a main effect for pre-treatment severity ($F_{2, 1178} = 5.87$, $p = .003$) indicated that percent difference varied based on severity of MVAS prior to treatment. Patients scoring the lowest at pre-treatment (mild group) averaged 17.76±29.96, those in the mid-range at pre-treatment (moderate) averaged 22.99±23.93, whereas those scoring the highest at pre-treatment (severe) averaged a change of 25.32±23.13 (Table 38). No significant interaction between outcome and pre-treatment level of severity was detected,

$F_{4, 1178} = .360$, $p = .837$. Based on these results, it was concluded that the MCID would be the same for patients in all pre-treatment categories.

Utilizing the gold standard MCID approach, with fair outcome considered the minimum acceptable outcome, the MCID (percent change) of the MVAS would be 18.34.

9.2 Million Effect Size Results

Table 39 presents effect size calculations computed separately for patients in the poor, fair, and good outcome groups. All patients experienced a large magnitude of change, irrespective of outcomes. Patients in the poor category had an effect size of 1.17, the fair category 1.03, and the good category 1.45. Ideally a measure will have a negative or small effect size for patients that do not experience a good outcome. This was not the case in our current study, suggesting that with respect to the composite outcome criteria and percent change, the Million was not a responsive measure. If percent change in MVAS is predictive of outcomes, than in a study in a comparable population, an effect size of at least 1.45 is necessary to obtain good objective outcomes.

9.3 Million Regression Analysis

Correlational analyses between the outcome and MVAS predictors indicated a small, but significant, positive relationship between MVAS percent difference and outcome ($r = .135$, $p < .001$). A small negative relationship was also detected between pre MVAS scores and outcome, suggesting that patients that are more severely impaired were more likely to have negative outcomes ($r = -.117$, $p < .001$). Post MVAS scores were negatively correlated with outcome ($r = -.203$, $p < .001$), indicating that patients with high MVAS following treatment may be at risk for poor outcomes. Strong

relationships among predictors were detected (see Table 40). To decrease the amount of multicollinearity, the interaction term was not included in the model based on the ANOVA analyses indicating no significant interaction effect. Following deletion of this variable, tolerance estimates were all above .20, thus pre MVAS, post MVAS, and MVAS were all included in the model, and decisions based on what predictors to eliminate were based on significance of each individual predictor.

There was a good model fit using the demographic variables alone, χ^2 (6, N=1187) = 50.048, $p < .001$ (Table 42). Following addition of the three MVAS variables, χ^2 (12, N=1187) = 102.318, $p < .001$, Nagelkerke R^2 = 10.2. Addition of the MVAS variables significantly improved the fit of the demographic only model, χ^2 (6, N=1187) = 52.27, $p < .001$.

To further evaluate the role of the individual predictors in the model, the predictors with the 2 highest p values were dropped (post MVAS, and gender), and the resulting chi-square tested against the full model. Following deletion of the two variables, χ^2 (8, N=1187) = 94.691, $p < .001$. The difference between the two models was not significant, χ^2 (4, N=1187) = 7.627, p = .106, indicating that prediction was not significantly decreased by dropping the two MVAS variables. To determine if any additional variables were providing redundant information, the variable with the highest p value was dropped from the model (LOD), and compared against the model with pre MVAS, MVAS difference, age, and LOD. Following deletion of LOD, χ^2 (6, N=1187) = 91.2, $p < .001$. The difference between the reduced and the previous model was not statistically significant, χ^2 (2, N=1187) = 3.491, p = .1746, indicating that LOD was not

adding to the prediction of the model. In the model with MVAS, MVAS difference, and age, all three variables were statistically significant, thus all were retained for the final model (see Table 42).

Overall classification was not impressive, with 0.3 % of the poor, 0 % of the fair, and 99.7 % of the good cases being correctly classified, for a total of 64.1% of cases being accurately predicted. In both models, cases were overwhelmingly classified as good.

Tables 43-44 present the regression coefficients, chi-square test, and odds ratios for significant coefficients. Patients in both the poor and the fair group were more likely to be older (1.06, $p < .001$, and 1.02, $p = .002$, respectively), and have more moderate levels of pre-treatment disability (1.03, $p < .001$, and 1.01, $p = .009$, respectively), than patients in the good group. In addition, people in both the poor and fair group were approximately 1 times more likely to have a smaller degree of change as compared to patients in the good group ($p < . 001$).

Overall, pre-treatment MVAS scores, percent difference scores, and age all provide significant information for predicting one year outcomes. By combining these variables to identify patients that are in the highest risk category, clinicians may utilize these criteria to select patients for a higher level of care.

9.4 Million Cross-Validation Results

In order to evaluate the replication of our results, mean percent change in MVAS and relevant effect size statistics were calculated in a cross-validation study. Table 45 presents the mean percent change for the MVAS for both the training and test set. A

difference approaching significance following correction for number of comparisons (p = .013) indicated that the mean percent change was different between training and test set for the poor sample. This is not surprising, given the large degree of variability in individual percent change, and the smaller sample size in the test set. The average change for the fair group, which is considered the MCID, is not statistically different, suggesting that this finding is stable. Table 46 presents the effect size calculations for the training and test set.

The MCID derived from the anchor-based approach in the training set was applied to the test to set to explore the accuracy of classification. Fair categories were collapsed with poor categories for ease of classification and interpretation. Classification results are presented in Table 47. If the MCID of 19 was used, based on the "fair" outcome group in the gold-standard approach, sensitivity is 78.3 and specificity is 52.5. In other words, out of patients predicted to have good outcomes, 78.3% actually have good outcomes, and 21.7% actually end up with a poor outcome. Of patients predicted to have a poor outcome, 47.5% actually do, and 52.5% do not. By increasing the MCID to 24% change, sensitivity decreases slightly to 76%, and specificity increases to 55.4%. Through cost-utility analysis, further information may be gained, as to whether application of such criteria will prove beneficial. Based on results from the regression analyses, classification may be improved somewhat if pre-treatment level of severity, and age are both taken into consideration when developing classification criteria.

9.5 Summary of Million Results

Table 48 presents relevant summary statistics for the MVAS analyses. The mean raw change for patients that had a fair outcome was 24.39±47.16, and the mean percent difference was 18.34±24.46. A student's paired t-test indicated a significant amount of pre-post change on the MVAS ($t_{1186} = 27.468$, $p < .001$), suggesting that the measure is *sensitive* to change in the current population. In addition, an effect size for this change of 1.3 indicates that the amount of change is of a large magnitude. These two statistics provide evidence that the measure provides adequate power to detect group differences. In addition, the current analyses indicate the percent difference, combined with pre-treatment level of severity and gender, is predictive of outcomes. More specifically, utilizing a MCID of 19, approximately ½ (47.5) of patients in the test set that ended up with a poor outcome were correctly identified. In a treatment situation, this subset could be targeted for further intervention. On the other hand, utilizing the above MCID, approximately 52.5% percent of patients would be incorrectly targeted and provided additional unnecessary treatment. Future cost-utility analyses will provide useful information as to whether utilizing such cut-offs will prove clinically useful. The current analyses of the MVAS shows promise in the area of responsiveness, however.

CHAPTER 10

SF-36 RESULTS

The SF-36 describes both the physical and mental components of health. There are 8 primary scales that form distinct physical and mental components, derived from factor analysis. Two component summary scales, the Physical Component (PHS) and Mental Component Summary (MHS) were created based on the factor analysis of correlations among the eight SF-36 scales (Ware, Kosinski et al. 1994). Both the PHS and MHS utilize a T score transformation so that they have a mean of 50 and a standard deviation of 10 in a normal population, with higher scores indicating better health.

Results are presented separately for PHS and MHS scores (section 10.1 and 10.2 respectively). Scores on these scales are inversely related with disability levels, thus high scores are associated with lower pain and disability. For ease of interpretation, percent difference scores were multiplied by -1, so that a positive difference score corresponds to an overall improvement in condition.

Following exploration of the data set, one extreme PHS percent difference outlier, and one extreme MHS percent difference outlier (both more than twice the magnitude of change as compared to any other case) were detected and identified as a data entry errors. In addition, one extreme (more than 40 times any other case) MHS post score was

identified and determined to be a data entry error. These 3 cases were dropped from all analyses, leaving a total of 627 cases.

10.1 Physical Health Component Score

10.1.1 Descriptive and ANOVA Physical Health Component Score Results

Table 49 presents the mean percent change in SF-36 PHS score for patients classified as having poor, fair, and good success. Percent change in SF-36 scores were not found to vary among outcome groups ($F_{2, 618} = 2.86$, $p = .05$), although this effect approached significance. Patients classified as having poor outcome averaged a percent difference of 7.69±15.96, whereas the fair outcome group averaged a difference of 7.96±13.37, and the good outcome group a difference of 8.34±11.62.

A main effect for pre-treatment severity ($F_{2, 618} = 60.94$, $p < .001$) was detected, however, indicating that percent difference varied based on severity of SF-36 PHS prior to treatment. Patients in the mild group averaged 15.49±12.34, those in the moderate group averaged 8.48±10.77, whereas those scoring the highest at pre-treatment (severe) averaged a change of 0.72±10.46. No significant interaction between outcome and pre-treatment level of severity was detected, $F_{4, 618} = 1.90$, $p = .10$, thus the MCID percent difference score was considered to be the same for all groups of pre-treatment severity.

Utilizing the gold standard MCID approach, with fair outcome considered the minimum acceptable outcome, the MCID percent change of the PHS on the SF-36 would be 7.96.

10.1.2 Physical Health Summary Component Results

Effect size calculations were computed separately for patients in the poor, fair, and good outcome groups, and are presented in Table 50. All effect sizes were considered large by Cohen's standards (poor = .84, fair = .90, good = 1.02), for a total effect size of .96. The effect size for patients that had poor outcomes was still positive, and large by Cohen's standards, evidence that the percent difference in PHS was not indicative of outcome groups.

10.1.3 Physical Health Summary Component Score Regression Analysis

Correlational analyses indicated that pre-treatment scores ($r = .150$, $p < .001$) and post-treatment scores were positively related to outcome ($r = .136$, $p < .001$), suggesting that patients with more physical impairment prior to treatment had worse outcomes (Table 51). No significant correlation was detected between percent difference and the outcome ($r = .032$, $p = .429$), indicating that percent difference on the Physical Health Summary Score would not be useful as an MCID for predicting outcomes on an individual basis. Despite this fact, logistic regression analysis was conducted to determine if pre or post PHS scores were predictive of outcome.

There was a good model fit using the demographic variables alone, χ^2 (6, N=627) = 40.113, $p < .001$ (Table 52). Following addition of the 2 PHS variables, χ^2 (10, N=627) = 55.077, $p < .001$, Nagelkerke $R^2 = 10.2$. Addition of the SF-36 variables significantly improved the fit of the demographic only model, χ^2 (4, N=627) = 14.964, $p = .005$. Tolerance estimates were all greater than .20, thus further decisions about which predictors to drop from the model were made with regards to the statistical value of the

individual predictors. The two variables with the largest p values were dropped (LOD and Post PHS), and the analysis re-run. Deletion of these two variables resulted in χ^2 (6, N=627) = 45.847, $p < .001$, and the difference between the models approached significance χ^2 (4, N=627) = 9.23, $p = .056$. The decision was made to eliminate both variables from the model for simplicity, resulting in a final model with age, gender, and pre PHS score.

Classification was not overly impressive, with only 2.7% of the poor, 0% of the fair, and 99.8% of the good accurately classified, for a total of 64.6% percent of cases accurately predicted. Overall, cases were overwhelmingly classified as good.

Table 53 presents the individual contribution of each of the variables in the final reduced model. Age and pre PHS scores were the only significant predictors of outcome, following alpha correction for number of predictors, $p < .01$. Tables 54-55 present the regression coefficients, chi-square test, and odds ratios for significant coefficients. Age varied significantly between the poor and good outcome groups, with people in the poor outcome group 1.04 times more likely to be older than patients in the good group, $p < .001$ (Table 54). In addition, patients classified as having a good outcome were 0.95 times more likely to have a mild level of pre-treatment severity as compared to those in the poor outcome group. Similar odds for patients in the fair as compared to the good group were also indicated, with patients in the fair group 1.03 times more likely to be older ($p = .011$), and 0.94 times more likely to have a more moderate level of pre-treatment severity ($p = .013$; Table 55).

In sum, although percent difference was not a significant predictor of outcome, results suggest that in combination with age, PHS pre-treatment level of severity is an important factor to consider when identifying patients that are at risk for poor outcomes.

10.1.4 Physical Health Summary Cross-Validation Results

Mean percent change in PHS, and relevant effect size statistics were calculated in a cross-validation study to explore the reliability of findings from the training set. Table 56 presents the mean percent change for the PHS for both the training and test set. No significant differences were detected between samples. Table 57 presents the effect size calculations for the training and test set. Due to the regression and correlation analyses, which indicated the PHS percent difference score was not independently predictive of outcome, classification results were not calculated in the current sample.

10.1.5 Summary of Physical Health Summary Results

Table 48 presents relevant summary values for MCID evaluation of PHS score for the SF-36. The MCID for the raw difference was 5.41, and the MCID for the percent difference was 7.96. Analyses did not support the application of an MCID percent difference score to identify patients at risk for poor outcomes. However, pre-treatment level of severity in combination with age was predictive of outcome. Important information may be gained by developing predictive categories utilizing these criteria, and evaluating the accuracy of classification upon application of these data. Future research should target this set of variables to develop predictive categories to enable clinicians to identify patients that are at risk for poor outcomes.

10.2 Mental Health Component Score Results

10.2.1 Descriptive and ANOVA Mental Health Summary Results

Descriptive statistics for gold standard derived percent differences are presented in Table 59. MHS SF-36 percent difference did not significantly vary among outcome groups ($F_{2, 618} = 2.247$, $p = .107$), indicating the percent difference would not likely provide adequate information to classify patients into poor and good outcome groups.

A main effect for pre-treatment severity indicated that percent difference varied based on score of MHS SF-36 prior to treatment ($F_{2, 618} = 97.50$, $p < .001$). As expected, patients more severely impaired experienced less change than those mildly impaired. Patients in the mild group averaged 18.56±12.51, those in the moderate group averaged 8.89±9.27, whereas those in the severe group averaged a change of -0.63±9.98. No significant interaction between outcome and pre-treatment level of severity was detected, $F_{4, 618} = 1.25$, $p = .287$, thus the percent change calculated utilizing the gold standard approach was the same for all three levels of pre-treatment severity.

Utilizing the gold standard MCID approach, with fair Outcome considered the minimum acceptable outcome, the MCID of the SF-36 would be 7.95 percent change.

10.2.2 Mental Health Summary Effect Size Calculations

Effect size calculations computed separately for patients in the poor, fair, and good outcome groups are presented in Table 60. The total effect size was .80, and was medium for the poor group (.69), and large for the fair (.74) and good (.84) groups. If the percent difference in MHS score was predictive of individual outcome status, than in a

comparable sample, a treatment effect size of at least .84 would be indicative of good outcomes.

10.2.3 Mental Health Summary Regression Analyses

Correlational analyses indicated that the only significant correlation between MHS variables and the outcome was post MHS scores ($r = .086$, $p = .031$, see Table 61). Thus, use of the MHS percent difference score as a predictor of outcome was not anticipated to provide adequate information to be useful. Regression analysis was run to evaluate the issue more closely. Initial tolerance estimates indicated at least one variable needed to be dropped from the model to obtain a tolerance of greater than 0.20 for all variables. Based on the results from the correlational analysis and ANOVA, the interaction between pre-treatment level of severity and percent difference was dropped.

There was a good model fit using the demographic variables alone, χ^2 (6, N=618) = 40.113, $p < .001$ (Table 63). Following addition of the two SF-36 variables, χ^2 (10, N=618) = 47.065, $p < .001$, Nagelkerke $R^2 = 8.8$. Addition of the SF-36 variables did not significantly improve the fit of the demographic only model, χ^2 (4, N=618) =6.952, p = .1384. These results indicate that the MHS summary score does not provide useful information in prediction of outcome status. Upon deletion of the two highest demographic predictors from the model, χ^2 (2, N=618) =29.563, $p <.001$, and the difference between the reduced demographic model and full demographic model was statistically significant, χ^2 (4, N=618) =10.55, p = .032. A follow-up model was run with only gender and age, and χ^2 (4, N=618) =35.895, $p < .001$. The difference between the two variable demographic model and the full demographic model was not statistically

significant, χ^2 (2, N=618) =4.285, p = .121. Thus, the final reduced model included age and gender (Table 64).

Classification with the final reduced model heavily favored good outcome classification (100% accurately classified), as compared to poor (1.4%) and fair (0.7%), for a total of 64.7% of cases accurately classified.

The contribution of each demographic predictor in the model is presented in Table 64, followed by beta weights, chi-square results, and odds ratios for individual group comparisons in Table 65-66. Patients classified as poor were 1.07 times more likely to be older than those classified as good, $p < .001$, whereas patients classified as fair were 1.03 times more likely to be older than those with good outcomes. Following correction for number of variables in the model (p = .025), gender was not quite statistically different among poor and good, or fair and good groups (p = .051) (Table 65).

10.2.4 Mental Health Summary Cross-Validation Results

Mean percent change for patients in the various outcome groups were calculated in the test to validate the results calculated in the training set. No significant differences were observed between training and test set (Table 66). Effect sizes for training and test set are presented in Table 67. Percent change in MHS SF-36 is not an adequate predictor of outcome status, thus classification results utilizing the MCID in the training test were not conducted.

10.2.5 Summary of Mental Health Summary Results

Table 68 presents summary statistics for the MHS MCID. Overall, the Mental Health Summary Component Score does not provide useful information for the classification of individual patients into risk categories for poor outcomes. There was an overall lack of correlation among the individual MHS variable and outcome, and this was confirmed by regression analysis. Age and gender may provide useful classification for patients at risk for poor outcomes.

CHAPTER 11

PDQ RESULTS

11.1 Descriptive and ANOVA Results for the Pain Disability Questionnaire

Descriptive statistics for gold standard derived percent differences are presented in Table 69. A 2-way between subjects ANOVA, with outcome and pre-treatment severity as independent variables, and percent difference as dependent variable, indicated that the PDQ percent difference did not vary among outcome groups ($F_{2, 254} = 1.456$, $p = .235$). Patients classified as having poor outcome averaged a percent difference of 20.59±20.21, whereas the fair outcome group averaged a difference of 26.0±22.14, and the good outcome group a difference of 27.41±25.32.

No main effect for pre-treatment severity ($F_{2, 254} = 0.121$, $p = .887$), indicated that percent difference did not vary based on PDQ score prior to treatment. In addition, no significant interaction between outcome and pre-treatment level of severity was detected, $F_{4, 254} = 1.914$, $p = .109$. The average percent difference for the fair outcome group would be the MCID for all levels of pre-treatment severity.

Utilizing the gold standard MCID approach, with fair outcome considered the minimum acceptable outcome, the MCID of the PDQ would be 26.0.

11.2 Pain Disability Questionnaire Effect Size Results

Effect size calculations were computed separately for patients in the poor, fair, and good outcome groups, and are presented in Table 70. All effect sizes were considered large by Cohen's standards (poor = 1.36, fair = 1.74, good = 1.44), for a total effect size of 1.46.

11.3 Pain Disability Questionnaire Regression Analysis

Correlational analyses were conducted on PDQ variables and outcome, as well as among all PDQ variables prior to regression analysis. Pre and post PDQ were significantly correlated with outcome, ($r = -.156$, $p < .001$, and $r = -1.97$, $p < .001$, respectively), suggesting that these variable may play an important role in prediction of outcome (Table 71). Not surprisingly, several of the PDQ variables were significantly correlated as well. Sequential regression analysis was conducted in an effort to explore the relationship among PDQ variables and outcome. Evaluation of tolerance levels with all four PDQ variables indicated severe violations of multicollinearity. Based on the results from the correlational and ANOVA, pre and post PDQ were included in the model and percent difference and the interaction term were dropped.

Overall, there was a good model fit using the demographic variables alone, χ^2 (6, N=263) = 30.013, $p < .001$ (Table 73). Following addition of the two PDQ variables, χ^2 (10, N=263) = 39.695, $p < .001$, Nagelkerke R^2 =17.0. Addition of the PDQ variables did significantly improve the fit of the demographic only model, χ^2 (4, N=263) =9.682, p = .046. Tolerance estimates indicated all variables met the greater than .20 criteria, thus the

variables with the largest p values were dropped from the model. Following deletion of gender and pre PDQ, χ^2 (6, N=263) = 29.461, p < .001. Deletion of both variables significantly reduced the prediction of the model, χ^2 (4, N=263) = 9.682 p = .046. Upon addition of LOD back in the model, χ^2 (6, N=263) = 34.280, p < .001, and the difference between the model of age, LOD, and post PDQ, was not statistically different from the model with age, LOD, pre PDQ, and post PDQ, χ^2 (4, N=263) = 5.415, p = .273. Thus age, LOD, and post PDQ were retained in the final model.

In the demographic only model, approximately 9.8% of the poor, 0% of the fair, and 9.8% of the good cases were accurately predicted. Approximately 12.2% of the poor, 0% of the fair, and 97.7% of the good were correctly classified, for a total of 66.9% percent of cases accurately predicted following addition of the PDQ variables. The addition of post PDQ improved the classification of cases into the poor group slightly, although overall the model grossly over-classified cases into the good category. The PDQ provided the best classification of all variables for poor outcome status.

Tables 51b-c present the regression coefficients, chi-square test, and odds ratios for significant coefficients. Age varied significantly between the poor and good outcome groups, with people in the poor outcome group 1.08 times more likely to be older than patients in the good group, p < .001 (Table 74). In addition, patients in the poor group were 1.02 times more likely to have more severe levels of disability as measured by the PDQ. No significant differences were identified between the fair and good group.

In sum, post PDQ scores, combined with age and LOD, predicted outcomes. Percent difference was not correlated with outcomes, and overall, analyses indicated that

application of MCID percent difference for the PDQ would not prove useful in application of identifying individuals at risk for program non-completion.

11.4 Pain Disability Questionnaire Cross-Validation Results

Percent Change in PDQ is not an adequate predictor of outcome status, thus an MCID calculated based on percent change has no clinical utility. Table 76 presents the average percent change associated with poor, fair, and good outcomes for the test. No significant differences were detected between percent difference scores calculated in the test and training set. Effect sizes are presented in Table 77.

11.5 Summary of Pain Disability Questionnaire Results

Table 78 presents a table of summary results for evaluation of PDQ MCID. ANOVA results indicated that percent difference PDQ did not vary based on outcome category. This finding was confirmed by the correlational analysis that indicated no relationship between the outcome and PDQ difference variable. PDQ post scores, combined with age and LOD, as a set predicted 1 year outcomes. The utility of developing post-treatment risk categories to identify patients at risk for poor outcomes should be evaluated further. Utilizing post-treatment PDQ scores combined with age, the PDQ had the best classification of patients in the poor outcome group of all measures evaluated.

CHAPTER 12

PI RESULTS

12.1 Descriptive and ANOVA Results for the Pain Intensity Scale

Descriptive statistics for gold standard derived percent differences are presented in Table 79. As anticipated, patients classified as having poor outcomes averaged the smallest percent difference (14.39±24.01), whereas the fair outcome group averaged a difference of 17.0±24.87, and the good outcome group a difference of 23.47±30.19. A 2-way between subjects ANOVA, with outcome and pre-treatment severity as independent variables and percent difference as dependent variable, indicated that the PI percent difference varied among outcome groups ($F_{2, 1898} = 16.647$, $p < .001$). This supports the sensitivity of the measure to detect differences at the group level.

A main effect for pre-treatment severity ($F_{2, 1898} = 24.221$, $p < .001$) indicated that percent difference varied based on severity of PI prior to treatment. The most severely affected patients experienced a greater magnitude of change as compared to less severely afflicted patients. Patients in the minor category averaged a percent change of 15.53±31.56, those in the moderate category averaged a percent change of 25.38±24.95, whereas those scoring the highest at pre-treatment (severe) averaged a change of 28.74±20.82. No significant interaction between outcome and pre-treatment level of

severity was detected, $F_{4, 1898} = .353$, $p = .842$, thus the same MCID was applied across all levels of pre-treatment severity.

Utilizing the gold standard MCID approach, with fair outcome considered the minimum acceptable outcome, the MCID of the PI (percent change) would be 17.0.

12.2 Pain Intensity Effect Size Results

All effect sizes were considered large by Cohen's standards (poor = .85, fair = .97, good = 1.14), for a total effect size of 1.06 (Table 80). If the percent difference is predictive of outcome category, than in studies evaluating efficacy of treatment outcomes, an effect size of at least 1 would be indicative of good 1-year outcomes.

12.3 Pain Intensity Regression Analysis

Prior to regression analysis, correlational analyses were conducted to explore the relationship among PI variables and outcome. Table 81 presents Spearman correlations among each of the PI variables and the composite outcome variable. Each of the four PI variables were significantly correlated with the outcome, with post correlated the highest, ($r = -.203$, $p < .001$), followed by pre, ($r = -.131$), PI difference ($r = .121$, $p < .001$), and the interaction between pre PI and PI difference ($r = .105$, $p < .001$). Overall, higher pre and post pain scores were associated with worse outcomes, and greater differences were associated with better outcomes. Tolerance estimates indicated a high degree of multicollinearity with all four PI variables, thus the model was run with the two PI variables with the highest correlation (pre and post).

There was a good model fit using the demographic variables alone, χ^2 (6, N=1907) = 75.042, $p < .001$ (Table 82). Following addition of two PI variables, χ^2 (10,

N=1907) = 169.886, p < .001, Nagelkerke R^2 = 10.4. Addition of the PI variables significantly improved the fit of the demographic only model, χ^2 (4, N=1907) = 94.844, p < .001. Following deletion of gender, the only non-significant variable in the model, χ^2 (8, N=1907) = 165.554, p < .001, and the difference in the model was not significant χ^2 (2, N=1907) = 4.332, p = .115. Deletion of LOD, the variable with the highest p value, resulted in χ^2 (8, N=1907) = 157.405, p < .001, and the difference in the model was significant, χ^2 (2, N=1907) = 8.149, p = .017. Thus, age, LOD, pre PI, and post PI were all retained in the final reduced model.

Classification for the demographic only model was poor, with only 0.5% of the poor, 0% of the fair, and 99.8% of the good being accurately classified, for a total of 64.1% percent of cases. In the final model, only 4.3 % of the poor, 0.6 % of the fair, and 99.2 % of the good cases were correctly classified, for a total of 64.2% of cases being accurately predicted. Cases were overwhelmingly classified as good.

Regression coefficients, chi-square test, and odds ratios for significant coefficients are presented in Tables 84-85. Age varied significantly between the poor and good outcome groups, as well as between the fair and good outcome groups. In addition, people in the poor group were between 1 and 1.3 times more likely to have a longer LOD and higher pre and post treatment PI scores, as compared to patients in the good group, p < .01 (Table 84). Patients in the fair group were also 1.15 times more likely to have a higher Post treatment PI score than patients in the good group (Table 85).

In sum, age and LOD are key factors in predicting outcome group. After controlling for differences in demographic variables, pre-treatment scores on PI did not

significantly predict outcome group. Of all the PI variables, post pain scores were the most highly correlated with outcome.

12.4 Pain Intensity Cross-Validation Results

Percent change in PI is not an adequate predictor of outcome status, thus an MCID calculated based on percent change has no clinical utility. Table 86 presents the average percent change associated with poor, fair, and good outcomes for the test. No significant differences were observed between the training and test set. Relevant effect size statistics for both sets are presented in Table 87.

12.5 Summary of Pain Intensity Results

Table 88 presents summary statistics for Pain Intensity MCID analysis. Percent difference in PI was not predictive of outcome in the ANOVA and correlational analysis. Pre and post PI were highly correlated with outcomes. In the regression analysis, both provided significant prediction of outcome status. Overall classification heavily favored patients in the good outcome category overall, however by decreasing the cut-off and accepting a lower overall classification rate, better specificity may be obtained. Pre and post PI may be utilized to identify patients that are at risk for poor outcomes.

CHAPTER 13

BDI RESULTS

13.1 Descriptive and ANOVA Results for Beck Depression Inventory

Descriptive statistics for gold standard derived percent differences are presented in Table 89. A 2-way between subjects ANOVA, with outcome and pre-treatment severity as independent variables, and percent difference as dependent variable, indicated that the BDI percent difference varied among outcome groups ($F_{2, 1966} = 17.937$, $p < .001$). Patients classified as having poor outcome averaged a percent difference of 25.09±34.74, whereas the fair outcome group averaged a difference of 29.64±36.08, and the good outcome group a difference of 37.81±39.94.

In addition, a main effect for pre-treatment severity ($F_{2, 1966} = 17.765$, $p < .001$) indicated that percent difference varied based on pre-treatment BDI score. Patients scoring the lowest at pre-treatment (mild) averaged 27.0±47.18, those in the mid-range averaged 37.68±34.11, whereas those in the severe category averaged a change of 39.96±30.25. No significant interaction between outcome and pre-treatment level of severity was detected, $F_{4, 1966} = .607$, $p = .658$, thus the same MCID was applied across all levels of pre-treatment severity.

Utilizing the gold standard MCID approach, with fair outcome considered the minimum acceptable outcome, the MCID of the BDI would be 29.64.

13.2 Beck Depression Inventory Effect Size Results

Table 90 presents effect size calculations computed separately for patients in the poor, fair, and good outcome groups. All effect sizes were considered medium by Cohen's standards (poor = .62, fair = .67, good = .76), for a total effect size of .72.

13.3 Beck Depression Inventory Regression Analysis

Prior to regression analysis, correlations were conducted between BDI Variables and the outcome variable. All four BDI variables (pre, post, percent difference, pre x percent difference) were significantly correlated with outcome (Table 91), suggesting that the variables may provide useful information in the prediction of outcome. Initial evaluation of tolerance indicated that inclusion of all four BDI variables would result in multicollineraity, and unreliable statistical values of the individual predictor variables. Thus, the interaction term was dropped based on the lack of significant interaction in the ANOVA. Re-evaluation of tolerance indicated that all tolerance levels were above the greater than 0.2 criteria, thus pre BDI, post BDI, and the percent difference were all assessed in the regression.

There was a good model fit using the demographic variables alone, χ^2 (6, N=1975) = 99.963, $p < .001$ (Table 93). Following addition of three BDI variables, χ^2 (12, N=1975) = 154.396, $p < .001$, Nagelkerke $R^2 = 9.2$. Addition of the BDI variables significantly improved the fit of the demographic only model, χ^2 (6, N=1975) = 54.433, $p < .001$. Two of the predictors in the reduced model, gender ($p = .066$), and BDI post ($p =$

.767) were not significant. Analyses with these dropped from the model resulted in χ^2 (8, N=1975) = 148.427, $p < .001$. The difference between this model and the previous one was not statistically different χ^2 (4, N=1975) = 5.969, p = .202, thus gender and BDI post were dropped from the model. All remaining variables were statistically significant at the $p < .001$ level, leading to a final model that included age, LOD, BDI pre and BDI difference.

Classification for the demographic only model was poor, as 2.2% of the poor, 0% of the fair, and 99.9% of the good were correctly classified, for a total of 64.9% percent of cases accurately predicted. Despite a significant improvement in fit with the addition of the BDI variables, overall classification was not drastically changed, with 4.4 % of the poor, 1.9 % of the fair, and 99.0 % of the good cases being correctly classified, for a total of 65.0% of cases being accurately predicted. In both models, cases were overwhelmingly classified as good.

The contribution of each individual predictor in the model is presented in Table 96. Age, LOD, pre BDI and BDI difference were all significant predictors in the model. Tables 95-96 present the regression coefficients, chi-square test, and odds ratios for significant coefficients. Age varied significantly between the Poor and Good Outcome groups, with people in the Poor Outcome group between 1.01 and 1.07 times more likely to be older than patients in the Good group, $p < .001$, and have a greater LOD, p = .004 (Table 95). In addition, people in the Poor group were 1 times more likely to have more severe BDI scores prior to treatment as compared to patients in the Good group, $p < .001$, and have a smaller percent difference in BDI (Table 96). Patients in the Fair group were

1.02 times more likely to be older than patients in the Good group, and 1.02 times more likely to have more severe Pre-treatment BDI scores and 0.99 times more likely to have smaller percent differences in BDI than those in the Good group (Table 96).

13.4 Beck Depression Inventory Cross-Validation Results

Table 97 presents the average percent change associated with Poor, Fair, and Good Outcomes for the test, followed by relevant effect size statistics in Table 62b. No significant differences were observed between the training and test set.

Percent Change in BDI was a significant predictor of Outcome status. In an effort to evaluate the accuracy of the MCID obtained in the current analysis, a classification analysis was conducted in the test set reserved for cross-validation. Table 99 presents the results from this analysis. Fair categories were collapsed with Poor categories for ease of classification and interpretation. Utilizing the average percent change for patients in the Fair Outcome group as the MCID (rounded up to greater than or equal to 30 difference points), sensitivity is 70.4% (correctly identify patients that have good outcomes) and specificity is 58.7% (correctly identifying patients that have poor outcomes). Increasing the criteria to the mean percent change in the training set for patients in the Good outcome group, sensitivity is 70% and specificity is 61.0%. Further evaluation of the efficacy and expense of providing additional treatment for at-risk patients is needed to determine if utilization of such criteria would be cost-effective.

13.5 Summary of Beck Depression Inventory Results

Table 100 presents summary statistics for the BDI MCID analysis. Percent difference in BDI varied based on Pre-treatment level of severity, with patients in the

most severe Pre-treatment category averaging the smallest change. All BDI variables were significantly correlated with Outcome, however, following regression analysis only Pre BDI and BDI Difference significantly predicted Outcome status. Utilizing MCID calculated as the average percent change for patients in the Fair outcome group, a subset of patients reserved for cross-validation were classified into predicted Poor and Good Outcome categories. Overall sensitivity was approximately 70%, and specificity 58.7%. An alternative to utilizing an important *difference,* patients may be identified that are at risk of Poor Outcomes prior to treatment. By identifying patients that fall within high risk categories at Pre-treatment, they may be treated for underlying psychopathology that may prevent success in the treatment program.

CHAPTER 14

DISCUSSION

The purpose of the current study was to make the quantitative change scores in health outcome measures more clinically meaningful, by relating percent change in chronic pain outcome measures, to objective socioeconomic outcome criteria. Ideally, patients with change scores below the average threshold of those with a "fair" outcome could be targeted for additional intervention. By providing supplementary treatment to at risk patients, a higher success rate for those least likely to obtain good outcomes may be realized. Specifically, we sought to explore the following: 1) whether percent change scores were predictive of objective outcome status, 2) whether the percent difference varied based on pre-treatment severity, 3) relate the effect sizes of measures to objective outcome criteria, and 4) evaluate the predictive ability of a minimum clinical important difference, derived from a gold-standard approach with objective socioeconomic outcome as the gold standard.

Results will be discussed with regards to each individual variable below, following a general discussion of the findings on demographic variables.

14.1 Demographics

Initial analyses of the individual training samples for differences in demographic variables among outcome groups identified three variables of interest: age, LOD, and

gender. Of all these variables, the finding on age was the most robust. Patients in the poor outcome group had an average older mean age as compared to patients in the good outcome group for all six samples. This finding is not surprising given the generally higher occurrence of chronic musculoskeletal pain found in the literature for older patients (Franklin, Haug et al,. 1994; Stewart, Sachs et al., 1996 ; Dempsey, Burdorf et al., 1997; Bendix, Bendix et al., 1998). Older patients have accumulated more spinal damage over their working lives, and are exposed to repeated microtrauma, which might account for an increased likelihood of functional impairment (Kumar, 1990). Additionally, the load-bearing capacity of the spine decreases with age, causing an increased likelihood that occupational demands will not be met (Dempsey et al., 1997).

In addition to being identified in the preliminary comparisons, age was a significant predictor for outcome in logistic regression analyses for all samples. To further explore the ability of age to predict outcome, the average of patients in the poor category (49.57±9.18) was utilized in the test set to explore accuracy of classification (Table 103). Sensitivity was 67.41, and specificity was 59.3. Combined with criteria that are more discriminating among patients with good outcomes, age could prove useful for identifying patients in need of additional attention following functional restoration. One reason that age may be a good predictor of outcome is that the composite outcome criterion was heavily weighted by work return/retention data. Patients that are older may find it more difficult to return to work due to age discrimination. In addition, they may have additional retirement resources or social security to rely upon for income, and be

less motivated to return to the work force. In fact, this has been one criticism of utilizing such data as a "gold standard" for measuring success in treatment.

Length of disability was identified in the preliminary analyses to vary among outcome groups for all six samples. After correction for number of predictors in the logistic regression, however, the variable did not provide adequate information to significantly predict outcome group. Previous research indicates a linear relationship between LOD and socioeconomic outcomes, (Sandstrom, 1986; Krause and Ragland, 1994; Jordan, Mayer et al., 1998) however, the magnitude of this relationship has been questioned. Some reports indicate that patients with a 1 year length of disability prior to rehabilitation have work return rates as low as 25% (McGill, 1968), while more recently Jordan and colleagues report rates as high as 80%. Factors affecting the differences include lack of systematic evaluation, and varying lengths of disability combined for assessment. Jordan and colleagues (1998) identified significant differences among patients with varying LODs, however, patients with longer LOD continued to have impressive work outcomes. Patients with a LOD of at least 18 months experienced work retention at an average of 72%, one year following treatment. These results are consistent with our current findings.

Gender was also found to be significantly different in the MVAS, SF-36, PI, and BDI samples, with more males in good outcome groups as compared to the poor outcome groups. These were the largest of the six samples evaluated, thus the magnitude of the effect for this demographic variable is smaller as compared to the age and LOD variables. Following correction for number of predictors in the logistic regression analyses, gender

was most often not identified as a significant predictor of outcomes. Although there is not a large body of literature regarding gender differences in functional status, studies do indicate that females exhibit a lower threshold for pain than males, resulting in higher subjective reports of pain intensity (Vallerand and Polomano, 2000; Sheffield, Biles et al., 2000). Thus, the findings of the current study appear to be consistent with the few research studies examining the effect of gender upon functional status.

14.2 ODI

The ODI is one of the most frequently used functional assessment questionnaires utilized to assess health related quality of life in patients with low back injuries (Beurskens, de Vet et al., 1995). As a result, a number of studies have evaluated the validity and reliability of the instrument (Ohnmeiss, 2000). Numerous researchers provide evidence that the instrument has good test-retest reliability (Fairbank, Couper et al., 1980; Gronblad, Hupli et al., 1993; Triano, McGregor et al., 1993), and fair internal consistency (Fairbank, Couper et al., 1980; Kopec, Esdaile et al., 1996; Fisher and Johnson, 1997). Fewer studies have been conducted regarding the Minimum Clinical Important Difference, or responsiveness.

A handful of studies were identified that claim reports of responsiveness for the ODI, with MCIDs ranging from 5.2-16.3 (Taylor, Taylor et al., 1999; Suarez-Almazor, Kendall et al., 2000; Lurie, Hanscom et al., 2001; Hagg, Fritzell et al., 2002). These estimates were all based on a raw pre to post difference, and are consistent with the current estimate of 12.87. Unfortunately, due to the large confidence intervals surrounding the average change for patients in the fair outcome category, the change in

ODI was not predictive of 1-year outcomes on an individual patient basis. This negates the use of the ODI percent change as a clinical tool for identifying patients that are at risk for poor outcomes.

Although no studies were identified that evaluated the capacity of the ODI in predicting patient change at the individual level, two studies were identified that explored responsiveness at the group level. Taylor and colleagues (1999) combined the effect size and patient self-report gold standard approach, and conducted a study to evaluate responsiveness of the ODI for patients receiving treatment for low back pain. Effect sizes were then calculated for patients that fell in "worse", "unchanged," and "better" categories. The authors concluded the ODI was responsive for patients that classified themselves as better and worse, because patients in the "improved" category had a large effect size (1.1), and patients in the "worse" category had a slight negative effect size (-.5). The scale was less responsive for patients that were unchanged, as the effect size was still moderate for this group (.4). Beurskens and colleagues (1996) utilized a similar methodology and report the ODI was responsive for patients that categorized themselves as "improved" (effect size .8) and "unchanged" (effect size -.4).

These findings for responsiveness of the ODI at the group level were not replicated in the current study. Utilizing a similar methodology, with objective criteria vs. subjective self-report as the gold standard, the ODI was not responsive at the group level for patients that experienced a "fair" or "poor" outcome, as the effect sizes remained large for both of these categories (.94, .86, respectively). This finding is a result of the large magnitude of change that is observed in all patients, irrespective of outcome. One

interpretation of these findings is that all patients report significant improvement from treatment, but a subset of these patients are not able to return to work, or discontinue healthcare use for alternative reasons. The fact that post-treatment ODI scores are predictive of outcomes suggests that this is not the case, and indeed patients in the poor outcome group are more functionally impaired as compared to patients that obtain good outcomes.

Overall, these results indicate that a significant improvement may be detected in the ODI upon treatment. However, a patient that improves 100%, as compared to a patient that improves only 10%, is no more likely to have good outcomes. Older patients that have elevated ODI scores following treatment, are at greater risk for poor outcomes, however. Criteria may be developed for post-treatment clinical use, to identify patients that fall in a "high risk" category (i.e., older than 50, Post ODI score of at least 30), so that patients at greatest risk for poor outcomes may be targeted for future intervention.

14.3 MVAS

The Million Visual Analog Scale was designed to measure pain intensity, and assess progress among patients with back pain (Million, Haavik-Nilsen et al., 1981). Although the measure won the Volvo award in Clinical Science in 1981, relatively few studies have been published regarding the psychometric properties of the instrument. The studies that have been published primarily deal with validity and reliability of the instrument, and not responsiveness as it applies to the MCID.

Beurskens and colleagues (1995) identified five studies that present data regarding responsiveness of the MVAS. Four of the five evaluate responsiveness as

significant pre to post change, with three of these presenting data that supports the sensitivity of the MVAS in detecting group differences (Million, Haavik-Nilsen et al., 1981; Million, Hall et al., 1982; Triano, McGregor et al., 1993), and one that suggests the scale is not sensitive to detecting differences in treatment groups for patients in a multidisciplinary treatment center (Cassisi, Sypert et al., 1989). The scale has also demonstrated a significant correlation with improvement on other objective measures (Hazard, Fenwick et al., 1989). Results from the current study regarding the sensitivity of the MVAS to detect pre to post change were consistent with the majority of these studies, and suggest the MVAS is sensitive to pre to post change, with a large effect size (1.3).

The current study sought to explore the ability of a MCID, calculated utilizing a gold-standard approach with objective outcome as an anchor, to predict patient outcome status. Significant correlations between pre-treatment, post-treatment, and percent difference scores confirmed that the MVAS was significantly related to the gold-standard anchor. Following regression analyses, both pre-treatment MVAS and percent difference scores significantly predicted outcome. The mean percent change for patients in the fair outcome group was utilized to classify patients in a cross-validation study at risk for poor outcomes. Previous reports of the MCID raw difference ranged from 24-40 difference points (Mayer et al, 1985; Gatchel, Mayer, Capra, Barnett & Diamond, 1986; Hazard et al., 1989)**,** and current results fall within this range (24.39±24.36). Classification results indicated that of the patients identified at risk for poor outcomes, only 47.5% actually resulted in poor outcome. Thus, utilizing this criterion, only ½ of the patients targeted for future intervention would actually be at risk for poor outcomes. Of those predicted to

have a good outcome, 78.3% actually experienced a good outcome, whereas 21.7% resulted in poor outcomes. Thus, utilizing this criterion, approximately 20% would "fall through cracks," and not receive additional intervention when it was needed. Through additional consideration of pre-treatment severity scores, improvement in classification may be realized.

Results of the current project are consistent with a recent study that aimed to explore the prediction of MVAS pre and post scores in outcome status (Anagnostis, Mayer et al., 2003). In a unique design, the authors divided patients into five different levels of pre-treatment severity, ranging from "no reported disability," to "extreme disability." Groups were compared on 1 year outcomes to determine if level of pre-treatment severity was related to other psychosocial variables, or 1 year outcomes (Anagnostis, Mayer et al., 2003). Moderate pretreatment scores were associated with a higher rate of post-rehabilitation health care, lower program completion rates, decreased work retention, and decreased work return. In addition to the strong correlation between pre-treatment level of severity and outcomes, post-treatment level of severity was also related to outcomes, patients that scored higher at post-treatment were at risk for worse outcomes. The authors suggested that pre-treatment MVAS scores may be used to classify patients into risk categories, so clinicians may identify patients at risk for poor outcomes.

Whether the application of such criteria will provide useful from an economic standpoint is dependent on how successful additional intervention would be in preventing

poor outcomes, and the cost of additional intervention. Future cost-utility analysis is needed to further explore these issues.

14.4 SF-36

The Short-Form 36 is one of the most widely used general health related quality of life measures (Brazier, Roberts et al., 2002). The general nature of the measure makes application among a variety of diseases possible. Both the reliability and validity of the SF-36 have been demonstrated in numerous studies, (Brazier, Harper et al., 1992; Ware, Kosinski et al., 1994; Gatchel, Polatin et al., 1998; Brazier, Roberts et al., 2002), although the evaluation of the responsiveness of the measure to detect individual differences has been less well researched.

Although no study was identified that evaluated responsiveness with regards to individual prediction, a variety of studies were identified aimed at assessing the responsiveness of the instrument at detecting group differences. Taylor and colleagues (1999) combined a distribution-based and anchor-based approach by calculating effect sizes for patients that assessed themselves as "worse", "unchanged", and "better" following treatment. Moderate effect sizes were achieved for all SF-36 scales for patients that assessed themselves as better, and negative effect sizes were observed for a majority of the scales for patients that considered themselves worse. The authors concluded, based on these findings, that the SF-36 is sensitive to detecting clinically important change. Unfortunately, in addition to the issues associated with subjective self-report of clinical improvement, the effect size measure assesses group change, as opposed to change at the individual level. While estimates of the magnitude of an effect are critical for the

evaluation of group differences, assessments of responsiveness with regards to individual change are critical for clinical application.

In a previous study conducted on a subset of the current dataset, Gatchel and colleagues (1998) provided evidence of significant improvement on all ten SF-36 scales following completion of a functional restoration program. The authors reported that in terms of clinical utility, however, large confidence intervals surrounding the average scores prevented utility of the scale at the individual patient level. In a follow-up study utilizing the same cohort, Gatchel and colleagues (1999) evaluated the relationship among Pre and Post treatment SF-36 scores, and 1-year socioeconomic outcomes. Better pre and post treatment SF-36 scores were more frequently associated with good outcomes; however post treatment scores, specifically the scores related to physical health, were more frequently associated with good outcomes than pre-treatment scores.

In the current study, utilizing a larger cohort of patients, a logistic regression analysis was utilized to evaluate the predictability of pre, post, and percent difference scores on 1-year outcomes. Only the Physical Component and Mental Component Summary score were included for analysis. The only variable found to significantly predict outcome was the pre-treatment Physical Component Score. This is in slight contrast to the earlier report of a higher correlation between post PHS scores and socioeconomic outcomes. The previous study that reports a smaller correlation between post PHS and outcomes was conducted on a relatively small sample, however, and included very few cases of poor outcomes.

Another notable finding with regards to the Mental Health Summary Scale is that patients with higher pre-treatment MHS scores were less likely to experience change on the MHS scale. Although the magnitude of change did not vary across outcome group, patients with more severe levels of disability did experience less change on the MHS scale as compared to patients in mild or moderate categories. This suggests that patients with higher levels of psychopatholgy are not likely to eliminate any underlying psychopathology during the course of treatment. Despite this fact, level of pre-treatment severity on the MHS was not predictive of outcome ($r = .007$, $p = .857$), which implies that even the patients scoring the highest on the MHS have an equal chance in obtaining good outcomes. It should be noted, however, that only patient completers were included in the current analysis. It has been demonstrated previously that patients scoring higher on the MHS scale are less likely to complete the program (Gatchel et al, 1999). It is possible that in the current sample patients with the highest MHS scores dropped out.

Overall, evidence in the current study does not support the use of the SF-36 for identification of patients at risk for poor outcomes. Thus, no minimum clinical important difference may be recommended for use in clinical assessment of high risk patients.

14.5 PDQ

The Pain Disability is a relatively new Health Related Quality of Life measure that was designed specifically for use in the chronic musculoskeletal disorder population (Anagnostis, Gatchel et al, 2004). More traditional measures of chronic pain are disease specific, and focus on low back pain (ODI), or spinal disorders (MVAS). The PDQ was designed for use in patients with upper extremity, lower extremity, and spinal disorders. The scale yields a functional component, psychosocial component, and total component score. In the current study, only the total component scores were analyzed.

Although very few studies have explored the reliability and validity of the PDQ, the initial analysis conducted by the authors indicated the scale had very good responsive properties. In comparison to the ODI, MVAS, and SF-36, the PDQ had the largest effect size with regards to pre to post change. This finding is consistent with our current analysis, with the PDQ obtaining the largest overall effect size of all measures (1.46, see table 102 for summary statistics). As discussed previously, however, a measures' ability to detect a large pre to post change is not necessarily indicative of the capacity to predicting patients that will have poor outcomes.

The current study evaluated the responsive properties of the PDQ, with the intent of utilizing a MCID to identify patients at risk for poor outcomes. Following logistic regression for evaluation of significant predictors of outcome, only post treatment PDQ scores provided adequate information to aid in the classification of individual patients.

The rather large magnitude of change documented by the PDQ may be indicative of the measures' ability to capture the full breadth of functional impairments involved in

chronic musculoskeletal disorders. Unlike the ODI, and MVAS, the measure includes a variety of psychosocial factors that are known to play an important role in the etiology and maintenance of chronic pain conditions. Despite the fact that the measure captures change in these important factors, change in PDQ is not predictive of 1 year outcomes, preventing the application of a MCID determination that may be applied in a clinical setting. Post-treatment PDQ combined with age provided the best classification of poor outcome groups however, indicating that the measure provides good responsive properties when utilizing post-treatment scores as compared to percent difference scores.

Additional research is needed to evaluate the functional component and social component independent from the total component summary score. In addition, a limiting factor in the current study was sample size. Of all the six samples evaluated in the current study, the PDQ sample was the smallest (n = 395), with only 41 cases in the poor outcome category. Additional studies should be conducted with a larger sample size.

14.6 PI

Visual analog scales of pain are utilized in the majority of treatment outcome studies to evaluate treatment efficacy (McGeary, Mayer et al., 2006). A number of studies have been published regarding the psychometric properties of visual analog scales of pain, and report that as whole, VAS measures are more sensitive than categorical measures of pain (Joyce, Zutski et al., 1975; Scott and Huskisson, 1976). Reliability and validity of Visual Analog Scales has been demonstrated (Carlsson, 1983; Sriwatanakul, Kelvie et al., 1983), however few studies have evaluated the MCID of Visual Analog

Scales, and none were identified that assessed the classification accuracy for identifying individual patients at risk for poor completion.

A handful of studies have reported the effect size of VAS scales for patients receiving various treatments. Beurskens and colleagues (1999) calculated mean pre to post changes, and effect sizes for patients that rated themselves as "improved" and "not improved." The effect size for the Visual Analog Scales for patients that had improved was 1.66, and for those that had not, 0.26. Based on the difference in effect sizes, they concluded that the VAS is a sensitive measure that is able to detect clinically important change. In the current study, the average change associated with patients with a fair outcome was 1.82±2.33, consistent with Beurskens reports of patients that categorized themselves as improved. Unfortunately, patients with poor outcomes had a difference of 1.56±2.44, which was not statistically different from those with either fair or good outcomes. However, patients with poor outcomes did have a smaller effect size (.66), as compared to those in the fair (1.12) and good categories (1.10).

More recently, Hagg and colleagues (2003) evaluated the clinical important difference in a VAS scale utilizing a gold standard approach, with patient global assessment of functioning as the anchor, following treatment for chronic low back pain. They calculated the MCID of the VAS as 18-19 units (the average change in patients that rated themselves as better following treatment on a scale from 0-100), which exceeded the 95% tolerance level of 15 units. In fact, compared to the ODI, the General Function Score, and the Zung Depression Scale, the VAS scale was the most responsive to change.

A change in 18-19 units on a 100 point scale is similar in magnitude to the 1.82±2.33 units observed in the current study on a VAS scale from 0-10.

A notable finding in the current analysis is that of all PI variables evaluated (pre, post, difference, and pre x difference), post-treatment pain ratings were the most highly correlated with 1-year socioeconomic outcomes ($r = .204$, $p < .001$). In fact, in the logistic regression analysis, only pre and post Pain Intensity Variables (combined with age) were significant predictors of outcome category. Percent difference in PI was significantly correlated with outcomes, ($r = .121$, $p < .001$), but did not provide any unique contribution to the model with age, pre, and post treatment scores. Development of classification criteria to identify patients at highest risk for poor outcomes may be more fruitful on post-treatment criteria, as opposed to identifying a minimum clinical important difference.

14.7 BDI

The Beck Depression Inventory was introduced in 1961 (Beck, Ward, Mendelson, Mock, and Erbaugh, 1961), and over the years has become one of the most widely used instruments for assessing depression in both normal populations and patients with a clinical depression diagnosis (Beck, Steer et al., 1988). A number of reviews have been published regarding the psychometric properties of the BDI (Beck, Rush et al., 1979; Beck, Steer et al., 1996).

The use of the instrument within the chronic pain population has been less well documented. The relationship between chronic pain and depression has been well documented, however (Romano and Turner, 1985; Banks and Kerns, 1996; Dersh,

Gatchel, Mayer, Polatin, and Temple, 2006), and the use of a depression measure in evaluation of treatment outcomes is not uncommon. No research was identified that assessed a Minimum Clinical Important Difference in a chronic pain population. However, responsive properties have been reported in alternative populations, and the BDI is a good discriminator both within depressive categories and between depressed and non-depressed patients (Beck, Steer, and Garbin, 1988).

The results indicated that percent change in the BDI was predictive of outcome status, with a strong favor for accurately classifying good outcomes. The MCID identified in the current population, utilizing the gold standard approach, was 29.64±36.08. Patients in the good outcome category were more likely to have larger percent change in BDI than patients in the poor outcome group. In addition to percent difference, pre-treatment level of severity was a significant predictor of outcome, with patients more severely affected at pre-treatment more likely to experience poor outcomes. Interestingly, those in the most severe pre-treatment category experienced, on average, no change from pre to post. This suggests that significant depressive symptoms preclude success in the functional restoration program. One interpretation of this finding is that patients scoring the most severe at pre-treatment are experiencing a more chronic level of depression, consistent with "trait" psychopathology. Patients that score moderate or mild scores on the BDI prior to treatment may be experiencing more of a "state" depression, and are more readily treated by the functional restoration approach. It has been suggested by Gatchel et al (1999), that patients with very severe pre-treatment depression be treated for their depression prior to admission into the functional restoration program.

14.8 General Discussion

Overall, evidence for predictability of percent difference in outcome scores was not strong for the majority of measures. Only three of the six measures evaluated had a significant correlation between the change score and outcome (Table 101). A significant correlation among outcome and change score is only a minimum requirement for application of MCID criteria in an individual patient population. In addition to a correlative relationship, the measure must provide adequate prediction of outcome (Testa, 1988; Samsa, Edelman, et al, 1999). In the current study, percent change significantly predicted outcome in only two of the variables: Million Visual Analog Scale, and the Beck Depression Inventory. The variables provided significant prediction of outcome status only in combination with pre-treatment scores and age.

The Million Visual Analog Scale was one of two scales that were significantly correlated with outcome based on pre, post, percent change, and pre x percent change in MVAS. The Million was designed to describe pain and disability in patients with back pain. Utilizing an analog format, the scale presents 15 items that ask patients about their pain, and how it interferes with their life. The significance of the percent difference variable indicates that not only are pre-treatment pain and functional levels indicative of success, but the amount of improvement a patient reports to experience is important as well. This supports use of an MCID for individual classification purposes for the MVAS. In the current sample, an average percent difference score of 18.34±24.46, and the raw difference of 24.39±24.36 was associated with fair outcomes. This is the first report the authors are aware of regarding a gold-standard linked clinically important difference in a

chronic musculoskeletal population. The finding is consistent with previous findings regarding significant association of MVAS scores and treatment outcomes (Mayer et al, 1985; Gatchel, Mayer, Capra, Barnett, and Diamond, 1986; Hazard et al, 1989, Gatchel et al, 2003). Following classification of patients in the test set into poor risk categories utilizing the MCID as indicated in the training set, classification was not overly impressive, with a sensitivity of 78.3% and specificity of 52.5%. The scale was most effective at predicting patients with good outcomes, as opposed to predicting patients with poor outcomes. This is illustrated by the better sensitivity of the scale as compared to the specificity. The logistic regression analysis indicates that classification may be improved upon inclusion of pre-treatment scores, and age into the categories.

The Beck Depression Inventory was the only other percent difference variable that was predictive of outcomes. The variable was significantly correlated with the outcome ($r = .130$, $p < .001$), and was predictive of outcome when combined with pre-treatment level of severity, age, and length of disability. The significance of the percent difference variable indicates that degree of improvement in depressive symptoms plays a key role in successful socioeconomic outcomes. In the training set, the average percent difference for patients in the fair outcome group was 29.64±36.08, and the average raw change was 6.87±9.74. No additional studies were identified that evaluated the clinical important difference for use in a chronic pain population; however, the current finding is consistent with research that has evaluated the responsive properties in alternative populations (Beck, Steer, and Garbin, 1988).

Despite only two scales including percent difference in the final regression model, a number of variables had significant correlations with *pre-treatment scores*. In fact, all but the Mental Component Summary Scale of the SF-36 had correlations of r = .10 or greater (Table 101). Following regression analyses, four of the six scales resulted in the inclusion of pre scores in the final model. SF-36 PHS summary scores were predictive of outcome when combined with age and gender. MVAS pre-treatment scores were significant when combined with percent difference and age. Pain Intensity was predictive when combined with age, LOD, and post scores, whereas BDI pre-treatment was predictive when combined with age, LOD, and percent difference. In sum, these results indicate that more significant than an "important difference", pre-treatment severity provides predictive information for outcomes. This finding is encouraging for clinicians, as it provides initial support for the development of risk categories prior to treatment. For example, with the BDI, patients with very severe levels of depression may be recommended to receive treatment for their depression, prior to admittance into the functional restoration program (Gatchel, 1999).

Post-treatment scores also provided superior prediction overall as compared to percent difference scores. Post treatment ODI scores were predictive of outcomes when combined with age, whereas post PDQ scores were predictive when combined with age and LOD, and Pain Intensity Scores were predictive when combined with age, LOD, and pre-treatment scores. In fact, out of all the models, the PDQ model was the best performer with regards to classification of poor outcome status (correctly classified 12.2%), with the ODI a runner up, with 7.3% of the poor outcomes correctly classified.

All other measures classified between 0 and 3% of poor outcomes. In addition, these two models had the highest estimates of total variance accounted for (Nagelkerke 17% for the PDQ, and 9.9% for the ODI final model). Overall, these two measures appeared to classify poor outcomes better than any other measure evaluated.

A secondary focus of the current study was to relate effect size calculations with objective outcome criteria. Effect size calculations were originally designed in an effort to describe the *magnitude* of a statistically significant change (Cohen, 1977). Cohen suggested criteria of .2 for a small effect size, .5 for a medium effect size, and .8 for a large effect size. Unfortunately, although these measures provide insight into the magnitude of a change, it is not linked to clinically important differences. Stated differently, the estimates are still *quantitative* in nature. By relating effect size calculations in the current study to objective outcome success, future studies may provide *clinical meaning* to a numerical estimate. The only variables with statistically significant correlations between difference and outcome were the MVAS, the PI, and BDI. Effect sizes associated with each of these variables are presented in Table 64. For the MVAS, the effect size for patients in the fair category was 1.03, and was 1.45 for patients in the good category. Ideally, a negative effect size, or in the least a smaller effect size, would be identified in patients with poor outcomes. Unfortunately, for the MVAS, the effect size for patients in the poor outcome group was larger than for the fair outcome group. This is likely an artifact of the large variability in the fair outcome group, and the much smaller sample size for the MVAS Fair group. For the PDQ, the poor group had the smallest effect size as compared to the good and poor, but the effect size was still a large

magnitude. A greater discrimination in effect size was observed for the PI sample. Patients in the poor outcome group had a moderate effect size (.66), as compared to those in the fair (1.12) and good (1.10) categories. Based on these results, in a similar population, a treatment effect of at least 1.10 would be associated with good 1-year socioeconomic criteria. For the BDI, the poor outcome group had an effect size of .62, the fair outcome group .67, followed by the good outcome group with an effect size of .76. Again, patients with worse outcomes experienced a smaller magnitude of change. Studies conducted with a similar population may roughly estimate that a moderate effect size would be associated with poor 1-year outcomes, and a large effect size with good 1-year outcomes.

One issue with generalizing outcome estimates based on effect sizes calculated in the current study, is that effect sizes are sample dependent. They are heavily influenced by the variability within a sample. The current population may be considered a relative homogenous population within the pain community. Patients receiving treatment at PRIDE are all chronic pain patients. Their pain conditions have affected their ability to work, social relationships, and general quality of life. Effect sizes will be very different for a sample that includes more acute pain conditions. Additional research is needed to further explore the relationship between effect sizes and more clinically interpretive applications of data.

In addition to identifying a minimum clinical important difference for relevant psychosocial variables, the current study also sought to explore whether percent difference scores varied based on pre-treatment. Previous research indicates that people

with more functional limitation and disability may have different MCID estimations as compared to people with less severe disability (Riddle, Stratford et al., 1998; Stratford, Binkley et al, 1998). Both these studies utilized an effect size approach to estimating clinically important differences, and were geared towards detecting group differences as opposed to individual differences. In the current study, with a gold-standard approach, none of the measures evaluated were found to have a significant interaction between pre-treatment level of severity and outcome status.

Overall, for the purpose of the prediction of at-risk patients, percent difference scores were only predictive of outcomes for the MVAS and BDI. However, additional pre and post scores did display responsive properties in the current study. Based on the higher specificity of the ODI and PDQ, and the higher sensitivity of the PI, MVAS, and BDI, combining these scales for predictive outcome may maximize sensitivity and specificity. Specifically, scales identified with the best potential are pre-treatment BDI, post-treatment PDQ, post-treatment PI, and pre and percent difference MVAS.

14.9 Limitations of the Present Study and Future Directions

One limitation of the current study was that the classification selected for use may be affected by discrepant sample sizes in the criterion variable. Logistic regression has gained popularity over the past decade as an alternative procedure to discriminate analysis for classification purposes. The procedure is more forgiving of normality and homogeneity of variance violations, and normally distributed errors need not be assumed. In addition, the independent variables may be binary, categorical, or continuous, and procedures exist for either binary or multinomial outcomes. The procedure was selected

for use in the current project for the flexibility in variance and normality violations, and the capacity to select a multinomial outcome category (in this case poor, fair, and good outcome groups). One downside of logistic regression is that it is geared towards maximizing the total classification rate, which may be biased by discrepant sample sizes. In every sample evaluated in the current study, outcome status was heavily biased towards good outcomes. On average, the good category made up 60-65% of each sample, the fair category 20-25%, followed by the poor category which made up 10-15%. The effect this had on overall classification rate was evident, as all classification results, despite overall significance of the model, heavily biased good outcomes. Results of classification based on alternative cut-off scores are readily available through creation of dummy classification variables, however. The overall classification rate may be sacrificed for an improvement in classification of poor outcomes. In the largest samples, such as the PI, and BDI, the effect of the unequal outcome distribution was minimized by the large sample size (total N = 1976 and 1949, respectively). Unfortunately, in the smaller sample sizes, such as the PDQ (total N = 263, total Poor = 41), the discrepant sample size may have affected significance of the overall model.

Another limiting factor was the limited use of the cross-validation test set. Reserving a sub-set of the total sample for use in a cross-validation study is one of the most powerful methods for verifying reliability of results. Numerous statistical comparisons were made in the current project, increasing the chance for erroneous type I and II errors. Through validation of results, we can be more confident in our results. Additional future research should maximize on this test set, and evaluate the accuracy of

utilizing various cut-off scores designed to identify patients based on pre and post treatment criteria, in addition to the percent difference criteria that were evaluated in the current study.

In addition, the use of a composite outcome variable is unconventional, and one downside of combining socioeconomic outcomes is that it limits the interpretation of individual outcome variables. For example, results from the current study indicate that overall, outcome status is predicted by post-treatment PDQ scores in combination with age. What this doesn't tell us, is if it affects work retention/return differentially from healthcare utilization. Due to the fact that the focus of this paper was on evaluating the application of an MCID in predicting outcomes, and not a treatment effect, the use of composite score is justified. Future studies may tease apart these effects to look at the differential effects on various outcomes.

14.10 Conclusions

The measurement of function, pain, and general health is a central component of today's health field. Health outcome measures are vital tools that are utilized by patients, clinicians, and researchers to provide a means of examining clinical, functional, and patient satisfaction. More specifically, Functional Assessment Questionnaires (FAQs) were developed as subjective self-assessment tools that measure disability, pain, and overall functioning. Psychometric evaluation of these tools provides empirical evidence of their validity, reliability, and the ability to detect "clinically meaningful change." Recently, more attention has been given to the latter of these three psychometric indices. Responsive properties of an instrument may be related to the ability of an instrument to

detect clinically meaningful group changes, or clinically meaningful individual changes. Results from the current study suggest that with regards to responsiveness, clinically important *change* is not always the best predictor of individual outcomes. In fact, post-treatment and pre-treatment scores of common psychosocial outcomes in pain research were often equally, or more strongly, correlated with outcome status. Future research should evaluate the accuracy of classification based on risk categories developed utilizing age, pre-treatment BDI and MVAS, post-treatment PDQ, and post-treatment PI.

APPENDIX A

TABLES

Table 1. Decision Criteria for Composite Outcome Variable in Cases with all 3 Outcome Measures Available

Work Outcome	Health Care Utilization	Surgery Outcome	Composite Outcome
Good	Good	Good	Good
Good	Fair	Good	Fair
Good	Poor	Good	Fair
Good	Good	Poor	Fair
Good	Fair	Poor	Fair
Good	Poor	Poor	Fair
Fair	Good	Good	Fair
Fair	Fair	Good	Fair
Fair	Poor	Good	Fair
Fair	Good	Poor	Fair
Fair	Fair	Poor	Poor
Fair	Poor	Poor	Poor
Poor	Good	Good	Poor
Poor	Fair	Good	Poor
Poor	Poor	Good	Poor
Poor	Good	Poor	Poor
Poor	Fair	Poor	Poor
Poor	Poor	Poor	Poor

Table 2. Sample sizes for ODI, MVAS, SF-36, PDQ, PI, and BDI Analyses
(TRT=treatment, Compls=Completers)

Group	TRT Compls	Compls missing outcome variable	Compls missing outcome and Difference Score	Total N	Cross-Valid-ation N (30%)	Training N (60%)
ODI (n, %) (1999-2004) N=1,042	n=829 (79.6)	n=102 (12.3)	n=251 (30.3)	n=476	n=135	n=341
MVAS (n, %) (1993-2002) N=2,527	n=2,163 (85.6)	n=243 (11.2)	n=205 (9.5)	n=1,715	n=528	n=1,187
SF36 (n, %) (1999-2004) N=1,904	n=1,502 (78.9)	n=174 (11.6)	n=423 (28.2)	n=905	n=275	n=630
PDQ (n, %) (2002-2004,) N=870	n=682 (78.4)	n=74 (10.9)	n=213 (31.2)	n=395	n=132	n=263
PI (n, %) (1992-2004) N=4,134	n=3,488 (84.4)	n=393 (9.5)	n=272 (7.8)	n=2,823	n=874	n=1,949
BDI (n, %) (1992-2004) N=4,134	n=3,488 (84.4)	n=393 (9.5)	n=350 (10.0)	n=2,745	n=828	n=1,976

Table 3. Frequency of Patients (Completers with Available Predictor and Criterion Data Only) That Were Categorized as Poor, Fair, and Good Success Each Set of Analyses

ODI	Poor	Fair	Good
Training Set n=345	n=41 (12.0)	n=86 (25.2)	n=214 (62.8)
Test Set n=131	n=15 (11.1)	n=43 (31.9)	n=77 (57.0)
Total N=476	n=56 (11.8)	n=129 (27.1)	n=291 (61.1)

MVAS	Poor	Fair	Good
Training Set n=1,187	n=91 (7.7)	n=334 (28.1)	n=762 (64.2)
Test Set n=528	n=48 (9.1)	n=148 (28.0)	n=332 (62.9)
Total N=1,715	n=139 (8.1)	n=482 (28.1)	n=1,094 (63.8)

Table 3-continued.

SF-36	Poor	Fair	Good
Training Set n=630	n=73 (11.6)	n=150 (23.8)	n=407 (64.6)
Test Set n=275	n=37 (13.5)	n=66 (24.0)	n=172 (62.5)
Total N=905	n=110 (12.2)	n=216 (23.9)	n=579 (64.0)

PDQ	Poor	Fair	Good
Training Set n=263	n=41 (15.6)	n=47 (17.9)	n=175 (66.5)
Test Set n=132	n=18 (13.6)	n=25 (18.9)	n=89 (67.4)
Total N=395	n=59 (14.9)	n=72 (18.2)	n=264 (66.8)

Table 3 - continued.

PI	Poor	Fair	Good
Training Set n=1,949	n=188 (9.6)	n=510 (26.2)	n=1,251 (64.2)
Test Set n=874	n=79 (9.0)	n=227 (26.0)	n=568 (65.0)
Total N=2,823	n=267 (9.5)	n=737 (26.1)	n=1,819 (64.4)

BDI	Poor	Fair	Good
Training Set n=1,976	n=181 (9.2)	n=517 (26.2)	n=1,278 (64.7)
Test Set n=828	n=84 (10.1)	n=210 (25.4)	n=534 (64.5)
Total N=2,804	n=265 (9.5)	n=728 (26.0)	n=1,811 (64.6)

Table 4. Demographic Variables for Training and Test Set, and Total ODI Groups

	Training Set	**Test Set**	**Total**	**T test p value or $\chi^2$and odds ratio**
N	341	135	476	
Age (X, SD)	46.5±9.7	45.8±9.5	46.3±9.6	N.S.
Gender (n/total available n, % male)	206/341 (60.4)	84/135 (62.2)	290/476 (60.9)	N.S.
Race (n/total available n, %)				N.S.
Caucasian	191/341 (56.0)	73/134 (54.5)	264/475 (55.6)	
African-American	77/341 (22.6)	39/134 (29.1)	116/475 (24.4)	
Hispanic	63/341 (18.5)	22/134 (16.4)	85/475 (17.9)	
Other	9/341 (2.6)	1/134 (<1.0%)	10/475 (2.1)	
Length of Disability (months)	18.5±21.1	17.1±17.3	18.1±20.1	N.S.

Table 5. Demographic Variables for Poor, Fair, and Good Success Groups in the ODI Training Sample

	Poor Success	**Fair Success**	**Good Success**	**Total Sample**	**ANOVA or χ^2 p value**
N	41	86	214	341	N/A
Age (X, SD)	51.7±10.1	47.6±8.7	45.0±9.6	46.5±9.7	**.001**
Gender (n/total available n, % male)	24/41 (58.5)	45/86 (52.3)	137/214 (64.0)	206/341 (60.4)	N.S.
Race (n/total available n, %)					N.S.
Caucasian	28/41 (68.3)	51/86 (59.3)	112/213 (52.6)	191/341 (56.0)	
African-American	7/41 (17.1)	20/86 (23.3)	50/213 (23.5)	77/341 (22.6)	
Hispanic	4/41 (9.8)	15/86 (17.4)	44/213 (20.7)	63/341 (18.5)	
Other	2/41 (4.9)	0	7/213 (3.3)	9/341 (2.6)	
Length of Disability (months)	23.3±24.1	21.7±26.3	16.2±17.7	18.5±21.1	**.038**

Table 6. Statistical Analyses of Demographic Variables for ODI Training Group

Age				
Group	Mean (SD)	df	F	p value
Poor	51.7±10.1	2, 338	9.61	**<.001**
Fair	47.6±8.7			
Good	45.0±9.6			

Gender				
Group	M/F (% Males)	df	X^2	p value
Poor	24, (58.5)	2	3.58	.167
Fair	45, (52.3)			
Good	137, (62.8)			

Race							
Group	Caucasian (%)	A.A. (%)	Hispanic (%)	Other (%)	df	X^2	p value
Poor	28, (68.3)	7, (17.1)	4, (9.8)	2, (4.9)	10	9.17	.516
Fair	51, (59.3)	20, (23.3)	15, (17.4)	0			
Good	112, (51.6)	50, (23.0)	44, (20.3)	7, (3.2)			

Table 6 - continued.

LOD				
Group	Mean (SD)	df	F	p value
Poor	23.3±24.1	2, 337	3.31	**.038**
Fair	21.7±26.3			
Good	16.2±17.7			

Table 7. Post Hoc and Planned Comparisons of Demographic Variables for ODI Training Group

Age			
Group	Mean (SD)	p value	CI
Poor vs. Fair	51.7±10.1 47.6±8.7	.06	
Poor vs. Good	51.7±10.1 45.0±9.6	.001	(2.9, 10.5)
Good vs. Fair	45.0±9.6 47.6±8.7	.07	

LOD			
Group	Mean (SD)	p value	CI
Poor vs. Fair	23.3±24.1	.916	
Poor vs. Good	21.7±26.3	.121	
Good vs. Fair	16.2±17.7	.104	

Table 8. Demographic Variables for Training and Test Set, and Total MVAS Groups

	Training Set	**Test Set**	**Total**	**T test or χ^2 p value**
N	1,187	528	1,715	
Age (X, SD)	44.1±10.0	44.0±9.7	44.0±9.1	N.S.
Gender (n/total available n, % male)	717/1,187 (60.4)	305/528 (57.8)	1,022/1,715 (59.6)	N.S.
Race (n/total available n, %)				N.S.
Caucasian	757/1,172 (64.6)	333/521 (63.9)	1,090/1,693 (64.4)	
African American	186/1,172 (15.9)	89/521 (17.1)	275/1,693 (16.2)	
Hispanic	209/1,172 (17.8)	93/521 (17.9)	302/1,693 (17.8)	
Other	20/1,172 (1.7)	6/521 (1.2)	26/1,693 (1.5)	
Length of Disability (months)	15.8±21.6	15.7±23.1	15.8±22.1	N.S.

Table 9. Demographic Variables for Poor, Fair, and Good Success Groups in the MVAS Sample (training set only)

	Poor Success	**Fair Success**	**Good Success**	**Total Sample**	**ANOVA or χ^2 p value**
N	91	334	762	1,187	
Age (X, SD)	49.3±8.7	45.2±10.0	42.9±9.9	44.1±10.0	**.000**
Gender (n/total available n, % male)	52/91 (57.1)	181/334 (54.2)	484/762 (63.5)	717/1187 (60.4)	**.012**
Race (n/total available n, %)					.084
Caucasian	73/90 (81.1)	213/332 (64.2)	471/750 (62.8)	757/1172 (64.6)	
African-American	8/90 (8.9)	52/332 (15.7)	126/750 (16.8)	186/1172 (15.9)	
Hispanic	8/90 (8.9)	61/332 (18.4)	140/750 (18.7)	209/1172 (17.8)	
Other	1/90 (1.1)	6/332 (1.8)	13/750 (1.7)	20/1172 (1.7)	
Length of Disability (months)	21.7±17.8	16.9±20.1	14.6±22.5	15.8±21.6	**.007**

Table 10. Statistical Analyses of Demographic Variables for MVAS Training Group

Age				
Group	Mean (SD)	df	F	p value
Poor	49.3±8.7	2, 1184	20.13	**<.001**
Fair	45.2±10.0			
Good	42.9±9.9			

Gender				
Group	M/F (% Males)	df	X^2	p value
Poor	57.1	2	8.89	**.012**
Fair	54.2			
Good	63.5			

Race							
Group	Caucasian (%)	A.A. (%)	Hispanic (%)	Other (%)	df	X^2	p value
Poor	81.1	8.9	8.9	1.1	10	16.6	.084
Fair	64.2	15.7	18.4	1.8			
Good	62.8	16.8	18.7	1.7			

Table 10 - continued

LOD				
Group	Mean (SD)	df	F	p value
Poor	21.7±17.8	2, 1184	5.04	**.007**
Fair	16.9±20.1			
Good	14.6±22.5			

Table 11. Post Hoc and Planned Comparisons of Demographic Variables for MVAS Training Group

Age			
Group	Mean (SD)	p value	CI
Poor vs. Fair	49.3±8.7 45.2±10.0	.001	(1.4, 6.8)
Poor vs. Good	49.3±8.7 42.9±9.9	.000	(3.8, 8.9)
Good vs. Fair	42.9±9.9 45.2±10.0	.001	(-3.8, -0.8)

Table 11 - continued.

Gender					
Group	M/F (% Males)	df	X^2	p	odds ratio
Poor vs. Fair	57.1 54.2	1	.251	.616	
Poor vs. Good	57.1 63.5	1	1.41	.234	
Good vs. Fair	63.5 54.2	1	8.46	**.004**	1.3 (1.2, 1.5)

LOD			
Group	Mean (SD)	p value	CI
Poor vs. Fair	21.7±17.8 16.9±20.1	.144	
Poor vs. Good	21.7±17.8 14.6±22.5	**.008**	(1.5, 2.7)
Good vs. Fair	14.6±22.5 16.9±20.1	.232	

Table 12. Demographic Variables for Training and Test Set, and Total SF-36 Groups

	Training Set	**Test Set**	**Total**	**T test or χ^2 p value**
N	630	275	905	
Age (X, SD)	45.8±9.4	46.2±9.7	45.9±9.5	.519
Gender (n/ total available n, % male)	328/630 (52.1)	145/275 (52.7)	473/905 (52.3)	.854
Race (n/total available n, %)				.768
Caucasian	347/629 (55.2)	148/273 (54.2)	495/902 (54.9)	
African-American	157/629 (25.0)	70/273 (25.6)	227/902 (25.2)	
Hispanic	110/629 (17.5)	50/273 (18.3)	160/902 (17.7)	
Other	15/629 (2.4)	5/273 (1.8)	20/902 (2.2)	
Length of Disability (LOD) (months)	17.7±18.6	18.2±19.3	17.9±18.8	.772

Table 13. Demographic Variables for Poor, Fair, and Good Success Groups in the SF-36 Sample (training set only)

	Poor Success	**Fair Success**	**Good Success**	**Total Sample**	**ANOVA or χ^2 p value**
N	73	150	407	630	
Age (X, SD)	50.5±9.4	47.0±9.5	44.5±9.1	45.8±9.4	**.000**
Gender (n/total available n, % male)	31/73 (42.5)	69/150 (46.0)	228/407 (56.0)	328/630 (52.1)	**.024**
Race (n/total available n, %)					.550
Caucasian	49/72 (68.1)	77/148 (52.0)	221/399 (55.4)	347/619 (56.1)	
African-American	16/72 (22.2)	39/148 (26.4)	102/399 (25.6)	157/619 (25.4)	
Hispanic	6/72 (8.3)	31/148 (20.9)	73/399 (18.3)	110/619 (17.8)	
Other	1/72 (1.4)	1/148 (0.7)	3/399 (0.8)	5/619 (0.8)	
Length of Disability (months)	23.3±22.1	19.3±21.0	16.1±16.7	17.7±18.6	**.005**

Table 14. Statistical Analyses of Demographic Variables for SF-36 Training Group

Age				
Group	Mean (SD)	df	F	p value
Poor	50.5±9.4	2, 627	14.78	**<.001**
Fair	47.0±9.5			
Good	44.5±9.1			

Gender				
Group	M/F (% Males)	df	X^2	p value
Poor	42.5	2	7.5	**.024**
Fair	46.0			
Good	56.0			

LOD				
Group	Mean (SD)	df	F	p value
Poor	23.3±22.1	2, 626	5.32	**.005**
Fair	19.3±21.0			
Good	16.1±16. 7			

Table 15. Post Hoc and Planned Comparisons of Demographic Variables for SF-36 Training Group

Age			
Group	Mean (SD)	p value	CI
Poor vs. Fair	50.5±9.4 47.0±9.5	.023	(0.4, 6.5)
Poor vs. Good	50.5±9.4 44.5±9.1	**<.001**	(3.2, 8.8)
Good vs. Fair	44.5±9.1 47.0±9.5	**.013**	(-4.6, -0.4)

Gender					
Group	M/F (% Males)	df	X^2	p	odds ratio
Poor vs. Fair	42.5 46.0	1	.248	.619	
Poor vs. Good	42.5 56.0	1	4.58	.032	1.7, (1.0, 2.9)
Good vs. Fair	56.0 46.0	1	4.42	.036	1.5, (1.0, 2.2)

Table 15 - continued.

LOD			
Group	Mean (SD)	p value	CI
Poor vs. Fair	23.3±22.1 19.3±21.0		
Poor vs. Good	23.3±22.1 16.1±16. 7		
Good vs. Fair	16.1±16. 7 19.3±21.0		

Table 16. Demographic Variables for Training and Test Set, and Total PDQ Groups

	Training Set	**Test Set**	**Total**	**T test or χ^2 p value**
N	263	132	395	
Age (X, SD)	46.7±10.2	46.8±8.9	46.7±9.7	.930
Gender (n/ total available n, % male)	130/263 (49.4)	78/132 (59.1)	208/395 (52.7)	.070
Race (n/total available n, %)				.581
Caucasian	139/262 (53.1)	76/131 (58.0)	215/393 (54.7)	
African-American	64/262 (24.4)	27/131 (20.6)	91/393 (23.2)	
Hispanic	55/262 (21.0)	24/131 (18.3)	79/393 (20.1)	
Other	4/262 (1.5)	4/131 (3.1)	8/393 (2.0)	
Length of Disability (LOD) (months)	20.2±19.4	22.7±23.0	21.0±20.6	.254

Table 17. Demographic Variables for Poor, Fair, and Good Success Groups in the PDQ Sample (training set only)

	Poor Success	**Fair Success**	**Good Success**	**Total Sample**	**ANOVA or χ^2 p value**
N	41	47	175	263	
Age (X, SD)	52.4±10.1	48.6±8.7	44.9±10.0	46.7±10.2	**.000**
Gender (n/total available n, % male)	16/41 (39.0)	23/47 (48.9)	91/175 (52.0)	130/263 (49.4)	.326
Race (n/total available n, %)					.882
Caucasian	19/41 (46.3)	29/47 (61.7)	91/174 (52.3)	139/262 (53.1)	
African-American	13/41 (31.7)	10/47 (21.3)	41/174 (23.6)	64/262 (24.4)	
Hispanic	9/41 (22.0)	8/47 (17.0)	38/174 (21.8)	55/262 (21.0)	
Other	0/41 (0.0)	0/47 (0.0)	4/174 (2.3)	4/262 (1.5)	
Length of Disability (months)	27.4±24.5	23.5±19.6	17.6±17.4	20.2±19.4	**.006**

Table 18. Statistical Analyses of Demographic Variables for PDQ Training Group

Age				
Group	Mean (SD)	df	F	p value
Poor	52.4±10.1	2,262	10.93	**<.001**
Fair	48.6±8.7			
Good	44.9±10.0			

Gender				
Group	M/F (% Males)	df	X^2	p value
Poor	39.0	2	2.243	.326
Fair	48.9			
Good	52.0			

Race							
Group	Caucasian (%)	A.A. (%)	Hispanic (%)	Other (%)	df	X^2	p value
Poor	46.3	31.7	22.0	0.0	10	5.13	.882
Fair	61.7	21.3	17.0	0.0			
Good	52.3	23.6	21.8	2.3			

Table 18 - continued

LOD				
Group	Mean (SD)	df	F	p value
Poor	27.4±24.5	2,262	5.238	**.006**
Fair	23.5±19.6			
Good	17.6±17.4			

Table 19. Post Hoc and Planned Comparisons of Demographic Variables for PDQ Training Group

Age			
Group	Mean (SD)	p value	CI
Poor vs. Fair	52.4±10.1 48.6±8.7	.169	
Poor vs. Good	52.4±10.1 44.9±10.0	**<.001**	(3.5, 11.5)
Good vs. Fair	44.9±10.0 48.6±8.7	.053	(-7.5, 0.0)

LOD			
Group	Mean (SD)	p value	CI
Poor vs. Fair	27.4±24.5 23.5±19.6	.590	
Poor vs. Good	27.4±24.5 17.6±17.4	**.009**	(2.0, 17.6)
Good vs. Fair	17.6±17.4 23.5±19.6	.153	

Table 20. Demographic Variables for Training and Test Set, and Total PI Groups

	Training Set	**Test Set**	**Total**	**T test or χ^2 p value**
N	1,949	874	2,823	
Age (X, SD)	44.5±10.0	43.8±9.8	44.3±9.9	.090
Gender (n/ total available n, % male)	1,091/1,941 (56.2)	491/874 (56.2)	1,582/2,815 (56.2)	.921
Race (n/total available n, %)				.309
Caucasian	1,212/1,926 (62.9)	534/864 (61.8)	1,746/2,790 (62.6)	
African-American	338/1,926 (17.5)	140/864 (16.2)	478/2,790 (17.1)	
Hispanic	344/1,926 (17.9)	178/864 (20.6)	522/2,790 (18.7)	
Other	32/1,926 (1.7)	12/864 (1.4)	44/2,790 (1.6)	
Length of Disability (LOD) (months)	16.2±20.5	15.9±19.1	16.1±20.1	.651

Table 21. Demographic Variables for Poor, Fair, and Good Success Groups in the PI Sample (training set only)

	Poor Success	**Fair Success**	**Good Success**	**Total Sample**	**ANOVA or χ^2 p value**
N	188	510	1,251	1,949	
Age (X, SD)	49.2±9.4	45.6±9.8	43.4±9.9	44.5±10.0	**.000**
Gender (n/total available n, % male)	93/188 (49.5)	271/510 (53.1)	727/1,251 (58.1)	1,091/1,949 (56.0)	**.027**
Race (n/total available n, %)					.150
Caucasian	124/185 (67.0)	339/506 (67.0)	749/1,235 (60.6)	1,212/1,926 (62.9)	
African-American	34/185 (18.4)	82/506 (16.2)	222/1,235 (18.0)	338/1,926 (17.5)	
Hispanic	25/185 (13.5)	74/506 (14.6)	245/1,235 (19.8)	344/1,926 (17.9)	
Other	2/185 (1.1)	11/506 (2.2)	19/1,235 (1.5)	32/1,926 (1.7)	
Length of Disability (months)	21..6±23.0	17.1±19.2	15.0±20.6	16.2±20.5	**.000**

Table 22. Statistical Analyses of Demographic Variables for PI Training Group

Age				
Group	Mean (SD)	Df	F	p value
Poor	49.2±9.4	2, 1948	32.984	**<.001**
Fair	45.6±9.8			
Good	43.4±9.9			

Gender				
Group	M/F (% Males)	Df	X^2	p value
Poor	49.5	2	7.218	**.027**
Fair	53.1			
Good	58.1			

Race							
Group	Caucasian (%)	A.A. (%)	Hispanic (%)	Other (%)	df	X^2	p value
Poor	67.0	18.4	13.5	1.1	12	16.981	.150
Fair	67.0	16.2	14.6	2.2			
Good	60.6	18.0	19.8	1.5			

Table 22 - continued

LOD				
Group	Mean (SD)	df	F	p value
Poor	21..6±23.0	2, 1948	32.984	**<.001**
Fair	17.1±19.2			
Good	15.0±20.6			

Table 23. Post Hoc and Planned Comparisons of Demographic Variables for PI Training Group

Age			
Group	Mean (SD)	p value	CI
Poor vs. Fair	49.2±9.4 45.6±9.8	**<.001**	(1.7, 5.6)
Poor vs. Good	49.2±9.4 43.4±9.9	**<.001**	(4.1, 7.7)
Good vs. Fair	43.4±9.9 45.6±9.8	**<.001**	(-3.4, -0.9)

Table 23 - continued

Gender					
Group	M/F (% Males)	df	X^2	p	odds ratio
Poor vs. Fair	49.5 53.1	1	.741	.389	
Poor vs. Good	49.5 58.1	1	4.984	**.026**	1.4 (1.0, 1.9)
Good vs. Fair	58.1 53.1	1	3.654	.056	

LOD			
Group	Mean (SD)	p value	CI
Poor vs. Fair	21.6±23.0 17.1±19.2	**.030**	(0.3, 8.5)
Poor vs. Good	21.6±23.0 15.0±20.6	**<.001**	(2.8, 10.3)
Good vs. Fair	15.0±20.6 17.1±19.2	.124	

Table 24. Demographic Variables for Training and Test Set, and Total BDI Groups

	Training Set	**Test Set**	**Total**	**T test or χ^2 p value**
N	1,976	828	2,804	
Age (X, SD)	44.1±9.9	44.7±10.0	44.3±9.9	.174
Gender (n/ total available n, % male)	1,106/1,976 (56.0)	466/828 (56.3)	1,572/2,804 (56.1)	.881
Race (n/total available n, %)				.566
Caucasian	1,220/1,952 (62.5)	515/820 (62.8)	1,735/2,772 (62.6)	
African-American	339/1,952 (17.4)	137/820 (16.7)	476/2,772 (17.2)	
Hispanic	363/1,952 (18.6)	155/820 (18.9)	518/2,772 (18.7)	
Other	30/1,952 (1.5)	13/820 (1.6)	43/2,772 (1.6)	
Length of Disability (LOD) (months)	15.9±19.6	16.8±21.4	16.2±20.1	.289

Table 25. Demographic Variables for Poor, Fair, and Good Success Groups in the BDI Sample (training set only)

	Poor Success	**Fair Success**	**Good Success**	**Total Sample**	**ANOVA or χ^2 p value**
N	181	517	1278	1,976	
Age (X, SD)	49.6±9.2	45.1±9.8	43.0±9.7	44.1±9.9	**.000**
Gender (n/total available n, % male)	87/181 (48.1)	267/517 (51.6)	752/1,278 (58.8)	1,106/1,976 (56.0)	**.002**
Race (n/total available n, %)					.670
Caucasian	110/177 (62.1)	337/512 (65.8)	773/1,263 (61.2)	1,220/1,952 (62.5)	
African-American	37/177 (20.9)	80/512 (15.6)	222/1,263 (17.6)	339/1,952 (17.4)	
Hispanic	27/177 (15.3)	87/512 (17.0)	249/1,263 (19.7)	363/1,952 (18.6)	
Other	3/177 (1.7)	8/512 (1.6)	19/1,263 (1.5)	30/1,952 (1.5)	
Length of Disability (months)	22.8±21.7	16.1±18.1	14.8±19.7	15.9±19.6	**.000**

Table 26. Statistical Analyses of Demographic Variables for BDI Training Group

Age				
Group	Mean (SD)	Df	F	p value
Poor	49.6±9.2	2, 1975	40.452	**<.001**
Fair	45.1±9.8			
Good	43.0±9.7			

Gender				
Group	M/F (% Males)	Df	X^2	p value
Poor	48.1	2	12.792	**.002**
Fair	51.6			
Good	58.8			

Race							
Group	Caucasian (%)	A.A. (%)	Hispanic (%)	Other (%)	df	X^2	p value
Poor	62.1	20.9	15.3	1.7	12	9.383	.670
Fair	65.8	15.6	17.0	1.6			
Good	61.2	17.6	19.7	1.5			

Table 26 - continued.

LOD				
Group	Mean (SD)	df	F	p value
Poor	22.8±21.7	2, 1975	13.5	**<.001**
Fair	16.1±18.1			
Good	14.8±19.7			

Table 27. Post Hoc and Planned Comparisons of Demographic Variables for BDI Training Group

Age			
Group	Mean (SD)	p value	CI
Poor vs. Fair	49.6±9.2 45.1±9.8	**<.001**	(2.5, 6.4)
Poor vs. Good	49.6±9.2 43.0±9.7	**<.001**	(4.8, 8.4)
Good vs. Fair	43.0±9.7 45.1±9.8	**<.001**	(-3.3, -0.9)

Table 27 - continued

Gender					
Group	M/F (% Males)	df	X^2	p	odds ratio
Poor vs. Fair	48.1 51.6	1	.687	.407	
Poor vs. Good	48.1 58.8	1	7.534	**.006**	1.3 (1.0, 1.6)
Good vs. Fair	58.8 51.6	1	7.770	**.005**	1.5 (1.1, 2.1)

LOD			
Group	Mean (SD)	p value	CI
Poor vs. Fair	22.8±21.7 16.1±18.1	**<.001**	(2.7, 10.6)
Poor vs. Good	22.8±21.7 14.8±19.7	**<.001**	(4.4, 11.6)
Good vs. Fair	14.8±19.7 16.1±18.1	.387	

Table 28. Mean ODI Percent Change for Patients Classified as Poor, Fair, and Good Success

	Poor Success (N=41)	**Fair Success (N=86)**	**Good Success (N=214)**	**Total (N=341)**
Mean % Change, SD	18.92±18.5	22.41±24.42	26.12±27.93	24.32±26.17
Mean % Change based on pre-treatment level of severity				
Mild	17.74±20.88	26.50±28.75	18.05±27.35	19.99±27.21
Moderate	15.5±14.07	16.08±22.17	29.14±27.20	24.0±25.43
Severe	21.93±20.04	25.80±21.04	33.07±27.32	29.47±25.05

Table 29. ODI Effect Size Calculations for Percent Difference for Poor, Fair, and Good Outcome Groups

Outcome Group	*Cohen's Effect Size*
Poor (N=41)	.86
Fair (N=86)	.94
Good (N=214)	1.00
Total (N=341)	.96

Table 30. Correlations among ODI Variables and Outcome Variable (Spearman rho correlation, p value)

Variable	*Outcome*
Pre ODI	**-.115 (.033)**
Post ODI	**-.181 (< .001)**
ODI Diff	.033 (.541)
Pre x ODI Diff	.010 (.847)

Table 31.Correlations Among ODI Variables (Pearson correlation, p value)

Variable	*Pre ODI*	*Post ODI*	*ODI Diff*	*Pre x ODI Diff*
Pre ODI	1			
Post ODI	**.455 (< .001)**	1		
ODI Diff	**.532 (< .001)**	**-.531 (< .001)**	1	
Pre x ODI Diff	**.664 (< .001)**	**-.305 (< .001)**	**.931(< .001)**	1

Table 32. Logistic Regression Analysis of Outcome as a Function of Demographic and ODI Variables

Variables	χ^2 *to Remove*	*df*	*Model* χ^2	*p value*	*Tolerance*
Demographic					
Age	15.33	2		**< .001**	.931
Gender	2.435	2		.296	
LOD	3.485	2		.175	.972
All Demographic Variables		6	26.33	**< .001**	
ODI Variables					
Pre ODI	2.059	2		.357	.851
Post ODI	5.220	2		.074	.864
All Variables		10	37.407	**< .001**	
Final Reduced Model					
Age	17.024	2		**< .001**	
Post ODI	10.128	2		**.006**	
		4	29.215	**<.001**	

Table 33. Logistic Regression Analysis of Outcome as a Function of Demographic and ODI Variables: Poor vs. Good

Variables	*B*	*Wald* χ^2*-test*	*df*	*p value*	*Odds Ratio*	*95% CI for Odds Ratio*
Age	.076	14.161	1	**<.001**	1.08	(1.04, 1.12)
Post ODI	.034	8.754	1	**.003**	1.03	(1.01, 1.06)

Table 34. Logistic Regression Analysis of Outcome as a Function of Demographic and ODI Variables: Fair vs. Good

Variables	*B*	*Wald* χ^2*-test*	*df*	*p value*	*Odds Ratio*	*95% CI for Odds Ratio*
Age	.028	4.116	1	**.042**	1.03	(1.0, 1.06)
Post ODI	.017	3.625	1	.057	1.02	(1.0, 1.04)

Table 35. Mean Percent Change in ODI for Patients Classified as Poor, Fair, and Good Success in Training and Test Sample (Mean % Change, SD)

	Training Set (N=341)	**Test Set** (N=135)	**p value**
Poor Success	18.92±18.5	13.76±20.76	.875
Fair Success	22.41±24.42	18.34±24.46	.849
Good Success	26.12±27.93	24.67±26.81	.729
Total	24.32±26.17	22.06±25.99	.881

Table 36. ODI Effect Size Calculations for Percent Difference for Poor, Fair, and Good Outcome Groups in Training and Test Sample

Outcome Group	**Training Set** (N=341)	**Test Set** (N=135)
Poor	.86	.72
Fair	.94	1.09
Good	1.00	1.02
Total	.96	1.0

Table 37. Summary Table of Relevant Statistics for ODI MCID

Outcome Group	*Mean Raw Difference*	*Mean Percent Difference*	*Effect Size*
Poor	13.85±14.07	18.92±18.5	.86
Fair	12.87±15.21	22.41±24.42	.94
Good	15.04±15.89	26.12±27.93	1.00
Total	14.35±15.50	24.32±26.17	.96

Table 38. Mean Percent Change in MVAS for Patients Classified as Poor, Fair, and Good Success

	Poor Success (N=91)	**Fair Success (N=334)**	**Good Success (N=762)**	**Total (N=1187)**
Mean % Change, SD	13.76±20.76	18.34±24.46	24.67±26.81	22.06±25.99
Mean % Change based on pre-treatment level of severity				
Mild	3.93±21.37	13.70±28.88	20.10±30.49	17.76±29.96
Moderate	14.90±19.19	20.14±24.60	25.10±23.87	22.99±23.93
Severe	16.46±20.93	20.61±19.40	29.55±24.41	25.32±23.13

Table 39. Effect Size Calculations for MVAS Percent Difference for Poor, Fair, and Good Outcome Groups

Outcome Group	*Cohen's Effect Size*
Poor (N=91)	1.17
Fair (N=334)	1.03
Good (N=762)	1.45
Total (N=1187)	1.30

Table 40. Correlations Among MVAS Variables and Outcome Variable (Spearman rho Correlation, p value)

Variable	*Outcome*
Pre MVAS	**-.117 (<.001)**
Post MVAS	**-.203 (< .001)**
MVAS Diff	**.135 (< .001)**
Pre x MVAS Diff	**.117 (<.001)**

Table 41. Correlations Among MVAS Variables (Pearson correlation, p value)

Variable	*Pre MVAS*	*Post MVAS*	*MVAS Diff*	*Pre x MVAS Diff*
Pre MVAS	1			
Post MVAS	**.279 (<.001)**	1		
MVAS Diff	**.203 (<.001)**	**-.725 (<.001)**	1	
Pre x MVAS Diff	**.317 (<.001)**	**-.691 (<.001)**	**.931 (<.001)**	1

Table 42 . Logistic Regression Analysis of Outcome as a Function of Demographic and MVAS Variables

Variables	*χ^2 to Remove*	*df*	*Model χ^2*	*p*	*Tolerance*
Demographic					
Age	24.968	2		**< .001**	.971
Gender	5.788	2		.055	
LOD	3.402	2		.182	.978
All Demographic Variables		6	50.048	**< .001**	
MVAS Variables					
Pre MVAS	13.661	2		**.001**	.585
Post MVAS	1.830	2		.400	.293
MVAS Diff	4.017	2		.134	.301
All Variables		12	102.439	**< .001**	
Final Reduced Model					
Age	29.571	2		**<.001**	.988
Pre MVAS	31.379	2		**<.001**	.953
MVAS Diff	32.230	2		**<.001**	.950
		6	91.20	**<.001**	

Table 43. Logistic Regression Analysis of Outcome as a Function of Demographic and MVAS Variables: Poor vs. Good

Variables	*B*	*Wald χ^2-test*	*df*	*p value*	*Odds Ratio*	*95% CI for Odds Ratio*
Demographic						
Age	.057	24.337	1	**< .001**	1.06	(1.04, 1.08)
MVAS Variables						
Pre MVAS	.030	25.688	1	**<.001**	1.03	(1.02, 1.04)
MVAS Diff	-.025	18.602	1	**<.001**	0.98	(0.96, 0.99)

Table 44. Logistic Regression Analysis of Outcome as a Function of Demographic and MVAS Variables: Fair vs. Good

Variables	*B*	*Wald χ^2-test*	*df*	*p value*	*Odds Ratio*	*95% CI for Odds Ratio*
Demographic						
Age	.021	9.659	1	**.002**	1.02	(1.01, 1.04)
MVAS Variables						
Pre MVAS	.008	6.755	1	**.009**	1.01	(1.0, 1.01)
MVAS Diff	-.011	15.320	1	<**.001**	0.99	(0.98, 1.0)

Table 45. Mean Percent Change in MVAS for Patients Classified as Poor, Fair, and Good Success in Training and Test Sample (Mean % Change, SD)

	Training Set (N=1,187)	**Test Set** (N=528)	**p value**
Poor Success	13.76±20.76	5.99±21.87	.037
Fair Success	18.34±24.46	16.71±22.47	.186
Good Success	24.67±26.81	23.16±26.68	.961
Total	22.06±25.99	19.79±25.64	.178

Table 46. MVAS Effect Size Calculations for Percent Difference for Poor, Fair, and Good Outcome Groups in Training and Test Sample

Outcome Group	**Training Set** (N=1,187)	**Test Set** (N=528)
Poor	1.17	.42
Fair	1.03	1.16
Good	1.45	1.21
Total	1.30	1.1

Table 47. Percent of Good Cases in the Test Set Classified Correctly Utilizing Various MCID Cut-points (n, %)

MCID Cut-off		*Prediction*		
	Actual		Yes	No
≥19		True	166 (78.3)	150 (47.5)
		False	46 (21.7)	166 (52.5)
	Actual		Yes	No
≥25		True	146 (76.0)	150 (44.6)
		False	46 (24.0)	186 (55.4)

Table 48. Summary table of relevant statistics for MVAS MCID

Outcome Group	*Mean Raw Difference*	*Mean Percent Difference*	*Effect Size*
Poor	21.65±30.87	13.76±20.76	1.17
Fair	24.39±24.36	18.34±24.46	1.03
Good	32.33±31.45	24.67±26.81	1.45
Total	29.28±36.72	22.06±25.99	1.30

Table 49. Mean Percent Change in SF-36 PHS for Patients Classified as Poor, Fair, and Good Success

Physical Component Summary (PHS)	**Poor Success (N=73)**	**Fair Success (N=150)**	**Good Success (N=404)**	**Total (N=627)**
Mean % Change, SD	7.69±15.96	7.96±13.37	8.34±11.62	8.18±12.6
Mean % Change based on pre-treatment level of severity				
Mild (n=193)	16.69±17.72	14.33±12.92	15.74±10.60	15.49±12.34
Moderate (n=235)	2.30±13.66	8.13±11.33	9.68±9.64	8.48±10.77
Severe (n=199)	0.05±9.30	-1.77±11.25	1.44±10.34	0.72±10.46

Table 50. Effect Size Calculations for SF-36 PHS Percent Difference for Poor, Fair, and Good Outcome Groups

Outcome Group	*Effect size*
Poor (N=73)	.84
Fair (N=150)	.90
Good (N=404)	1.02
Total (N=376)	.96

Table 51. Correlations among SF-36 PHS Variables and Outcome Variable (Spearman rho Correlation, p value)

Variable	Outcome
Pre PHS SF-36	**.150 (< .001)**
Post PHS SF-36	**.136 (<.001)**
PHS SF-36 Diff	.032 (.429)
Pre x PHS SF-36 Diff	.053 (.189)

Table 52. Correlations among SF-36 PHS Variables (Pearson correlation, p value)

Variable	Pre PHS SF-36	Post PHS SF-36	PHS SF-36 Diff	Pre x PHS SF-36 Diff
Pre PHS SF-36	1			
Post PHS SF-36	**.205 (< .001)**	1		
PHS SF-36 Diff	**-.579 (<.001)**	**.658 (<.001)**	1	
Pre x PHS SF-36 Diff	**-.491 (<.001)**	**.710 (<.001)**	**.929 (<.001)**	1

Table 53 . Logistic Regression Analysis of Outcome as a Function of Demographic and SF-36 PHS Variables

Variables	*χ^2 to Remove*	*df*	*Model χ^2*	*p*	*Tolerance*
Demographic					
Age	17.721	2		**<.001**	.936
Gender	6.501	2		.039	
LOD	5.312	2		.070	.956
All Demographic Variables		6	40.113	**<.001**	
PHS SF-36 Variables					
Pre SF-36 PHS	8.675	2		.013	.952
Post SF-36 PHS	4.623	2		.099	.946
All Variables		10	55.077	**<.001**	
Final Reduced Model					
Age	25.087	2		**< .001**	.990
Gender	7.166	2		.028	
Pre SF-36 PHS	9.952	2		**.007**	.990
		6		**< .001**	

Table 54. Logistic Regression Analysis of Outcome as a Function of Demographic and SF-36 PHS Variables: Poor vs. Good

Variables	*B*	*Wald χ^2-test*	*df*	*p value*	*Odds Ratio*	*95% CI for Odds Ratio*
Demographic						
Age	.068	21.357	1	**<.001**	1.07	(1.04, 1.10)
Gender	.552	4.346	1	.037		
PHS SF-36 Variables Pre SF-36 PHS	-0.056	5.703	1	**.017**	0.95	(0.90, 0.99)

Table 55. Logistic Regression Analysis of Outcome as a Function of Demographic and SF-36 PHS Variables: Fair vs. Good

Variables	*B*	*Wald χ^2-test*	*df*	*p value*	*Odds Ratio*	*95% CI for Odds Ratio*
Demographic						
Age	.027	6.456	1	**.011**	1.03	(1.01, 1.05)
Gender	.408	4.386	1	.036		
Pre SF-36 PHS	-0.043	6.105	1	**.013**	0.96	(0.93, 0.99)

Table 56. Mean Percent Change in SF-36 PHS for Patients Classified as Poor, Fair, and Good Success in Training and Test Sample (Mean % Change, SD)

	Training Set (N=627)	**Test Set** (N=275)	**p value**
Poor Success	7.69±15.96	2.82±12.15	.553
Fair Success	7.96±13.37	8.52±13.32	.596
Good Success	8.34±11.62	12.12±13.69	.702
Total	8.18±12.6	10.01±13.74	.444

Table 57. SF-36 PHS Effect Size Calculations for Percent Difference for Poor, Fair, and Good Outcome Groups in Training and Test Sample

Outcome Group	**Training Set** (N=627)	**Test Set** (N=275)
Poor	.84	.5
Fair	.90	.7
Good	1.02	1.38
Total	.96	1.1

Table 58. Summary Table of Relevant Statistics for SF-36 PHS MCID

Outcome Group	*Mean Raw Difference*	*Mean Percent Difference*	*Effect Size*
Poor	5.04±8.85	7.69±15.96	.84
Fair	5.41±9.27	7.96±13.37	.90
Good	5.79±8.22	8.34±11.62	1.02
Total	5.62±8.55	8.18±12.6	.96

Table 59. Mean Percent Change in SF-36 MHS for Patients Classified as Poor, Fair, and Good Success

Mental Component Summary (MHS)	**Poor Success (N=73)**	**Fair Success (N=150)**	**Good Success (N=404)**	**Total (N=627)**
Mean % Change, SD	7.49±12.26	7.95±14.66	10.09±14.28	9.28±14.17
Mean % Change based on pre-treatment level of severity				
Mild	14.31±11.86	18.37±15.27	19.46±11.35	18.56±12.51
Moderate	8.33±8.51	6.43±10.79	9.89±8.63	8.89±9.27
Severe	0.29±11.27	-1.49±9.93	-0.50±9.78	-0.63±9.98

Table 60. Effect size calculations for SF-36 MHS Percent Difference for Poor, Fair, and Good Outcome Groups

Outcome Group	*Effect Size*
Poor (N=73)	.69
Fair (N=150)	.74
Good (N=404)	.84
Total (N=627)	.80

Table 61. Correlations among SF-36 MHS Variables and Outcome Variable (Spearman rho Correlation, p value)

Variable	*Outcome*
Pre MHS SF-36	.007 (.857)
Post MHS SF-36	**.086 (.031)**
MHS SF-36 Diff	.071 (.076)
Pre x MHS SF-36 Diff	.072 (.071)

Table 62. Correlations among SF-36 MHS Variables (Pearson correlation, p value)

Variable	*Pre MHS SF-36*	*Post MHS SF-36*	*MHS SF-36 Diff*	*Pre x MHS SF-36 Diff*
Pre MHS SF-36	1			
Post MHS SF-36	**.301(< .001)**	1		
MHS SF-36 Diff	**-.666(< .001)**	**.491(< .001)**	1	
Pre x MHS SF-36 Diff	**-.571(< .001)**	**.581(< .001)**	**.945(< .001)**	1

Table 63. Logistic Regression Analysis of Outcome as a Function of Demographic and SF-36 MHS Variables

Variables	χ^2 *to Remove*	*df*	*Model* χ^2	*p*	*Tolerance*
Demographic					
Age	23.698	2		**<.001**	.952
Gender	5.962	2		.051	
LOD	4.832	2		.089	.956
All Demographic Variables		6	40.113	**<.001**	
MHS SF-36 Variables					
Pre SF-36 MHS	0.644	2		.725	.906
Post SF-36 MHS	6.889	2		.032	.906
All Variables		10	47.065		
Final Reduced Model					
Age	28.229	2		**<.001**	
Gender	6.332	2		.042	
		4	35.895	**< .001**	

Table 64. Logistic Regression Analysis of Outcome as a Function of Demographic and SF-36 MHS Variables: Poor vs. Good

Variables	*B*	*Wald* χ^2*-test*	*df*	*p value*	*Odds Ratio*	*95% CI for Odds Ratio*
Demographic						
Age	0.071	23.565	1	**<.001**	1.07	(1.04, 1.10)
Gender	0.513	3.813	1	.051	1.67	(1.0, 2.8)

Table 65. Logistic Regression Analysis of Outcome as a Function of Demographic and SF-36 MHS Variables: Fair vs. Good

Variables	*B*	*Wald* χ^2*-test*	*df*	*p value*	*Odds Ratio*	*95% CI for Odds Ratio*
Demographic						
Age	0.029	7.606	1	**.006**	1.03	(1.01, 1.05)
Gender	0.380	3.865	1	.049	1.46	(1.0, 2.14)

Table 66. Mean Percent Change in SF-36 MHS for Patients Classified as Poor, Fair, and Good Success in Training and Test Sample (Mean % Change, SD)

	Training Set (N=630)	**Test Set** (N=275)	**p value**
Poor Success	7.49±12.26	5.63±19.49	.533
Fair Success	7.95±14.66	6.97±16.68	.596
Good Success	10.09±14.28	9.66±15.55	.702
Total	9.28±14.17	8.47±16.41	.444

Table 67. SF-36 MHS Effect Size Calculations for Percent Difference for Poor, Fair, and Good Outcome Groups in Training and Test Sample

Outcome Group	**Training Set** (N=630)	**Test Set** (N=275)
Poor	.69	.34
Fair	0.74	.60
Good	.84	.78
Total	.80	.67

Table 68. Summary Table of Relevant Statistics for SF-36 MHS MCID

Outcome Group	*Mean Raw Difference*	*Mean Percent Difference*	*Effect Size*
Poor	6.36±10.52	7.49±12.26	.69
Fair	6.83±11.98	7.95±14.66	0.74
Good	8.09±10.85	10.09±14.28	.84
Total	7.59±11.10	9.28±14.17	.80

Table 69. Mean Percent Change in PDQ for Patients Classified as Poor, Fair, and Good Success

	Poor Success (N=41)	**Fair Success (N=47)**	**Good Success (N=175)**	**Total (N=263)**
Mean % Change, SD	20.59±20.21	26.0±22.14	27.41±25.32	26.10±24.08
Mean % Change based on pre-treatment level of severity				
Mild	22.38±28.67	33.58±31.71	21.10±29.16	23.21±29.52
Moderate	16.0±10.07	25.37±20.27	33.26±23.21	29.99±22.25
Severe	21.91±19.69	20.61±11.85	28.19±20.22	25.03±18.76

Table 70. Effect Size Calculations for PDQ Percent Difference for Poor, Fair, and Good Outcome Groups

Outcome Group	*Cohen's Effect Size*
Poor (N=41)	1.36
Fair (N=47)	1.74
Good N=175)	1.44
Total (N=263)	1.46

Table 71. Correlations Among PDQ Variables and Outcome Variable (Spearman rho Correlation, p value)

Variable	*Outcome*
Pre PDQ	**-.156 (.011)**
Post PDQ	**-.197 (.001)**
PDQ Diff	.113 (.068)
Pre x PDQ Diff	.068 (.270)

Table 72. Correlations among PDQ Variables (Pearson correlation, p value)

Variable	*Pre PDQ*	*Post PDQ*	*PDQ Diff*	*Pre x PDQ Diff*
Pre PDQ	1			
Post PDQ	**.402**	1		
PDQ Diff	**.166**	**-.798**	1	
Pre x PDQ Diff	**.343**	**-.688**	**.920**	1

Table 73. Logistic Regression Analysis of Outcome as a Function of Demographic and PDQ Variables

Variables	*χ^2 to Remove*	*df*	*Model χ^2*	*p*	*Tolerance*
Demographic					
Age	16.542	2		**<.001**	.935
Gender	2.262	2		.323	
LOD	5.275	2		.072	.951
All Demographic Variables		6	30.013	**<.001**	
PDQ Variables					
Pre PDQ	2.537	2		.281	.983
Post PDQ	3.026	2		.220	.983
All Variables		10	39.695	**<.001**	
Final Reduced Model					
Age	15.623	2		**< .001**	
LOD	4.819	2		.090	
Post PDQ	7.554	2		**.023**	
		6	34.280	**< .001**	

Table 74. Logistic Regression Analysis of Outcome as a Function of Demographic and PDQ Variables: Poor vs. Good

Variables	*B*	*Wald χ^2-test*	*df*	*p value*	*Odds Ratio*	*95% CI for Odds Ratio*
Demographic						
Age	0.075	12.946	1	**<.001**	1.08	(1.04, 1.12)
LOD	0.017	4.101	1	.043	1.02	(1.0, 1.03)
Post PDQ	0.018	7.037	1	**.008**	1.02	(1.01, 1.03)

Table 75. Logistic Regression Analysis of Outcome as a Function of Demographic and PDQ Variables: Fair vs. Good

Variables	*B*	*Wald χ^2-test*	*df*	*p value*	*Odds Ratio*	*95% CI for Odds Ratio*
Demographic						
Age	0.034	3.628	1	.057		
LOD	0.013	2.297	1	.130		
Post PDQ	0.007	1.269	1	.260		

Table 76. Mean Percent Change in PDQ for Patients Classified as Poor, Fair, and Good Success in Training and Test Sample (Mean % Change, SD)

	Training Set (N=263)	**Test Set** (N=132)	**p value**
Poor Success	20.59±20.21	25.11±18.44	.256
Fair Success	26.0±22.14	28.50±19.18	.460
Good Success	27.41±25.32	30.97±23.08	.240
Total	26.10±24.08	29.71±21.76	.089

Table 77. PDQ Effect Size Calculations for Percent Difference for Poor, Fair, and Good Outcome Groups in Training and Test Sample

Outcome Group	**Training Set** (N=263)	**Test Set** (N=132)
Poor	1.36	1.9
Fair	1.74	2.2
Good	1.44	1.73
Total	1.46	1.8

Table 78. Summary Table of Relevant Statistics for PDQ MCID

Outcome Group	*Mean Raw Difference*	*Mean Percent Difference*	*Effect Size*
Poor	32.17±26.80	20.59±20.21	1.36
Fair	36.83±24.01	26.0±22.14	1.74
Good	19.85±29.00	27.41±25.32	1.44
Total	36.44±29.49	26.10±24.08	1.46

Table 79. Mean Percent Change in PI for Patients Classified as Poor, Fair, and Good Success

	Poor Success (N=186)	**Fair Success (N=497)**	**Good Success (N=1224)**	**Total (N=1907)**
Mean % Change, SD	14.39±24.01	17.0±24.87	23.47±30.19	20.89±28.53
Mean % Change based on pre-treatment level of severity				
Mild	7.70±27.37	9.75±27.65	18.2±32.79	15.53±31.56
Moderate	17.61±24.03	20.84±20.61	28.95±26.33	25.38±24.95
Severe	18.74±14.36	28.10±20.61	32.49±21.74	28.74±20.82

Table 80. Effect Size calculations for PI Percent Difference for Poor, Fair, and Good Outcome Groups

Outcome Group	*Cohen's Effect Size*
Poor (N=186)	.85
Fair (N=497)	.97
Good (N=1224)	1.14
Total (N=1907)	1.06

Table 81. Correlations Among PI Variables and Outcome Variable (Spearman rho Correlation, p value)

Variable	*Outcome*
Pre PI	**-.131(<.001)**
Post PI	**-.204(<.001)**
PI Diff	**.121(<.001)**
Pre x PI Diff	**.105(<.001)**

Table 82. Correlations Among PI Variables (Pearson Correlation, p value)

Variable	*Pre PI*	*Post PI*	*PI Diff*	*Pre x PI Diff*
Pre PI	1			
Post PI	**.294(<.001)**	1		
PI Diff	**.303(<.001)**	**-.758(<.001)**	1	
Pre x PI Diff	**.404(<.001)**	**-.714(<.001)**	**.920(<.001)**	1

Table 83. Logistic Regression Analysis of Outcome as a Function of Demographic and PI Variables

Variables	*χ^2 to Remove*	*df*	*Model χ^2*	*P*	*Tolerance*
Demographic					
Age	53.427	2		**<.001**	.983
Gender	4.332	2		.115	
LOD	8.073	2		.018	.983
		6	75.042	**<.001**	
All Demographic Variables					
PI Variables					
Pre PI	8.151	2		.017	
Post PI	65.194	2		**<.001**	
All Variables		10	169.886	**<.001**	
Final Reduced Model					
Age	55.020	2		**< .001**	
LOD	8.148	2		**.017**	.915
Pre PI	9.236	2		**.010**	.915
Post PI	64.326	2		**< .001**	
		8	165.551	**< .001**	

Table 84. Logistic Regression Analysis of Outcome as a Function of Demographic and PI Variables: Poor vs. Good

Variables	*B*	*Wald χ^{2-} test*	*df*	*p value*	*Odds Ratio*	*95% CI for Odds Ratio*
Demographic						
Age	0.058	47.826	1	**<.001**	1.06	(1.04, 1.08)
LOD	0.009	8.675	1	**.003**	1.01	(1.0, 1.02)
PI Variables						
Pre PI	0.137	7.006	1	**.008**	1.14	(1.03, 1.25)
Post PI	0.273	44.042	1	**<.001**	1.32	(1.21, 1.43)

Table 85. Logistic Regression Analysis of Outcome as a Function of Demographic and PI Variables: Fair vs. Good

Variables	*B*	*Wald χ^2-test*	*df*	*p value*	*Odds Ratio*	*95% CI for Odds Ratio*
Demographic						
Age	0.021	14.545	1	**<.001**	1.02	(1.01, 1.03)
LOD	0.004	2.023	1	.155		
PI Variables						
Pre PI	0.059	3.837	1	.050	1.06	(1.0, 1.13)
Post PI	0.145	30.017	1	**<.001**	1.15	(1.10, 1.22)

Table 86. Mean Percent Change in PI for Patients Classified as Poor, Fair, and Good Success in Training and Test Sample (Mean % Change, SD)

	Training Set (N=1,949)	**Test Set** (N=874)	**p value**
Poor Success	14.39±24.01	8.51±26.59	.291
Fair Success	17.0±24.87	17.92±25.13	.509
Good Success	23.47±30.19	22.77±29.65	.901
Total	20.89±28.53	30.65±28.56	.969

Table 87. PI Effect Size Calculations for Percent Difference for Poor, Fair, and Good Outcome Groups in Training and Test Sample

Outcome Group	**Training Set** (N=1,949)	**Test Set** (N=874)
Poor	.85	.66
Fair	.97	1.12
Good	1.14	1.10
Total	1.06	1.07

Table 88. Summary Table of Relevant Statistics for PI MCID

Outcome Group	*Mean Raw Difference*	*Mean Percent Difference*	*Effect Size*
Poor	1.56±2.44	14.39±24.01	.85
Fair	1.82±2.33	17.0±24.87	.97
Good	2.17±2.47	23.47±30.19	1.14
Total	2.02±2.44	20.89±28.53	1.06

Table 89. Mean Percent Change in BDI for Patients Classified as Poor, Fair, and Good Success

	Poor Success (N=181)	**Fair Success (N=516)**	**Good Success (N=1278)**	**Total (N=1975)**
Mean % Change, SD	25.09±34.74	29.64±36.08	37.81±39.94	34.51±38.77
Mean % Change based on pre-treatment level of severity				
Mild	11.19±40.74	22.03±42.99	30.45±48.79	27.0±47.18
Moderate	25.23±29.62	31.93±33.21	41.62±34.36	37.68±34.11
Severe	34.63±30.57	35.15±29.33	43.27±30.22	39.96±30.25

Table 90. Effect Size Calculations for BDI Percent Difference for Poor, Fair, and Good Outcome Groups

Outcome Group	*Cohen's Effect Size*
Poor (N=181)	.62
Fair (N=517)	.67
Good (N=1278)	.76
Total (N=1976)	.72

Table 91. Correlations Among BDI Variables and Outcome Variable (Spearman rho Correlation, p value)

Variable	*Outcome*
Pre BDI	**-.102 (<.001)**
Post BDI	**-.189 (<.001)**
BDI Diff	**.130 (<.001)**
Pre x BDI Diff	**.066 (.003)**

Table 92. Correlations Among BDI SF-36 Variables (Pearson correlation, p value)

Variable	*Pre BDI*	*Post BDI*	*BDI Diff*	*Pre x BDI Diff*
Pre BDI	1			
Post BDI	**.447(<.001)**	1		
BDI Diff	**.161(<.001)**	**-.578(<.001)**	1	
Pre x BDI Diff	**.582(<.001)**	**-.349(<.001)**	**.713(<.001)**	1

Table 93. Logistic Regression Analysis of Outcome as a Function of Demographic and BDI Variables

Variables	χ^2 *to Remove*	*df*	*Model* χ^2	*p*	*Tolerance*
Demographic					
Age	68.080	2		**<.001**	.978
Gender	5.444	2		.066	
LOD	8.840	2		.012	.970
All Demographic Variables		6	99.963	**<.001**	
BDI Variables					
Pre BDI	14.990	2		**<.001**	.546
Post BDI	.531	2			.372
BDI Diff	13.519	2			.451
All Variables		12	154.396	**<.001**	
Final Reduced Model					
Age	70.160	2		**< .001**	.978
LOD	8.912	2		**.012**	.970
Pre BDI	33.515	2		**< .001**	.966
BDI Diff	34.736	2		**< .001**	.973
		8		**< .001**	

Table 94. Percent of People in the Mild, Moderate, and Severe Pre-treatment BDI with Poor, Fair, and Good Outcomes

Outcome	*Poor* (n=181)	*Fair* (n=516)	*Good* (n=1278)
Pre-treatment level of severity			
Mild (N, %) n=717	52 (7.3)	175 (24.4)	490 (68.3)
Moderate (N, %) n=643	54 (8.4)	170 (26.4)	419 (65.2)
Severe (N, %) n=615	75 (12.2)	171 (27.8)	369 (60.0)

Table 95. Logistic Regression Analysis of Outcome as a Function of Demographic and BDI Variables: Poor vs. Good

Variables	*B*	*Wald χ^{2}-test*	*df*	*p value*	*Odds Ratio*	*95% CI for Odds Ratio*
Demographic						
Age	.068	61.382	1	**<.001**	1.07	(1.05, 1.09)
LOD	.010	8.165	1	**.004**	1.01	(1.0, 1.02)
BDI						
Variables	.040	27.060	1	**<.001**	1.04	(1.03, 1.06)
Pre BDI	-.011	20.861	1	**<.001**	0.99	(0.99, 0.99)
BDI Diff						

Table 96. Logistic Regression Analysis of Outcome as a Function of Demographic and BDI Variables: Fair vs. Good

Variables	*B*	*Wald χ^{2}-test*	*df*	*p value*	*Odds Ratio*	*95% CI for Odds Ratio*
Demographic						
Age	.022	16.098	1	**<.001**	1.02	(1.01, 1.03)
LOD	.002	0.262	1	.608	1.0	(1.0, 1.01)
BDI						
Variables	.020	14.576	1	**<.001**	1.02	(1.01, 1.03)
Pre BDI	-.006	19.763	1	**<.001**	0.99	(0.99, 1.0)
BDI Diff						

Table 97. Mean Percent Change in BDI for Patients Classified as Poor, Fair, and Good Success in Training and Test Sample (Mean % Change, SD)

	Training Set (N=1,976)	**Test Set** (N=828)	**p value**
Poor Success	25.09±34.74	26.48±31.16	.541
Fair Success	29.64±36.08	26.15±37.82	.184
Good Success	37.81±39.94	34.76±42.05	.175
Total	34.51±38.77	31.74±40.19	.098

Table 98. BDI Effect Size Calculations for Percent Difference for Poor, Fair, and Good Outcome Groups in Training and Test Sample

Outcome Group	**Training Set** (N=1,976)	**Test Set** (N=828)
Poor	.62	.69
Fair	.67	.65
Good	.76	.66
Total	.72	.66

Table 99. Percent of Good Cases in the Test Set Classified Correctly Utilizing Various MCID Cut-points (n, %)

MCID Cut-off		*Prediction*		
			Yes	No
≥30	*Actual*	True	290 (70.4)	172 (41.3)
		False	122 (29.6)	244 (58.7)
≥38	*Actual*	True	226 (70.0)	197 (29.0)
		False	97 (30.0)	308 (61.0)

Table 100. Summary Table of Relevant Statistics for BDI MCID

Outcome Group	*Mean Raw Difference*	*Mean Percent Difference*	*Effect Size*
Poor	7.11±10.49	25.09±34.74	.62
Fair	6.87±9.74	29.64±36.08	.67
Good	7.69±10.81	37.81±39.94	.76
Total	7.42±10.51	34.51±38.77	.72

Table 101. Summary Table of MCID Results for all Measures

Measure	*Correlation with Outcome (r, p value)*	*Regression p value*
ODI		
Pre	**-.115 (.033)**	N.S.
Post	**-.181 (<.001)**	**.006**
Diff	.033 (.541)	N.S.
PrexDiff	.010 (.847)	N.S.
MVAS		
Pre	**-.117 (<.001)**	**<.001**
Post	**-.203 (<.001)**	N.S.
Diff	**-.135 (<.001)**	**<.001**
PrexDiff	**.117(<.001)**	N.S.
SF-36		
PHS		
Pre	**.150 (<.001)**	**.007**
Post	**.136 (<.001)**	N.S.
Diff	.032 (.429)	N.S.
PrexDiff	.053 (.189)	N.S.
MHS		
Pre	.007 (.857)	N.S.
Post	**.086 (.031)**	N.S.
Diff	.071 (.076)	N.S.
PrexDiff	.072 (.071)	N.S.
PDQ		
Pre	**-.156 (.011)**	N.S.
Post	**-.197 (.001)**	**.023**
Diff	.113 (.068)	N.S.
PrexDiff	.068 (.270)	N.S.
PI		
Pre	**-.131 (<.001)**	**.010**
Post	**-.204 (<.001)**	**<.001**
Diff	**.121 (<.001)**	N.S.
PrexDiff	**.105 (<.001**)	N.S.
BDI		
Pre	**-.102 (<.001)**	**<.001**
Post	**-.189 (<.001)**	N.S.
Diff	**.130 (<.001)**	**<.001**
PrexDiff	**.066 (<.001)**	N.S.

Table 102. Summary Table of Relevant Means, Standard Deviations, and Effect Sizes for All Measures (*, MCID raw difference, **MCID percent difference)

Measure	*Mean Raw Difference*	*Mean Percent Difference*	*Effect Size*
ODI			
Poor	13.85±14.07	18.92±18.5	.86
Fair	**12.87±15.21***	**22.41±24.42****	.94
Good	15.04±15.89	26.12±27.93	1.0
Total	14.35±15.50	24.32±26.17	.96
MVAS			
Poor	21.65±30.87	13.76±20.76	1.17
Fair	**24.39±47.16***	**18.34±24.46****	1.03
Good	32.33±31.45	24.67±26.81	1.45
Total	29.28±36.72	22.06±25.99	1.30
SF-36			
PHS			
Poor	5.04±8.85	7.69±15.96	.84
Fair	**5.41±9.27***	**7.96±13.37****	.90
Good	5.79±8.22	8.34±11.62	1.02
Total *MHS*	5.62±8.55	8.18±12.6	.96
Poor			
Fair	6.36±10.52	7.49±12.26	.69
Good	**6.83±11.98***	**7.95±14.66****	.74
Total	8.09±10.85	10.09±14.28	.84
	7.59±11.10	9.28±14.17	.80
PDQ			
Poor	32.17±26.80	20.59±20.21	1.36
Fair	**36.83±24.01***	**26.0±22.14****	1.74
Good	19.85±29.00	27.41±25.32	1.44
Total	36.44±29.49	26.10±24.08	1.46
PI			
Poor	1.56±2.44	14.39±24.01	.66
Fair	**1.82±2.33***	**17.0±24.87****	1.12
Good	2.17±2.47	23.47±30.19	1.10
Total	2.02±2.44	20.89±28.53	1.07
BDI			
Poor	7.11±10.49	25.09±34.74	.62
Fair	**6.87±9.74**	**29.64±36.08****	.67
Good	7.69±10.81	37.81±39.94	.76
Total	7.42±10.51	34.51±38.77	.72

Table 103. Percent of Good Cases in the Test Set Classified Correctly Utilizing Various MCID Age Cut-points (n, %)

MCID Cut-off		*Prediction*		
			Yes	No
≥49	*Actual*	True	358 (67.41)	121 (40.7)
		False	173 (32.59)	176 (59.3)

REFERENCES

American Psychological Association (1985). Standards for education and psychological testing. Washington D. C., American Psychological Association.

Anagnostis, C., R. J. Gatchel, et al. (2004). "The pain disability questionnaire (PDQ); A new psychometrically sound measure for chronic musculoskeletal disorders." Spine **29**: 2290-2302.

Anagnostis, C., T. G. Mayer, et al. (2003). "The Million Visual Analog Scale: Its utility for predicting tertiary rehabilitation outcomes." Spine **28**: 1-10.

Anderson, G. B. J., M. H. Pope, et al. (1991). Occupational Low Back Pain: Assessment, Treatment, and Prevention. Epidemiology and Cost. M. H. Pope, G. B. J. Anderrson, J. W. Frymoyer and D. B. Chaffin. St. Louis, Mosby Year Book.

APA, A. P. A. (2000). Diagnostic and statistical manual of mental disorders. Washington, DC, American Psychiatric Press, Inc.

Atlas, S. and R. A. Deyo (2001). "Evaluating and managing acute low back pain in the primary care setting." Journal of General Internal Medicine **16**: 120-131.

Banks, S. M. and R. D. Kerns (1996). "Explaining the high rates of depression in chronic pain: A diathesis-stress framework." Psychological Bulletin **119**(1): 95-110.

Beaton, D., M. Boers, et al. (2002). "Many faces of the minimal clinically important difference (MCID): a literature review and directions for future research." Current Opinion in Rheumatology **14**: 109-114.

Beaton, D., C. Bombadier, et al. (2001). "Looking for important change/difference in studies of responsiveness." The Journal of Rheumatology **28**(2): 400-405.

Beck, A. (1967). Beck Depression Inventory.

Beck, A. T., A. J. Rush, et al. (1979). Cognitive Therapy of Depression. New York, Guilford Press.

Beck, A. T., R. A. Steer, et al. (1996). Beck Depression Inventory Manual. San Antonio. TX, Psychological Corporation.

Beck, A. T., R. A. Steer, et al. (1988). "Psychometric properties of the Beck Depression Inventory: Twenty-five years of evaluation." Clinical Psychology Review **8**: 77-100.

Bendix, A. E., T. Bendix, et al. (1998). "A prospective, randomized 5-year follow-up study of functional restoration in chronic low back pain patients." Eur.Spine J. **7**(2): 111-119.

Beurskens, A. J., H. C. de Vet, et al. (1999). "A patient-specific approach for measuring functional status in low back pain." J.Manipulative Physiol Ther. **22**(3): 144-148.

Beurskens, A. J., H. C. de Vet, et al. (1995). "Measuring the functional status of patients with low back pain. Assessment of the quality of four disease-specific questionnaires." Spine **20**(9): 1017-1028.

Bigos, S. J., M. C. Battie, et al. (1991). "A prospective study of work perceptions and psychosocial factors affecting the report of back injury." Spine **16**(1): 1-6.

Blancett, S. and D. L. Flarey, Eds. (1998). Health Care Outcomes: Collaborative, Path-Based Approaches. Gaithersburg, Maryland, Aspen.

Blyth, F. M., L. M. March, et al. (2001). "Chronic pain in Australia: a prevalence study." Pain **89**: 127-134.

Boersma, K. and S. J. Linton (2006). "Psychological processes underlying the development of a chronic pain problem: A prospective study of the relationship between profiles of psychological variables in the fear-avoidance model in disability." Clin J Pain **22**: 160-166.

Bowling, A. (1991). Measuring Health: A review of quality of life measures. Buckingham, Biddles Limited, Guildford and King's Lynn.

Brady, S., T. Mayer, et al. (1994). "Physical progress and residual impairment quantification after functional restoration. Part II: Isokinetic trunk strength." Spine **19**(4): 395-400.

Brady, W., J. Bass, et al. (1997). "Defining total corporate health and safety costs--significance and impact." Journal of Occupational and Environmental Medicine **39**: 227-231.

Brazier, J., J. Roberts, et al. (2002). "The estimation of a preference-based measure of health on the SF-36." Journal of Health Economics **21**: 271-292.

Brazier, J. E., R. Harper, et al. (1992). "Validating the SF-36 health survey questionnaire: New outcome measure for primary care." British Medical Journal **305**: 160-164.

Capra, P., T. G. Mayer, et al. (1985). "Adding psychological scales to your back pain assessment." The Journal of Musculoskeletal Medicine **2**: 41-52.

Carlsson, A. (1983). "Assessment of chronic pain: I. Aspects of the reliability and validity of the visual analog scale." Pain **16**: 87-101.

Cassisi, J. E., G. W. Sypert, et al. (1989). "Independent evaluation of a multidisciplinary rehabilitation program for chronic low back pain." Neurosurgery **25**: 877-83.

Cohen, J. (1988). Statistical Power Analyses for the Behavioral Sciences. Hillsdale, Hive and Landen.

Cornwall, A. and D. C. Donderi (1988). "The effect of experimentally induced anxiety on the experience of pressure pain." Pain **35**(1): 105-113.

Cousins, M. J. and I. Power (1999). Acute and postoperative pain. Textbook of Pain. P. Wall and R. Melzack. Boston, Science Press.

Crombez, G., J. W. S. Vlaeyen, et al. (1999). "Pain-related fear is more disabling than pain itself: Evidence on the role of pain-related fear in chronic back pain disability." Pain **80**(1-2): 329-339.

Cronbach, L. J. and L. Furby (1970). "How should we measure "change"-Or should we?" Psychol Bull **74**: 68-80.

Curtis, L., T. G. Mayer, et al. (1994). "Physical progress and residual impairment quantification after functional restoration. Part III: Isokinetic and isoinertial lifting capacity." Spine **19**: 401-5.

Dempsey, P. G., A. Burdorf, et al. (1997). "The influence of personal variables on work-related low-back disorders and implications for future research." Journal of Occupational and Environmental Medicine **39**: 748-759.

Dersh, J., R. J. Gatchel, et al. (2006). "Prevalence of psychiatric disorders in patients with chronic disabling occupational spinal disorders." Spine **31**: 1156-1182.

Dersh, J., R. J. Gatchel, et al. (2002). "Prevalence of psychiatric disorders in patients with chronic work-related musculoskeletal pain disability." Journal of Occupational and Environmental Medicine **44**: 459-488.

Dersh, J. A. (2000). Comprehensive Evaluation of Psychopathology in Chronic Musculoskeletal Pain Disability Patients. Division of Psychology. Dallas, TX, The University of Texas Southwestern Medical Center at Dallas.

Dersh, J. A., R. J. Gatchel, et al. (2001). "Chronic spinal disorders and psychopathology: Research findings and theoretical considerations." The Spine Journal **1**: 88-94.

Deyo, R. A. (1988). "Measuring the functional status of patients with low back pain." Archives of Physical Medicine & Rehabilitation **69**: 1044-53.

Deyo, R. A., G. Andersson, et al. (1994). "Outcome measures for studying patients with low back pain." Spine **19**: 2032S-36S.

Deyo, R. A., D. Cherkin, et al. (1991). "Cost, controversy, crisis: Low back pain and the health of the public." Annual Review of Public Health **12**: 141-156.

Deyo, R. A. and A. K. Diehl (1988). "Psychosocial predictors of disability in patients with low back pain." Journal of Rheumatology **15**: 1557-1564.

Fairbanks, J. C., J. Couper, et al. (1980). "The Oswestry low back pain disability questionnaire." Physiotherapy **66**: 271-273.

Fardon, D. (1997). Differential diagnosis of low back disorders. Principles of classification. The adult spine: Principles and Practice. J. Frymoyer, T. B. Ducker, N. Hadleret al. Philadelphia, Lippincott-Raven**:** 1745-1768.

Fayers, P. M. and D. Machen (2001). Quality of Life: Assessment, Analysis, and Interpretation. New York, John Wiley and Sons, LTD.

Finger, S. (1994). Origins of Neuroscience. New York, Oxford University Press.

Fishbain, D. A., R. Cutler, et al. (1997). "Chronic pain-associated depression: Antecedent or consequence of chronic pain? A review." Clinical Journal of Pain **13**: 116-137.

Fishbain, D. A., M. Goldberg, et al. (1986). "Male and female chronic pain patients categorized by DSM-III psychiatric diagnostic criteria." Pain **26**: 181-197.

Fisher, K. and M. Johnson (1997). "Validation of the Oswestry low back pain disability questionnaire, its sensitivity as a measure of change following treatment and its relationship with other aspects of the chronic pain experience." Physiotherapy Theory and Practice **13**: 67-80.

Fitzpatrick, R., A. Fletcher, et al. (1992). "Quality of life measures in health care. I: Applications and issues in assessment." Br. Med. J **305**: 1074-1077.

Flores, L., R. J. Gatchel, et al. (1997). "Objectification of functional improvement after nonoperative care." Spine **22**(14): 1622-33.

Franklin, G., J. Haug, et al. (1994). "Outcome of lumbar fusion in Washington state worker's compensation." Spine **17**: 1897-904.

Garcy, P., T. G. Mayer, et al. (1996). "Recurrent or new injury outcomes after return to work in chronic disabling spinal disorders. Tertiary prevention efficacy of functional restoration treatment." Spine **21**(8): 952-9.

Gatchel, R. J. (1991). Early development of physical and mental deconditioning in painful spinal disorders. Contemporary Conservative Care for Painful Spinal

Disorders. T. G. Mayer, V. Mooney and R. J. Gatchel. Philadelphia, Lea & Febiger**:** 278-289.

Gatchel, R. J. (1993). Psychophysiological disorders: Past and present perspectives. Psychophysiological Disorders: Research in Clinical Applications. R. J. Gatchel and E. B. Blanchard. Washington, American Psychological Association Press.

Gatchel, R. J. (1996). Psychological disorders and chronic pain: Cause and effect relationships. Psychological Approaches to Pain Management: A Practitioner's Handbook. R. J. Gatchel and D. C. Turk. New York, Guilford**:** 33-52.

Gatchel, R. J. (2001). Research outcomes compendium. LaGrange, IL, North American Spine Society.

Gatchel, R. J. (2004). "Comorbidity of chronic mental and physical health disorders: The biopsychosocial perspective." American Psychologist **59**: 792-805.

Gatchel, R. J., J. P. Garofalo, et al. (1996). "Major psychological disorders in acute and chronic TMD: An initial examination of the "chicken or egg" question." Journal of the American Dental Association **127**: 1365-1374.

Gatchel, R. J., T. G. Mayer, et al. (1986). "Million Behavioral Health Inventory: Its utility in predicting physical function in patients with low back pain." Archives of Physical Medicine & Rehabilitation **67**(12): 878-82.

Gatchel, R. J., T. G. Mayer, et al. (1999). "The association of the SF-36 Health Status Survey with one-year socioeconomic outcomes in a chronically disabled spinal disorder population." Spine **24**: 2162-2170.

Gatchel, R. J., T. G. Mayer, et al. (2006). "The pain disability questionnaire: relationship to one-year functional and psychosocial rehabilitation outcomes." J.Occup.Rehabil. **16**(1): 75-94.

Gatchel, R. J., Y. Peng, et al. (2007). "The biopsychosocial approach to chronic pain: scientific advances and future directions." Psychological Bulletin **In Press**.

Gatchel, R. J., P. B. Polatin, et al. (1994). "Psychopathology and the rehabilitation of patients with chronic low back pain disability." Archives of Physical Medicine and Rehabilitation **75**: 666-670.

Gatchel, R. J., P. B. Polatin, et al. (1998). "Use of the SF-36 health status survey with a chronically disabled back pain population: Strengths and limitations." Journal of Occupational Rehabilitation **8**: 237-246.

Greenough, C. G., L. J. Taylor, et al. (1994). "Anterior lumbar fusion: A comparison of noncompensation patients with compensation patients." Clinical Orthopaedics and Related Research **300**: 30-37.

Gronblad, M., M. Hupli, et al. (1993). "Intercorrelation and test-retest reliability of the Pain Disability Index and the Oswestry Disability Questionnaire and their correlation with pain intensity in low back pain patients." Clinical Journal of Pain **9**: 189-195.

Guereje, O., M. Von Korff, et al. (1998). "Persistent pain and well-being: A World Health Organization study in primary care." JAMA **280**(2): 145-151.

Hagg, O., P. Fritzell, et al. (2002). "The clinical importance of changes in outcome scores after treatment for chronic low back pain." Eur Spine J **12**: 12-20.

Hays, R. and J. M. Woolley (2000). "The concept of clinically meaningful difference in health-related quality of life research." Pharmacoeconomics **18**(5): 419-423.

Hazard, R., J. W. Fenwick, et al. (1989). "Functional restoration with behavioral support. A one-year prospective study of patients with chronic low-back pain." Spine **14**: 157-61.

Huber, D. L. and M. Oermann (1998). The Evolution of Outcomes Management. Health Care Outcomes: Collaborative, Path-Based Approaches. S. Blancett, Flarey, D. L. Gaithersburg, Maryland, Aspen**:** 3-11.

Jaeschke, R., J. E. Singer, et al. (1989). "Measurement of Health Status." Controlled Clinical Trials **10**(407-415).

Jones, A. and R. Zachariae (2004). "Investigation of the interactive effects of gender and psychological factors on pain response." British Journal Of Health Psychology **9**(Pt 3): 405-418.

Jordan, K. D., T. G. Mayer, et al. (1998). "Should extended disability be an exclusion criterion for tertiary rehabilitation? Socioeconomic outcomes of early versus late functional restoration in compensation spinal disorders." Spine **23**(19): 2110-2116.

Joyce, C., D. Zutski, et al. (1975). "Comparison of fixed interval and visual analogue scales for rating chronic pain." Europ. J. Clin Pharmacol. **8**: 415-420.

Karnofsky, D. A. and J. H. Burchenal, Eds. (1947). The clinical evaluation of chemotherapeutic agents in cancer. Evaluation of chemotherapeutic agents. New York, Columbia University Press.

Katon, W., K. Egan, et al. (1985). "Chronic pain: Lifetime psychiatric diagnoses and family history." American Journal of Psychiatry **142**: 1156-1160.

Katz, J. N., M. G. Larson, et al. (1992). "Comparative measurement sensitivity of short and longer health status instruments." Med Care **30**: 917-925.

Kazis, L. E., J. J. Anderson, et al. (1989). "Effect sizes for interpreting changes in health status." Med Care **27**(Suppl. 3): S178-79.

Kirshner, B. and G. H. Guyatt (1985). "A methodological framework for assessing health indices." J Chron Dis **38**(1): 27-36.

Kopec, J. A. (2000). "Measuring functional outcomes in persons with back pain." Spine **25**: 3110-4.

Kopec, J. A., J. M. Esdaile, et al. (1995). "The Quebec Back Pain Disability Scale: Measurement properties." Spine **20**: 341-352.

Kopec, J. A., J. M. Esdaile, et al. (1996). "The Quebec Back Pain Disability Scale: Conceptualization and development." Journal of Clinical Epidemiology **49**: 151-161.

Krause, N. and D. R. Ragland (1994). "Occupational disability due to low back pain: A new interdisciplinary classification used in a phase model of disability." Spine **19**: 1011-1020.

Kulkarni, A. V. (2006). "Distribution-based and anchor-based approaches provided different interpretability estimates for Hydrocephalus Outcome Questionnaire." Journal of Clinical Epidemiology **59**: 176-184.

Kumar, S. (1990). "Cumulative load as a risk factor for low-back pain." Spine **15**: 1311-1316.

Kuritzky, L. and D. J. Carpenter (1995). "The primary care approach to low back pain." Primary Care Representatives. **1**: 29-38.

Linton, S. J. (2000). "A review of psychological risk factors in back and neck pain." Spine **25**(9): 1148-1156.

Lurie, J., B. S. Hanscom, et al. (2001). Outcome measures in spine patients: what is clinically important change? 28th Annual Meeting of ISSLS. Edinburgh.

Marketdata Enterprises (1995). Chronic Pain Management Programs: A Market Analysis. Valley Stream, NY, Author.

Marx, R., P. Hudak, et al. (1997). Development of an Upper Extremity Outcome Measure: The "DASH" (Disabilities of the Arm, Shoulder and Hand). San Francisco, CA, American Academy of Orthopedic Surgeons.

Mayer, T., R. Gatchel, et al. (1994). "A male incumbent worker industrial database. Part III: Lumbar/cervical functional testing." Spine **19**(7): 765-70.

Mayer, T., R. Gatchel, et al. (1999). "Outcomes comparison of treatment for chronic disabling work-related upper extremity disorders." Journal of Occupational and Environmental Medicine **41**: 761-770.

Mayer, T., R. J. Gatchel, et al. (1994). "A male incumbent worker industrial database. Part I: Lumbar spinal physical capacity." Spine **19**(7): 755-761.

Mayer, T., R. J. Gatchel, et al. (1994). "A male incumbent worker industrial database. Part II: Cervical spinal physical capacity." Spine **19**(7): 762-4.

Mayer, T., M. J. McMahon, et al. (1998). "Socioeconomic outcomes of combined spine surgery and functional restoration in workers' compensation spinal disorders with matched controls." Spine **23**(5): 598-605; discussion 606.

Mayer, T., P. Pope, et al. (1994). "Physical progress and residual impairment quantification after functional restoration. Part I: Lumbar mobility." Spine **18**: 389-94.

Mayer, T. G., C. Anagnostis, et al. (2002). "Impact of functional restoration after anterior cervical fusion on chronic disability in work-related neck pain." Spine J. **2**(4): 267-273.

Mayer, T. G. and R. J. Gatchel (1988). Functional restoration for spinal disorders: The sports medicine approach. . Philadelphia, Lea & Febinger.

Mayer, T. G., R. J. Gatchel, et al. (2001). "Effect of Age on Outcomes of Tertiary Rehabilitation for Chronic Disabling Spinal Disorders." Spine **26**.

Mayer, T. G., R. J. Gatchel, et al. (1985). "Objective assessment of spine function following industrial injury: A prospective study with comparison group and one-year follow-up." Spine **10**: 482-493.

Mayer, T. G., R. J. Gatchel, et al. (1987). "A prospective two-year study of functional restoration in industrial low back injury. An objective assessment procedure [published erratum appears in JAMA 1988 Jan 8;259(2):220]." JAMA **258**(13): 1763-1767.

Mayer, T. G., R. J. Gatchel, et al., Eds. (2000). Occupational Musculoskeletal Disorders: Function, Outcomes and Evidence. Philadelphia, Lippincott Williams & Wilkins.

Mayer, T. G., R. J. Gatchel, et al. (2006). Postinjury Rehabilitation/Management. Interventions, Controls, and Applications in Occupational Ergonomics. W. S. Marras, Karwowski, W. New York, CRC Press.

McCracken, L. M., R. T. Gross, et al. (1996). "The assessment of anxiety and fear in persons with chronic pain: a comparison of instruments." Behavioral Research and Therapy **34**(11/12): 927-933.

McGeary, D. D., T. G. Mayer, et al. (2006). "High pain ratings predict treatment failure in chronic occupational musculoskeletal disorders." J.Bone Joint Surg.Am. **88**(2): 317-325.

McHorney, C. and A. Tarlov (1995). "Individual patient-monitoring in clinical practice: are available health status surveys adequate?" Qual Life Res **4**: 293-307.

McMahon, M. J., R. J. Gatchel, et al. (1997). "Early childhood abuse in chronic spinal disorder patients. A major barrier to treatment success." Spine **22**(20): 2408-15.

Melhorn, J. M. and P. Gardner (2004). "How we prevent prevention of musculoskeletal disorders in the workplace." Clinical Orthopaedics And Related Research(419): 285-296.

Melzack, R. (2001). "Pain and the neuromatrix in the brain." J Dent Educ **65**: 1378-1382.

Million, R., J. Haavik-Nilsen, et al. (1981). "Evaluation of low back pain and assessment of lumbar corsets with and without back supports." Annals of the Rheumatic Diseases **40**: 449-454.

Million, R., W. Hall, et al. (1982). "Assessment of the progress of the back-pain patient 1981 Volvo Award in Clinical Science." Spine **7**(3): 204-212.

Nafe, J. P. (1934). The pressure, pain and temperature sense. Handbook of general experimental psychology. C. A. Murchison. Wooster, MA, Clark University.

National Safety Council (2000). Injury Facts. Washington, DC, Author.

Norman, G. R., P. W. Stratford, et al. (1997). "Methodological problems in the retrospective computation of responsivenes to change: the lesson of Cronbach." J Clin Epidemiol **50**: 869-879.

Nunnally, J., Ed. (1975). The study of change in evaluation research: Principles concerning measurement, experimental design and analysis. Handbook of Evaluation Research. Beverly Hills, Sage.

Ohnmeiss, D. D. (2000). Oswestry back pain disability questionnaire. Compendium of Outcome Instruments for Assessment and Research of Spinal Disorders. R. J. Gatchel. LaGrange, IL, North American Spine Society.

Ong, K. S. and S. B. Keng (2003). "The biological, social, and psychological relationship between depression and chronic pain." Cranio. **21**(4): 286-294.

Papageorgiou, A., T. V. Macfarlane, et al. (1997). "Psychosocial factors in the workplace-do they predict new episodes of low back pain? : Evidence from the South Manchester Back Pain Study." Spine **22**(10): 1137-1142.

Polatin, P. B., R. Kinney, et al. (1993). "Psychiatric Illness and Chronic Low Back Pain: The Mind and the Spine-Which Goes First?" Spine **18**: 66-71.

Praemer, A., S. Furnes, et al. (1992). Musculoskeletal conditions in the United States. Rosemont, AAUS.

Reid, S., L. Haugh, et al. (1997). "Occupational low back pain: recovery curves and factors associated with disability." J Occup Rehabil **7**: 1-14.

Riddle, d. L., P. W. Stratford, et al. (1998). "Sensitivity to change of the Roland-Morris Back Pain Questionnaire: Part 2" Physical Therapy **78**(11): 1197-1207.

Roland, M. and J. Fairbank (2000). "The Roland-Morris Disability Questionnaire and the Oswestry Disability Questionnaire." Spine **25**(24): 3115-3124.

Romano, J. M. and J. A. Turner (1985). "Chronic pain and depression: Does the evidence support a relationship?" Psychological Bulletin **97**: 18-34.

Roth, J. H., R. S. Richards, et al. (1994). "Endoscopic carpal tunnel release." Canadian Journal of Surgery **37**: 189-193.

Roy, R., M. Thomas, et al. (1984). "Chronic pain and depression: A review." Comprehensive Psychiatry **25**: 96-105.

Samsa, G., D. Edelman, et al. (1999). "Determining clinically important differences in health status measures." Pharmacoeconomics **15**(2): 141-155.

Sandstrom, J. (1986). "Clinical and social factors in rehabilitatoin of patients with chronic low back pain." Scandinavian Journal of Rehabilitation Medicine **18**: 35-43.

Schultz, I. Z., A. W. Stowell, et al. (2007). "Models of Return to Work for Musculoskeletal Disorders." J Occup Rehabil.

Scott, J. and E. Huskisson (1976). "Graphic representation of pain." Pain **2**: 175-184.

Seferlis, T., G. Nemeth, et al. (2000). "Prediction of functional disability, recurrences, and chronicity after 1 year in 180 patients who required sick leave for acute low-back pain." Journal of Spinal Disorders **13**: 470-477.

Sheffield, D., P. L. Biles, et al. (2000). "Race and sex differences in cutaneous pain perception." Psychosomatic Medicine **62**(4): 517-523.

Sinclair, D. C. (1955). "Cutaneous sensation and the cotrine of specific nerve energies." Brain **78**: 584-614.

Skovron, M. L. (1992). "Epidemiology of low back pain." Baillieres Clinical Rheumatology **6**(559-73).

Spitzer, W. O., F. E. LeBlanc, et al. (1987). "A scientific approach to the assessment and management of activity-related spinal disorders: A monograph for clinicians; Report of the Quebec Task Force on Spinal Disorders." Spine **75**: 53-559.

Sriwatanakul, K., W. Kelvie, et al. (1983). "Studies with different types of visual analog scales for measurement of pain." Clin Pharmacol Ther **34**: 234-39.

Stewart, G., B. L. Sachs, et al. (1996). "Patient outcomes after reoperation on the lumbar spine." Journal of Bone & Joint Surgery – American Volume(78): 706-711.

Stratford, P. W., J. M. Binkley, et al. (1998). "Sensitivity to change of the Roland-Morris Back Pain Questionnaire: Part 1." Physical Therapy **78**(11): 1186-1196.

Suarez-Almazor, M. E., C. Kendall, et al. (2000). "Use of health status measures in patients with low back pain in clinical settings. Comparison of specific, generic, and preference-based instruments." Rheumatology **39**(783-790).

Subramanian, A., A. Desai, et al. (2006). "Changing trends in US injury profiles: revisiting non-fatal occupational injury statistics." J Occup Rehabil **16**(1): 123-55.

Taylor, S. J., A. E. Taylor, et al. (1999). Responsiveness of common outcome measures for patients with low back pain.

Terwee, C. B., F. W. Dekker, et al. (2003). "On assessing responsiveness of health-related quality of life instruments: guidelines for instrument evaluation." Qual Life Res **12**(4): 349-62.

Testa, M. A. (1987). "Interpreting quality-of-life clincial trial data for use in the clinical practice of antihypertensive therapy." J Hypertension **5**(Suppl 1): S9-S13.

Testa, M. A. and D. C. Simonson (1996). "Assessment of quality of life outcomes." N.Engl.J.Med **334**(835-840).

Triano, J. J., M. McGregor, et al. (1993). "A comparison of outcome measures for use with back pain patients: Results of a feasibility study." Journal of Manipulative & Physiological Therapies **16**: 67-73.

Turk, D. C. (1999). "The role of psychological factors in chronic pain." Acta Anaesthesiologica Scandinavica **43**(885-888).

Turk, D. C., J. P. Robinson, et al. (2004). "Prevalence of fear of pain and activity in fibromyalgia syndrome patients." Journal of Pain **5**: 483-490.

U.S. Census Bureau (1996). Statistical Abstract of the United States: 1996. Washington, DC, United States Bureau of the Census.

U.S. Department of Labor (2000). Lost-worktime injuries and illnesses: Characteristics and resulting time away from work, 1998, Publication USDL 00-115. Washington D.C., U.S. Department of Labor.

U.S. Department of Labor (2005). Lost worktime injuries and illnesses in 2005, Publicaiton USDL 06-1816. Washington D.C., U.S. Department of Labor.

Vallerand, A. H. and R. C. Polomano (2000). "The relationship of gender to pain." Pain Management Nursing **1**(3S).

Verhaak, P. F. M., J. J. Kerssens, et al. (1998). "Prevalence of chronic benign pain disorder among adults: A review of the literature." Pain **77**: 231-239.

Vlaeyen, J., Kole-Snijders, et al. (1995). "The role of fear of movement/(re)injury in pain disability. *Journal of Occupational Rehabilitation*." **5**: 235-252.

Von Korff, M. and G. Simon (1996). "The relationship between pain and depression." British Journal of Psychiatry **Supplement 1996**(30): 101-108.

Vowles, K. E., M. J. Zvolensky, et al. (2004). "Pain-related anxiety in the prediction of chronic low-back pain distress." J Behav Med **27**(1): 77-89.

Wallenstein, S. and R. Houde (1975). The clinical evaluation of analgesic effectiveness. Methods of Narcotics Research. S. Ehrenpreis and A. Neidle. New York, Marcel Dekker**:** 127-145.

Ward, N. G., V. L. Bloom, et al. (1982). "Psychobiological markers in coexisting pain and depression: toward a unified theory." J.Clin.Psychiatry **43**(8 Pt 2): 32-41.

Ware, J. E., M. Kosinski, et al. (1994). SF-36 Physical and Mental Health Summary Scores: A User's Manual. Boston, The Health Institute, New England Medical Center.

Ware, J. E. and C. D. Sherbourne (1992). "The MOS 36-Item Short-Form Health Survey (SF-36). I. Conceptual framework and item selection." Medical Care **30**: 473-483.

Weddell, G. (1955). "Somesthesis and the chemical senses." Annual Review of Psychology **6**: 119-136.

Wells, G., D. Beaton, et al. (2001). "Minimal clinically important differences: review of the methods." The Journal of Rheumatology **28**(2): 406-412.

Wennberg, J. (1990). "Outcomes research, cost containment and the fear of health care rationing." N.Engl.J.Med **323**: 1202-1204.

Wolfe, F., H. A. Smythe, et al. (1990). "The American College of Rheumatology 1990 Criteria for the Classification of Fibromyalgia. Report of the Multicenter Criteria Committee." Arthritis and Rheumatism **33**(2): 160-72.

Wright, A., T. G. Mayer, et al. (1999). "Outcomes of disabling cervical spine disorders in compensation injuries. A prospective comparison to tertiary rehabilitation response for chronic lumbar spinal disorders." Spine **24**(2): 178-183.

Wyrwich, K. W., N. A. Nienaber, et al. (1999). "Linking clinical relevance and statistical significance in evaluating intra-individual changes in health-related quality of life." Med Care **37**(5): 469-478.

BIOGRAPHICAL INFORMATION

Hilary Wilson was born in Greenfield, Massachusetts and spent the majority of her childhood growing up in Fort Worth, Texas. She attended Cornell College in Mt. Vernon, Iowa in the fall of 1996 to pursue a B.A. in Psychology. Upon graduation in the Fall of 1999, she spent 4 years working in the disability field, in both a residential and vocational setting. Her interest in biological mechanisms of disability led her to pursue a Masters of Experimental Psychology in Behavioral Neuroscience, under the guidance of Dr. Perry N. Fuchs. Her Masters work introduced her to the world of pain research, and following completion of her M.S. in Experimental Psychology, she began research with Dr. Robert Gatchel and Dr. Tom Mayer at a functional restoration program for patients with chronic work-related injuries. She is currently pursuing her doctoral degree in Experimental Psychology under the guidance of Dr. Robert Gatchel, and upon completion of the program plans to continue an academic career, with a focus on health outcomes research.