CLASSIFYING CHRONIC LOWER BACK PAIN GROUPS USING A TIME SERIES MODEL OF LIFTING

by

Jill Christina Slaboda

BS, University of Pittsburgh, 2001

MS, University of Pittsburgh, 2004

Submitted to the Graduate Faculty of

School of Engineering in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

University of Pittsburgh

2007

UNIVERSITY OF PITTSBURGH

SCHOOL OF ENGINEERING

This dissertation was presented

by

Jill Christina Slaboda

It was defended on

March 29, 2007

and approved by

Rory Cooper, Distinguished Professor, Health and Rehabilitation Sciences and Bioengineering Departments

Amro El-Jaroudi, Professor, Electrical and Computer Engineering Department

Mark Miller, Associate Professor, Mechanical Engineering Department

Mark Redfern, Professor, Bioengineering Department

Dissertation Co-Director: Thomas Rudy, Professor, Anesthesiology Department

Dissertation Co-Director: J. Robert Boston, Professor, Electrical and Computer Engineering Department

CLASSIFYING CHRONIC LOWER BACK PAIN GROUPS USING A TIME SERIES MODEL OF LIFTING

Jill Christina Slaboda, Ph.D

University of Pittsburgh, 2007

A classification procedure was developed that uses hidden Markov models (HMMs) to identify sub-groups within a chronic lower back pain (CLBP) patient population based on their time series of lifting patterns during a repetitive lifting task. Based on clinical observations of a repetitive lifting task, our approach assumed that the patient population was composed of two groups: one group that performed lifts more similar to controls than to other patients and another group that lifted differently from control subjects. Two HMMs were designed to describe the repetitive lifting data, one derived from the control subject data and one derived from the CLBP subject data. The HMMs were designed based on the results of a data reduction procedure that reduced and combined the multidimensional lifting parameters into discrete lifting patterns using factor analysis and cluster analysis.

Simulation studies were performed to demonstrate that the HMMs could reliably identify subjects from one group that were intentionally mislabeled as the other group. When the HMMs were applied to clinical data, 35 of the 81 CLBP subjects were classified to the control HMM and 46 were classified to the CLBP HMM. For the control group, 46 of 53 control subjects were classified to the control HMM and only seven were classified to the CLBP HMM. The CLBP groups were found to use different lifting patterns during the task. The CLBP subjects that were classified to the CLBP HMM were found to use a lifting pattern that involves slow, controlled

movements. Self-reported measures of the two groups of CLBP subjects were compared and self-reported pain intensity, pain severity and perceived self-efficacy found to be statistically different. The CLBP subjects that were classified to the CLBP HMM reported higher pain intensity and pain severity, and lower self-efficacy suggesting that the CLBP population is heterogeneous and that the HMM classification procedure can successfully identify two meaningfully different sub-groups of CLBP patients.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

1.0 INTRODUCTION

Patients with the same clinical diagnoses are often assumed to be a homogeneous group. In reality, these patients are heterogeneous and identifying differences between the patients may suggest diverse treatment options that can improve treatment effectiveness. For example, Delitto et al. identified sub-groups of patients with acute low back pain [1]. When these patients received treatment that was specialized to their symptoms assessed by clinical examination, the patients had a greater return to work status and lower self-reported disability compared to acute low back pain patients who were assigned to a single standardize treatment protocol [2].

Classifications of chronic lower back pain (CLBP) patients have been found based on the patients' responses to self-reported measures and on clinical examinations. Turk and Rudy identified three groups of CLBP patients based on responses to questionnaires related to pain psychosocial aspects, responses of patient's significant other to their pain and frequency of general activities [3, 4]. O'Sullivan classified CLBP patients to one of two groups based on posture of the lower back and pain reports of subjects [5]. Dunn et al found four classifications of low back pain patients with latent class analysis of self-reported pain measures reported over a year [6]. These studies indicate the heterogeneity of CLBP patients but are limited in that the classifications relied on patient's perception or the clinician's experience. The objective of this project is to develop an unbiased classification procedure of CLBP patients based on lifting patterns performed during a repetitive lifting task.

Repetitive lifting tasks have been reported to initiate fatigue, which causes distinct changes in body segmental motion [7-8], posture [9] and flexibility [10] in pain-free control subjects. The presence of CLBP causes adaptations in motion in addition to those caused by fatigue. Several studies have compared motion of pain-free controls and CLBP subjects and have found differences in flexibility [11], spinal loading [12], muscle activation [13], segmental motion [14], jerk [15] and lifting strategies over time [16]. For example, Lavivierie et al. showed that chronic lower back pain patients (CLBP) demonstrated less lumbar flexion than pain-free controls during a flexion task [17]. Marras et al found differences in spinal loading with CLBP subjects, who experienced 26% more compression and 75% more shear than control subjects for a static exertion task [18]. Kankaanpää et al observed that CLBP patients showed greater fatigability of the gluteal maximus muscle during a flexion-extension task [19]. Oddsson and De Luca found that patients perform at a lower maximum voluntary contraction of paraspinal muscles than controls and the presence of pain causes a redistribution of the activation behavior between synergistic muscles of the lumbar back [13].

In addition to differences in types of motion and muscle activation patterns during specific tasks, several studies have found temporal differences in motion patterns when control subjects are compared to CLBP subjects during a lifting task. Rudy et al. examined changes in lifting parameters over the duration of a repetitive lifting task by separating task time into three phases of early, middle and late [16]. CLBP subjects were found to modify lifting parameters in the early to middle phases of the task, while control subjects made modifications throughout all task phases. Similar results were found in Slaboda et al., who analyzed jerk at the shoulder in control and CLBP subjects performing the same repetitive lifting task [15]. CLBP subjects were found to perform lifts with lower jerk values than controls. Over task time, CLBP subjects increased jerk

from the early to middle phases while control subjects increased jerk throughout all phases of the task. These studies have shown that CLBP and pain-free control subjects demonstrate different motion patterns and different temporal characteristics of motion during tasks such as lifting. However, the differences in motion among CLBP subjects have received little research attention despite evidence in the psychosocial literature that shows the heterogeneity of CLBP group.

The present investigation was based on two conclusions drawn from the repetitive lifting study: first, CLBP subjects and controls as groups use different temporal lifting motion patterns during the task and second, based on clinical observations, the CLBP subjects are not homogeneous. Several of the CLBP subjects appear to use a lifting style that is very similar to controls while other CLBP subjects perform lifting styles that are very different. *Our goal was to develop an objective procedure to identify a sub-group of CLBP subjects that performed more like controls from the CLBP population.* The procedure used hidden Markov models (HMM) to provide classification of CLBP subjects based on their time series of lifting patterns. HMMs have previously been used to describe the time series of lifting patterns of CLBP subjects and control subjects during a repetitive lifting task [20-21]. The HMM describing control subjects' motion was found to have a topology that included more transitions than the HMM topology that described lifting motion of CLBP subjects, suggesting that control and CLBP subjects demonstrate different lifting patterns over time and HMMs can be used to describe these differences.

In this project, a database that contains measures of psychological, medical and physical functioning of CLBP subjects from a previously conducted clinical study was used. During the clinical study, CLBP subjects completed a series of self-reported measures to assess psychological functioning, and CLBP subjects and control subjects completed a repetitive lifting

task to assess physical functioning. The body motion of the subjects during each lift was described by one of five lifting patterns, as determined from a data reduction procedure. The procedure used factor analysis and cluster analysis to combine the multidimensional motion parameters into lifting patterns. Two HMMs, one to describe the lifting pattern sequences of control subjects and the other to describe the lifting pattern sequences of CLBP subjects, were designed. The classification procedure assigned each subject to the HMM that had a higher probability of describing the subject's sequence of lifting patterns. The CLBP subjects that are assigned to the control HMM were identified as the sub-group of CLBP subjects that perform lifts more similar to control subjects.

The possibility of using HMM classification approach to identify CLBP groups was evaluated with a simulation study. The simulation was performed to assess reliability because all CLBP subjects were classified in one group and no information about the CLBP sub-groups was available. The simulations were conducted to determine how reliably the HMMs can detect lifting sequences that are classified to the wrong group and classify them to the appropriate group. Since the HMM classification procedure was successful in detecting mislabeled subjects in the simulation study, the procedure was applied to the clinical data. The self-reported measures of the CLBP groups were compared to determine whether the HMM classification produced could identify two meaningful different groups within a sample of CLBP subjects.

This thesis is organized into six chapters. Chapter 2 provides a literature review of previous lifting studies comparing controls and CLBP subjects, classification of CLBP patients, cluster analysis methods and the use of HMMs to identify groups within a clinical population. The third chapter is the methods chapter and it describes the protocol of the clinical study including the self-reported measures and the lifting task. The data reduction procedure and the design of the

HMMs are also described in the methods chapter. Chapter 4 describes the simulation studies that were performed to determine whether the HMMs can reliably identify sequences to the correct model when the sequences are mislabeled to the wrong group. The results of the simulation studies are described in Chapter 5. The methods of HMM classification procedure on the clinical data, the results of the HMM classification procedure and the discussion which interprets the results are described in Chapter 6.

2.0 BACKGROUND

This section will review literature that has shown differences in the lifting patterns of controls and CLBP subjects, identified groups within the chronic pain population, shown that hidden Markov models can be applied to clinical time series and methods of validating cluster solutions.

2.1 PREVIOUS LIFTING STUDIES

The lifting parameters that were used in this dissertation to describe lifting patterns were used previously and have shown significant differences in motion between CLBP subjects and control subjects. This section will review previous research related to lifting differences during a repetitive lifting task that were performed at the University of Pittsburgh Pain Evaluation and Treatment Institute.

The relative motion of the hip and knee angles during repetitive lifting task was investigated to determine whether control subjects and CLBP subjects perform lifts differently [14]. Relative motion was described with a coordination index, derived from hyperbolic tangent curves that were fit to the hip and knee angle data as functions of lift time. Each curve was described by four parameters: starting angle, ending angle, risetime and midpoint time. The

risetime was defined as the time required for the angle to decrease from 88% to 12% of the total change in angle. The midpoint time defined the time after the beginning of the lift at which the hip or knee angle had completed half of its range of motion. The difference between the hip and knee risetime and the difference between the hip and knee midpoint described relative movement of the two joints. The difference is zero when a subject moves the hip and knees at the same speed, resulting in synchronous motion. In this case, both angles reach the midpoint at the same time. A negative midpoint difference occurs when the hips straighten before the knees. A positive risetime difference indicates that the knee angle is changing faster than the hip angle, producing a longer hip risetime than knee risetime.

Pain-free controls were found to move the hip and knee asynchronously when they initiated the lift, but the hip and knee angles reach full extension simultaneously at the end of the lift. Controls either moved the knees earlier and the hip moved faster or the hip moved earlier and the knees moved faster in order to end together. For this lifting style, the midpoint and risetime differences are opposite in sign and the hip and knee motion produce a coordinated ending. CLBP patients used a "guarded" lifting pattern in which the hip and knee moved synchronously when patients initiated the lift but the hip and knees angles finished motion at different times. In this lifting style, the midpoint and risetime differences are the same sign and the hip and knee are not coordinated at the end of the lift. The guarded lifting style can result from contracting both agonist and antagonist muscles of a joint and has been suggested as a mechanism to avoid or minimize pain during movement [22].

The coordination index was used to assess the impact of a rehabilitation program on lifting motion of CLBP subjects. The coordination index showed that CLBP patient’s pre- and post-treatment index changed significantly, showing more coordinated endings between the hip

and knee post-treatment. The post-treatment coordination indices were not significantly different from those observed in controls. The test-retest reliability of the coordination index was found to be high at 0.76 [23].

In addition to the coordination index, similar motion pattern differences were found in parameters of starting posture, lift duration and a work index [14]. The starting posture measures whether the subject performed a torso lift or a squat lift. CLBP patients pre-treatment were found to use more of a squat lift than controls, and post-treatment patients showed a greater squat lifting style. Lift duration is the time required for the subject to perform a lift. Control subjects performed lifts faster than CLBP patients pre- and post-treatment. Post-treatment testing resulted in decreased lift duration of CLBP patients when compared with pre-treatment testing. Work index was defined as the number of lifts performed multiplied by the weight lifted. The work index was greater for control subjects than CLBP patients. Treatment showed a 71% increase in the work index from pre-treatment values. However the post-treatment values never approached the work indices of control subjects [23].

Differences in the hip midpoint, knee midpoint, hip-knee midpoint difference, starting posture, and lift duration over task time were investigated [16]. Lifts were grouped as early, middle, or late phase, based on the individual subject's number of repetitions, to minimize the effects of wide variations among subjects in the number of lifts performed. To determine whether the control and CLBP subjects changed lifting parameters differently over time, repeated-measures ANOVA were used to assess differences between experimental group and changes over time. Significant differences between the motion of the lifting patterns over time were found for starting posture, hip, midpoint, knee midpoint and lift duration. Starting posture changed over time for CLBP patients demonstrating a greater knee angle as the task progressed

from the early to the middle phase. Control subjects consistently produced greater hip flexion than patients throughout the task. The hip and knee midpoints showed dissimilar changes over task time for controls compared to patients. Control subjects increased both hip and knee midpoints as the task progressed but patients increased midpoint from early to middle phase and then stabilized from the middle to the late phase. Lift duration demonstrated the same changes over task time for both groups, in which speed increased as the task progressed from early to middle phase and then speed stabilized. Controls performed faster lifts than patients for all phases.

HMMs have been used to describe the temporal changes in the angle parameters of pain-free controls and CLBP subjects during a repetitive lifting task [20]. The HMMs were designed from the *DISSOLVE* algorithm that determined the simplest model structure for discrete HMMs [21,24]. This algorithm iteratively removed state transitions and/or states from a trained fully-connected HMM until the model had only a single state. The Baum-Welch algorithm was used to train the HMMs. During training, the initial estimates of HMM parameters were based on the training data and the lifitng sequences could start in any state within the HMM. The algorithm was validated with simulation studies [21].

The *DISSOLVE* algorithm was applied to a fully-connected HMM that described the lifting parameters of CLBP patients and control subjects performing a repetitive lifting task to determine the appropriate HMM for each group. The three lifting parameters, risetime difference, midpoint difference and the difference between the starting hip angle and starting knee angle, were calculated for each lift that the subjects completed. For each group, the parameters were normalized to the range of the values and then vector quantized into 32 clusters. A 3-state HMM was found to be the appropriate HMM for the control HMM and CLBP HMM.

The control HMM had two interstate transitions: from state 2 to state 1 and from state 2 to state 3. The CLBP HMM had one isolated state (state 1) and one interstate transition: from state 2 to state 3.

Although the angle parameters from the hyperbolic tangent equation showed differences between groups of CLBP patients and controls during lifting, these parameters relied on small differences between the timing of the hip and knee angle motion and were often difficult to interpret. A search for a more robust measure to differentiate lifting patterns between CLBP patients and control lead to jerk [15,25], which is defined as the rate of change of acceleration or the third derivative of position [26]. Jerk was applied to the shoulder of control and CLBP subjects to describe motion of the subjects when performing a repetitive lifting task. Since jerk was calculated as the third derivative of displacement, it was a noisy measure and smoothing methods were necessary to obtain an estimate of jerk. A simulation study was performed to assess the performance of the smoothing methods to estimate jerk of a known trajectory with additive correlated noise. Mean-squared-error was used as a measure of performance and was calculated between the jerk estimates of the noise-free known trajectory and the smoothed trajectory. The two smoothing methods that were assessed in the simulation were Woltring's generalized cross-validation hepatic spline [27] and Wells and Winter's method of low-pass filtering [28-29].

The results of the simulation showed that Woltring's spline produced the best estimates of jerk. This method was applied to a database of control and CLBP subjects' lifting data to calculate jerk at the shoulder. Derivatives of shoulder displacement were calculated using differentiation of the spline coefficients, and root-means-square (rms) amplitude of jerk was used for comparison. Lifts were divided into phases of early, middle or late based on the number of

repetitions completed by the subject. Average values of rms jerk during a lift were computed at each of the task phases. Significant group differences were found for rms jerk. CLBP patients were found to perform lifts with lower jerk values than controls and as the task progressed, rms jerk increased for both groups. A group-by-phase interaction was significant. After completion of a rehabilitation program, CLBP patients performed lifts with greater rms jerk. In general, patients performed lifts with lower jerk values than controls, suggesting that pain impacts lifting style.

2.2 CLASSIFICATION SYSTEMS OF LOW BACK PAIN PATIENTS

Previous classifications of pain patients have been found based on self-reported pain and psychological measures, clinical examination and motion patterns. This section will review the classification methods of previous studies and identify the measures that separate pain patients into different classifications.

Turk and Rudy developed a multiaxial classification of temporomandibular disorders (TMD) patients based on responses to the West-Haven Multidimensional Pain Inventory (MPI), which assesses pain-relevant psychosocial aspects, responses of the significant other to the patient's pain and frequency of common activities [3-4]. Subjects were asked to complete the MPI questionnaire and cluster analysis was applied to the responses. From the cluster solution, three distinct groups of TMD patients were found: Dysfunctional group, characterized by higher levels of pain, life interferences, emotional distress and functional limitation; Interpersonally distressed group, characterized by lower levels of social and personal support; and Adaptive copers group, characterized by lower levels of pain, functional limitation, and emotion distress.

To validate the classifications, additional self-reported measures that were not included in the formation of the classifications were compared between the three groups. The results showed that the Dysfunctional group reported significantly greater pain severity and improved mood, the Interpersonally Distressed reported lower levels of support and the Adaptive cope group reported higher levels of perceived control when compared to the two other classifications [4]. Turk and Rudy extended the MPI classification to a CLBP sample and a chronic headache sample and found that both samples demonstrate the same sub-groups as found with TMD patients [30].

The clinical utility of the classifications were shown in a study that administered the MPI questionnaire to TMD patients before, immediately after and six months after a standard treatment protocol for TMD [31]. The MPI classifications of TMD subjects were found to have differential responses to the treatment. Among the classifications, the Dysfunctional group showed the greatest improvement on measures of pain intensity, perceived impact of TMD on their lives, depression and negative thoughts. In each group, the greatest improvement was found on the measures that defined the groups at pre-treatment, indicating the importance of tailoring treatment components to specific characterization of the groups.

A recent study identified sub-groups of patients with musculoskeletal pain based on self-reported measures of disability, self-efficacy, pain intensity, fear of movement/(re)injury, and catastrophizing [32]. Subjects were asked to respond to self-reported measures and from these responses, three groups were identified. The first group is the High self-efficacy-Low fear-avoidance sub-group and subjects in this group reported low levels of pain intensity, disability, fear of movement/(re)injury, catastrophizing and high levels of self-efficacy. The second group was labeled Low self-efficacy-Low fear-avoidance and the subjects in this group reported high levels of pain intensity and disability, and low levels of fear of movement/(re)injury,

catastrophizing and self-efficacy. The last group was labeled the Low self-efficacy-High fear-avoidance and this sub-group was characterized as reporting high levels of pain intensity, disability, fear of movement/(re)injury, and catastrophizing, and low levels of self-efficacy. A binary measure of work status of the subjects were compared between the clusters and found to be significantly different. Subjects assigned to the High self-efficacy-Low fear-avoidance sub-group reported working significantly more frequently than the subjects in the other two sub-groups. The authors concluded that the differences in the working status of the sub-groups indicate the utility of the sub-groups and suggest that different treatment approaches are necessary for the sub-groups [32].

A latent class analysis was applied to the self-reported measures reported by low back pain patients over a period of six months to identify groups with different pathways of back pain [6]. Subjects were recruited from a primary care clinic and received standard care. The subjects were asked to respond to questionnaires related to pain intensity, disability, and the psychosocial measures at baseline, monthly for a period of six months and at one year. Four pathways were defined and these pathways were persistent mild pain, recovering, severe chronic pain, and fluctuating pain. The persistent mild group reported moderate levels of pain intensity and disability. The recovering groups reported no to little back pain, good psychosocial status and low disability ratings. The severe chronic group reported high levels of pain intensity, poor psychosocial status and high levels of disability. The fluctuating group varied from high to moderate pain intensity, moderate levels of disability and poor psychosocial status. At one year follow-up, the persistent mild group and recovering group had improved on all measures while the severe chronic groups and fluctuating group showed little to no improvement. The authors

suggest that these four groups provide information about the course of low back pain and could possible form a basis for intervention [6].

O'Sullivan identified two groups of CLBP patients based on self-reported measures of pain and physical examination of the patient's spinal motion [5]. The CLBP patients were assigned to either the flexion pattern group or the active extension pattern group based on the physical therapists assessment and the patient's self-reported pain. The CLBP patients assigned to the flexion pattern reported aggravating of symptoms with movements and postures involving the flexion of the lower lumbar spine, difficulty maintaining neutral lordosis with the tendency to flex the lumber spine, and pain relief in spinal extension [5]. The active extension pattern reported difficulty performing extension motion and the symptoms reported by this group were opposite of those associated with the flexion group. The characteristics of the active extension group included aggravating of symptoms with movements and postures involving the extension of the lower lumbar spine, excess of lordosis with posture and sitting, and pain relief in spinal flexion [5].

The CLBP flexion pattern group and active extension pattern group were compared to control subjects in separate studies to determine whether there were any biomechanical differences between the groups during an unsupported sitting task. The biomechanical parameters used to describe body motion during the task were the posture of the spine, measured with reflective markers located on the spine [33], and activation patterns of the trunk muscles, measured with surface electromyography of the superficial trunk muscles [34]. Control subjects were found to activate low back muscles while maintaining a neutral position of the spine during sitting. The CLBP flexion pattern group had decreased muscle activity of the lower back muscles [34] and had greater lumbar flexion [33]. The CLBP active extension pattern group co-contracted

the stabilizing muscles [34] and had hyperlordotic posture [33]. Significant differences in posture and muscle activation between control subjects and CLBP subjects were only found when the CLBP subjects were classified into the sub-groups of flexion pattern and active extension pattern, indicating the importance of identifying sub-groups within the CLBP population [34].

Delitto et al. developed a classification system that assigned treatment protocols to lower back pain subjects based on physical examination and self-reported measures [1]. Patients were assigned into four classifications and each classification had a different treatment protocol. If patients reported less pain in performing extension and greater pain when performing flexion or vice versa, the patient was assigned to the specific exercise groups. The treatment protocol of this group involved exercises that were directed to the exercises in which the patient reported less pain, i.e. for less pain in extension, patients were given specific exercises related to extension. The manipulation groups consisted of patients with recent report of back pain, no symptoms of pain below the knee and lumbar segmental hypomobility. The treatment protocol of this group was manipulation of the lumbosacral spine that involved the therapists delivering a force to the pelvis. The stabilization group included patients with frequent previous episodes of back pain, greater straight leg raise range of motion, aberrant motions and lumbar hypermobility. The treatment protocol of this group consisted of trunk muscle strengthening and stabilization exercises. The traction group consisted of patients with nerve root compression and treatment involved mechanical or auto traction [35].

The treatment classification was further investigated by Fritz et al in a randomized controlled study to determine effectiveness of the treatment [2]. The study assigned patients to either a standardized treatment protocol or specialize treatment classifications described by Delitto [1]. Treatment outcomes measures included the Oswestry disability rating [36] and the

return work status of the patients. The Oswestry disability rating is a self-reported measure that asks subjects to rate their perceived difficulty in performing activities of daily life such as walking, sitting, standing, sleeping and personal care. The results showed that when patients received treatment specialized to their symptoms these patients had a greater return to work status and lower self-reported disability compared to acute low back pain patients who were assigned to standardize treatment protocol.

Wrigley et al. used principal component analysis to distinguish the lifting techniques between healthy subjects that develop low back pain (LBP) and those who do not and to compare the principal component analysis approach to traditional parameter-based approach [37] The study used a database that contained the lifting motion parameters of subjects without LBP or with LBP without medical attention recruited from a nylon production plant and monitored for 2 years to assess LBP status. The lifting parameters were from a lifting task that asked subjects to lift a 15 kg box from the floor to shoulder height for five repetitions, and the motion of the trunk and the box were tracked with markers during the lifts. The parameters of acceleration, velocity and displacement of the box and spinous processes of T1, S1 and L1, trunk compression and shear, and the moments at spinous processes of T1, S1 and L1 were calculated, resulting in 16 parameters. All parameter data were analyzed with principal component analysis.

Six principal components were found in the data and principal component scores were calculated and compared between the groups. The traditional parameter-based approach involved the calculation of peak, time to peak, minimum, time to minimum and mean values of each of the displacement waveforms, resulting in 48 parameters. The results showed significant differences in the principal component scores of box vertical velocity, T1 acceleration, T1 and S1 moments and trunk compression between subjects that developed LBP and those that did not. No

significant differences were found between the groups for the traditional parameter-based approach. These results suggest that subjects that develop LBP used different lifting patterns, especially in motion related to the trunk kinematics and placement of the box, than subjects that do not develop LBP and also demonstrate the utility of using principal component analysis to describe motion of displacement waveforms during a repetitive lifting task.

Bishop et al. applied neural networks to identify LBP patients from control subjects based on trunk motion [38]. Two neural networks were designed. The first neural network assessed whether the subject was a LBP patient or a control. Once the subject was assigned as a LBP patient, a second neural network was assigned the patient to pain classification based on back motion. The neural networks were designed to describe back motion during several repetitions of five tasks: flexion/extension, axial rotation, lateral bending, clockwise circumduction and counterclockwise circumduction. A triaxial goniometer tracked the motion during the tasks and the features of velocity, shape and symmetry of the displacement waveform were calculated.

The pain classifications of the LBP subjects and control subjects were based on the Quebec Task Force questionnaire which assigned classification from the patient's pain history, pain complaints, clinical examination and complementary studies such as magnetic imaging or electromyogram [39]. These classifications were no LBP (a zero on the Quebec Task force questionnaire), subjective complaints with pain radiation, subjective complaints without pain radiation, objective signs, postoperative and non-specific low back pain. To train the neural networks, the features of the back motion data were separated into a training sample and a test sample with equal distribution of the pain groups in both samples. The results showed that the neural network classified the subjects in the test sample as either a LBP or control at 86% accuracy rate. The second neural network could distinguish the LBP pain classification at 65%

accuracy rate. The authors concluded that neural networks provided good discrimination between LBP subjects and control subjects.

The studies reviewed in this section showed that CLBP subjects are heterogeneous population and sub-groups can be identified based on the patient's self-reported measures and clinical evaluations. These classifications showed that CLBP sub-groups are significantly different on measures of self-reported pain and psychological distress and for measures of back posture and muscle activation of the lumbar back during tasks such as sitting. In addition, complex models have been able to identify LBP subjects from control subjects based on lifting motion and lumbar back motion, indicating that it is possible to use these types of models to identify sub-groups of CLBP subjects.

2.3 HIDDEN MARKOV MODELS AND CLINICAL DATA

Hidden Markov models (HMMs) have been used extensively in research to describe time series data. The most common areas of research that use HMMs are speech research [40-42] and bioinformatics [43-44] but other research areas such as animal migration and behavior [45], human behavior and psychology [46], gait [47] and machinery wear [48] have used these models. This section will review literature that pertains to application of HMMs to clinical data.

Yu et al applied a Markov model to assess treatment effectiveness of four treatment protocols based on the self-reported outcome measures of LBP patients [49]. The subjects were randomized to either medical care, medical care with physical therapy, chiropractic clinic or a chiropractic clinic with physical therapy. Patients were asked to choose one of four statements to

rate their perceived improvement: a lot better, a little better, about the same and worse. The self-reported measures were collected at two, four and six weeks, and 12 and 18 months post-treatment. A polytomous logistic regression model with a Markov structure was used to model the data. The results showed that patients assigned to the chiropractic treatment or medical doctor plus physical therapy treatment were more likely to report that symptoms were better than the patients assigned to the medical doctor treatment only. The patients receiving the chiropractic treatment with physical modalities were less likely than the patients receiving only chiropractic treatment to report that their symptoms were worse. The authors concluded that the Markov model provided a straightforward procedure to compare treatment modalities based on transitions and likelihood probabilities.

Hidden Markov models were applied to accelerometer data that classified the physical activity of subjects performing four distinct tasks of walking on a treadmill, walking up-hill on a treadmill, working at a computer and vacuuming [50]. The purpose of this study was to evaluate whether HMMs or a quadratic discriminant analysis (QDA) model could more accurately identify an activity given the accelerometer data than the traditional method of using cut-off values. In the cut-off values, a linear regression model is used to relate accelerometer data and physiological variable (such as VO_2), and the output corresponds to certain physical activities. During each activity, an accelerometer was placed on the subject's hip and recorded acceleration for a total of seven minutes.

For each activity, an ergodic 3-state Poisson distribution HMM was trained with Baum-Welch [51] algorithm. The sequences of the accelerometer data were classified to HMMs using a leave-one-out method that excluded a sequence and trained the models with the remaining data. The excluded data were then classified to one of the activity HMMs. This process continued until

all sequences were classified to a HMM. On average, the HMMs were found to better classify the activities than the QDA models. Both the QDA and HMMs were found to perform better than the cut-off method, which misclassified the walking uphill and the vacuuming activities. The authors concluded that HMM and QDA were improved methods for determining the physical activity levels than the traditional models.

Wong et al designed a HMM to evaluate the performance of a seizure detection algorithm [52]. The intracranial electroencephalogram (IEEG) signals of five patients with mesial temporal sclerosis were recorded for 515 hours and within that data were 29 seizures. The IEEG signals were entered into a seizure detection algorithm and a binary output of the algorithm interpreted the IEEG signals as either negative for seizure activity, termed baseline, or positive for seizure activity, termed detected. In addition to the algorithm output, the IEEG signals were marked by an electroencephalographer who interpreted the IEEG signal for the onset and termination of the seizure. The total number of classes was three with 1 denoting baseline, 2 denoting detection and 3 denoting a seizure. A fully-connected 3-state discrete HMM was trained with the Baum-Welch algorithm. The results showed that 17 of the 29 seizures were detected with the HMM that described the algorithm outputs. The authors concluded that the HMM approach provides a tool for designing and validating prediction algorithms [52].

HMMs were used to identify subjects with pregnancy disorders based on the time series of blood pressure [53]. The blood pressure of fifteen pregnant women with pregnancy-induced hypertension and 34 pregnant women with preeclampsia were recorded over a 30 minute period. For each group, HMMs were designed to describe the systolic blood pressure data over the 30 minutes. Several HMMs that varied in the number of states from 5, 10, and 15 states were trained resulting in a total of 36 models. All models were finite state HMMs with ergodic topologies and

trained with the Baum-Welch algorithm. The sequences of each subject was tested against the two HMMs and classified to a model based on the likelihood probability. A table of the number of correct and incorrect classifications for each HMM classification was constructed and a Fisher's exact statistic was calculated to assess the significance of the classifications.

The results of the study showed that 5-state HMMs could significantly classified sequences and correctly identify 80% of the preeclampsia subjects and 62% of the pregnancy-induced hypertension subjects. The 10-state HMMs also showed significant classifications and correctly classified 91% of preeclampia subjects and 47% of the pregnancy-induced hypertension subjects. The 15-state HMMs did not identify significant classifications and classified all subjects to the preeclampsia HMM. The authors concluded that either the 5-state or 10-state HMMs could sufficiently characterize the different blood pressure variations in patients with preeclampsia and pregnancy-induced hypertension, indicating the problems in identifying the appropriate number of states in HMMs. The author concluded that the significant classifications of subjects with HMMs suggest the etiology of the pregnancy disorders is dissimilar [53].

Cooper and Lipsitch evaluated whether structured HMMs were appropriate model to describe hospital infection data of three classes of pathogens: methicillin-resistant Staphylococcus aureus (MRSA), vanomycin-resistant enterococci (VRE) and third generation cephalosporin-resistant Gram-negative rods (R-GNR) [54]. The authors defined structured HMMs as models that incorporate epidemic process of infection. In these types of infections, asymptomatic individuals carry the pathogen without developing an infection while another proportion of individuals within the sample develop the infection. The structured HMM is based on a Poisson distribution and includes the epidemic process of the infections with variables for

the rate of transmission, rate patients are discharged, and the probability that the patient is already carrying the pathogen.

Data of MRSA infections were collected from a 10 hospital bed ward over 45 months and data of VRE and R-GNR were collected from a 16 hospital bed ward over 42 months. Three models were evaluated to determine the appropriate model for each of the pathogens. These models were a Poisson distribution model, 2-state HMMs with a Poisson distribution and structured 2-state HMMs with a Poisson distribution. Simulated data of 1000 monthly infections were generated from each model and statistics were calculated to evaluate the fit of the model to the data. The results showed that the structured HMM was the most appropriate HMM for the MRSA and VRE data. The Poisson distribution model was found as the appropriate model for the R-GNR data. The authors concluded that the new approach of the structured HMMs is an improvement over standard Poisson distributions models especially for MRSA and VRE data.

2.4 CLUSTER ANALYSIS

The HMMs in this project were designed from a data reduction procedure, which is described in Chapter 3. The data reduction procedure was applied to the lifting parameters and involves factor analysis and cluster analysis. Since clustering methods are exploratory and will produce a solution even if the solution is incorrect, the reliability and validity of the cluster solution must be assessed [55]. Invalid cluster solutions can occur when the number of clusters in the data is arbitrarily chosen. For instance, if the number of clusters is underestimated, information could be lost due to merging of clusters.

Milligan and Cooper suggested seven steps to cluster analysis process based on a review of the literature [56]. These steps are: (1) select entities to cluster, (2) select variables of the entities to cluster, (3) decide whether to standardize the data, (4) chose a similarity or dissimilarity measure that separates the data into clusters, (5) chose cluster analysis method, (6) determine number of clusters in data and (7) interpret, test and replicate the cluster solution. The first two steps of the process are related to the researcher's study design. Aldenderfer and Blashfield emphasized that selecting variables is not trivial and should ideally be selected within the context of explicitly stated theory that will support the classifications [57].

Once the variables have been selected, the next step is to decide whether to standardize the variables in the data. If the scales and/or magnitudes of the variables are different, the cluster analysis will be biased by the higher magnitude variables. To avoid bias, the data may be standardized to the normal distribution or the data can be normalized to the range of magnitudes of the variables. The fourth step is to choose a similarity measure and this choice is usually based on the data. The four common measures are correlation coefficients, distance measures, association coefficients and probabilistic similarity coefficient. A distance measure was used in this thesis. The distance measure calculates the distance between the data and each of the cluster centriods. The data is assigned to the cluster with the minimum distance.

The fifth step is to determine the type of cluster analysis method to apply to the data. The two types that were used in this thesis were agglomerative hierarchical and iterative partitioning. Agglomerative hierarchical cluster analysis was used to determine the number of clusters in the data, and partitioning cluster analysis was used to separate the data into disjoint clusters. In the hierarchical cluster analysis, clusters are initially formed for each observation and then two clusters are joined until only a single cluster that contains the entire dataset is formed. There are

four common rules of joining the clusters during hierarchical cluster analysis: single linkage, complete linkage, Ward's minimum distance, and average linkage. In this thesis, Ward's minimum distance was used.

Techniques to determine the appropriate number of clusters in a sample were investigated by Milligan and Cooper with a simulation study [55]. The study created data sets that each contained 50 data points, either 2, 3, 4 or 5 distinct non-overlapping clusters and were embedded in either 4, 6, 8 dimensional Euclidean space. In addition to these experimental variations, two other factors were varied in the simulation data to assess the performance of the stopping rules. The first factor was the number of data points that were contained in each cluster and this factor had three conditions: (1) equal number of points in all clusters, (2) one cluster must contain 10% of the data points, and (3) one cluster must contain 60% of the data points. The addition of the third factors resulted in 36 cells of conditions. Since the cells were replicated 3 times, a total of 108 data sets were used in the study.

The final factor that was tested in the simulation involved the dissimilarity measure used to separate the clusters during the hierarchical cluster analysis. The four methods that were assessed were Ward's minimum variance, group average, complete linkage and single linkage. This increased the design of the simulation studies to 432 test conditions. The stopping rules were evaluated by whether the rule could detect the true number of clusters in the data. Two external criterions: the adjusted Rand statistic [58] and the Jaccard index [59]. Only stopping rules that were automatic and method independent were assessed in this study.

The simulation evaluated the utility of 30 different stopping methods to detect the correct number of clusters in the data for the different conditions in the hierarchical cluster analysis of the data. The simulation study found that the pseudo F statistic of Calinski and Harabasz [60]

and $J_e(2)/J_e(1)$ rule of Duda and Hart [61] were the two best methods to detect the correct number of clusters within a sample. The Calinski and Harabasz index is calculated as the

$$\left[trace\, B/(k-1)\right]/\left[trace\, W/(n-k)\right]$$

In this equation, k is the number of clusters in the solution, n is the total number of items, B is the cross product matrix of between-cluster-sum-of-squares and W is the cross product matrix of the pooled within-cluster-sum-of-squares. The maximum value of the statistic indicated the number of clusters in the data. The Je(2)/Je(1) index is the sum of squared errors within the clusters when the data is separated into two clusters divided by the sum of squared errors when only one cluster is present. If the ratio is smaller than the critical value, the hypothesis that one cluster is in the data is rejected. The critical value used by Milligan and Copper was 3.20. The data was partitioned into clusters until the hypothesis was first rejected.

The results of the simulation showed that pseudo F statistic identified the correct number of clusters in 390 of the 432 data sets and the Je(2)/Je(1) identified the correct number of clusters in 388 of the 432 data sets. The number of errors with the pseudo F statistic was consistent over the varying number of clusters in the data, and the Je(2)/Je(1) statistic had more errors when the true cluster number was 2 clusters. Although these statistics performed well and this study provides validated stopping rules, the authors emphasize that the findings might be data dependent.

McIntyre and Blashfield proposed a method of evaluating the reliability of cluster solutions [62]. The method randomly separates the data into two equal samples: a training sample and a test sample. The training sample is clustered into disjoint clusters and the statistics of the cluster solution is determined, such as the mean and the standard deviation of the clusters. To assess the reliability of the cluster solution, cluster analysis is applied twice to the test sample, and the two cluster assignments are compared with a kappa statistic. The first cluster assignment of the test

sample is determined by assigning the test sample data to the clusters of the training sample. In this analysis, the cluster means and standard deviations do not change and the test data is assigned to the nearest cluster based on the cluster statistics of the training sample. The second cluster assignment of the test sample is determined by directly clustering the test sample data. The two different cluster solutions of the test sample are compared with a kappa statistic to assess reliability of the cluster solution. Kappa statistics of 1 to greater than 0.8 indicate excellent reliability, from 0.8 to greater than 0.6 indicate substantial reliability, from 0.6 to greater than 0.4 indicate moderate reliability, from 0.4 to greater than 0.2 indicate fair reliability and kappa statistics from 0.2 to 0 indicate slight to no reliability [63].

The methods of determining the number of clusters in the data and assessing reliability of the cluster solution that were reviewed in this section are used in Chapter 3.

3.0 METHODS

A procedure to identify sub-groups of CLBP subjects based on their time series data during a repetitive lifting task was developed in this project and is described in this chapter. The Chapter is separated into 3 sections. The first section describes the protocol of the clinical study on which the present study is based, the second section describes a data reduction procedure that was used to determine lifting patterns and the last section describes the design of the hidden Markov models.

3.1 CLINICAL STUDY PROTOCOL

The data used for this project was from a database that contains lifting parameters, medical findings, and self-reported measures collected during a clinical study conducted at the University of Pittsburgh Medical Center Pain Evaluation and Treatment Institute [14-16]. During the clinical study, subjects completed three evaluations: medical, psychological and functional capacity. The medical evaluation involved a general medical screening and MRI images of the spine of CLBP subjects. The psychological evaluation assessed perceived disability,

psychosocial measures, pain frequency and intensity, and cognitive coping abilities based on the subject's response to several questionnaires.

The functional capacity evaluation assessed physical functioning and involved a repetitive lifting task that had been designed to quantify motion of CLBP patients and pain-free controls. During the lifting task, subjects repeatedly lift a resistant load and reflective markers track segmental motion. Several lifting parameters were calculated from the motion data of each lift resulting in a time series of parameters for each subject. In this section, the subject demographics, protocols of the medical evaluation, psychological evaluation and functional capacity test as well as the lifting parameters calculated during the repetitive lifting task are described.

3.1.1 Subjects

One hundred thirty-four subjects, 53 pain-free control and 81 CLBP patients, participated in the repetitive lifting study. Subjects were defined as CLBP subjects if the subject reported having pain everyday or almost everyday for the past three month that was of moderate or greater intensity. The age of the subjects ranged from 36-63 years. The average age of control subjects was 34.5 yrs (standard deviation = 11.8 yrs.) and CLBP subjects was 37.8 years (standard deviation = 10.1 years). Both groups were approximately matched for gender. In the control group, 29 subjects were female and 24 subjects were male. In the CLBP group, 38 subjects were male and 43 were female. All subjects gave written informed consent as approved by the University of Pittsburgh Biomedical Institutional Review Board before the start of the study. In the CLBP patient group, all of the 81 patients had a history of prolonged back pain with mean

pain duration of 4.1 years (standard deviation = 5.4 yrs.) The pain-free control group was composed of adult volunteers with no medical history or current complaint of back pain. College students and competitive athletes were excluded from the control group. A physician and physical therapist evaluated all subjects to determine his or her ability to participate in the repetitive lifting task.

3.1.2 Medical Evaluation

The medical evaluation obtained information about the subjects' general health and pain history. For CLBP subjects, the second part of the medical evaluation consisted of magnetic resonance imaging (MRI) of the axial and sagittal sections of the back. CLBP subjects were scanned on 1.5T superconducting magnets. A standardize scoring system that assigns linear weights to the medical procedures according to the relevance in diagnosing chronic pain patients was used. Twenty-three medical procedures were listed in the medics form and diagnosis of each procedure was assigned a score [64]. For example, in neurological examination, the doctor chose either no abnormality, non-specific abnormality or significant abnormality with the scores of 0, 1 and 2, respectively. There was the possibility that one or more of the procedures would not have been performed on the patients. To account for lack of data, a proportion of the number of abnormalities to the number of procedures was calculated. The proportions were transformed into a logit score and combined with the linear weights of the procedures to compute a weighted logit score of pathology.

3.1.3 Psychological Evaluation

In the psychological evaluation, CLBP subjects answered a series of standardized self-reported inventories that assessed pain-relevant psychosocial aspects, perceived disability, self-efficacy, cognitive coping ability and frequency of daily activities. These inventories include the Pain Behavior Checklist [65], West Haven-Yale Multidimensional Pain Inventory (MPI) [3-4], Task Self-efficacy [66], Coping Strategy questionnaire [67], and Oswestry Low Back Disability Scale [36]. None of the measures were collected on control subjects because the measures assess pain impacts and the control subjects are pain-free.

The Pain behavior checklist asked subjects the frequency that they performed certain pain behaviors. These pain behaviors were separated into categories of facial/ audible expressions of distress, distorted posture, negative affect and avoidance of activity [65]. The facial/audible expressions category consisted of facial grimacing, sighing, moaning, and clenching teeth. The distorted posture category were limping, walking with distorted gait, moving extremely slow, moving in a guarded or protective fashion, sitting with a rigid posture, stooping while walking, frequently shifting posture and supporting, rubbing, or holding affected body area. The negative affects category were irritability, requesting to be excused from activities, seeking help in ambulating and questions such as why did this happen to me. The avoidance of activities consisted of taking analgesic medication on a schedule, using prosthetic devices, lying down frequently during the day and avoidance of physical activity.

The West-Haven Multidimensional Pain Inventory (MPI) assessed pain-relevant psychosocial aspects, responses of the significant other to the patient's pain and frequency of common activities [3-4]. CLBP subjects completed the three sections of the MPI. The first

section contains five sub-sections that assess the pain patients (1) report of pain severity, (2) perceptions of how pain interferes with their lives, (3) appraisals of the amount of support received from significant others, (4) perceived life control and (5) affective distress. In each of the questions of the section 1, subjects were asked to rate their response on a scale of 0 (none) to 6 (extreme). The second section of the MPI assessed the responses of significant others to the patients' pain with three sub-sections: (1) Punishing Response, (2) Solicitous Responses and (3) Distracting Responses. Subjects were asked to assess how frequently their spouse or significant other used a type of response when the subject is in pain on a scale of 0 (never) to 6 (very often). The last section of the MPI asked CLBP subjects how often on a scale of 0 (never) to 6 (very often) that the subject completed 19 common activities. The 19 activities are grocery shopping, gardening, mowing the lawn, washing the dishes, going to the movies, playing cards or games, visiting friends, house cleaning, working on the car, riding in car or bus, visiting relatives, preparing meal, washing car, taking a trip, going to park or beach, doing laundry, household repair and engaging in sexual activity.

The responses to the nine sections of the MPI were analyzed with the MPI computer program version 3 [68]. In this version, the cluster analysis method was not used to assign subjects to groups since this method resulted in errors when subjects did not fit into one of the discrete clusters. The MPI program assigns subjects to either the dysfunctional or interpersonally distress composite scores based on their responses and these scores are continuous values [68]. The MPI composite scores were based on factor analysis of 6,545 heterogeneous chronic pain patients.

Perceived self-efficacy is one's belief that one can perform a specific task [69]. A task self-efficacy questionnaire asked subjects to rate on a scale ranging from 0% (very uncertain) to

100% (very certain) how confidently they felt they could perform a task for a period of time or number of repetitions [66]. For example, a subject were asked if he or she could walk on a level surface for (a) 1 minute, (b) 5 minutes, (c) 15 minutes, (d) 30 minutes, (e) 45 minutes, (f) 60 minutes and (g) longest time they could walk, and to rate how confident they were in their answer on confidence scale of 0% to 100%. Another question asked the subject if they could lift (a) 10 pounds, (b) 20 lbs, (c) 30 lbs, (d) 40 lbs, (e) 50 lbs, (f) maximum weight, and how confident they were in their answers. The other tasks were remain standing, sitting in a comfortable chair, walk up steps, how many times they could lift their self-reported maximum weight, and turning a wheel with effort ranging from 10-50 lbs at 10 lb intervals.

The Coping Strategies Questionnaire assessed the cognitive coping strategies that patients used when experiencing pain [67]. The questionnaire asked patients to rate on a scale of 0 (never) to 6 (very often) how often the patients experienced certain thoughts or feelings when the pain was very severe. These feelings include frustration, worry, irritability, anger, anxiety and catastrophizing statements such as I think about whether life is worth living.

The Oswestry low back disability rating scale is a 10 item questionnaire that measured the patient's perceived level of disability due to pain during activities of daily life including walking, lifting, personal care, sitting, standing, sleeping, traveling, sexual activity and social life [36]. Subjects chose one of six answer statements to describe how pain has impacted to subjects' ability to manage everyday life. For example, the personal care statements are (1) I can look after myself normally without causing extra pain, (2) I can look after myself normally but it causes extra pain, (3) It is painful to look after myself and I am slow and careful, (4) I need some help but manage most of my personal care, (5) I need help everyday in most aspects of self care, and

(6) I do not get dressed, wash with difficulty, and stay in bed. Higher scores on the Oswestry disability rating scale indicate greater levels of perceived disability.

3.1.4 Functional Capacity Evaluation

The functional capacity evaluation assessed the physical functioning of the subjects with the Jan van Breemen examination and a repetitive lifting task. The Jan van Breemen examination involved a self-report of pain intensity and functional status, and involved an examination of the subject's spinal mobility by a physical therapist [70]. The repetitive lifting task measured the subject's motion and physical endurance. Both control and CLBP subjects completed the repetitive lifting task.

3.1.4.1 Jan van Breemen examination

CLBP subjects were asked to respond to a set of standardized questions that concerned physical functioning in relation to their back pain in the past week. There are three sections to the Jan Van Breemen: pain scores, functional status, and spinal mobility exam. The pain scores section asked subjects to rate on a scale of 0 (no pain at all) to 10 (unbearable pain) how much backache the patient suffered during the past week (1) in general, (2) at night (3) during the first hour of the morning, (4) during sitting, (5) during walking, and (6) during standing. The functional status scores section asked subjects to rate on a scale of 0 (very bad/ impossible) to 10 (very good/ normal) how well during the past week he or she was able to (1) carry, (2) walk, (3) sit, (4) lift, (5) stand, (6) to outdoors, (7) sleep, (8) perform household and hobby activities and (9) perform occupational activities in relation to their back problem.

The spinal mobility exam measured lumbar flexion index and lumbar flexion/extension index and was measured by a physical therapist. For the flexion index, the physical therapist marked the position of lumbosacral junction and the position of a point at 15 cm above the lumbosacral junction when the subject was standing. The subject was then instructed to bend forward to touch the floor with his or her fingertips while keeping the knees straight. The therapist measured the distance between the marked points on the back of the subject as the subject was bent forward. The flexion index is the distance between the two marked positions when the subject bends forward subtracted by the original distance of 15 cm. For the flexion/extension index, the same marked positions on the subject's back are used and the subject is asked to lean backwards. The difference between the two markings when the subject is leaning backwards is measured. The flexion/extension index is calculated as the sum of the distance of the marked positions when the subjects is in flexion and the distance between the marked positions when in the subject is in extension subtracted by 15 cm (the distance between the marked positions at standing).

3.1.4.2 Repetitive lifting task Protocol

The lifting task required control subjects and CLBP subjects to repeatedly lift a handle attached to a resistant load located 13 inches from the ground to waist height. The BTE Work Simulator (Baltimore Therapeutic Equipment Company, Baltimore, MD, USA) provided the resistant force for the up-phase of the lift. Subjects performed lifts for a total of twenty minutes with a fifteen second rest interval between each of the lifts, during which the subject returned to the standing position. Four hemispheric infra-red reflective markers, placed on the left side of the body, tracked joint motion throughout the lifting task. Markers placement was on the ankle, apex of the patella, greater trochanter of the femur, and acromion of the shoulder. Figure 1 shows a picture

of a subject lifting the handle of the BTE work simulator and the placement of the reflective markers during the lifting task.

Figure 1: Subject is lifting the handle of the BTE work simulator. The reflective markers located on the subject's joints tracked motion during the lifting task.

The resistant force applied during the lifting task was equal to 40% of each subject's maximum voluntary static strength. Maximum voluntary static strength was measured with a force gauge (Chatillon Muscle Strength Dynamometer, Sammons Preston, Bolingbrook, IL, USA) attached to a platform. Subjects were instructed to assume a bilateral symmetrical leg lift position with the forearm in supination and the handle of the force gauge adjusted to knee height. The subject was then instructed to pull on the force gauge for approximately four seconds. This process was repeated three times with a fifteen second rest period between each attempt, during which the subject was instructed to return to a standing position. The average static strength of the three trials was calculated and 40% of the average static strength was used as the resistant force in the lifting task.

Before the start of the repetitive lifting task, each subject was given the opportunity to practice the lift without resistant load applied to the handle in order to become familiar with the task. Once the subject was comfortable, the resistant load was applied and the repetitive lifting task began. Throughout the experiment the subject was given no verbal or visual feedback concerning performance. The task was terminated: (1) if the subject felt physically unable to continue, (2) if the experimenter stopped the task due to unsafe body biomechanics or (3) the time limit was reached.

Subjects were asked to rate their pain intensity at the beginning of the functional capacity evaluation (baseline rating), after the static strength task and at the end of the dynamic lifting task. The pain rating is a self-reported measure that asked the subject to rate their pain on a scale from 0 (no pain) to 10 (extreme pain) [71].

3.1.4.3 Instrumentation

The BTE work simulator provided the resistance for each subject to work against during the repetitive lifting task. The work simulator is a computerized device that maintains constant force on the handle during the lift. The force transmission occurs at the handle by a rope connected through a pulley system. The starting height, waist level (ending height) and force were programmed into the work simulator before beginning the task. A series of tones instructed the subject when to lift and when to lower the handle. A high tone indicated that the rest period was over and the subject was to perform a lift. A low tone indicated that the handle had reached waist height and the subject could return the handle to the holder. A second, lower tone indicated that the subject had placed the handle in the holder, initiating the rest period. The BTE software (Baltimore Therapeutic Equipment Company, Baltimore, MD, USA) allowed the subject to perform the lifts at his or her own pace. The work simulator recorded the force and handle velocity at a sampling rate of 50 samples per second.

Motion Analysis Model 110 Video Processor using Expert Vision Software (Motion Analysis Corporation, Santa Rosa, CA, USA) and an NEC TI-23A CCD camera with LED ring-light tracked the retro-reflecting markers attached to the subject. The motion analysis system tracked the markers at 30 frames per second during the up-phase of the lift by detecting the marker boundaries. A schematic of the instrumentation used during the repetitive lifting task is shown in Figure 2.

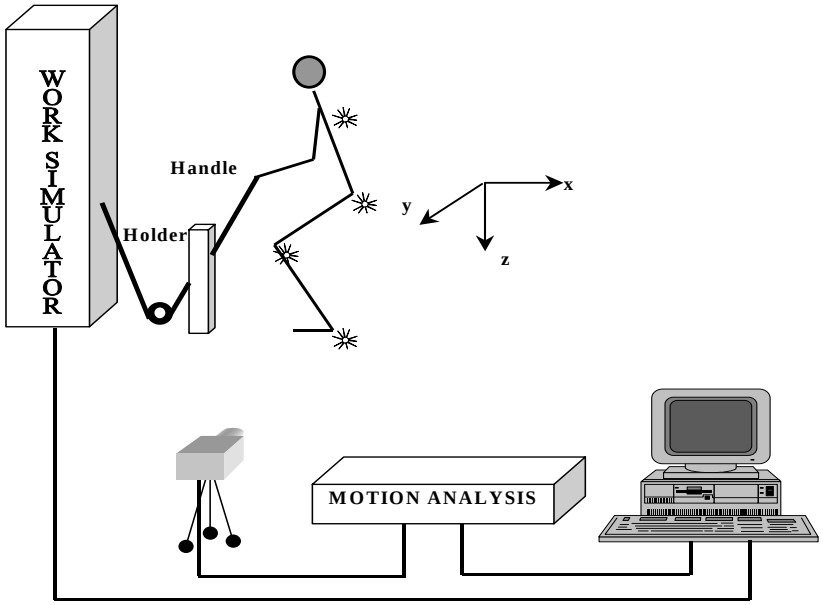


Figure 2: Schematic of the instrumentation used in the repetitive lifting task

3.1.4.4 Lifting Parameters

The lifting parameters calculated to describe motion during the lifts were starting posture, lift duration, hip and knee midpoint, hip and knee risetime, midpoint difference, risetime difference, starting knee angle, starting hip angle, rms jerk, maximum jerk and time when maximum jerk occurred during the lift. These measures describe basic body biomechanics used by the subject for each lift.

A two-dimensional three-segment biomechanical model was constructed from the motion of the four joint markers [72]. The three segments were defined as the shank, thigh and trunk. From the model, joint angles were defined from the segment angles as shown in Figure 3. The knee angle was defined by the angle between the shank and thigh segments, and the hip angle was defined by the angle between the thigh and trunk segments. Full extension was defined as

zero degrees. The starting knee angle and the starting hip angle were the values of these angles at the start of each lift.

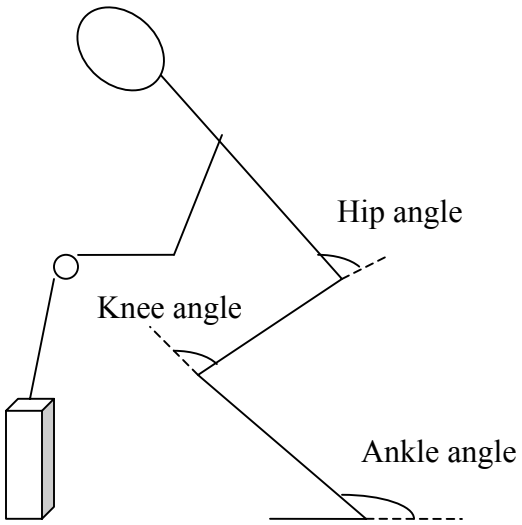


Figure 3: Three-segment model used to describe motion of the subjects during the task. The figure shows the orientation of the ankle angle, knee angle and hip angle.

The index of starting posture was derived as starting hip angle minus starting knee angle divided by the starting hip angle, creating an index that ranged from -1.0 to 1.0 [14]. Values approaching -1.0 indicate a starting posture characterized by a squat lift in which the back is kept vertical and the hip and knees are flexed at the start of the lift. Values near 1.0 indicate a torso style lift in which the back and hips are flexed and the knees are kept straight at the start of the lift. Values around zero indicate a freestyle lift in which the back, hip, and knees are flexed. Lift duration is the time during which the BTE work simulator applied resistance to the handle.

There were two parameters that described the timing of motion of the hip and knee angles: midpoint and risetime. In order to calculate these parameters, the knee and hip angles were fit to

hyperbolic tangent equations to describe changes of body angles as a function of time [14, 23]. The hyperbolic tangent equation described motion of the hip, knee and ankle angles using four parameters: midpoint, risetime, starting angle and ending angle [14]. The midpoint is defined as the time after the beginning of the lift at which the body angle had completed half of its range of motion. The risetime is the time required for the angle to decrease from 88% to 12% of the total change in angle. Midpoint difference is the difference between the hip midpoint and the knee midpoint. Risetime difference is the difference between the hip risetime and the knee risetime. The parameters of risetime, midpoint, risetime difference and midpoint difference were normalized by total lift duration to eliminate differences due to lift duration. An example of a knee angle trajectory for lift performed by a control subject is shown in Figure 4. The hyperbolic tangent parameters of starting angle, midpoint, risetime and the ending angle are labeled on the knee angle trajectory in the figure.

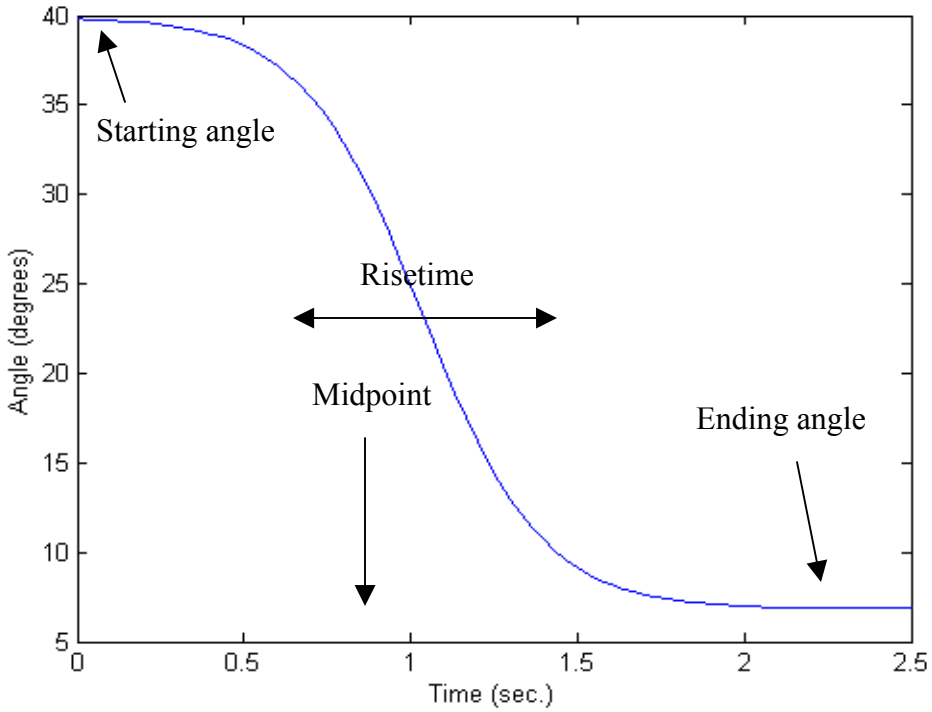


Figure 4: Knee angle plotted versus lift time. The hyperbolic tangent equation was fit to the knee angle displacement using the parameters of midpoint, risetime, starting angle and ending angle, and these parameters are labeled on the knee angle trajectory.

Three measures related to jerk were used to describe lifting patterns: rms jerk, maximum jerk, and time when maximum jerk occurs in the lift. Since jerk is the third derivative of position, it is a noisy measure and smoothing methods are necessary to reduce noise. Several smoothing methods, including Woltring's generalized cross validation spline [27], Wells and Winter method of low-pass filtering [28-29], and the use of the hyperbolic tangent model [14] were assessed [25]. The hyperbolic tangent model was used to calculate jerk in this project because it provided smoother estimates that allowed for maximum jerk to be identified on the jerk waveforms. The spline and filtering methods produced jerk waveforms that were too variable to measure maximum jerk.

The hyperbolic tangent model used the equation described in the previous paragraph to generate the angles of the ankle, knee and hip. For each lift, the midpoint, risetime, starting value and lift duration of each angle were entered into the hyperbolic tangent equation and the body angles were calculated. Position data of the body segments were calculated using body segment link lengths and the calculated angles. Body segment link lengths were determined as the magnitude of the difference between the distal marker position and the proximal marker position at each time point during the lift. For example, the length of the lower leg was the difference between the knee marker and the ankle marker for a given subject. The link lengths were smoothed with an 8^{th} order polynomial before the position data were calculated.

Jerk was calculated as the magnitude of the third derivative of the shoulder displacement. Derivatives were estimated using central finite differences. Before differentiation, the displacement data were extrapolated by one second at the beginning and at the end of the data to eliminate numerical errors associated with finite differences. These extrapolated points were eliminated from the third derivative, and a root-mean-squared (rms) measure of jerk was calculated for each lift performed. The hyperbolic tangent model was used in this project to calculate jerk because, as shown in Figure 5, the model produced smoother jerk profiles than spline and low-pass filtering, and these smoother profiles permitted characterization of maximum jerk. It was not feasible to characterize maximum jerk with the jerk profiles calculated with splines or low-pass filtering because variability was too large. The jerk profiles when smoothing with splines or filtering were found to oscillate over lift time, resulting in waveforms that contained multiple peaks as shown in Figure 5. Although both of these jerk waveforms (top graphs in Figure 5) contain one peak with a larger magnitude than the other peaks in the waveform, it was not clear whether the larger magnitude peak was appropriate to use as a

measure of maximum jerk. In addition, several of the waveforms calculated with the low-pass filtering and splines contained multiple peaks or oscillations that were approximately equal in magnitude, making it difficult to identify maximum jerk. Since the hyperbolic tangent model produced smoother profiles in which maximum jerk could be easily identified, it was used to obtain estimates of jerk for these data.

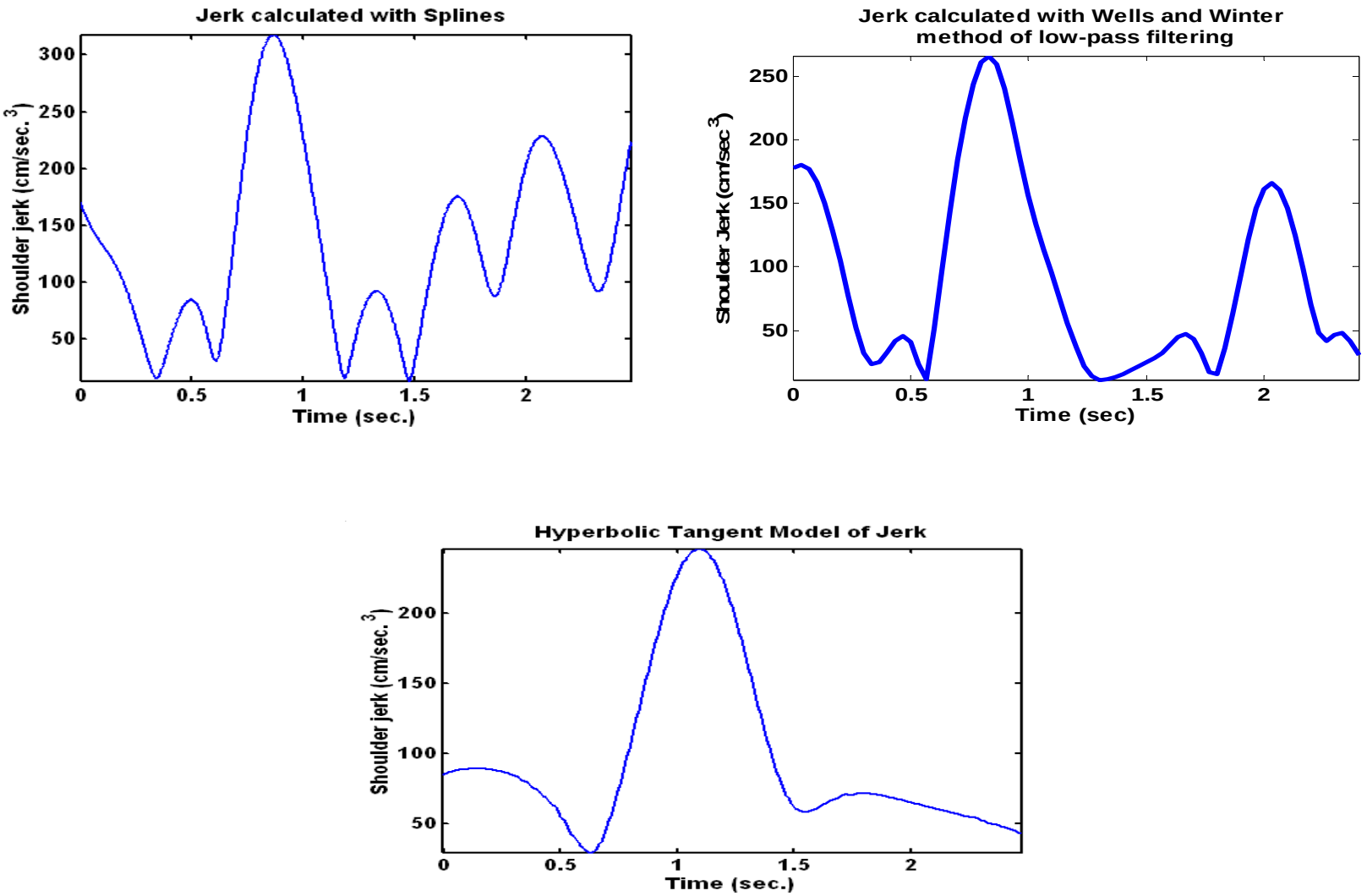


Figure 5: Shoulder jerk plotted versus lift duration. The top right graph is jerk calculated with hepatic spline, the top left graph is jerk calculated with Well and Winter's filtering method and the bottom center graph is jerk calculated with hyperbolic tangent equation

All of the lifting parameters were calculated for each lift that a subject completed and these parameters are listed in Table 1. The total number of lifts completed was 4762 (Controls 3437 lifts & CLBP 1325 lifts). Each of the parameters was checked for outliers, violations of heteroscedasticity and violations of non-linearity using histograms and normality plots in SYSTAT 11 [73]. The parameters of rms jerk, maximum jerk and lift duration violated the heteroscedasticity assumption, and a logarithm transformation was performed to normalize these parameters.

Table 1: List of the 13 lifting parameters that were calculated to describe motion during each lift

Lifting parameters		
Starting hip angle	hip midpoint	RMS jerk
Starting knee angle	knee midpoint	Maximum jerk
Starting posture index	hip risetime	Time at maximum jerk
	knee risetime	lift duration
	risetime difference	
	midpoint difference	

3.1.5 Treatment Protocol

Each of the CLBP patients was enrolled in an intensive rehabilitation program after completing the study protocol. The CLBP patients attended the Pain Evaluation and Treatment Institute daily for 8 hours for a 3 ½ week period. The rehabilitation program focused on discussing pain and the impact of pain on the individual's life to educate patient's to mange their pain. Patients are taught stress management skills, relaxation techniques, problem-solving, distraction skills, coping strategies, body mechanics, physical exercises to increase endurance, and flexibility [74].

Treatment included a combination of group and individual physical, occupational, and psychological therapies. During treatment, subjects did not have any training on the BTE work simulator. Once the rehabilitation program was completed, the CLBP patients completed the self-report evaluation and functional capacity evaluation using the same protocol.

3.2 DATA REDUCTION PROCEDURE

A data reduction procedure was developed to combine the multidimensional lifting parameters into a discrete set of lifting patterns. The procedure used factor analysis to reduce the number of parameters describing motion during the task and k-means cluster analysis to assign each lift to a cluster that describes a lifting pattern. Both of these methods are described in this section.

3.2.1 Factor Analysis

Each lift that a subject performed during the repetitive lifting task was described by the 13 different parameters, and each of the parameters changed over task time. The lifting parameters were highly variable within and between the subjects and between lifts. In examining the parameters, it appeared that redundancy might be a problem, since several of the parameters were describing the same motion but in different areas of the body. For example, posture at the beginning of the lift is described with starting knee angle, starting hip angle and a starting posture index. To reduce the redundant parameters and combine the multi-dimensional parameters into discrete lifting patterns, a data reduction procedure was applied to the lifting

parameters. The procedure used factor analysis to reduce the redundant parameters to four factors, and k-means cluster analysis to assign each lift to one of five clusters based on the four factor scores of the lift.

Pearson correlations were calculated between all parameters. If correlations between two parameters were high, only one of the parameters was included in the factor analysis. The only two parameters with high correlation were maximum jerk and rms jerk (Pearson correlation of 0.999), and rms jerk was retained in the matrix. Factor analysis was applied to the correlation matrix of the lifting parameters using principal component analysis with a Varimax rotation [75]. Since CLBP subjects completed fewer lifts than control subjects, the factor structure could be biased by the control data. To avoid bias, the groups were separated and factor analysis was applied to the control data and CLBP data separately to obtain an invariant factor structure, i.e. a factor structure that was not biased by group membership.

The number of factors extracted was determined as the number of eigenvalues of the correlation matrix that were greater than one [76]. The parameters of risetime difference, hip angle risetime, knee angle risetime and starting posture were eliminated from the factor analysis because these parameters loaded on multiple factors or did not load sufficiently well on any factor for one group or both groups. Once an invariant factor structure was found, the lifting parameters of the groups were combined and factor analysis was applied to the entire data set in order to calculate standard normal factor scores that are standardized to the mean and standard deviation of the entire sample of lifts. The invariant factor structure that was found for both group contained four factors with the same parameters loading on the each of the factors. All of the parameters had loadings of 0.75 or greater for a single factor and loadings of less than 0.45 for all other factors when factor analysis was applied to the entire data set as shown in Table 2,

indicating that a simple structure was obtained. The four factors were found to explain 89% of the variance in the data.

The factors were labeled to describe the lifting parameters that loaded on each of the four factors. The parameters of maximum jerk time, knee angle midpoint and hip angle midpoint had large loadings on the first factor. Since these parameters describe the time in the lift when certain motions occurred, the first factor was named the timing factor. The second factor was labeled as the posture factor because the parameters of starting knee angle and starting hip angle had high loading on the second factor and these parameters describe position of the hip and knee at the beginning of the lift. The third factor had only midpoint difference with a large loading on this factor. Since midpoint difference describes the relative timing of the knee and hip angle motion, the factor was named synchrony. The final factor contained the parameters of rms jerk and lift duration and was named the speed factor.

Table 2: Loading of the lifting parameters on the four factors

Lifting parameter	Factor 1 Timing of motion	Factor 2 Starting posture	Factor 3 Synchrony	Factor 4 Speed
hip angle midpoint	**0.956**	0.075	0.142	0.198
maximum jerk time	**0.943**	0.024	0.041	0.200
knee angle midpoint	**0.888**	0.031	0.402	0.180
Starting hip angle	0.070	**0.831**	0.096	0.067
Starting knee angle	0.020	**0.791**	0.175	0.023
midpoint difference	0.021	0.071	**0.987**	0.003
root-mean-squared jerk	0.134	0.175	0.053	**0.952**
lift duration	0.347	0.071	0.075	**0.904**

3.2.2 Cluster Analysis

A k-means clustering algorithm was applied to the factors scores of each lift to transform lifts into one of several clusters that described lifting patterns. The number of clusters in the factor scores data were determined with the pseudo F [60] and pseudo T^2 statistics (transformed $J_e(2)/J_e(1)$ statistic of Duda and Hart [61]). These statistics were calculated with hierarchical clustering using Ward's minimum-variance distance (algorithm CLUSTER in SAS) for cluster numbers ranging from 1 to 15. The statistics were then plotted versus cluster number as shown in Figure 6. The number of clusters was determined as the number that produced a peak in the pseudo F statistic combined with a small value of the pseudo T^2 statistic [77]. Using this criterion, Figure 6 indicates that an appropriate number of clusters for these data is five.

The FASTCLUS k-means algorithm in SAS software assigned each lift to one of the five clusters based on the factor scores of the lift. The initial estimates, or seeds, of the clusters are the first few observations in the file. The SAS program assigns each lift to the nearest seed that minimizes the Euclidean distance between the four factor scores of the lift and cluster seeds, resulting in the formation of temporary clusters. Once all observations are assigned to a cluster, the cluster means are calculated and used as the new seeds. The SAS program then uses the new seed and assigns the observations to the nearest cluster based on the Euclidean distance. This process repeats until changes in the seeds are small.

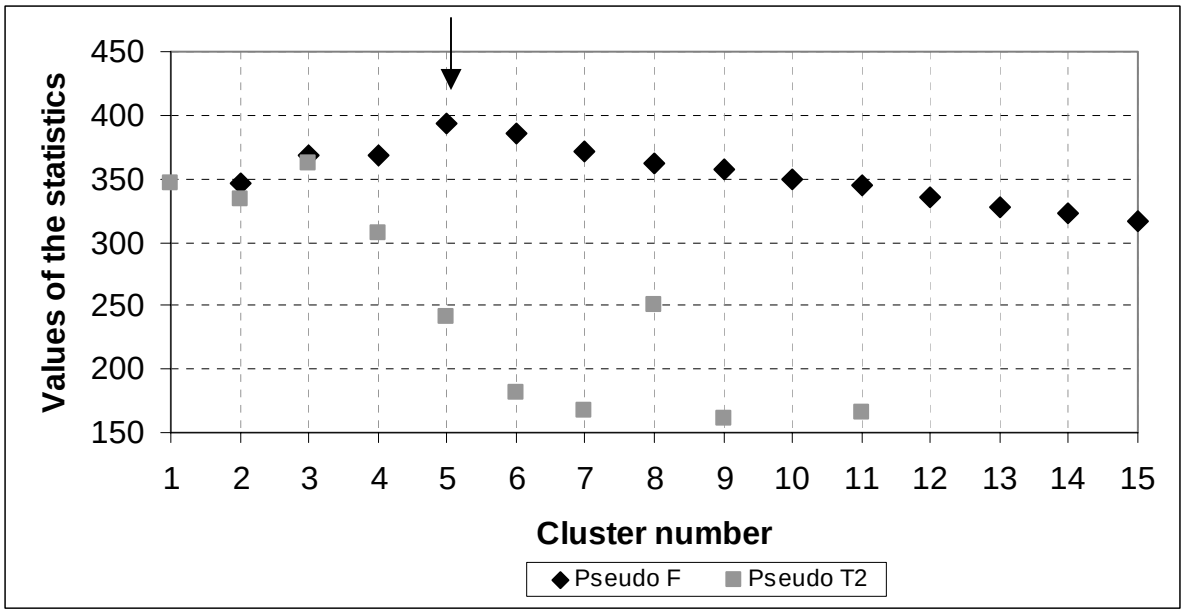


Figure 6: Plot of pseudo F and pseudo T^2 statistic versus cluster number. Squares indicate the pseudo T^2 statistic and the diamonds are the pseudo F statistic

Reliability of the cluster solution was assessed based on the methods described by McIntyre and Blashfield [62]. All of the lifts were randomly separated into two equal samples of control and CLBP lifts: training sample and test sample. The FASTCLUS k-means algorithm assigned each lift of the training sample to a cluster based on Euclidean distances and the five cluster centroids of the training sample were calculated. Each lift in test sample was then assigned to the nearest neighbor cluster based on the cluster centroids of the training sample, and the cluster solution of the test sample was determined. To assess reliability of the solution, the FASTCLUS algorithm was applied to the test sample without any constraints on the centroids and each lift was assigned to one of five cluster based on Euclidean distances. The two cluster solutions of the test sample were compared, using a kappa statistic to determine reliability.

The kappa statistic comparing the cluster solutions of the test sample was 0.95, which indicates that the clusters have excellent reliability. A plot of cluster assignment versus group

membership for all the data are shown in Figure 7. The biggest difference in the cluster compositions are in clusters 2, 3 and 5 that contain more control lifts than CLBP subjects and cluster 1 that contain more CLBP subject's lifts than control subjects.

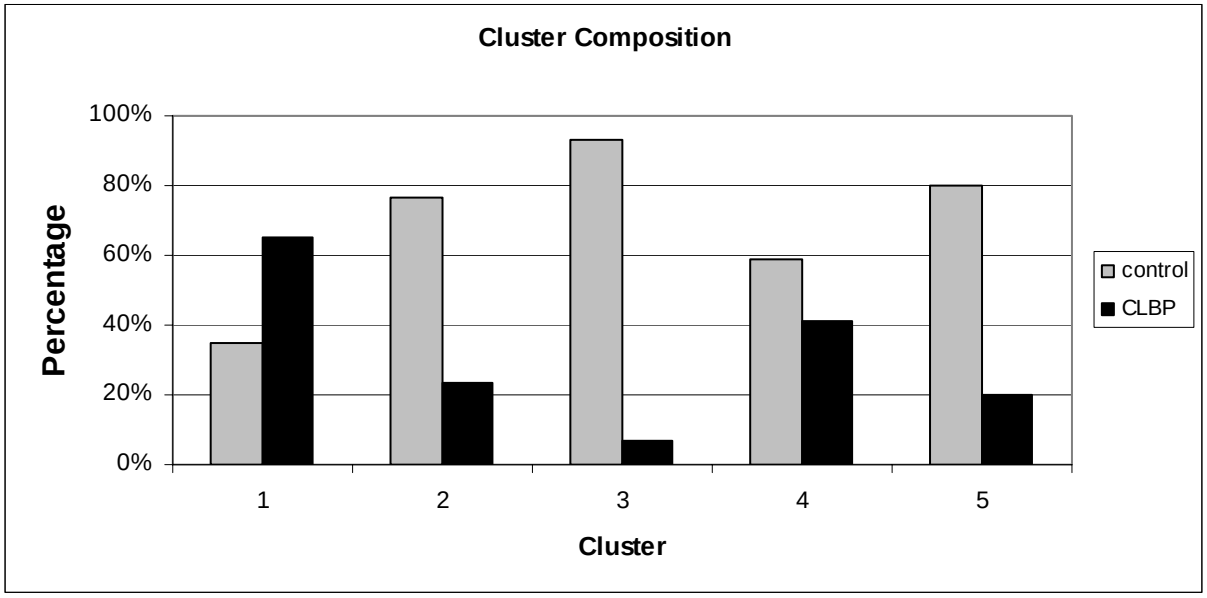


Figure 7: Cluster composition of the lifts for each group. Black bars are CLBP lifts and gray bars are control lifts

3.3 HIDDEN MARKOV MODELS

Since the temporal changes in several of the parameters were found to be significantly different between the groups, the HMMs were designed using the data reduction procedure to describe the temporal sequence of lifting patterns of individual subjects. HMMs can combine the temporal changes in a number of parameters and produce one sequence to describe lifting patterns of individual subjects. The phase separation model (early, middle and late phases), that was

previously used to describe the lifting parameters, created a separate model for each parameter that described the temporal changes in the parameter for groups of control and CLBP subjects [15-16]. The phase separation model ignored much of the data because only a few of the lifts in a trial were required to produce the model of each parameter. The HMMs can represent all of the lifts performed during the task in the sequence in which the lifting patterns were performed. The HMMs were constructed from the results of a data reduction procedure in order to simplify the structure of the models.

3.3.1 Design

Two HMMs were designed to describe lifting patterns of each group. The control HMM describes the lifting pattern time series sequence of controls and the CLBP HMM describes the lifting pattern time series sequence of CLBP subjects. The five clusters were defined as the output of the HMMs, and each state of the HMMs had a non-zero probability of generating each of the clusters. In the design of the HMMs, the output (or clusters) of the HMMs were called tokens. The input to the HMMs is the subject's lift sequence or token sequence, which was the sequence of cluster assignments (1 to 5) for the lifts performed by the subject.

The Baum-Welch algorithm [40-42, 51], with the multiple independent observation sequence modification [40-41], was used to train the HMMs. The modification was applied because each HMM describes the overall group lifting patterns during the task using the independent lift sequences of the subjects within that group. The hmmdecode program and a modified version of the hmmtrain program in the Statistical Toolbox of MatLab [78] were used to calculate the Baum-Welch algorithm. All of the control data were used to train the control

HMM, and all of the CLBP data were used to train the CLBP HMM. The training algorithm was iterated until the estimates of the transition, token and likelihood probabilities converged (difference between estimates of less than 1×10^{-6}). State 1 was designated as the starting state of all the sequences.

A common problem in mathematical modeling is determining the appropriate parameters to include in the model so that the model adequately describes the data. For HMMs, these parameters include the number of states, the topology and the initial conditions of the transition and token probability matrices. The initial probability matrices were chosen to be normalized uniform distributions, because they do not impose any structure on the HMM, resulting in probability matrices that are determined by training data. Defining the number of states and topology of the HMMs was a more difficult problem. One solution was to design the HMM with a fully-connected topology and three states since Vasko's previous work [21] determined this was the appropriate number of states to describe these data. There were several disadvantages to this solution. A fully-connected topology has a large number of parameters that require large training data sets to properly define the parameters of the HMM and more importantly, the fully-connected HMM have a more severe local minimum problem than a constrained HMM [44]. It is unclear whether three states would be appropriate for the data since there were differences between Vasko's work and this thesis in the data reduction to obtain tokens and the number of parameters used to describe lifting patterns.

In order to design HMMs with the appropriate topology and number of states, Vasko's et al. pruning algorithm [21] and the Viterbi algorithm [79] were used. The pruning algorithm iteratively removed state transitions from a fully-connected HMM until a single state HMM was obtained. Metrics were calculated at each step in the algorithm and used to identify the simplest

HMM. Since the pruning algorithm did not reduce the number of states in the HMMs for these data as was found in Vasko's work, the pruning algorithm was applied to fully-connected 4-state, 3-state and 2-state HMM of both groups to identify the appropriate topology for all possible variations in the number of states. The number of states of an HMM is called the order. The Viterbi algorithm was applied to the resulting HMMs from the pruning procedure to evaluate the appropriate order of the HMMs for each group.

3.3.2 Topology

The pruning procedure simplified the HMMs by removing state transitions from the model. At each step in the procedure, one interstate transition was set to zero in the initial transition probability estimates and the HMMs were trained with subject data. A zero state transition in the initial estimate resulted in a zero state transition probability in the trained HMM. The non-zero transitions in the initial transition estimates were normalized uniform estimates, which required the rows of the transition matrix to sum to one. For example, in the 4-state HMM, if transition from state 1 to state 2 was zero then all other transitions out of state 1 to another state would be equal to 1/3. None of the state self-transitions were pruned since eliminating self-transitions would create a null state as an intermediate step between states and this state would not be essential to the HMM topology. A state was removed from the HMM only when all of the interstate transitions to the state were pruned, and the procedure ended when a single state HMM was reached. The procedure was performed separately on the 4-state, 3-state and 2-state HMMs of the CLBP data and the control data.

At each of the steps in the pruning procedure, multiple trials were performed to determine which of the interstate transitions to remove without greatly impacting likelihood that the model fit the data. In each trial, a different interstate transition was eliminated, HMMs were trained and the likelihood probability was calculated. For example, step 1 pruned one transition from the fully-connected HMM. In the first trial of pruning step 1, the transition from state 1 to state 2 was pruned, the HMM was trained and the likelihood probability was calculated. In the second trial of step 1, the transition from state 1 to state 3 was removed, the HMM was trained and the likelihood probability was calculated. This process continued until all possible state transitions were individually pruned. Once all possible trials were completed, the HMMs were compared and the pruned HMM with the largest logarithm of likelihood probability of fitting the data was chosen as the most likely model for that step. The HMMs per step are described in the results.

3.3.3 Local minimum problem

A local minimum problem occurred when training the fully-connected HMM for all variations in the number of states. As stated previously, the initial estimates of the transition and token probabilities were normalized uniform estimates. The results of training showed that the token probabilities of the trained HMM were the same for all states and transition probabilities were the same as the initial estimates upon convergence. The logarithm of the likelihood was large and convergence to 10^{-6} was reached in 3 iterations, suggesting that the models were at a local minimum. Increasing the convergence threshold or requiring the training algorithm to complete 500 iterations did not change the model parameters. To avoid the local minimum problem,

weighted probability estimates of the transition probability were used to train only the fully-connected HMM.

The weighted initial estimate of the transition probabilities favored self-transitions i.e., for the 3-state HMM the state self-transitions had a probability of 0.5 and interstate transitions had a probability of 0.25. The initial estimate of the token probability matrix was kept as an uniform distribution, i.e. for all tokens the probability of observing a given token from a particular state was 0.2. When weighted initial estimates of the transition probability were used, the parameters of the HMMs did change from the initial estimates and the likelihood probability was larger than when the HMMs were trained with uniform initial estimates.

3.3.4 Metrics

Several metrics were calculated to determine the parsimonious HMM topology. These metrics were logarithm of likelihood probability [40-42], entropy measure of token distribution [21] and the Kullback-Leibler (K-L) distance measures [80]. The appropriate HMM topology would reduce the number of non-zero transitions without causing a large decrease in model likelihood, increase in entropy or increase in K-L distance measure.

A graph of the metrics versus the pruned HMM was constructed to evaluate the HMMs resulting from each step of the pruning procedure. According to Vasko, the parsimonious HMM is identified as the HMM that occurred before a large decrease or increase in the metrics [21]. For example, the logarithm of likelihood probability was one of the metrics used to evaluate the topology of the HMMs. The likelihood probability is the probability that a sequences was generated by a HMM given the model parameters. The natural logarithm of the likelihood

probability was used to prevent underflow of the computer floating point representation since this probability is typically very small [40]. Figure 8 shows an example of logarithm of likelihood probability calculated at each step in the pruning algorithm. When a state transition was removed from the HMM, the likelihood probability of the model was reduced. If the removal of the state transition did not alter the ability of the HMM to model the data, the likelihood probability plot remained approximately constant as seen in Figure 8 for the reduction in topology from HMM full to HMM #4. When the pruning causes a large decrease in likelihood probability, Vasko indicated that the pruned HMM did not adequately model the data and indicated that the pruned HMM occurring before the substantial decrease was the simplest model topology. According to the criterion, the parsimonious HMM in Figure 8 is HMM #4. This same criterion was used on the metric graphs to determine the parsimonious HMM for these data.

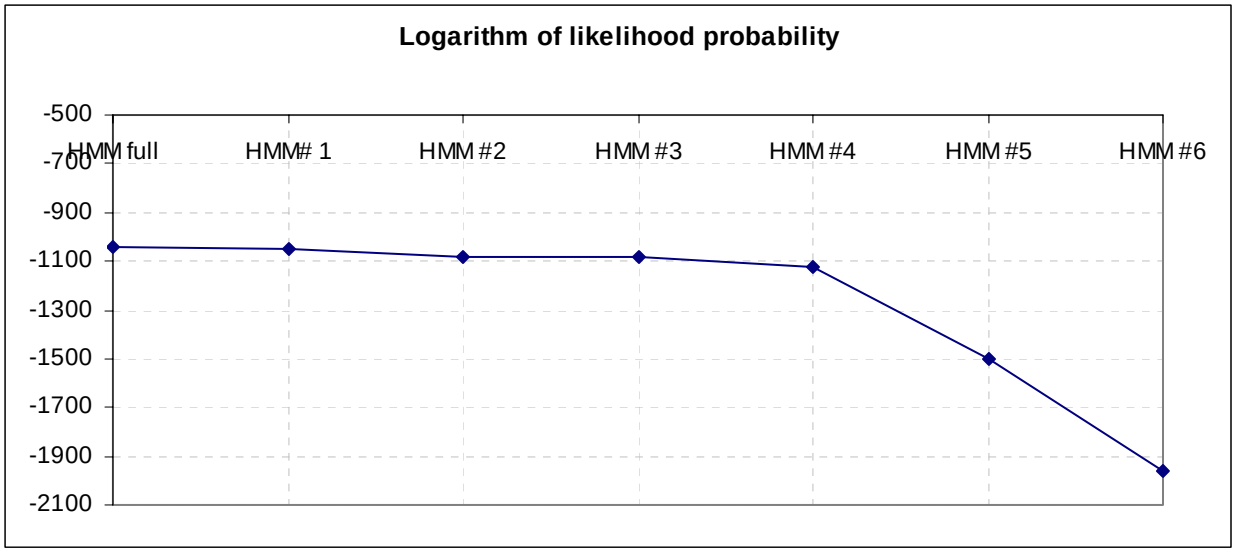


Figure 8: Example of the change in logarithm of likelihood probability as the topology of the HMM is reduced. The y-axis is the values of the logarithm of likelihood probability and the x-axis is the reduced topology HMM at each step in the pruning procedure. Likelihood probability is approximately constant from HMM full to HMM #4 and then decreases substantially at HMM #5, indicating that HMM #4 is the appropriate HMM.

The likelihood probability was calculated using the forward algorithm [40]. The forward algorithm is

1. Initialization:

$$\alpha_t(i) = \pi_i b_i(o_1), \quad 1 \le \mathrm{i} \le \mathrm{N}$$

2. Recursion

$$\alpha_{t+1}(j) = \left[\sum_{i=1}^{N} \alpha_t(i) a_{ij} \right] b_j(o_{t+1}), \quad 1 \le \mathrm{t} \le \mathrm{T\text{-}1},\ 1 \le \mathrm{j} \le \mathrm{N}$$

3. Termination

$$\Pr(O \mid \lambda) = \sum_{i=1}^{N} \alpha_T(i)$$

where N = total number of states and T = length of the sequence.

Since each of the lifting sequences was an independent sequence observation, the likelihood probability of the group HMM was calculated as

$$\Pr(O \mid \lambda) = \prod_{m=1}^{M} \Pr(O^m \mid \lambda)$$

where M is the total number of sequences.

The entropy of the token probability matrix was used as a measure of token distribution overlap between the states as model topology is reduced [21]. Entropy increases when the token probability becomes broader, which occurs as the topology is reduced. The simplest HMM topology was the HMM that occurred before a substantial increase in entropy. Entropy is calculated as

$$E = \frac{1}{\tau} \sum_{j=1}^{N} H(j) N(j)$$

$$H(j) = \sum_{k=1}^{K} b_j(k) \log_2 \left(\frac{1}{b_j(k)} \right)$$

where N(j) is the number of times state j is occupied and τ is the data length.

The Kullback-Leibler (K-L) distance is a measure of the average discriminating information contained per observation token between two HMMs [80]. The measure compares the distance between a true probability distribution to an arbitrary probability distribution which is usually an approximation of the true probability distribution. The statistical properties of two HMMs are compared using the K-L measure by calculating the difference between the likelihood probabilities that either of the two models fit a sequence. The K-L measure [80] is calculated as

$$D(\lambda_1 \| \lambda_2) \approx \frac{1}{T} \left[\log \Pr(O_T \mid \lambda_1) - \log \Pr(O_T \mid \lambda_2) \right]$$

In this calculation, random sequences are generated from model 1 and tested against both models. The likelihood probability is calculated using the forward algorithm.

In order to obtain good estimates of distance, sequences must have large number of observations and a large number of sequences have to be tested. To avoid the large computational cost associated with testing large observation sequences, an estimate of the Kullback-Leibler (K-L) measure was used. The estimated measure was based on the experiments of Liang et al.[81], who presented a closed form approximation of K-L distance measure. The K-L distance measure estimate was calculated as

$$D(\lambda_1 \| \lambda_2) = \sum_{i=1}^{M} \sum_{j=1}^{M} r_{1i} c_{1ij} (\log c_{1ij} - \log c_{2ij})$$

where r is the stationary observation distribution vector for a stationary HMM and is defined as $r' = v'B$. The variable v is called the stationary distribution vector and is defined as the normalized eigenvector that corresponds to eigenvalue of the state-transition probability matrix

that has a value of one. The variable c_{ij} is the observation-transition probability distribution matrix and is defined as $c_{ij} = \sum_{k=1}^{N} \sum_{p=1}^{N} \frac{v_k b_k(i) a_{kp} b_p(j)}{r_i}$

The K-L measure was calculated to compare the probability distance between the observable 5-state Markov model and the HMMs from the pruning procedure ($D(\lambda_1||\lambda_2)$). The fully-connected observable Markov model (OMM) was used in the K-L measure calculation as a reference model in order to compare statistical properties of the pruned HMMs. The K-L measures increased as the model complexity was reduced. The K-L measures were graphed and the simplest HMM topology was chosen as the HMM that occurred before a substantial increase in the K-L measures.

The OMM was constructed with five states corresponding to the five tokens. Each state emitted a single token with probability of 1 and zero probability for all other tokens. The state transition probability was determined by summing the transition from state S_i to state S_j divided by the total number of transitions out of state S_i. The parameters of the OMM for controls are shown in Table 3 and the parameters of the OMM for the CLBP are shown in Table 4.

Table 3: Values of the transition and token probabilities for the control observable Markov model

Parameters of the Control Observable Markov Model					
Control Transition Probability					
	Transition to state 1	Transition to state 2	Transition To state 3	Transition to state 4	Transition to state 5
In state 1	0.5628	0.1005	0.0804	0.0503	0.2060
In state 2	0.0209	0.7496	0.0820	0.0119	0.1356
In state 3	0.0116	0.0494	0.7880	0.0552	0.0958
In state 4	0.0244	0.0183	0.1280	0.7358	0.0935
In state 5	0.0333	0.0959	0.1019	0.0525	0.7164
Control Token Probability					
	Token 1	Token 2	Token 3	Token 4	Token 5
State 1	1	0	0	0	0
State 2	0	1	0	0	0
State 3	0	0	1	0	0
State 4	0	0	0	1	0
State 5	0	0	0	0	1

Table 4: Values of the transition and token probabilities for the CLBP observable Markov model

Parameters of the CLBP Observable Markov Model					
CLBP Transition Probability					
	Transition to state 1	Transition to state 2	Transition To state 3	Transition to state 4	Transition to state 5
In state 1	0.8560	0.0131	0.0131	0.0812	0.0366
In state 2	0.0222	0.8778	0.0389	0	0.0611
In state 3	0.0411	0.0959	0.5616	0.1096	0.1918
In state 4	0.0687	0	0.0305	0.8550	0.0458
In state 5	0.0535	0.0535	0.0453	0.0864	0.7613
CLBP Token Probability					
	Token 1	Token 2	Token 3	Token 4	Token 5
State 1	1	0	0	0	0
State 2	0	1	0	0	0
State 3	0	0	1	0	0
State 4	0	0	0	1	0
State 5	0	0	0	0	1

3.3.5 Results of pruning procedure

The pruning procedure reduced the topology of the 4-state, 3-state and 2-state fully connected HMMs. For both subject groups, there were 12 steps of the procedure for the 4-state fully-connected HMM pruning, 6 steps of the procedure for the 3-state fully-connected HMM pruning and 2 steps of the procedure for the 2-state fully-connected HMM pruning. The first step pruned one state transition from the fully-connected HMM, the second step pruned two transitions (the transition pruned with the first step plus an addition transition) and so forth until the final step resulted in a one state HMM. At each step, several trials were performed to determine which transition to remove from the HMM.

The models converged when the difference between iterations of the transition, token and likelihood probabilities were all less than 10^{-6}. The number of iterations needed to reach convergence ranged from 153 to 3 iterations. Training of the fully-connected HMM required more iterations to reach convergence than any other of the HMMs. The likelihood probability was calculated for all iterations of the Baum-Welch algorithm and graphed for each pruned HMM. The likelihood probability and the parameters of the HMM did change from the initial conditions suggesting that the convergence criterion was sufficient to train the HMM with subject data. As example, the logarithm of likelihood versus iteration of the training of the fully-connected 3-state CLBP HMM is shown in Figure 9.

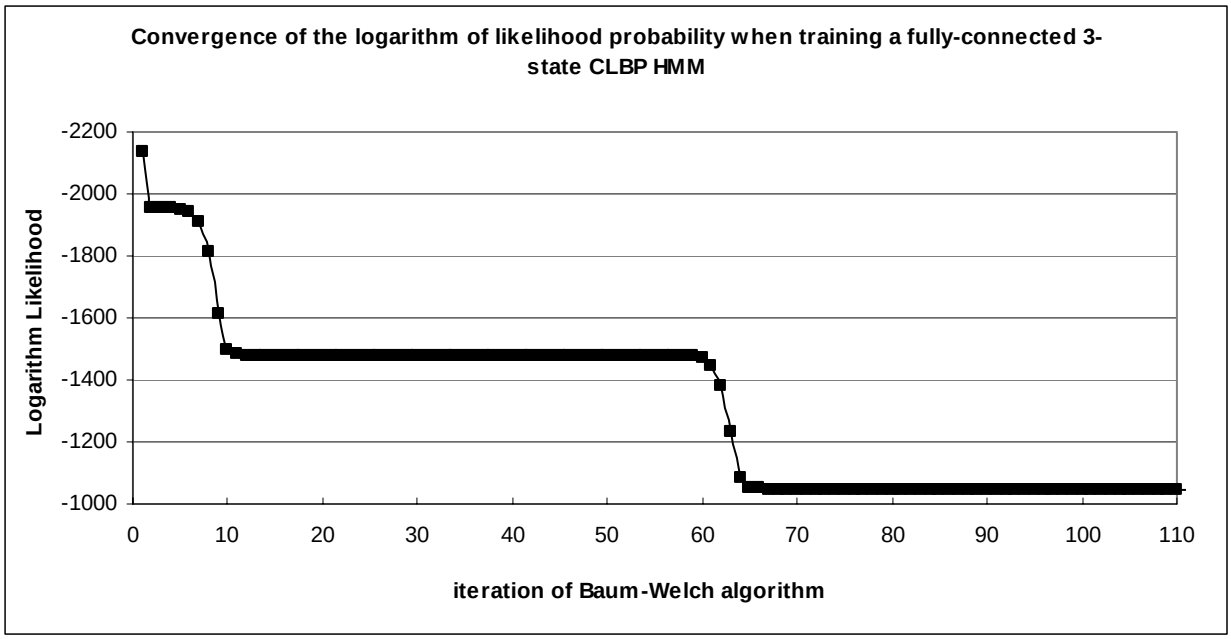


Figure 9: Logarithm of likelihood plotted versus iterations of the Baum-Welch algorithm to reach convergence for a fully-connected 3-state CLBP HMM

3.3.5.1 4-state HMMs: Control HMMs

The metrics were calculated for pruned HMM and plotted to compare the different HMMs. For each group, three graphs were constructed to describe the metrics. Figure 10 shows the graph of the three metrics for the 4-state control HMMs with the plot of likelihood probability in the top graph, the plot of entropy in the middle graph, and the K-L measure plotted versus pruned HMM in the bottom graph. Figure 11 is a schematic of the 4-state control HMM that was chosen as the appropriate HMM. The transition and token probabilities the HMMs chosen as the most appropriate HMM for pruned 4-state control HMM are shown in Table 5. The graph of the metrics, schematics of the HMM and tables of the HMM parameters for the 4-state CLBP HMM (Figures 12-13, Table 6), 3-state HMMs of CLBP and control data (Figures 14-17, Tables 7-8),

and the 2-state HMM of the CLBP data and control data (Figures 18-20, Table 9) are presented in the same format.

For organizational purposes, the pruned HMMs were named for the number of the zero transitions, i.e. HMM #1 corresponds to a HMM trained when one transition was zero; HMM #2 corresponds to a trained HMM when two interstate transition were zeros and so forth. The exception to the naming convention was the fully connected HMMs, which were named HMM full.

Two pattern of change in the metrics were observed. In the first pattern, all three metrics clearly indicated the HMM with the appropriate topology was the HMM that occurred before a substantial increase or decrease in a metric. In the second pattern, likelihood probability and entropy measures gradually changed as the HMMs were reduced and the K-L measure plot showed substantial change. In this case, the entropy and likelihood metrics suggested two possible models, while the K-L measure clearly indicated one model as the most appropriate reduced HMM.

The likelihood and entropy plots for the 4-state control HMM were approximately constant from HMM #2 to HMM #8, slightly decreases from HMM #8 to HMM #9 and then begins to decrease (Figure 10). The K-L measure was appropriately constant from HMM full to HMM #8, and then increased. The simplest topology for a 4-state control HMM was chosen as HMM #8 based on the K-L measure. A diagram of the HMM #8 is shown in Figure 11 and the parameters are listed in Table 5.

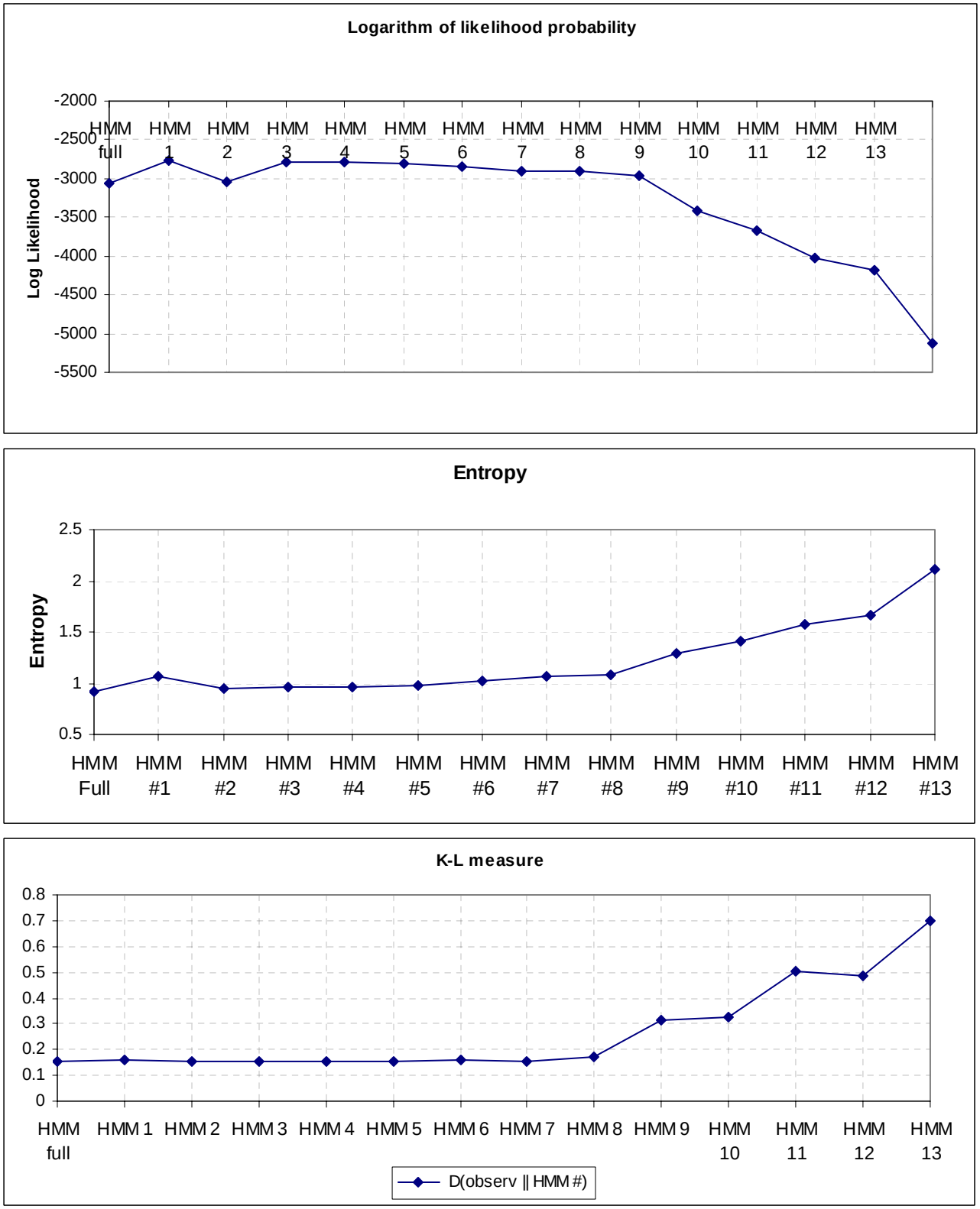


Figure 10: Logarithm of likelihood probability (top), entropy (middle) and K-L measures (bottom) for pruning of the fully-connected 4-state control HMM.

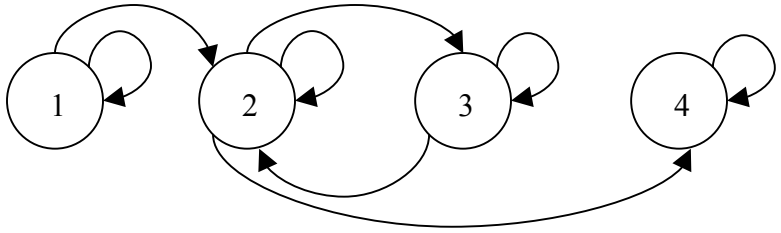


Figure 11: Diagram of the control 4-state HMM

Table 5: Values of the transition and token probability of HMM #8

Parameters of the trained control HMM #8					
Control Transition Probability					
	Transition to state 1	Transition to state 2	Transition to state 3	Transition to state 4	
In state 1	0.9690	0.0310	0	0	
In state 2	0	0.9332	0.0281	0.0387	
In state 3	0	0.0098	0.9902	0	
In state 4	0	0	0	1	
Control Token Probability					
	Token 1	Token 2	Token 3	Token 4	Token 5
State 1	0.1317	0.0967	0.0354	0.0431	0.6931
State 2	0.0479	0.0269	0.1165	0.7409	0.0679
State 3	0.0053	0.8947	0.0397	0.0017	0.0587
State 4	0.0136	0.0347	0.8305	0.0267	0.0944

3.3.5.2 4-state HMMs: CLBP HMMs

The likelihood probability and the entropy metrics when pruning the 4-state CLBP HMM were approximately constant for HMM full to HMM #8, slightly changed from HMM #8 to HMM #9 and then substantially changed (Figure 12). The K-L distance plot is appropriately constant from HMM full to HMM #9 and then substantially increases. From this plot, HMM #9 was chosen as the most appropriate HMM based on the K-L measure. The diagram of the HMM #9 is shown in Figure 13 and the parameters are shown in Table 6.

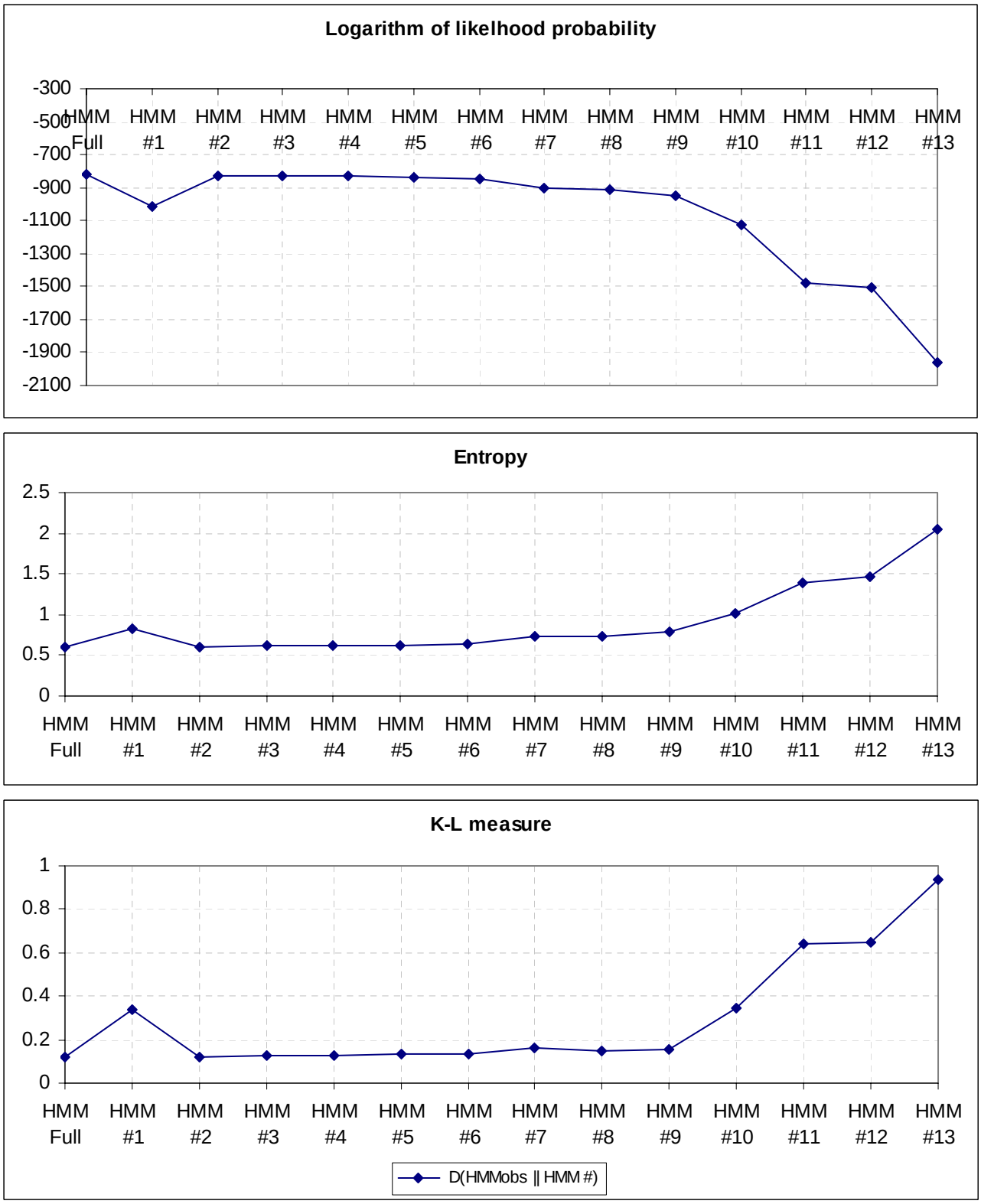


Figure 12: Logarithm of likelihood probability (top), entropy (middle) and K-L measures (bottom) for pruning of the fully-connected 4-state CLBP HMM.

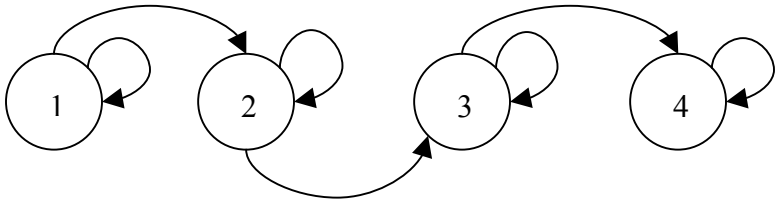


Figure 13: A diagram of the 4-state CLBP HMM

Table 6: Values of the transition and token probabilities of HMM #9

Parameters of the trained CLBP HMM #9					
CLBP Transition Probability					
	Transition to state 1	Transition to state 2	Transition to state 3	Transition To state 4	
In state 1	0.9294	0.0706	0	0	
In state 2	0	0.9809	0.0191	0	
In state 3	0	0	0.9708	0.0292	
In state 4	0	0	0	1	
CLBP Token Probability					
	Token 1	Token 2	Token 3	Token 4	Token 5
State 1	0.8833	0.0124	0.0197	0.0302	0.0544
State 2	0.1035	0.0000	0.0433	0.8167	0.0365
State 3	0.0040	0.0471	0.1570	0.0402	0.7516
State 4	0.0000	0.9281	0.0289	0.0000	0.0430

3.3.5.3 3-state HMMs: Control HMMs

The likelihood and the entropy measures showed gradual changes when the HMM was reduced but the K-L distance plot indicated that the most appropriate HMM is HMM #4 (Figure 14). The K-L distance measure shows very slight increase in the K-L measure from HMM #1 to HMM #4 and then a substantial increase at HMM #4 to HMM #6. Based on these results, HMM #4 was chosen. A diagram of HMM # 4 is shown in Figure 15 and the parameters of this HMM are shown in Table 7.

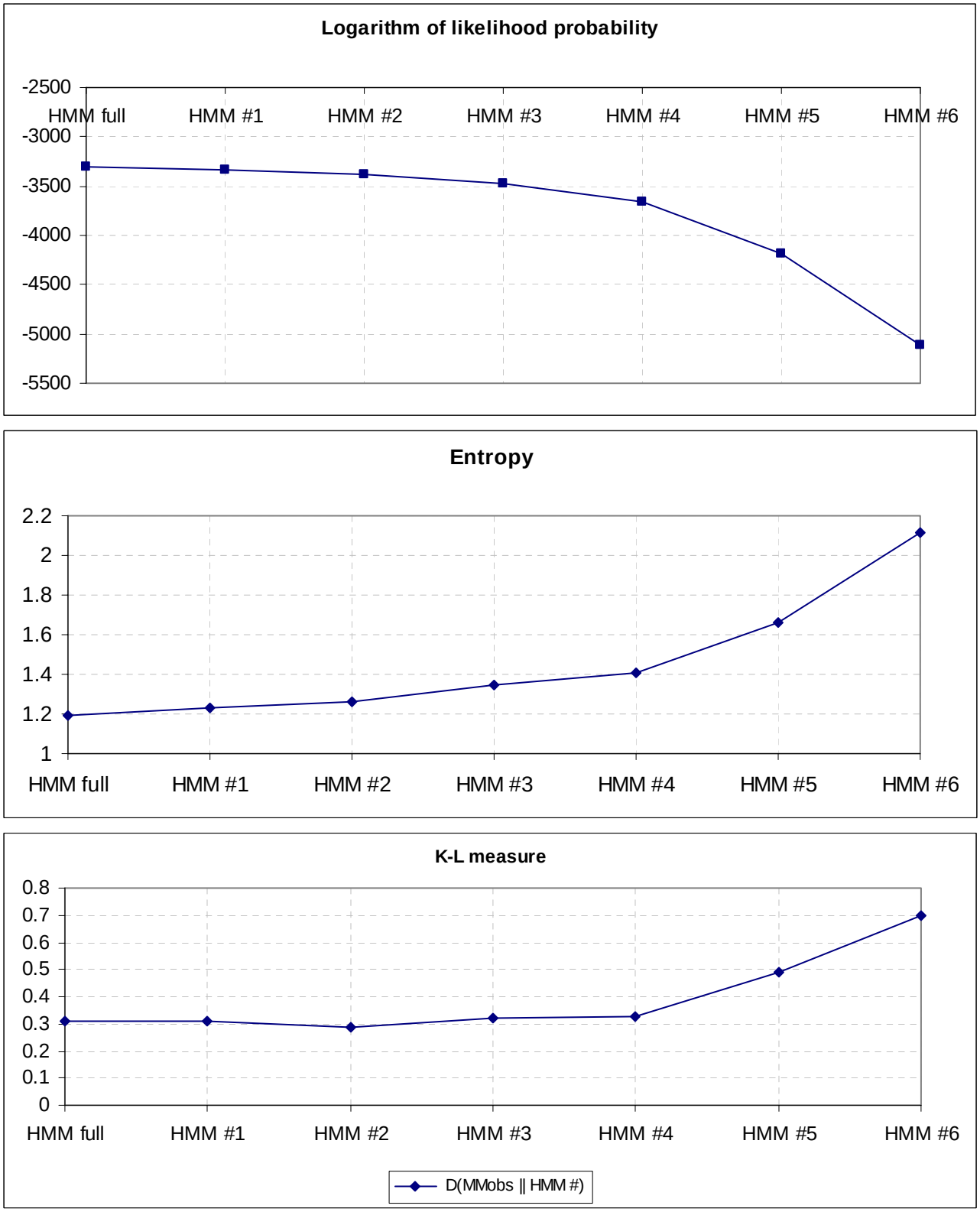


Figure 14: Logarithm of likelihood probability (top), entropy (middle) and K-L measures (bottom) for pruning of the fully-connected 3-state control HMM.

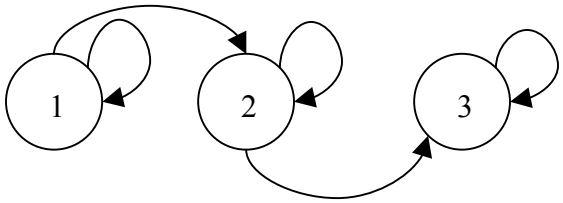


Figure 15: Diagram of the 3-state control HMM

Table 7: Values of the transition and token probability of HMM #4

Parameters of the trained Control HMM #4					
Control Transition Probability					
	Transition to state 1		Transition to state 2		Transition to state 3
In state 1	0.9759		0.0241		0
In state 2	0		0.9507		0.0493
In state 3	0		0		1
Control Token Probability					
	Token 1	Token 2	Token 3	Token 4	Token 5
State 1	0.1254	0.1567	0.0618	0.0370	0.6192
State 2	0.0393	0.0205	0.1400	0.7361	0.0642
State 3	0.0014	0.3118	0.6084	0.0161	0.0624

3.3.5.4 3-state HMMs: CLBP HMMs

The simplest topology can be easily identified in the plots of all three metrics (Figure 16). The metrics are approximately constant from HMM full to HMM #4 and then substantially decreases or increase at the HMM #5. The HMM with the most appropriate topology is HMM #4. The diagram of HMM #4 is shown in Figure 17 and the parameters are listed in Table 8.

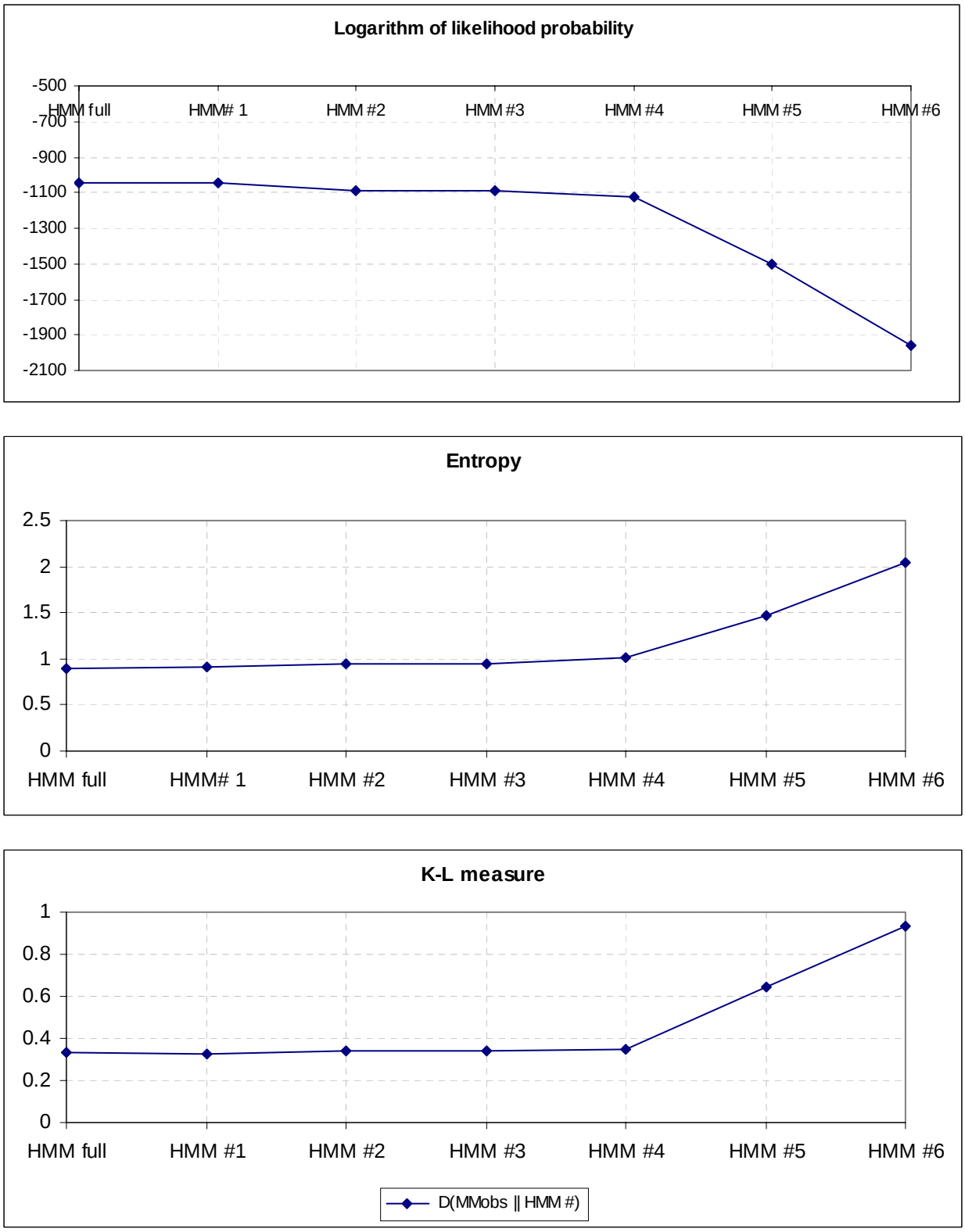


Figure 16: Logarithm of likelihood probability (top), entropy (middle) and K-L measures (bottom) for pruning of the fully-connected 3-state CLBP HMM.

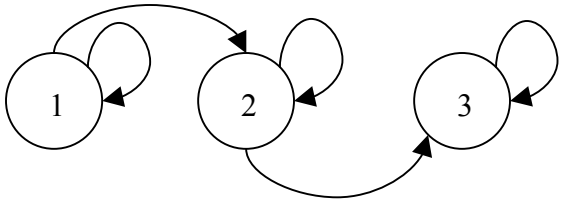


Figure 17: Diagram of 3-state CLBP HMM

Table 8: Values of the transition and token probabilities of HMM #4

Parameters of the trained CLBP HMM #4					
CLBP Transition Probability					
	Transition to state 1	Transition to state 2	Transition to state 3		
In state 1	0.9294	0.0706	0		
In state 2	0	0.9749	0.0251		
In state 3	0	0	1		
CLBP Token Probability					
	Token 1	Token 2	Token 3	Token 4	Token 5
State 1	0.8854	0.0122	0.0196	0.0302	0.0526
State 2	0.0948	0	0.0508	0.7902	0.0641
State 3	0.0023	0.4232	0.1016	0.0053	0.4676

3.3.5.5 2-state HMMs: Control and CLBP HMMs

The results of the pruning procedure applied to the 2-state fully-connected HMM of control subjects are shown in Figure 18 and of CLBP subjects are shown in Figure 19. For both groups, all three metrics indicate that HMM #1 is the most appropriate topology of 2-state control HMM and 2-state CLBP HMM. A diagram of the topology of both HMMs is shown in Figure 20. The parameters of control and CLBP HMM are shown in Table 9.

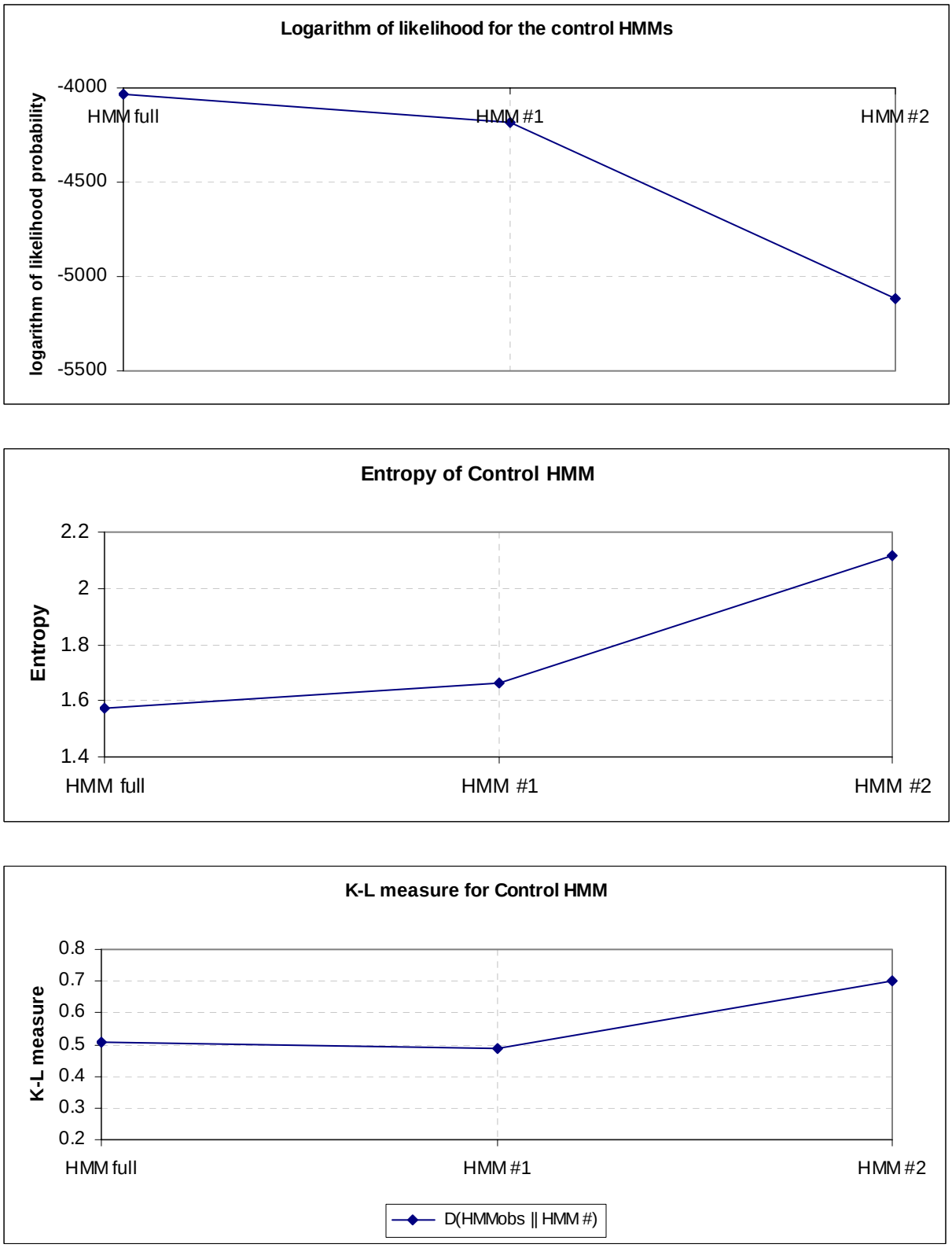


Figure 18: Logarithm of likelihood probability (top), entropy (middle) and K-L measures (bottom) for pruning of the fully-connected 2-state control HMM.

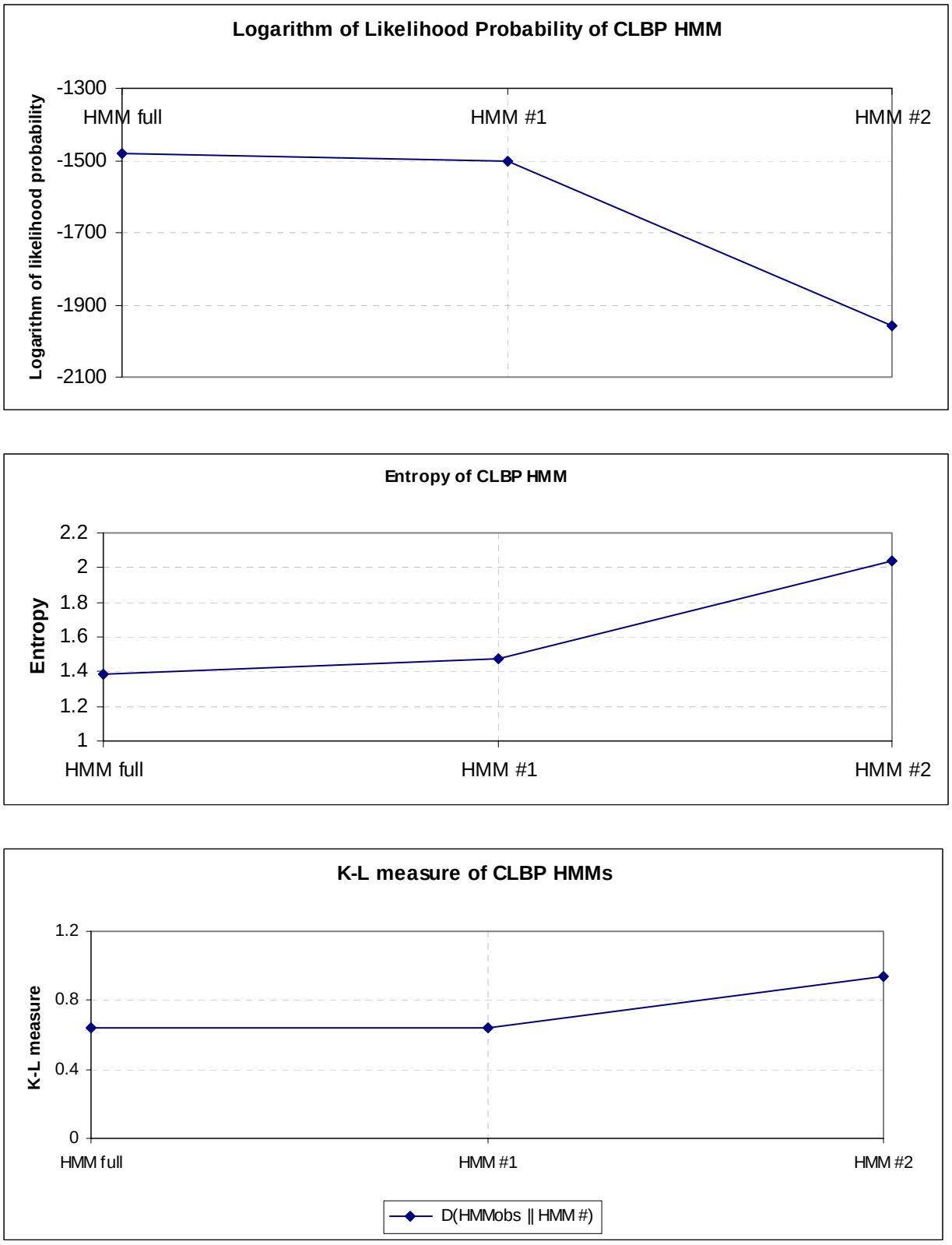


Figure 19: Logarithm of likelihood probability (top), entropy (middle) and K-L measures (bottom) for pruning of the fully-connected 2-state CLBP HMM.

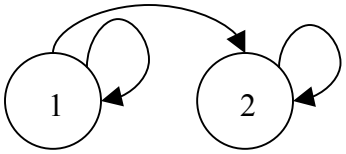


Figure 20: Diagram of the 2-state control HMM and the 2-state CLBP HMM

Table 9 : Values of the transition and token probabilities of control HMM #1 and CLBP HMM #1

Parameters of the trained Control HMM #1					
Control Transition Probability					
	Transition to state 1		Transition to state 2		
In state 1	0.9768		0.0232		
In state 2	0		1		
Control Token Probability					
	Token 1	Token 2	Token 3	Token 4	Token 5
State 1	0.1340	0.1066	0.0803	0.0563	0.6227
State 2	0.0026	0.2684	0.4729	0.2110	0.0450
Parameters of the trained CLBP HMM #1					
CLBP Transition Probability					
	Transition to state 1		Transition to state 2		
In state 1	0.9409		0.0591		
In state 2	0		1		
CLBP Token Probability					
	Token 1	Token 2	Token 3	Token 4	Token 5
State 1	0.8335	0.0067	0.0141	0.1082	0.0376
State 2	0.0167	0.2109	0.0813	0.4186	0.2724

3.3.6 Order estimation method

When the pruning algorithm was applied to the fully-connected HMMs, the metrics identified HMMs that contained the same number of states as the fully-connected model. In Vasko's thesis, the pruning algorithm reduced the number of states and transitions and identified a HMM that

contained fewer states than the fully-connected HMM. Since the pruning algorithm did not reduce the order of the HMMs, the Viterbi algorithm was used to determine the appropriate number of states in the HMM to describe each group data by assessing the frequency that the states were used by the data. Although this is a heuristic technique, it is similar to the technique of model surgery [43-44] that has been used in bioinformatics.

The model surgery technique applies either the forward-backward or the Viterbi algorithm to the trained profile HMMs to determine how frequently a transition is used by the training data. A profile HMM is a model that consists of many modules. Each module contains a match state, delete state and an insert state with only forward transitions to each state. Surgery eliminates a module if a match state is used by less than a certain fraction of the sequences. If an insert state is used by more than a certain fraction of the sequences, than the module is expanded [44]. The technique was found to work well in identifying the appropriate parameters for a HMM that describes protein domains and was used for these data. For these data, the HMM chosen to describe the group data was the model in which a majority of the sequences utilized all the states in the model. The details and results of the algorithms when applied to these data are described below.

The Viterbi algorithm was used to determine the most likely state path of each subject through the 4-state, 3-state and 2-state HMM resulting from the pruning procedure for each group. The frequency that the states were occupied by the subject's sequences was calculated from the most likely state path of all the subject sequences within the group. For these data, the HMM chosen to describe the group data was the model in which a majority of the sequences utilized all the states in the model. Since the reduced HMMs are temporal and the sequences all started in state 1, the sequences could only either stay in the initial state or transition in one

direction to the next state. If the last state was occupied by only a few or none of the sequences, then the state was considered to be not essential.

Table 10 shows the results of the frequency calculation for the control HMMs. The frequency calculation indicates that the 3-state HMM is the most appropriate HMM for the control data because the states in the 3-state HMM were approximately equally occupied by the control subjects. The 4-state HMM appeared to overfit the data because only 4% of the control subjects' sequences used a four state path through the model suggesting that the fourth state is not necessary to describe the control data. Since the states in the 3-state HMM were occupied by a large percentage of the subject sequences, reducing the HMM to a 2-state HMM would ignore a considerable amount of information by forcing the subjects into either a single state path or a two state path. From these results, the 3-state HMM was determined to be the appropriate HMM to describe the control time series data.

Table 10: Percentage of control subject sequences that occupied each of the states in the 4-state, 3-state and 2-state control HMMs

Frequency that states were occupied				
HMMs	state 1	state 2	state 3	state 4
Control 4-state HMM	31%	29%	37%	2%
Control 3-state HMM	39%	27%	33%	
Control 2-state HMM	49%	51%		

A 4-state HMM appeared to overfit the CLBP data since very few subjects occupied all four states when transitioning through the model as shown in Table 11. For the 3-state HMM, the frequency calculation showed that less than 15% of the CLBP subjects occupied the third state suggesting that a third state may not be necessary to describe the CLBP data. The results of the

frequency calculation for the 2-state HMM showed that about 1/3 of the CLBP subjects occupied the last state of the model. To determine whether the 3-state HMM or the 2-state HMM was the most appropriate HMM, reliability of these models was assessed with the second simulation study as described in Appendix A. The simulation compared the 3-state control HMM to the 3-state CLBP HMM and to the 2-state CLBP HMM. The results showed that the 2-state CLBP HMM was more reliable than the 3-state CLBP HMM and the 2-state HMM was chosen as the most appropriate HMM to describe the CLBP time series data.

Table 11: Percentage of CLBP subject sequences that occupied each of the states in the 4-state, 3-state and 2-state CLBP HMMs

Frequency that states were occupied				
HMMs	state 1	state 2	state 3	state 4
CLBP 4-state HMM	57%	25%	15%	4%
CLBP 3-state HMM	57%	31%	12%	
CLBP 2-state HMM	73%	27%		

3.3.7 Final HMMs

A 3-state temporal HMM was chosen to describe the control HMM and a 2-state temporal HMM was chosen to describe the CLBP HMM. The parameters of the control HMM and the CLBP HMM were shown in Table 7 and Table 9, respectively.

4.0 RELIABILITY OF HMM CLASSIFICATION PROCEDURE

The data reduction procedure described in Chapter 3 assigned each of the lifts that the subjects performed during the repetitive lifting task to a cluster that described a lifting pattern. From the results of the data reduction procedure, a temporal 2-state CLBP HMM was designed to describe the lifting patterns of CLBP subjects during the repetitive lifting task and a temporal 3-state control HMM was designed to describe the lifting patterns of the control subjects during the repetitive lifting task. The possibility of using these HMMs to identify sub-groups of CLBP subjects was evaluated with simulation studies. The methods and results of the simulation studies that were performed to determine reliability of the HMM classification procedure are described in this chapter.

4.1 SIMULATION STUDIES

Two simulation studies were conducted to determine how reliably the HMMs can identify sequences that are generated from a particular model, and to determine how reliably the HMMs can detect lifting sequences that are classified to the wrong group and classify the sequences to the appropriate group. A kappa statistic was used to assess reliability in both simulation studies and was calculated as the observed probability of agreement subtracted from the probability of

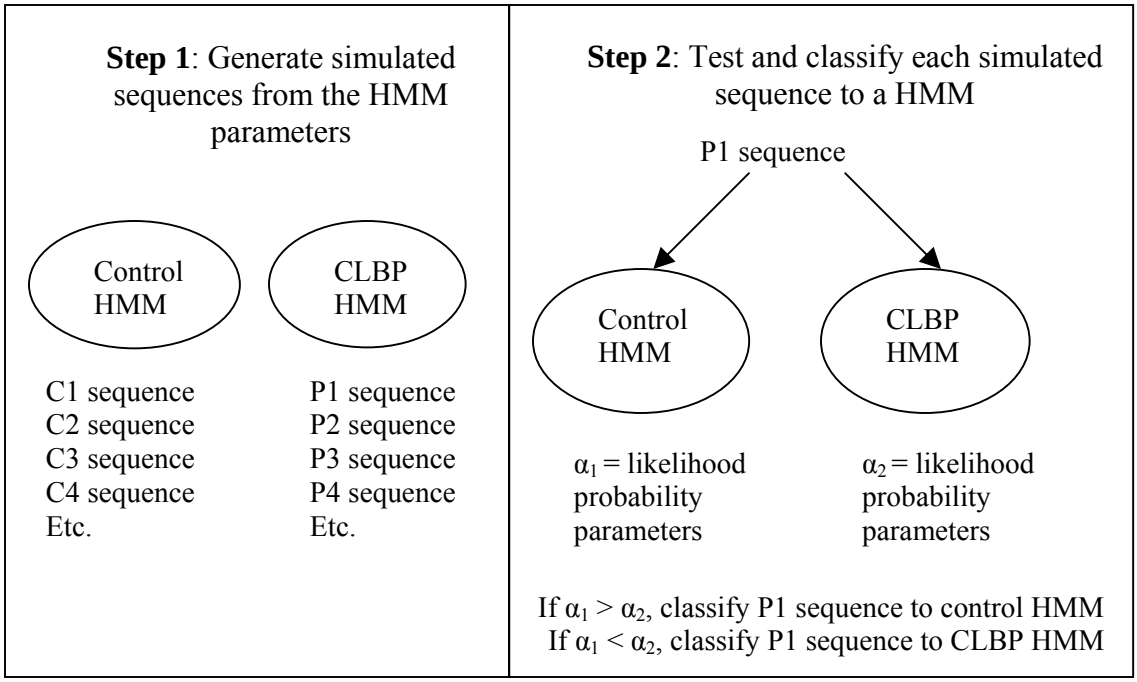
chance agreement divided by one minus the probability of chance agreement. The kappa statistic was chosen because it assesses rater agreement after adjusting for chance agreement [82]. Values of kappa greater than 0.8 indicate excellent reliability, between 0.8 and 0.6 indicate substantial reliability, between 0.6 and 0.4 indicate moderate reliability, between 0.4 and 0.2 indicate fair reliability and kappa statistics lower than 0.2 indicates slight to no reliability [63]. In the simulation studies, the HMMs were considered reliable when kappa was greater than 0.8.

4.1.1 First Simulation Study

The first simulation study was constructed to determine whether the control HMM and CLBP HMM can identify lifting sequences that were generated from a particular model. The study used the trained transition probability and token probability matrices of the control HMM and CLBP HMM to generate control simulated lifting sequences and CLBP simulated lifting sequences, respectively. All simulated lifting sequences started in state 1. To evaluate the influence of the number of sequences and the length of the sequences on reliability, the number of simulated sequences in each group was varied in increments of 10 from 20 sequences (10 simulated from control HMM and 10 simulated from the CLBP HMM) to 120 sequences (60 from control HMM and 60 from the CLBP HMM) and the number of lifts in the sequences was varied from 6 to 80 lifts.

The maximum sample size (120) and range of lifts in each sequence were selected to approximately match the clinical data. Each simulated sequence was tested against both HMMs and classified to the model with the greater likelihood probability that the sequence was observed given the model parameters. Once all of the sequences were classified, a kappa statistic was

calculated to determine how reliably the HMMs classified lifting sequences that were generated from a particular model. A schematic of the steps in the first simulation study and the calculation of kappa are shown in Figure 21.



Step 3: Calculate kappa as $$\kappa = \frac{p_o - p_c}{1 - p_c} = \frac{\left[\left(\frac{N_{control} + N_{clbp}}{total}\right) - \sum_{i=1}^{2} a_i b_i\right]}{1 - \sum_{i=1}^{2} a_i b_i}$$

	Classified to control HMM	Classified to CLBP HMM	
Generated from control HMM	$N_{control}$	$M_{control}$	$N_{control} + M_{control}$ $a_1 = \frac{N_{control} + M_{control}}{total}$
Generated from CLBP HMM	M_{clbp}	N_{clbp}	$M_{clbp} + N_{clbp}$ $a_2 = \frac{M_{clbp} + N_{clbp}}{total}$
	$N_{control} + M_{clbp}$ $b_1 = \frac{N_{control} + M_{clbp}}{total}$	$M_{control} + N_{clbp}$ $b_2 = \frac{N_{clbp} + M_{control}}{total}$	Total

Figure 21: Schematic of the steps in the first simulation study. The first step is to generate simulated sequences. The second step is to test the sequences against both HMMs and assign the sequence to a model based on likelihood probability. In step 3, a kappa statistic is calculated after all sequences have been tested and assigned to a HMM.

The MatLab program hmmgenerate [78] was used to generate simulated sequences from the control HMM and CLBP HMM. The program begins with all sequences starting in state 1. A random number was generated from an uniform distribution, and this number was compared to the cumulative transition probability matrices of the current state. If the random number was greater than the cumulative transition probability of staying in the state, the sequences transitioned to another state. If the random number was less than the cumulative probability of all interstate transitions out of the current state, the sequence remained in the current state. The program resets the current state depending on the type of transition (probabilities of a different state were considered if interstate transition occurred and the probabilities of the same state were considered if a self-transition occurred) and evaluated another random number against the cumulative transition probability. For example, the start state is state 1 and the random number 0.95 is generated. The cumulative transition probability of state 1 is [0.8, 0.92, 1]. Based on this probability, the sequence would transition to state 3. The program sets the current state to state 3 and the simulated state transition sequence is 13. The next random number is 0.62. The cumulative transition probability of state 3 is [0, 0.5, 1]. The sequences would stay in state three and the updated simulated state transition sequence is 133. The process was repeated until the designated length was reached.

Once the simulated state transition sequences were determined, simulated observation sequences were generated using the cumulative token probability and another set of random numbers. The cumulative token probability of the state defined by the simulated state transition sequence was compared to the random number. The program emitted the token that was less than the random number but greater than the previous token. For example, if a random number greater

than the cumulative probability of token 1 and token 2 but less than the probability of token 3, token 3 was emitted.

4.1.2 Second Simulation Study

The second simulation study assessed whether the HMMs can reliably identify lifting sequences to the appropriate model when some of the lifting sequences are intentionally mislabeled. A small number of simulated lifting sequences were switched between the groups to create intentionally mislabeled sequences. For example, lifting sequences generated from the CLBP HMM were labeled as lifting sequences generated from control HMM and lifting sequences generated from the control HMM were labeled as lifting sequences generated from CLBP HMM. A modified version of the jackknife method [83] was used to train the HMMs. The modified jackknife was used because it permitted classification of each subject's lifting sequence to a model without introducing bias associated with classifying sequences that were used to train the HMM. For each HMM, the modified jackknife method excluded one of the simulated lifting sequences and trained the HMM with the remaining sequences. The excluded lifting sequence was classified to one of the HMMs based on the likelihood probabilities. This process continued until all of the lifting sequences had been tested. This subject was then assigned to one of the models based on the largest likelihood probability.

The number of intentionally mislabeled sequences was varied equally between the groups (e.g. 1 simulated sequence generated from the control HMM and 1 simulated sequence from the CLBP HMM were both mislabeled) from 4 to 64 (4% to 50% of the total sample size) in increments of 5 (9%). At each increment in the number of intentionally mislabeled sequences,

the HMMs were retrained with the modified jackknife method and all of the sequences were classified to a HMM based on the likelihood probability. A kappa statistic assessed reliability. A schematic of the steps in the second simulation study is shown in Figure 22.

The sample size of the simulated sequences in the second simulation was chosen as 108 (54 CLBP simulated sequences and 54 control simulated sequences) to approximately match the data of the clinical study. The simulated lifting sequences were generated using the same methods as in the first simulation with a minor modification. The lengths of the sequences were varied to match the clinical data. Half of the CLBP sequences were randomly chosen to have a sequence length that varied from 6 lifts to 20 lifts and the remaining half were randomly chosen to have a length that varied from 21 lifts to 80 lifts. All of the control simulated sequences were randomly assigned to a length that varied from 30 lifts to 80 lifts.

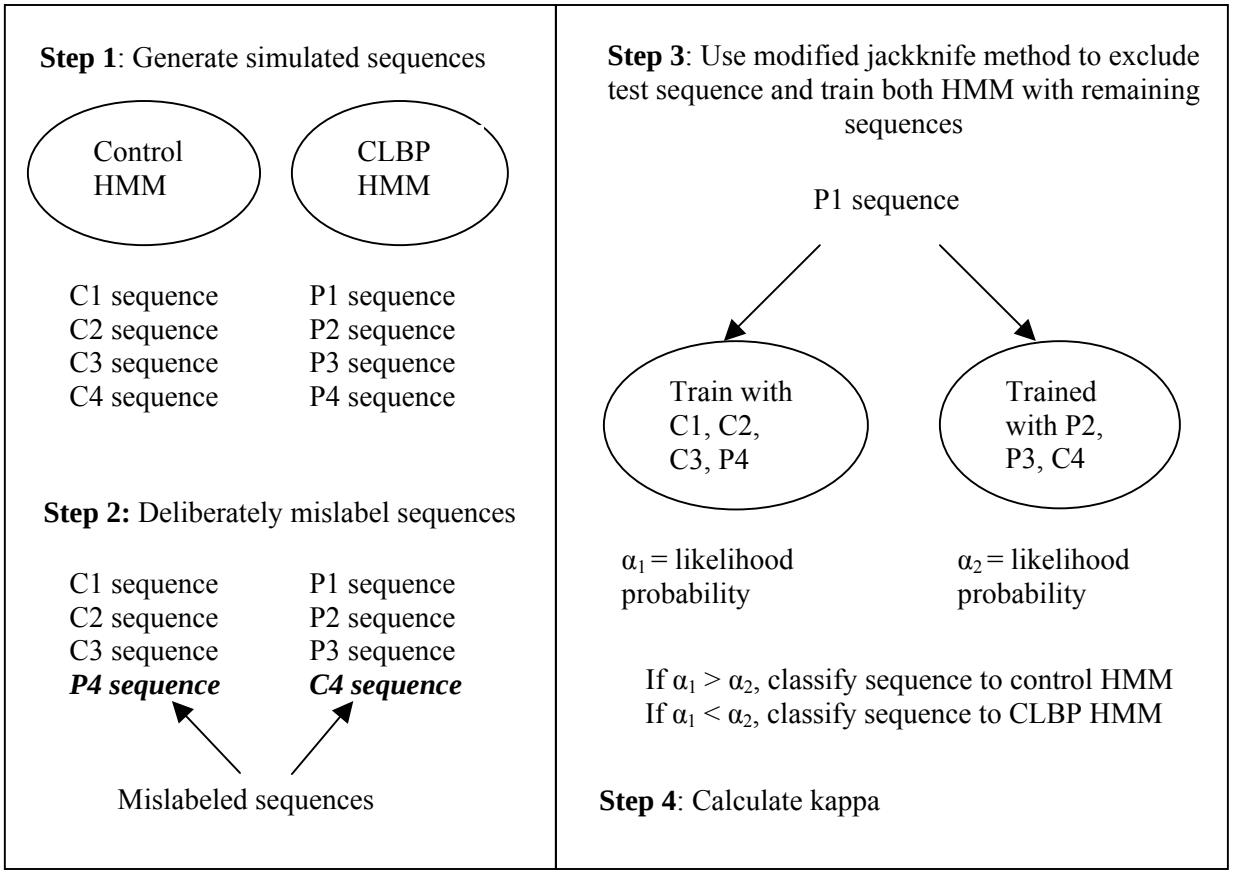


Figure 22: Schematic of the steps in the second simulation study. The first step is to generate simulated sequences from the HMMs. Simulated sequences are then intentionally mislabeled to the wrong group in the second step. In the step 3, the modified jackknife method is used to test the sequences against the HMMs. The sequence is assigned to a model based on likelihood probability. Once all sequences have been assigned to a HMM, a kappa statistic is calculated in step 4.

4.2 RESULTS OF SIMULATION STUDIES

A kappa value of greater than 0.8 was chosen as the criterion for determining whether the HMMs were reliable since this value of kappa indicates excellent reliability. The number of errors associated with a kappa value of 0.8 (when the sample size is 108) is 10 classification errors, which corresponds to an error rate of approximately 10%.

In the first simulation, the length of the lifting sequences and sample size were varied. Kappa was calculated for comparison of simulated sequences that had equal length and equal sample size. For all sample sizes and sequences with more than 7 lifts, kappa was greater than 0.8 as shown in Figure 23. Sequences with 45 or more lifts produced kappa values of 1 and sequences shorter than 45 lifts produced a kappa statistic than range from 0.8 to 1. These results suggest that the number of lifts in the sequences have a greater effect on reliability of the HMM to identify the sequence to the correct model than the sample size.

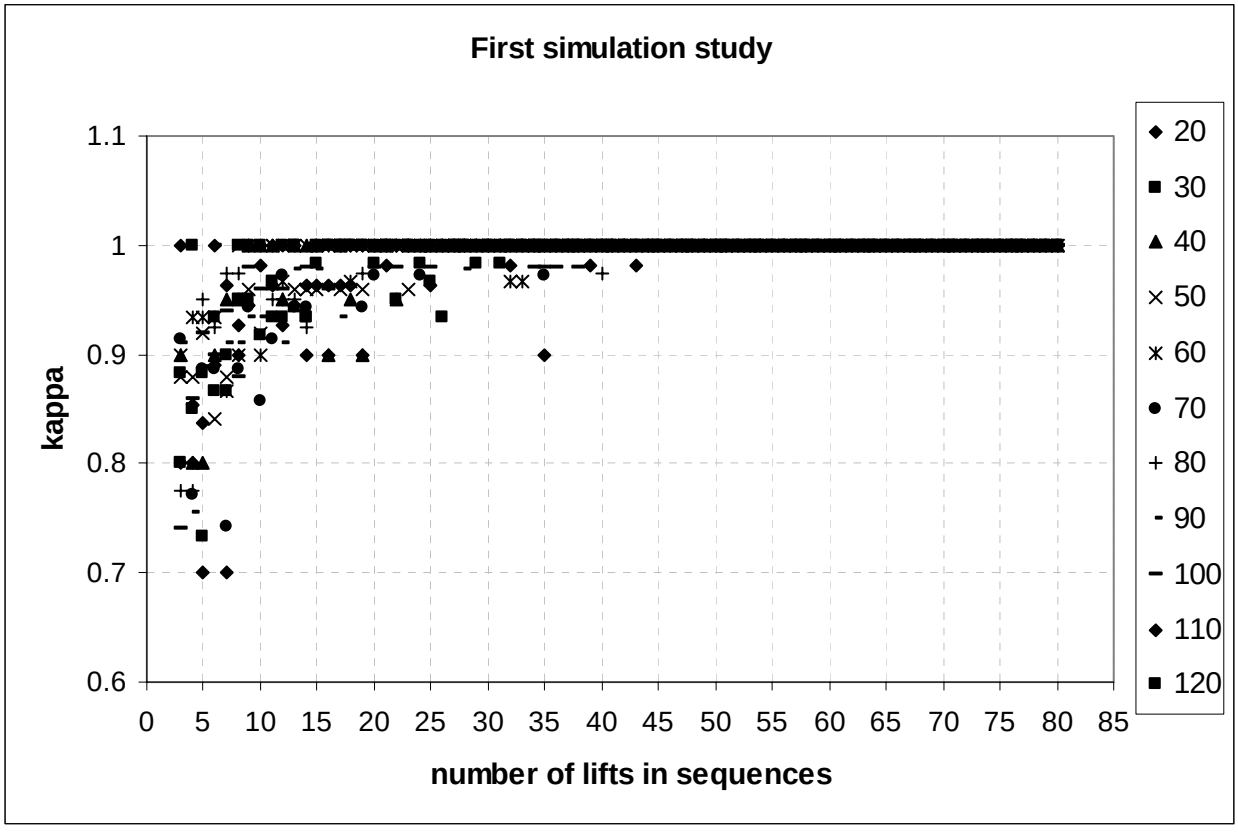


Figure 23: Kappa versus number of lifts in the simulated sequences calculated during the first simulation study. The symbols represent different sample sizes.

The second simulation determined how reliably the HMMs can identify lifting sequences that were generated from a particular model when a percentage of the simulated sequences were intentionally mislabeled. The HMMs were found to have excellent reliability when 4% to 41% of the data was intentionally mislabeled, as shown in Figure 24. This statistic corresponded to a total classification error of 1 to 10 simulated sequences (Table 11). As the percentage of mislabeled sequences increased, more classification errors were found in the mislabeled sequences than in the correctly in labeled sequences.

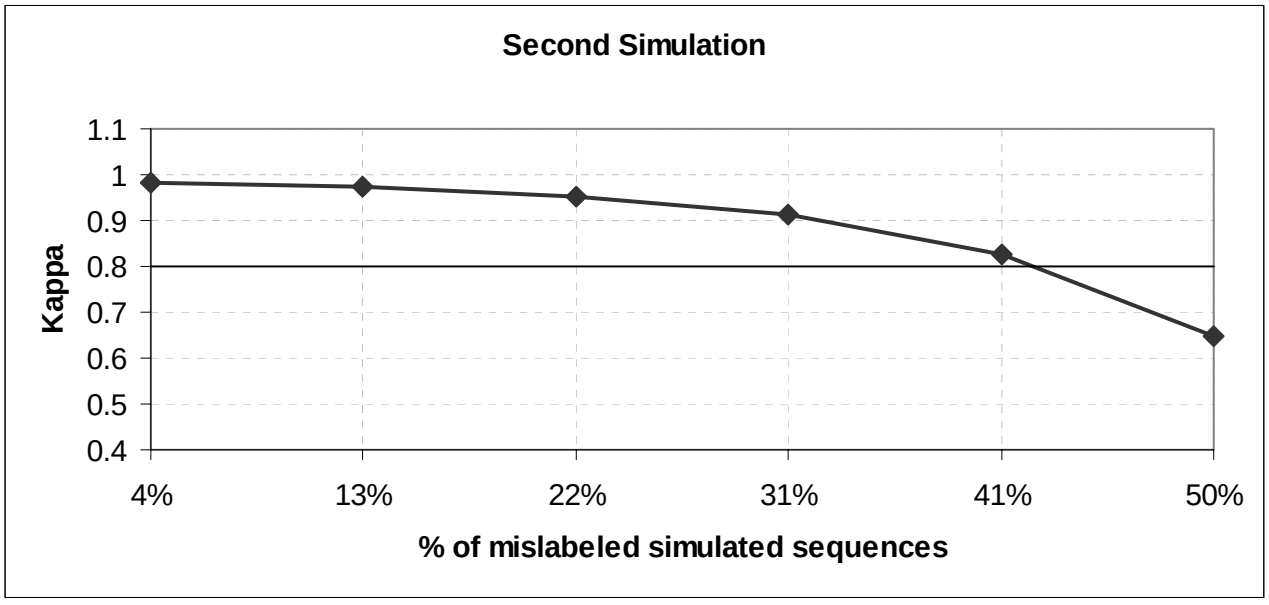


Figure 24: Kappa plotted versus percentage of intentionally mislabeled simulated sequences calculated during the second simulation study. Line at kappa = 0.8 indicates the cut-off value of reliability. HMMs can reliably identify sequences to the correct model when kappa is greater than 0.8.

Table 12: Kappa values, classifications errors in mislabeled and correctly labeled simulated sequences and error rate for each percentage of mislabeled simulated sequences during the second simulation study are listed.

Second Simulation					
Percentage of mislabeled sequences	Kappa values	# errors in correctly labeled sequences	# errors in intentionally mislabeled sequences	Total classification Errors	Error rate
4%	0.983	1	0	1	1%
13%	0.972	1	1	2	1%
22%	0.954	2	1	3	2%
31%	0.914	3	2	5	4%
41%	0.824	3	7	10	9%
50%	0.648	5	14	19	18%

The HMMs were able to reliably identify sequences to the correct model (i.e., the HMM that generated the simulated sequence) when the sequences contained more than 7 lifts and when 41% or less of the data was intentionally mislabeled to the wrong group. Based on these results, the HMMs were applied to the clinical data.

5.0 HMM CLASSIFICATIONS OF CLINCAL DATA

The CLBP subjects and control subjects were classified to either the control HMM or the CLBP HMM. Two CLBP groups (CLBP subjects classified to the CLBP HMM and the CLBP subjects classified to the control HMM) and two control groups (control subjects classified to the control HMM and control subjects classified to the CLBP HMM) were found. The HMM classification method was also applied to the lifting sequences of the CLBP subjects after the subjects completed treatment. This chapter describes the results of the HMM classification when applied to the clinical data, the statistical analyses performed to test the hypotheses and the results of the statistical analyses.

5.1 ADDITIONAL DATA

A subset of the data (54 CLBP subjects and 51 control subjects) was used to design the HMMs since not all of the data was available when the original models were designed. The total sample size within the clinical database is 81 CLBP subjects and 53 control subjects. The database also contained the post-treatment lifting data of 51 CLBP subjects. A procedure was developed to incorporate the additional data based on the statistics of the original sample. Each of the lifting parameters of the subject's lifts was checked for outliers, violations of heteroscedasticity and

violations of non-linearity using histograms and normality plots in SYSTAT 11. As found in the original sample, lift duration and rms jerk were found to violate the heteroscedasticity assumption and logarithm transformations of the lift duration and rms jerk were performed. The logarithms of lift duration and rms jerk were used in all analyses.

The factor coefficients that were determined from the factor structure and described in Chapter 3 were used to calculate the standard normal factor scores of the additional data. The additional data were normalized to the mean and standard deviation of the lifting parameters from the original sample, and Z scores were calculated for each of the lifting parameters. The factor scores were then calculated from the matrix multiplication of the factor coefficient matrix and the Z scores of each lift of the additional data.

Once the factor scores of each lift were calculated, the lifts were clustered based on the four factor scores of the lift. The statistics of the cluster solution of the original data were used to assign the factor scores of the additional data to a cluster, i.e. the additional data factor scores were assigned to the nearest cluster defined by the original data. The additional data set was added to the original data set and a control subject file and a CLBP subject file was created that contained all subjects' sequences. This same procedure was used to create the lifting sequences of the post-treatment CLBP subjects.

5.2 STATISTICAL ANALYSES

The lifting sequences of the control subjects and the CLBP subjects at pre-treatment were classified to a HMM using the modified jackknife method. The modified jackknife excludes one subjects and trains the control HMM and CLBP HMM with the remaining data. The excluded subject is then classified to either the control HMM or the CLBP HMM based on the logarithm of the likelihood probability.

This section is separated into two sub-sections. The first sub-section describes the statistical analyses performed to test the hypothesis that the HMM classification procedure can identify two groups of CLBP subjects. Specifically, the CLBP subjects classified to the control HMM will be different from the CLBP subjects classified to the CLBP HMM for one or more self-reported measures, medical findings and /or functional capacity measures.

The second sub-section describes the statistical analysis performed to assess whether treatment outcomes can be predicted and if CLBP subjects improved body mechanics after treatments by changing HMM classification at post-treatment assessment. The hypotheses of these data were: (1) the group-by-treatment interaction will be significant, indicating that treatment outcomes can be predicted based on the pre-treatment HMM classification of the CLBP subjects, and (2) the CLBP subjects that changed HMM classification after treatment will be significantly different from the those CLBP subjects that did not change HMM classification on one or more self-reported and/or functional capacity measures.

5.2.1 Pre-treatment CLBP subjects and control subjects

Several measures of the CLBP subjects were compared between the two CLBP groups to determine whether the HMM classification procedure can identify meaningfully different sub-groups of CLBP subjects. These measures include self-reported measures, medical findings, functional capacity measures and pain intensity rating during the functional capacity evaluation. Before these measures were compared, the demographics of the groups were compared to determine whether the composition of the subjects in the groups were similar. The statistical tests performed to assess differences between the two CLBP groups are described in the following paragraphs.

The demographics of the CLBP subjects that were classified to the CLBP HMM and the CLBP subjects classified to the control HMM were compared with analysis of variance (ANOVA) models and chi-squared statistics for the nominal and ordinal demographical measures, respectively. The continuous demographical measures are pain duration in years and age of the subjects in years. The ordinal demographical measures are gender, ethnicity, education, martial status, employment status, pain etiology, pain frequency in days/week, pain frequency in hours per day, and number of surgeries. Any of the demographics that were significantly different between the CLBP subjects classified to the CLBP HMM and the CLBP subjects classified to control HMM were treated as covariates.

The measures from the psychological, medical and functional capacity evaluations were compared between the two CLBP groups. The measures were assigned to a priori domains in order to combine similar measures and to account for correlations between the measures. These

domains were psychosocial, pain, self-efficacy, disability, cognitive, medical, lifting and spinal mobility. Table 13 lists the measures that were included in each of the domains.

Table 13: List of self-reported and functional capacity measures within each of the domains

Domains of self-reported and functional capacity measures		
Pain Domain	***Cognitive Domain***	***Medical Domain***
MPI : Pain Intensity	Coping strategies: emotionality	Medics score
Jan van Breemen: Pain Intensity	Coping strategies: anxiety	Body Mass Index
Spinal Mobility Domain (cm)	***Psychosocial Domain***	***Lifting Domain***
Jan van Breemen: Flexion	MPI Dysfunctional score	Number of lifts
Jan van Breemen: Flexion/Extension	MPI Interpersonally Distressed score	Static strength
Disability Domain	***Self-efficacy Domain***	
MPI : General Activities	Task self-efficacy	
Oswestry Disability rating		
Jan van Breemen: Walking speed		
Jan van Breemen: Functional status		
Pain behavior checklist		

The statistical significance of the psychosocial, pain, medical, spinal mobility, cognitive, and lifting domains were assessed with multivariate analysis of variance (MANOVA) models since each of these domains contains variables that are measured on different dimensions. The only exception was the self-efficacy domain that contained a single measure, for which an ANOVA was performed to assess group differences. The p-values resulting from the MANOVA determined whether any of the measures within the domain were significantly different between the CLBP groups (CLBP subjects assigned to the CLBP HMM and the CLBP subjects assigned to the control HMM). If the p-value was significant, follow-up ANOVAs were performed on the

individual measures within the domains to determine which measure were statistically different between the CLBP groups. A step-wise discriminant function analysis was performed on the measures that were significant based on the follow-up ANOVAs and these tests identified the measures that best separated the CLBP groups. Table 14 summarizes the statistical tests used to compare all measures between HMM classifications.

Effect sizes were calculated to compare the magnitude of the differences between the two CLBP groups on each of the self-reported measures. The effect size index assesses the size of the significant difference between two populations after adjusting for the magnitude of the variable and can be interpreted as the number of standard deviations that separates two groups. The larger the effect size of a dependent variable, the more likely a statistical significance will be attained and the greater the statistical power. Effect size is calculated as the difference in the means divided by a pooled standard deviation. An effect size range of 0-0.32 is small, 0.33-0.55 is medium and 0.56-1.2 is large [84].

During the functional capacity evaluation, CLBP subjects were asked to rate their pain intensity using the pain rating scale at the start of the functional capacity testing (baseline), after static lifting task and at the end of the repetitive lifting task. Differences in the pain intensity ratings between the two CLBP groups were assessed using repeated measures ANOVA. The independent measure was CLBP group and the dependent repeated measures were the pain rating at the three time points during the functional capacity evaluation.

Table 14: Summary of the statistical tests performed to compare the HMM classification groups

Statistical test	Dependent Variables	Independent Variables
Chi-square	Token distributions	Control subjects classified to control HMM and CLBP subjects classified to control HMM
MANOVA	Pain, Disability, Cognitive, Lifting, Spinal Mobility and Medical Domains	Two CLBP groups
Follow-up ANOVA when MANOVA is significant	Each measure within the domain	Two CLBP groups
ANOVA	Self-efficacy measure	Two CLBP groups
Effect size	All measures	Two CLBP groups
Repeated measures ANOVA	Three pain ratings during functional capacity evaluation	Two CLBP groups

5.2.2 Post-treatment CLBP subjects

This section describes the statistical analyses performed to verify that the measures can detect changes due to treatment program, to determine whether treatment outcomes can be predicted based on pre-treatment HMM classification and to identify differences in the self-reported measures of the CLBP subjects that changed HMM classification after treatment and those CLBP subjects that did not change HMM classification.

To determine whether the self-reported and functional capacity measure can detect treatment effects in CLBP patients, paired t-tests and effect sizes were calculated to compare the pre-treatment and post-treatment values of each measure for all CLBP subjects. The paired t-test assessed differences due to treatment and since multiple variables were compared, the Bonferroni correction, which controls for Type I errors, was used to assess significance of the t-

tests. The effect sizes were calculated as the difference between the pre-treatment mean value and post-treatment mean value of all CLBP subjects divided by the pooled standard deviation.

In the statistical analysis of the post-treatment self-reported and functional capacity measures performed to test the hypotheses, the same domains that are described in Table 12 were used. The only exceptions were the exclusion of the pain behavior checklist from the disability domain and the medical domain from the statistical analyses as these measures were not part of the post-treatment assessment. The number of lifts and the weight lifted during the lifting task were multiplied to produce a variable called work. These variables were combined because the sample size of the treatment data was reduced and overparameterization of the data was a concern.

To determine whether treatment outcomes can be predicted based on the HMM classifications at pre-treatment, the self-reported measures at pre-treatment and post-treatment were compared. Doubly repeated measures MANOVAs were used to assess whether the pain, disability, cognition and spinal mobility domains were significantly different. This statistical test was performed separately for each domain and was chosen because the measures within these domains are multidimensional and completed at two different assessments: pre-treatment and post-treatment. The dependent variables were the measures within a domain and the independent variables were the group assignment. The p-values resulting from the statistical test determined whether any of the measures within the domain were significantly different between the groups, due to treatment and group-by-treatment interaction. Follow-up ANOVAs were performed on the significant results of the MANOVA to assess which of the measures within the domain were significantly different. A significant group-by-treatment interaction indicated that the measure(s) could be used to predict treatment outcomes.

Separate repeated measures ANOVAs were performed for the domains of self-efficacy and work, since both of these domains contain a single measure for each assessment. The dependent variables were either the self-efficacy measure or work, and the independent variables were the pre-treatment HMM classifications of the CLBP subjects. The p-values determined whether the measures were significantly different between the groups and due to treatment. Effect sizes were calculated for all measures to compare the two CLBP groups at pre-treatment and post-treatment assessments.

The post-treatment classifications of the CLBP subjects that were assigned to the CLBP HMM at pre-treatment were examined to determine whether these subjects changed classification after treatment. The subjects were either labeled as a changer or non-changer when the subject was classified to a HMM at the post-treatment assessment. The subject was labeled as a changer when the CLBP subject changed HMM classification at post-treatment assessment, i.e. a CLBP subject was assigned to the CLBP HMM at the pre-treatment assessment and the same subject was assigned to the control HMM at post-treatment assessment. A CLBP subject was labeled as a non-changer when the subject was classified to the CLBP HMM at both pre-treatment and post-treatment assessments.

The CLBP subjects that were classified to the control HMM were also examined to determine whether these subjects changed HMM classification after treatment. Since it is unlikely that these CLBP subjects would perform worse after treatment, none of the CLBP subjects classified to the control HMM was expected to be classified to the CLBP HMM at the post-treatment assessment.

The same statistical tests used to assess whether treatment outcome can be predicted (doubly repeated measures MANOVA and repeated measures ANOVA) were used to determine

whether the changers and non-changers were different in any of domains. In addition, the effect sizes were calculated for all of the self-reported measures. Table 15 shows a summary of the statistical tests performed to assess whether treatment outcome can be predicted and to assess differences between changers and non-changers.

The pain intensity ratings reported at baseline, after the static strength task and after the repetitive lifting task for both the pre-treatment and post-treatment assessment were compared between changers and non-changers. Repeated measures ANOVA were used to assess for significant differences between treatment assessments and changes between pain ratings from baseline, after the static strength task and at end of the repetitive lifting task. The independent variable was the group assignment of changer or non-changer and the dependent variables were pain ratings at the three time points for both the pre-treatment and post-treatment functional capacity evaluations. The p-values resulting from the ANOVA determined statistical significance between the changers and non-changers, between the three time points, between treatment assessments and the interactions.

Table 15: Summary of the statistical tests that were performed on the pre-TX and post-TX CLBP data

Statistical test	Dependent Variables	Independent Variables
Effect size	All measures	CLBP pre-TX and post-TX
Repeated measures ANOVA	Three pain ratings during functional capacity evaluation at pre-TX and post-TX	CLBP groups assigned at pre-TX
Doubly repeated measures MANOVA	Pain, Disability, Cognitive, Lifting, and Spinal mobility Domains at pre-TX and post-TX	CLBP groups assigned at pre-TX
Follow-up ANOVA when MANOVA was significant	Each measure within the domain at pre-TX and post-TX	CLBP groups assigned at pre-TX
Effect size	All measures	CLBP groups assigned at pre-TX
Doubly repeated measures MANOVA	Pain, Disability, Cognitive, Lifting, and Spinal mobility Domains at pre-TX and post-TX	Changers and non-changers
Follow-up ANOVA when MANOVA was significant	Each measure within the domain at pre-TX and post-TX	Changers and non-changers
Effect size	All measures	Changers and non-changers
Repeated measures ANOVA	Three pain ratings during functional capacity evaluation at pre-TX and post-TX	Changers and non-changers

5.3 RESULTS

The results of the HMM classification procedure when applied to the clinical data and the statistical tests assessing whether the CLBP subjects that were classified to the control HMM and the CLBP subjects that were classified to the CLBP HMM were different on the self-reported, medical findings, and functional capacity measures are described in the first sub-section.

The second sub-section describes the results of the statistical tests to assess whether treatment outcomes can be predicted based on HMM classification at pre-treatment and whether

CLBP subjects classified to the CLBP HMM changed HMM classification after treatment. For all statistical analyses, a p-value of 0.05 or less indicated statistical significance.

5.3.1 HMM classifications of CLBP subjects at pre-treatment and control subjects

The classification procedure identified 35 CLBP subjects to the control HMM and 46 CLBP subjects to the CLBP HMM. In the control group, 46 control subjects were classified to the control HMM and 7 control subjects were classified to the CLBP HMM. Since the number of control subjects that were classified to the CLBP HMM is small, this sample was not considered a sub-group.

For each of the four groups, the token distributions were plotted with a histogram as shown in Figures 25, 26, 27, 28 respectively. The CLBP subjects that were classified to the control HMM frequently used all of the tokens except token 1. The CLBP subjects classified to the CLBP HMM used token 1 in 75% of the lifts. The control subjects that were classified to the control HMM frequently performed lifts that were associated with tokens 2, 3, 5. The control subjects classified to the CLBP HMM appeared to use all of the tokens almost equally except for token 3.

A chi-squared statistic showed that the token distributions were significantly different when comparing the control subjects that were classified to the control HMM and the CLBP subjects classified to the control HMM ($\chi = 301.3$, $p = 0.0001$). The significant chi-squared result indicates that even though both groups were classified to the same model, the frequency of the lifting patterns were different between the CLBP subjects classified to the control HMM and the control subjects classified to the control HMM.

Since the CLBP subjects that were classified to the control HMM used significantly different distribution of lifting patterns than the control subjects classified to the control HMM, the possibility of using two CLBP models to describe the CLBP subjects was investigated. Two HMMs were designed to describe the lifting patterns of the two CLBP groups and the simulation studies were performed to assess whether the HMMs could reliably identify mislabeled sequences. The results are described in Appendix B and showed that the CLBP subjects were still classified to the control HMM. Due to the unreliability of the HMMs, CLBP HMM and the control HMM were used to classify the CLBP subjects at post-treatment assessment.

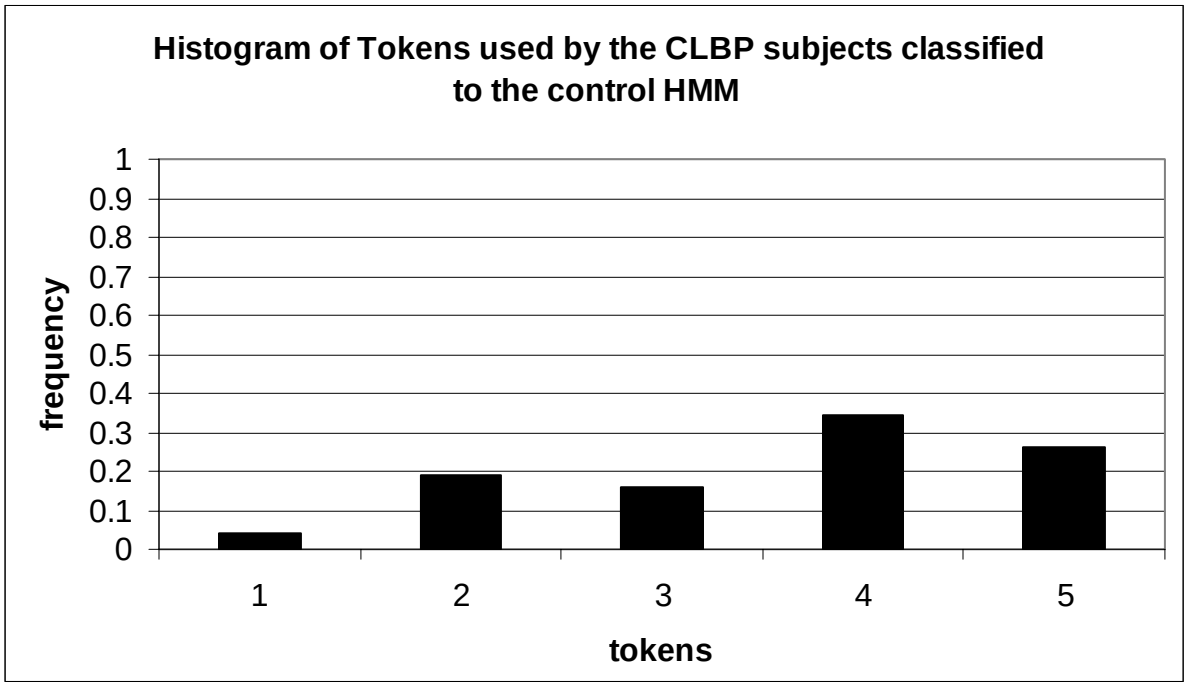


Figure 25: Histogram of the tokens used by the CLBP subjects classified to the control HMM

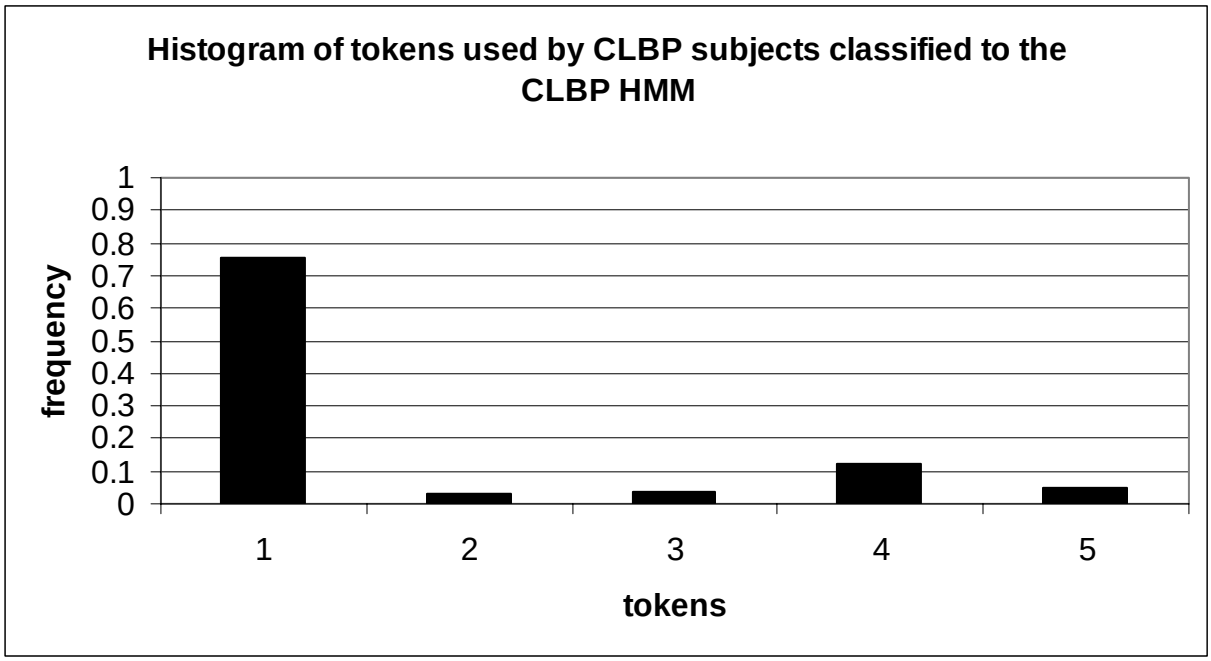


Figure 26: Histogram of the tokens used by the CLBP subjects classified to the CLBP HMM

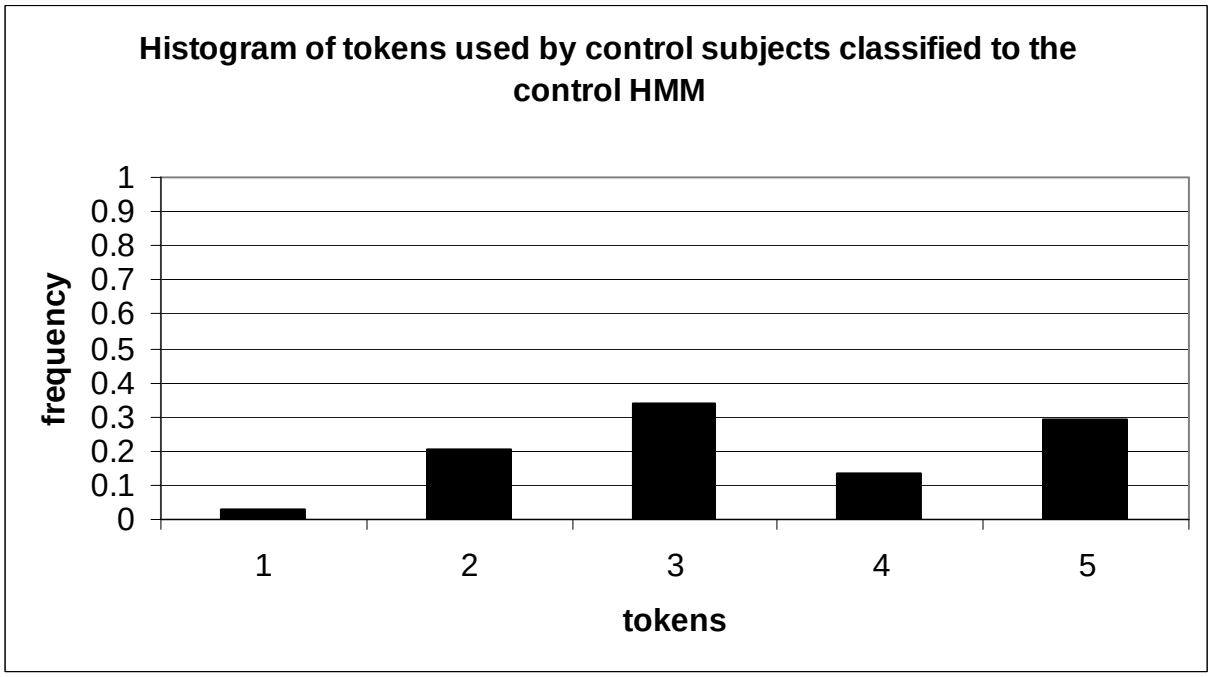


Figure 27: Histogram of the tokens used by the control subjects classified to the control HMM

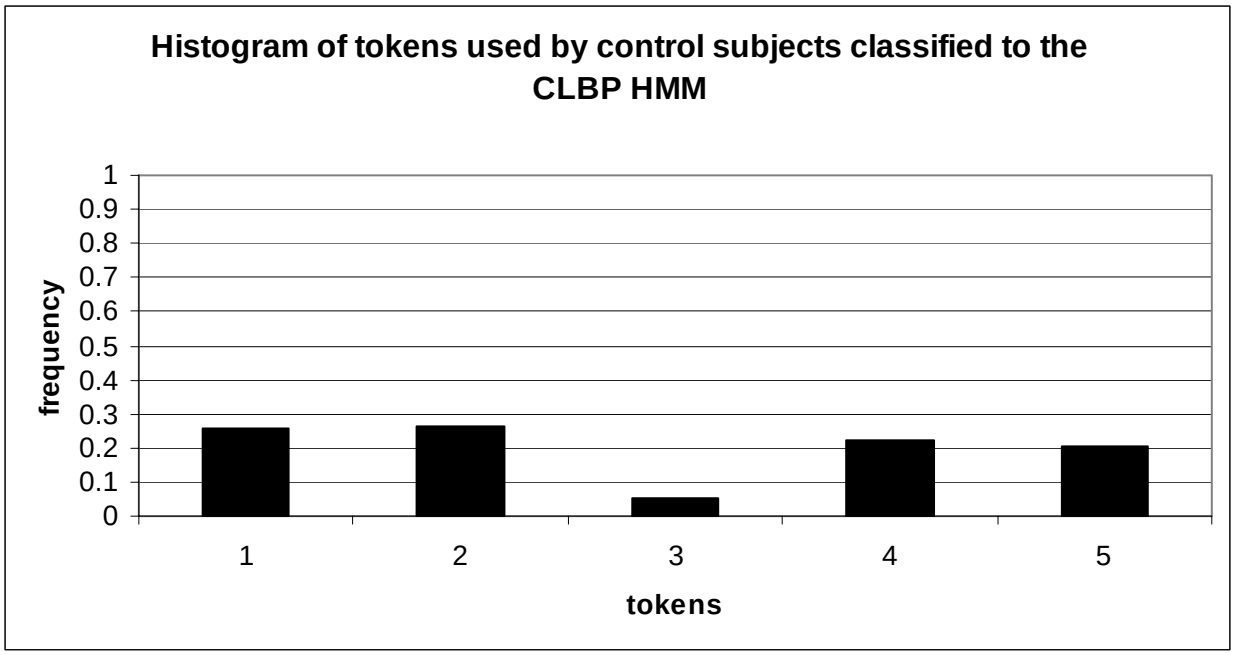


Figure 28: Histogram of the tokens used by the control subjects that were classified to the CLBP HMM

The demographics of the two CLBP groups were compared to determine whether these groups were different for any of these variables. The p-values from the ANOVA and chi-squared tests are shown in Table 16. None of the demographics were significantly different between the CLBP subjects classified to the control HMM and the CLBP subjects classified to the CLBP HMM.

Table 16: Average values (standard deviations) or percentage of subjects in each of the demographical variables and corresponding p-values are listed below.

		CLBP fit CLBP HMM	CLBP fit control HMM	p-value
Pain Duration (years)		3.72 (5.13)	4.52 (5.69)	0.523
Age (years)		36.36 (9.24)	39.66 (11.03)	0.162
Gender	Males	55%	63%	0.439
	Females	45%	36%	
Ethnicity	White	67%	77%	0.247
	African American	33%	19%	
	Other	0%	3%	
Education	12^{th} grade or less	20%	16%	0.589
	High school graduate	34%	31%	
	Trade/ technical school	25%	31%	
	Some college	16%	16%	
	College degree or higher	5%	6%	
Marital status	Single	25%	22%	0.851
	Separated/divorced	20%	28%	
	Married	50%	44%	
	Widowed	5%	6%	
Employment status	Full-time (over 30 hrs per week)	18%	32%	0.255
	Part-time (less than 30 hrs week)	3%	4%	
	Unemployed	0%	4%	
	Retired	11%	0%	
	Working part-time because of pain	8%	8%	
	Unemployed because of pain	55%	44%	
	Retired early because of pain	0%	4%	
	Other	5%	4%	
How pain began?	Accident at work	63%	65%	0.765
	Accident at home	7%	3%	
	Following surgery or illness	5%	3%	
	Pain just began	7%	16%	
	Other	19%	13%	
Pain hours	0-4 hours	2%	3%	0.206
	4-8 hours	10%	3%	
	8-12 hours	20%	7%	
	more than 12 hours	68%	87%	
Surgery	Never	73%	58%	0.265
	Once	14%	13%	
	Twice	14%	19%	
	More than twice	0%	10%	

CLBP subjects classified to the control HMM and CLBP subjects classified to the CLBP HMM were compared to determine whether the groups were different on any of the measures. The MANOVAs showed that the domains of pain ($p = 0.007$), lifting ($p = 0.001$), and self-efficacy ($p = 0.013$) were statistically significant when comparing the CLBP subjects classified to the control HMM and the CLBP subjects classified to the CLBP HMM. Follow-up ANOVAs indicate that within the lifting domain, the number of lifts completed ($p = 0.0001$) was significant and in the pain domain, MPI section of pain severity ($p = 0.025$) and pain intensity ($p = 0.005$) from the Jan van Breemen questionnaire were statistically significant between the two CLBP groups. There was only one measure in the self-efficacy domain and an ANOVA showed that the task self-efficacy measure ($p = 0.013$) was significantly different between the CLBP groups. The CLBP subjects that were classified to the control HMM were found to perform more lifts, reported lower levels of pain intensity and pain severity, and had greater task self-efficacy than the CLBP subjects classified to the CLBP HMM. The mean values, standard deviation and effect sizes along with the corresponding p-values of the statistical tests are shown in Table 17.

A stepwise discriminant function analysis was performed to determine which measures best separated the CLBP subjects classified to the control HMM from the CLBP subjects classified to the CLBP HMM. The discriminant analysis found that number of lifts, pain severity, pain intensity and self-efficacy were all independent contributors to CLBP group separation. The number of lifts and pain intensity from the Jan van Breemen were the two most discriminating measures between the CLBP subjects that were classified to the control HMM and the CLBP subjects classified to the CLBP HMM. The order that the measures entered into the discriminant function model are shown in the last column of Table 17.

A large effect size (greater than 0.55) was found for the measures of pain intensity from the Jan van Breemen examination, number of lifts performed during the lifting task, MPI Dysfunctional composite score and self-efficacy. A moderate effect size (between 0.33-0.55) was found for pain severity from MPI, the measures from the Coping Strategies, functional status and walking speed from the Jan van Breemen examination and static lifting strength. The remaining measures showed small effect sizes (between 0-0.32). Effect sizes comparing the two groups of CLBP subjects for each of the measures are listed in Table 17.

Table 17: Average values (standard deviation) and effect size calculation for comparison of the two CLBP groups are listed below. P-values and discriminant function analysis entry of the measure into the model are also shown. Bold indicate significant p-values and italic indicates domain.

Measures	**Means (SD) for CLBP groups**				
	CLBP subjects fit control HMM	CLBP subjects fit CLBP HMM	Effect size	p-values	DFA entry
Sample Size	35	46			
Pain Domain				**0.007**	
MPI : Pain Severity	4.47 (0.71)	4.89 (0.86)	0.535	**0.025**	3
Jan van Breemen: Pain Intensity	5.57 (1.72)	6.66 (1.56)	0.665	**0.005**	2
Psychosocial Domain				0.100	
MPI Dysfunctional composite score	57.27 (9.24)	62.60 (10.30)	0.547	0.035	
MPI Interpersonally Distressed composite score	39.12 (13.36)	38.10 (11.74)	0.081	0.737	
Cognitive Domain				0.119	
Coping strategies: emotionality	3.21 (1.32)	3.85 (1.31)	0.487	0.039	
Coping strategies: worrying	4.13 (1.35)	4.62 (1.18)	0.387	0.095	
Disability Domain				0.122	
MPI : General Activities	2.08 (0.9)	1.77 (0.79)	0.367	0.068	
Oswestry Disability rating	51.09 (14.43)	51.06 (14.43)	0.002	0.991	
Jan van Breemen: Walking speed	35.29 (7.95)	41.63 (15.89)	0.426	0.057	
Jan van Breemen: Functional status	4.50 (1.62)	3.91 (1.15)	0.532	0.123	
Pain behavior checklist	6.91 (3.60)	7.77 (3.94)	0.228	0.238	
Spinal Mobility Domain (cm)				0.662	
Jan van Breemen : Flexion	5.19 (4.01)	5.21 (1.41)	0.007	0.975	
Jan van Breemen: Flexion/Extension	6.10 (2.18)	6.48 (1.88)	0.187	0.414	
Medical Domain				0.665	
Medics scale	-0.15 (0.63)	0.003 (0.86)	0.205	0.383	
Body Mass Index	27.65 (5.9)	28.04 (6.77)	0.062	0.812	
Static lifting Domain				**0.002**	
Number of lifts	35.06 (23.28)	18.61 (16.39)	0.829	**0.001**	1
Static strength	84.31 (54.47)	60.09 (57.68)	0.432	0.121	
Self-efficacy Domain	3.42 (0.91)	2.88 (0.91)	0.593	**0.013**	4

The results of the repeated measure ANOVA showed that during the functional capacity evaluation the CLBP subjects classified to the control HMM reported significantly lower pain intensity ratings than CLBP subjects classified to the CLBP HMM ($p = 0.0001$). Both CLBP groups reported increased pain intensity ratings from the baseline to the end of the lifting task ($p = 0.0001$). There was no significant group-by-time interaction. The mean, standard deviation and p-values of the pain intensity during the functional capacity evaluation are listed in Table 18.

Table 18: Average values (standard deviations) of the pain intensity ratings of the CLBP subjects at the three time points during the functional capacity evaluation. P-values from the repeated measures ANOVA assessing differences in group, time and group-by-time interaction are listed.

HMM classifications	Average Pain Ratings			P-values		
	Baseline	After static lifting task	After dynamic lifting task	Group	Time	Group – by-time
CLBP classified to CLBP HMM	5.5 (2.13)	6.54 (2.08)	7.72 (1.57)	0.0001	0.0001	0.931
CLBP classified to Control HMM	3.91 (2.71)	4.83 (2.60)	6.06 (2.20)			

The number of lifts performed during the task was significantly different between the two CLBP groups. To assess whether the subjects could be separated into groups based only on the number of lifts completed, a histogram of the number of lifts was constructed for both CLBP groups (Figure 29). The CLBP subjects that were classified to the CLBP HMM performed fewer lifts more frequently than the CLBP subjects classified to the control HMM but the distributions overlap. It is does not appear from the histograms that a cut-off or threshold values could be used to clearly separate the CLBP subjects into groups.

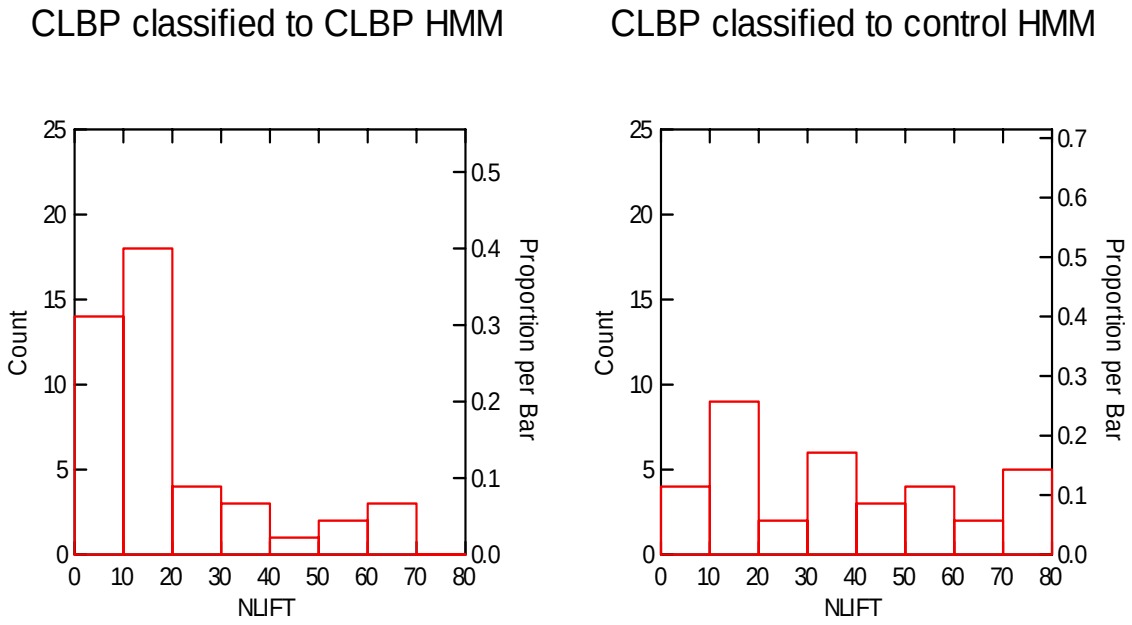


Figure 29: Histogram of the number of lifts performed during the repetitive lifting task for CLBP subjects classified to the CLBP HMM and the CLBP subjects classified to the control HMM.

5.3.2 Classification of the CLBP subjects at post-treatment assessment

The results of the statistical analyses to verify that the self-reported measures can detect treatment effects, to determine whether treatment outcome can be predicted from pre-treatment HMM classification and to determine whether changers and non-changers are different are discussed in this sub-section.

To verify that the self-reported and functional capacity measures can detect the effect of treatment in CLBP subjects, effect sizes and paired t-tests comparing the measures of all CLBP subjects at pre-treatment and post-treatment were calculated (Table 19). A large effect size was found for MPI section of pain severity, MPI Dysfunctional composite score, Coping Strategies

questions of emotionality and anxiety, Oswestry disability rating, walking speed, self-efficacy, and functional status questions from the Jan van Breemen. Moderate effect sizes were found for MPI section of General Activities, Jan van Breemen sections of pain intensity, and flexion/extension index. Since 14 variables were compared between the CLBP subjects, Bonferroni correction was used to determine significance of the t-tests and a p-value of 0.0005 or lower was considered statistically significant. Significant p-values were found for all measures except MPI Interpersonally Distressed composite score and task self-efficacy. The results of the effect sizes and p-values indicate that the self-reported and functional capacity measures can measure treatment effects.

Table 19: Average values (standard deviation) of the measures at pre-TX and post-TX assessments for all CLBP subjects. Effect size calculations and p-values from the t-tests comparing the pre-TX and post-TX means of all CLBP subjects are shown. Bold indicates significant p-values and large effect sizes.

Measure	Pre-TX mean (SD)	Post-TX mean (SD)	EFFECT SIZE	p-value significant when $< 5x10^{-4}$
MPI section: Pain Severity	4.64 (0.83)	3.59 (1.25)	**1.01**	**$1x10^{-6}$**
MPI section: General Activities	1.93 (0.86)	2.32 (0.81)	0.467	**0.0002**
MPI Dysfunctional score	60.08 (10.39)	48.91 (8.54)	**1.18**	**$1x10^{-9}$**
MPI Interpersonally Distressed score	36.50 (12.05)	39.55 (10.87)	0.266	0.001
Coping strategies: Emotionality	3.66 (1.38)	2.44 (1.38)	**0.837**	**$2.5x10^{-6}$**
Coping strategies: Anxiety	4.48 (1.32)	3.40 (1.41)	**0.791**	**$1x10^{-6}$**
Oswestry disability rating	52.04 (14.90)	39.39 (11.93)	**0.943**	**$6.2x10^{-7}$**
Walking Speed	38.64 (13.92)	32.81 (7.90)	**0.534**	**$1.5x10^{-6}$**
Jan van Breemen Questions: Pain Intensity	6.07 (1.79)	5.16 (1.95)	0.487	**0.00047**
Jan van Breemen Questions: Functional status	4.02 (1.41)	5.48 (1.55)	**0.986**	**$3x10^{-9}$**
Jan van Breemen Exam: Lumbar flexion index	4.86 cm (1.42)	5.30 cm (1.31)	0.325	**0.0003**
Jan van Breemen Exam: Flexion/extension index	6.26 cm (1.80)	7.11 cm (1.89)	0.461	**$1x10^{-5}$**
Task Self-Efficacy	3.01 (0.94)	4.37 (3.44)	**0.621**	0.0069
Work of dynamic lifting task	704 (1060)	1086 (1360)	0.316	**0.00048**

To determine whether it is possible to predict treatment outcome based on pre-treatment HMM classification, the pre-treatment and post-treatment values of the measures were compared. A total of 54 subjects had pre-treatment and post-treatment data and of these 54 subjects, 30 CLBP subjects were assigned to the CLBP HMM and 24 CLBP subjects were classified to the control HMM at the pre-treatment assessment. The p-values for the group-by-treatment interaction from the doubly repeated measures MANOVAs were non-significant for all

domains and measures, indicating that treatment outcomes cannot be predicted from the pre-treatment HMM classifications (Table 20).

Although the HMM classifications could not predict treatment outcomes, significant differences were found between the two CLBP groups. Group differences and treatment effects were significant for the pain domain and for the spinal mobility domain. Follow-up ANOVAs showed that within the pain domain, the pain intensity from the Jan van Breemen examination had a significant group difference ($p = 0.004$) and treatment effect ($p = 0.0001$). CLBP subjects that were classified to the control HMM reported lower pain intensity than the CLBP subjects that were classified to the CLBP HMM and for both groups, pain intensity decreased after treatment. In the spinal mobility domain, the flexion task from the Jan van Breemen examination was significantly different between the CLBP groups ($p = 0.011$) and for treatment assessment ($p = 0.001$). The CLBP subjects classified to the control HMM had greater range of motion during flexion than the CLBP subjects classified to the CLBP HMM. In addition to pain intensity and flexion, self-efficacy were also found to be significantly different between the groups ($p = 0.014$). The CLBP subjects that were classified to the control HMM reported higher levels of perceived self-efficacy at the post-treatment assessment than CLBP subjects classified to the CLBP HMM. For both groups, self-efficacy increased after treatment ($p = 0.005$).

Effect sizes of the measures at post-treatment comparing the CLBP subjects classified to the control HMM and the CLBP subjects classified to the CLBP HMM were calculated. The measures of pain intensity, flexion and the self-efficacy had large effect sizes. Medium effect sizes were found for walking speed, pain severity from the MPI, MPI Dysfunctional composite score and for the two measures of the Coping Strategies. The mean values, standard deviation, effect sizes and corresponding p-values from the statistical tests are listed in Table 20.

Table 20: Average values (standard deviations) and effect size calculations of the measures comparing the two groups of CLBP subjects found in pre-treatment assessment. P-values from the MANOVA assessing differences between groups, treatment and group-by-treatment interaction are shown. Bold indicates significant p-values

Measures	**Means (SD) for groups and TX**				**p-values**		
	TX	CLBP subjects fit CLBP HMM	CLBP subjects fit control HMM	Effect size	Group (G)	TX (T)	G x T
Sample Size		**30**	**24**				
Pain Domain					**0.015**	**0.0001**	0.976
MPI Pain Severity	Pre Post	4.81 (0.85) 3.83 (1.26)	4.33 (0.76) 3.16 (1.25)	0.596 0.534	0.056	**0.0001**	0.84
Jan van Breemen: Pain Intensity	Pre Post	6.62 (1.66) 5.75 (1.95)	4.99 (1.75) 4.15 (1.77)	0.956 0.860	**0.004**	**0.001**	0.884
Psychosocial Domain					0.339	**0.0001**	0.434
MPI Dysfunctional	Pre Post	62.60 (10.30) 50.55 (7.79)	57.27 (9.24) 46.85 (9.16)	0.546 0.437	0.014	**0.0001**	0.236
MPI Interpersonally Distressed	Pre Post	38.10 (11.74) 39.46 (10.87)	39.12 (13.36) 39.70 (11.19)	0.081 0.022	0.772	**0.002**	0.602
Cognitive Domain					0.213	**0.0001**	0.927
Coping strategies: emotionality	Pre Post	3.92 (1.39) 2.74 (1.43)	3.46 (1.37) 2.16 (1.38)	0.333 0.413	0.083	**0.0001**	0.784
Coping strategies: worrying	Pre Post	4.67 (1.28) 3.66 (1.49)	4.33 (1.31) 3.14 (1.36)	0.263 0.365	0.108	**0.0001**	0.703
Disability Domain					0.623	**0.0001**	0.195
MPI : General Activities	Pre Post	1.85 (0.86) 2.41 (0.78)	1.98 (0.86) 2.22 (0.83)	0.151 0.236	0.513	**0.0001**	0.394
Oswestry Disability rating	Pre Post	51.67 (15.47) 40.69 (11.25)	53.26 (15.08) 38.48 (13.22)	0.104 0.181	0.543	**0.0001**	0.319
Jan van Breemen: Walking speed	Pre Post	41.78 (16.71) 33.95 (8.21)	33.25 (6.80) 30.51 (6.55)	0.726 0.466	0.382	**0.0001**	0.432
Jan van Breemen: Functional status	Pre Post	3.73 (1.06) 5.50 (1.60)	4.29 (1.47) 5.32 (1.58)	0.443 0.113	0.640	**0.0001**	0.085
Spinal Mobility Domain (cm)					**0.012**	**0.0001**	0.460
Jan van Breemen : Flexion	Pre Post	5.13 (1.20) 5.68 (1.05)	4.81 (1.45) 4.92 (1.12)	0.242 0.700	**0.011**	**0.001**	0.263
Jan van Breemen : Flexion/Extension	Pre Post	6.26 (1.78) 7.28 (1.65)	6.51 (1.85) 7.14 (2.06)	0.138 0.075	0.418	**0.0001**	0.669
Self-Efficacy Domain	Pre Post	2.78 (0.87) 3.59 (1.05)	3.40 (0.97) 5.54 (5.43)	0.674 0.602	**0.014**	**0.005**	0.227

In the pre-treatment assessment classification, 46 CLBP subjects were classified to the CLBP HMM and of these 46 subjects, only 30 completed the lifting task after treatment. In the post-treatment assessment classification, fifteen of the 30 CLBP subjects were assigned to the CLBP HMM and these subjects were labeled as non-changers. Fifteen CLBP subjects were assigned to the control HMM at post-treatment assessment classification and these subjects were labeled as changers.

A total of 35 CLBP subjects were classified to the control HMM at pre-treatment and of these subjects, 22 CLBP subjects had post-treatment data. Twenty of the 22 CLBP subjects were classified to the control HMM at post-treatment assessment. Two subjects were classified to the CLBP HMM at the post-treatment and these subjects were not considered a group due to the small number of subjects. Figure 30 displays a diagram of the CLBP subject's assignments at pre-treatment and post-treatment.

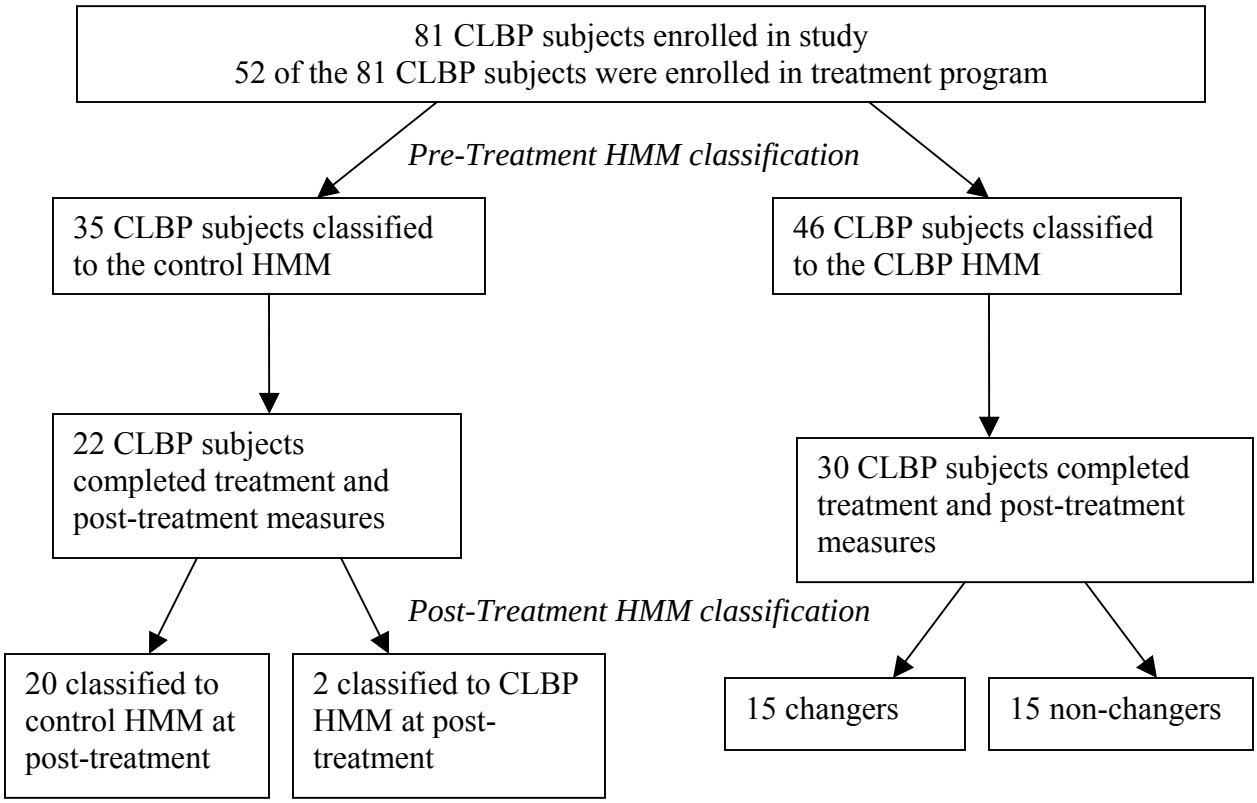


Figure 30: Diagram of the sample size of CLBP subjects at pre-treatment and post-treatment HMM classification.

The demographics of the changers and non-changers were investigated to determine whether the group compositions were different. None of the p-values were significant from the ANOVA and chi-squared test indicating that the groups were similar for age, pain duration, gender, ethnicity, education, martial status, employment status, how the pain began, pain frequency, or number of surgeries. The p-values comparing the demographics are shown in Table 21.

Table 21: Average values (standard deviations) or percentage of changer and non-changers in each of the demographical variables. P-values from the statistical tests comparing the demographics of the two groups are listed below.

		Non-changers	Changers	p-values
Pain Duration (years)		3.96 (5.92)	4.43 (6.24)	0.833
Age (years)		37.07 (8.84)	35.8 (10.19)	0.719
Gender	Males	40%	47%	0.713
	Females	60%	53%	
Ethnicity	White	71%	73%	0.909
	African American	29%	27%	
Education	12^{th} grade or less	20%	13%	0.587
	High school graduate	47%	27%	
	Trade/ technical school	20%	27%	
	Some college	13%	27%	
	College degree or higher	0%	7%	
Marital status	Single	33%	0%	0.102
	Separated/divorced	13%	27%	
	Married	47%	67%	
	Widowed	7%	7%	
Employment status	Full-time (over 30 hrs per week)	13%	7%	0.140
	Part-time (less than 30 hrs week)	0%	7%	
	Retired	0%	20%	
	Working part-time because of pain	20%	0%	
	Unemployed because of pain	47%	47%	
	Other	7%	0%	
How pain began?	Accident at work	57%	67%	0.270
	Accident at home	0%	13%	
	Following surgery or illness	14%	0%	
	Pain just began	14%	0%	
	Other	14%	20%	
Pain hours	0-4 hours	0%	7%	0.303
	4-8 hours	7%	0%	
	8-12 hours	31%	21%	
	more than 12 hours	54%	71%	
Surgery	Never	80%	80%	1.00
	Once	7%	7%	
	Twice	13%	13%	

The distributions of lifting tokens of the changers and non-changers at post-treatment are shown in Figures 31 and 32. Comparing the figures, it appears that non-changers use token 1 more frequently than the changers. The token distribution of the changers was compared to the distribution of the CLBP subjects that were classified to the control HMM at pre-treatment and at post-treatment assessments. A chi-squared statistic showed that the CLBP subjects classified to the control HMM had significantly different frequencies of lifting patterns than the changers (χ = 198.44, p = 0.0001). Figure 33 shows the token distribution of the CLBP subjects that were classified to the control HMM at pre-treatment and post-treatment.

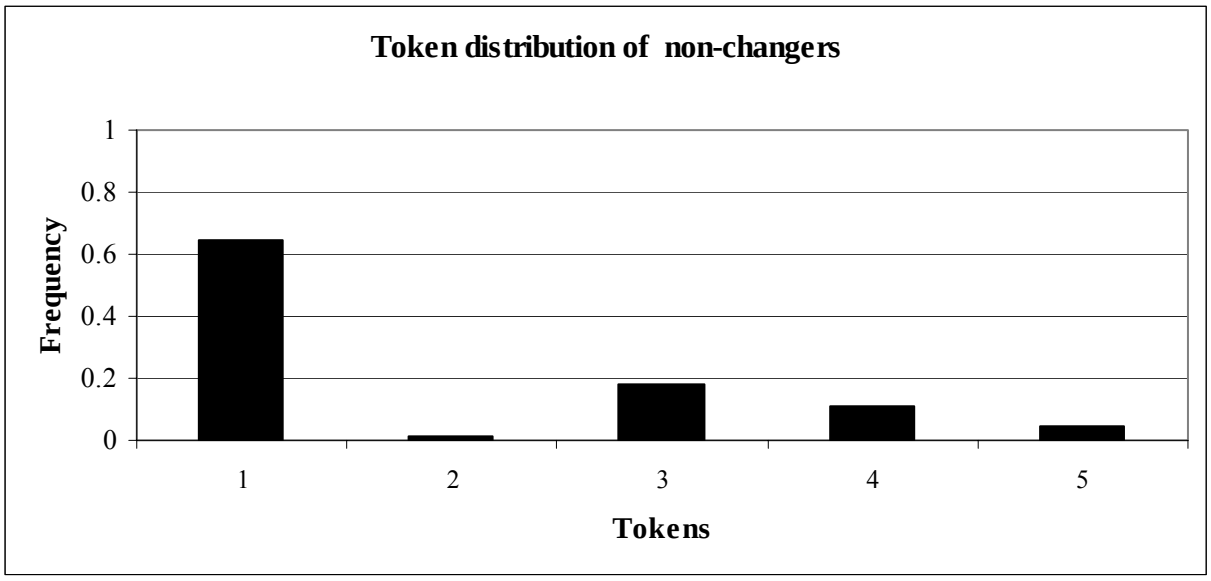


Figure 31: Histogram of the tokens used by the non-changers at post-treatment assessment

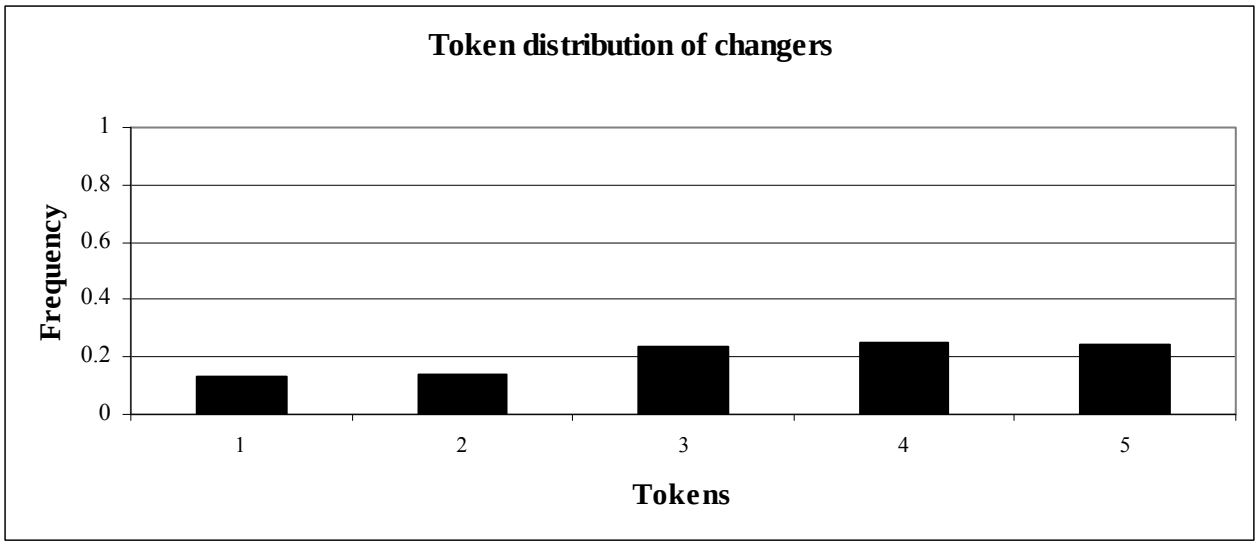


Figure 32: Histogram of the tokens used by the changers at post-treatment assessment

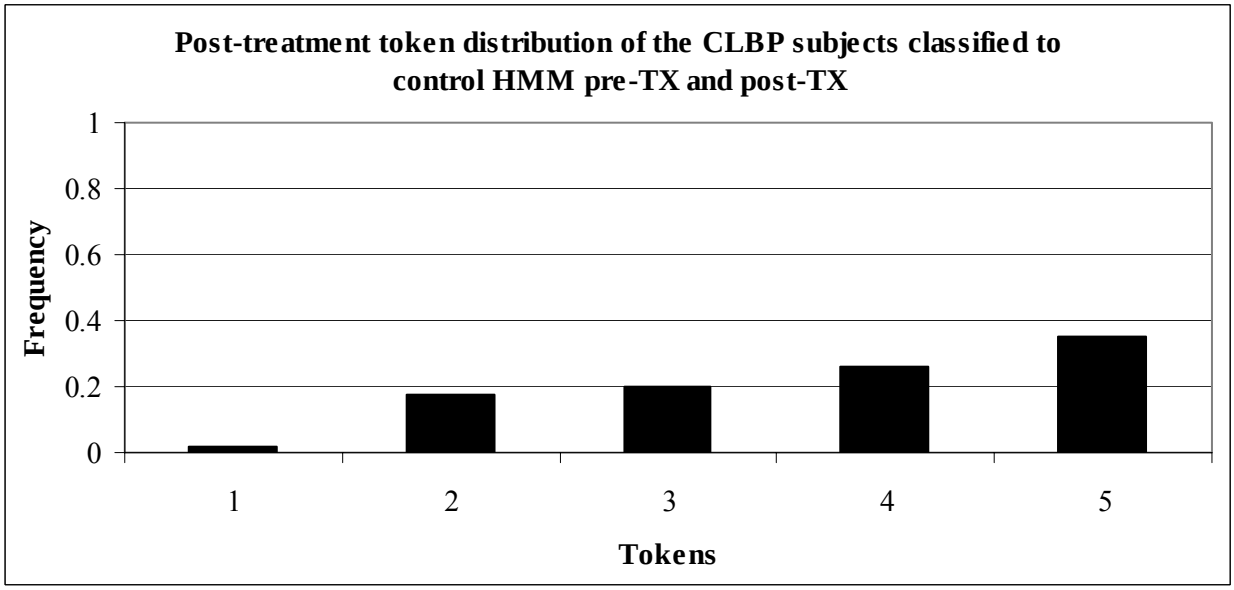


Figure 33: Histogram of the tokens used by the CLBP subjects that were assigned to the control HMM at pre-treatment and post-treatment assessment. This distribution corresponds to the post-treatment assessment.

The doubly repeated measure MANOVA found the only domain significantly different between the changers and the non-changers was spinal mobility ($p = 0.006$). A follow-up ANOVA showed that the changers and non-changers were significantly different for the flexion/extension task from the Jan van Breemen examination ($p = 0.001$). The CLBP changers had greater mobility when performing flexion\extension motion than the CLBP non-changers. For both groups, mobility significantly increased after treatment ($p = 0.0001$). For all of the domains and measures, the treatment effect was significant as shown in Table 22. There were no significant group-by-treatment interactions.

The effect sizes of the post-treatment measures comparing the changers and non-changers were calculated and are shown in Table 22. A large effect was found for walking speed, MPI Dysfunctional composite score, flexion/extension mobility, flexion mobility and for both measures from the Coping Strategies. The measures of MPI pain severity, Jan van Breemen pain intensity, work during the lifting task and MPI Interpersonally Distress composite score had medium effect sizes.

Table 22: Average values (standard deviation) and effect size calculations of the measures at pre-TX and post-TX assessment for changers and non-changers are shown. P-values from the MANOVA assessing differences between the groups, treatment and group-by-treatment interaction are also shown. Bold indicates significant p-values

Measures	**Means (SD) for post-treatment groups and assessments (pre/post)**				**MANOVA Results p-values**		
	TX	Non-changer	Changer	Effect size	Group (G)	TX (T)	G x T
Sample size		15	15				
Pain Domain					0.394	**0.001**	0.459
MPI : Pain Severity	Pre Post	4.89 (0.78) 4.16 (1.24)	4.73 (0.93) 3.53 (1.25)	0.187 0.506	0.221	**0.0001**	0.244
Jan van Breemen: Pain Intensity	Pre Post	6.89 (1.75) 6.14 (1.84)	6.34 (1.58) 5.36 (2.05)	0.33 0.401	0.197	**0.007**	0.970
Psychosocial Domain					0.694	**0.0001**	0.780
MPI Dysfunctional	Pre Post	64.22 (10.94) 52.87 (7.85)	61.58 (11.39) 48.08 (7.66)	0.236 0.618	0.389	**0.0001**	0.633
MPI Interpersonally Distressed	Pre Post	36.52 (15.04) 38.08 (11.80)	36.50 (11.97) 41.66 (10.54)	0.001 0.321	0.922	**0.006**	0.592
Cognitive Domain					0.755	**0.003**	0.134
Coping strategies: emotionality	Pre Post	3.88 (1.33) 3.20 (1.50)	3.96 (1.49) 2.13 (1.27)	0.057 0.773	0.450	**0.001**	0.062
Coping strategies: worrying	Pre Post	4.58 (1.49) 4.10 (1.41)	4.76 (1.09) 3.26 (1.49)	0.14 0.579	0.522	**0.001**	0.046
Disability Domain					0.303	**0.0001**	0.877
MPI: General Activities	Pre Post	1.83 (0.92) 2.07 (0.72)	1.86 (0.82) 2.40 (0.82)	0.034 0.429	0.269	**0.0001**	0.346
Oswestry Disability rating	Pre Post	53.59 (16.37) 41.86 (11.6)	49.87 (14.92) 39.60 (11.2)	0.238 0.198	0.399	**0.001**	0.912
Jan van Breemen: Walking speed	Pre Post	42.50 (13.9) 36.21 (6.94)	41.06 (19.59) 31.69 (8.99)	0.086 0.567	0.222	**0.001**	0.882
Jan van Breemen: Functional status	Pre Post	3.96 (0.89) 5.68 (1.68)	3.50 (1.20) 5.32 (1.55)	0.44 0.223	0.542	**0.0001**	0.922
Spinal Mobility Domain (cm)					**0.006**	**0.0001**	0.324
Jan van Breemen : Flexion	Pre Post	4.91 (1.37) 5.36 (1.17)	5.36 (0.99) 5.99 (0.83)	0.381 0.63	0.061	**0.001**	0.617
Jan van Breemen : Flexion/Extension	Pre Post	5.69 (1.77) 6.34 (1.58)	6.82 (1.65) 8.21 (1.12)	0.661 1.39	**0.001**	**0.0001**	0.564
Self-Efficacy Domain	Pre Post	2.74 (0.96) 3.55 (1.15)	2.82 (0.80) 3.63 (0.98)	0.091 0.075	0.801	**0.0001**	0.991
Work	Pre Post	308 (291) 408 (387)	304 (286) 816 (1385)	0.014 0.46	0.466	**0.037**	0.285

The pain intensity ratings reported during the functional capacity evaluation were not significantly different between changers and non-changers. The pain ratings were significantly different between the three time points ($p = 0.0001$) and from pre-treatment to post-treatment assessment ($p = 0.006$). The treatment-by-time interaction was significant ($p = 0.002$) and the group-by-time-by-treatment interaction was significant ($p = 0.0001$). The average pain ratings of the changers and non-changers for the three time points during the functional capacity evaluation are shown in Figure 34 and 35 respectively.

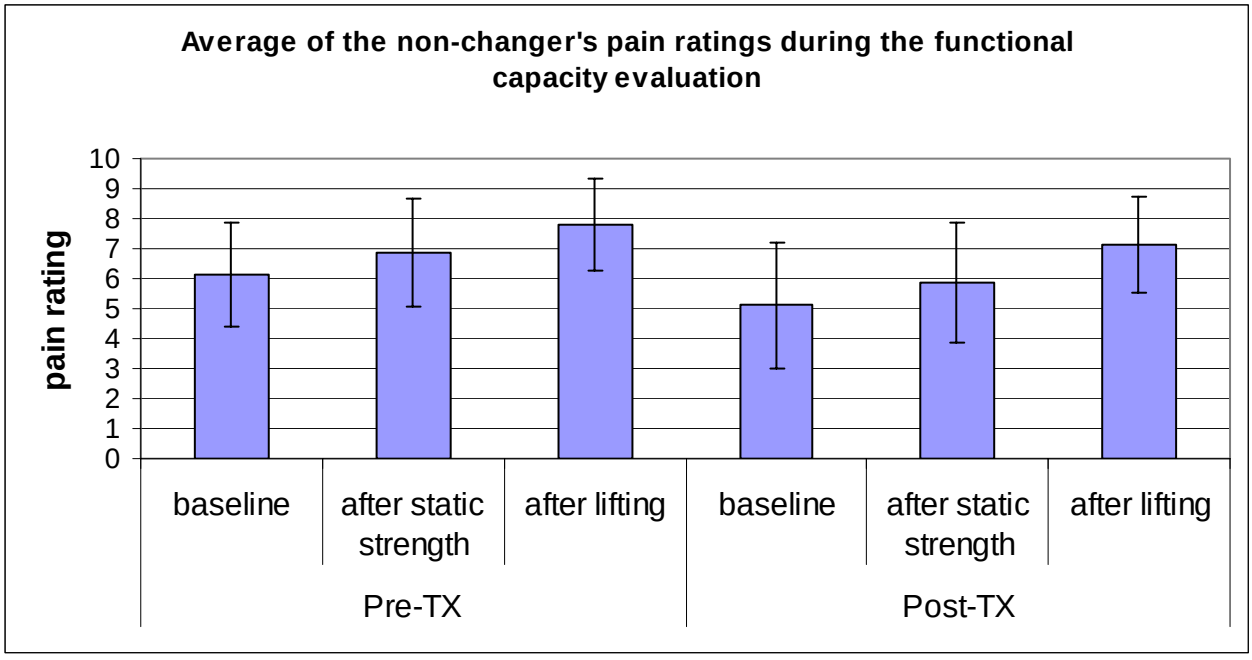


Figure 34: Average pain ratings with standard deviations error bars of the non-changers at each time point during the functional capacity evaluation

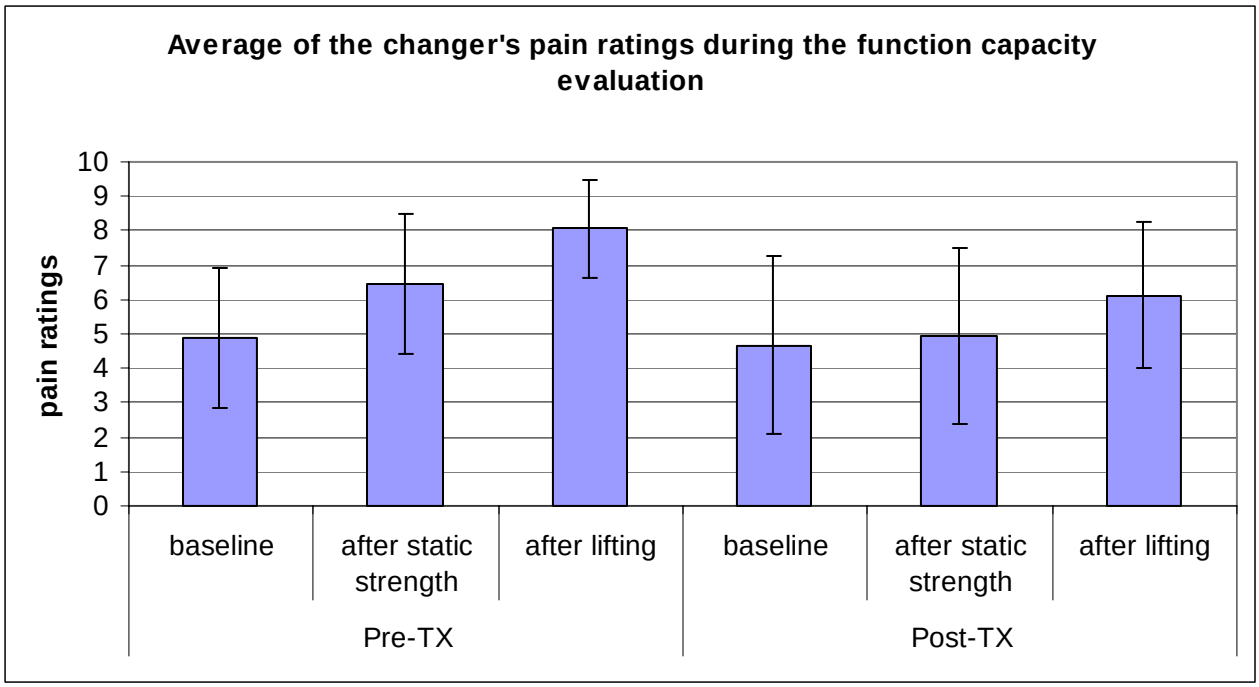


Figure 35: Average of pain ratings with standard deviation error bars of the changers at each time point during the functional capacity evaluation

6.0 DISCUSSION

A classification procedure to identify sub-groups within a population of CLBP subjects based on their lifting patterns was developed. Simulation studies have showed excellent reliability, suggesting HMMs can be applied to clinical time series data to reliably identify a group of CLBP subjects that perform the lifting task similar to control subjects. The HMMs identified 35 CLBP subjects that lifted more similar to control subjects than to the other CLBP subjects and 46 CLBP subjects that lifted differently from control subjects. The two CLBP groups were significantly different for pain intensity, pain severity, self-efficacy and the number of lifts performed during the lifting task, suggesting that CLBP population is heterogeneous and that the HMMs can successfully identify two meaningful different CLBP groups. Most of the control subjects were classified to the control HMM, signifying that the control group is relatively homogenous and the HMMs could correctly identify the lifting sequences of control subjects. A few control subjects were classified to the CLBP HMM but this group was too small to be considered as a sub-group. The analysis of the post-treatment data indicated that treatment outcomes could not be predicted from pre-treatment HMM classification. However, the HMM classification procedure was able to identify those CLBP subjects that change lifting patterns after treatment and those who did not, possibly demonstrating that the HMM can identify those CLBP patients who improve body mechanics after treatment. This result suggests that HMMs could be used as a research tool to evaluate the effectiveness of treatment protocols.

The HMMs were designed based on the results of a data reduction procedure that used factor analysis and k-means cluster analysis to reduce the multidimensional lifting parameters into lifting patterns. The lifts of each of the subjects were assigned to one of five clusters and each cluster was associated with a lifting pattern. When the clusters were examined, it appeared that the cluster solution did effectively discriminate lifting patterns of controls and CLBP subjects since a majority of the lifts contained in some of the clusters were performed by one of the groups. The cluster solution was also found to be highly reliable but could not be used to identify groups of CLBP subjects because of overlap.

The cluster solution identified five lifting patterns that provide a global description of body motion during a lift. Previously, individual lifting parameters (i.e. lift duration) have been compared between controls and CLBP subjects and inferences about the means of the parameters were made to distinguish lifting motions of control and CLBP groups [7-8,15-16]. For example, CLBP subjects as a group have been shown to perform slower lifts, use lower jerk and a more squat starting posture when compared to a group of controls. A limitation in computing averages of the parameters over the tasks is that variability can be large and differences between CLBP subjects cannot be detected due to the large variability. This same result was found in Wrigley et al., who could not differentiate between the lifting techniques of subjects that developed low back pain from those who did not when comparing the average or peak of lifting parameters [37]. The authors were able to differentiate lifting techniques when principal component analysis was applied to the displacement waveforms and the component scores were compared between the groups. The advantage of identifying the five lifting patterns in this project is that each lift can be assigned to a lifting pattern and used in the HMMs to distinguish motion between CLBP

subjects instead of relying on comparisons between individual parameter averages of a CLBP group and a control group.

HMMs were used to describe the lifting patterns of the control subjects and CLBP subjects during the lifting task. To determine the appropriate topology of these HMMs, a pruning procedure based on Vasko's *DISSOLVE* algorithm [21] was used. Comparing the methods used to construct the HMMs in this thesis to Vasko's methods, several differences are notable. These differences include data reduction, initial conditions of transition and token probabilities, starting state of the sequences and the parameters used to describe the lifting task. Despite these differences, the HMMs found in this thesis are very similar to those designed by Vasko, suggesting that the HMMs are reproducible.

Both the control HMM and the CLBP HMM were trained with a modified jackknife method to allow for the classification of subjects to a model. Typically, HMMs are trained with a training set and evaluated with a test set to determine the reliability of the HMM to model a particular time series. For these data, we had no prior knowledge of lifting patterns that separate the CLBP subjects into two groups (i.e. CLBP subjects that lifted similar to controls and those that lifted different from controls) and therefore could not separate the data into training and test sets. The modified jackknife method was used because it permitted classification of the sequences to a model without introducing bias associated with classifying a sequence that was used to train the HMM. The method excluded one sequence and trained both HMMs with the remaining sequences. The excluded sequence was then classified to either the control HMM or the control HMM. Two simulation studies were conducted and a kappa statistic was used to assess whether the HMMs can identify sequences to the appropriate model. The clinical data was simulated by intentionally mislabeling several of the simulated lifting sequences to the wrong

group (control sequence labeled as a CLBP sequences and vice versa) and using a modified jackknife method to train the HMMs. The HMMs trained with the modified jackknife approach were found to be reliable in the second simulation study and were able to identify CLBP groups when the HMMs were applied to the clinical data, indicating that the modified jackknife method is a valid technique to train HMMs.

The simulation studies found that the HMMs had excellent reliability when the lifting sequences contained more than 7 lifts and when 41% or fewer of the simulated sequences were intentionally mislabeled. When 50% of the data was mislabeled, 19 classification errors were found. This result while large was better than expected and could suggest that the length of the lifting sequences may have contributed to reliability. The length of the simulated sequences was chosen to approximately match the clinical data. On average, the simulated sequences from the CLBP HMM were shorter than the simulated sequences from the control HMM because control subjects completed more lifts than CLBP subjects during the repetitive lifting task. In addition, the number of states in the HMMs was different, with a 2-state HMM describing the CLBP data and a 3-state HMM describing the control data. The longer sequences are more likely to make more transitions through a HMM than shorter sequences and would possibly have a greater probability of being generated from a HMM with more state transitions than a HMM with fewer state transitions. Thus the shorter CLBP simulated sequences may have been biased to the 2-state CLBP HMM and the longer control simulated sequences to the 3-state control HMM in the simulation, resulting in the higher reliability.

The classification procedure identified a group of CLBP subjects that perform lifts similar to control subjects and a group of CLBP subjects that perform lifts very differently from control subjects. These groups were found to have significantly different lift token distributions.

To relate each token to a lifting pattern, the tokens were examined and a description of the lifting pattern of each token was obtained. The descriptions corresponded to the factor score(s) that had the highest mean value(s). The factor scores were interpreted by calculating the means of the lifting parameters that were contained in the particular factor score label. For example, for token 1, the synchrony factor and the speed factor had higher mean values than the other factors. The synchrony factor is composed of the midpoint difference, and the speed factor is composed of lift duration and rms jerk. The mean values of midpoint difference, rms jerk and duration of the lifts assigned to token 1 were calculated, and the description of the lifting pattern of this token is a slow, unsynchronized angle motion with the hip moving faster than the knee, and having low jerk. An unsynchronized knee and hip angle motion describes a lift for which the hip and knee angles reach the midpoint of motion at different times, resulting in a non-zero midpoint difference. A positive midpoint difference indicates that the knees are moving faster than the hip and a negative midpoint difference indicates the hips are moving faster than the knees.

The lifting pattern associated with token 2 is a more torso style starting posture with the knees moving faster than the hips. A torso lift is a lift that begins with the back bent and the knees at approximately full extension. Token 3 is a faster, high jerk lift with synchronized hip and knee motion and indicates a higher momentum lifting pattern. The lifting pattern associated with token 4 is a more squat starting posture with the knees moving faster than the hips. A squat lift is a lift that begins with the knees fully bent and the back straight. Token 5 was labeled as a lifting pattern in which the midpoint of the angles motion occurred at the same time but early in the lift time. This type of lifting pattern suggests that the subject is moving in two movements, the lower extremities move faster initially and then the upper extremities move faster later in the lift.

The CLBP subjects that were classified to the CLBP HMM frequently performed the lifting pattern associated with token 1. Since jerk is the rate change of acceleration and can be related to muscle force, the lower values of jerk for this type of lift suggests that the CLBP subject is not using all possible muscle force to lift the load but may instead be co-contracting antagonistic muscles (guarded motion) to possibly restrain motion that could produce further pain exacerbation [22]. Based on this lifting pattern, the CLBP subjects that were classified to the CLBP HMM are referred to as guarded CLBP lifters. The CLBP subjects that were classified to the control HMM are called the high performing CLBP lifters.

Although the high performing CLBP lifters were classified to the control HMM, these subjects were found to use different lifting patterns when compared to control subject classified to the control HMM. The biggest difference between the high performing CLBP lifters and the control subjects is in the tokens that describe lifting patterns related to starting posture and jerk. The high performing CLBP lifters were more likely to perform lifts that started in a more squat posture than controls. CLBP patients are often told to lift with their knees in rehabilitation programs. Since a majority of the CLBP subjects are likely to have had prior treatment before entering the study, the rehabilitation instructions could explain this lifting pattern difference. The other difference in the high jerk lifting pattern may be related to the significant difference in the amount of weight lifted during the repetitive lifting task, since controls lifted almost twice as much weight than CLBP subjects. It is reasonable to assume that the greater weight would require more muscle force. Since jerk is related to the rate of change of force, lifting a heavier weight would result in a higher jerk lifting style than lifting a lighter weight.

The guarded lifters performed significantly fewer lifts during the lifting task than high performing CLBP lifters, suggesting that number of states in the HMM may have biased the

classifications of the subjects to a particular HMM. To evaluate whether CLBP subjects could be separated into groups based on length of the sequences, a histogram of the number of lifts performed by each CLBP group was constructed and compared. The histograms revealed that although the guarded lifters performed fewer lifts than the high performing CLBP lifters, the distributions overlap. In addition to the histograms, a discriminant function analysis found that number of lifts, two measures of pain intensity and self-efficacy were all significant independent contributors to the differences between the two CLBP groups. These results indicate that the CLBP subjects could not be easily separated into groups based only on the number of lifts performed during the task.

During the functional capacity evaluation, the guarded CLBP lifters reported significantly higher levels of pain intensity than the high performing CLBP lifters. For both CLBP groups, pain intensity ratings increased from baseline to the end of the repetitive lifting task. Even though high performing CLBP lifters reported increase in pain intensity, they performed more lifts and used several different lifting patterns during the lifting task. The guarded CLBP lifters' perception of greater pain may be one of the reasons that these subjects used a constrained lifting pattern in a majority of the lifts and decided to quit the lifting task before the time limit was reached.

The guarded CLBP lifters also reported higher levels of pain severity and pain intensity and lower self-efficacy than high performing CLBP lifters, suggesting that these measures may have an impact on body motion and endurance during the lifting task. These results are consistent with previous studies that have found correlations between self-reported measures and performance on physical functioning tasks. Rudy et al. found self-efficacy expectations and perceived emotional and physical health were significant predictors of subjects' performance on a lifting

task in a sample of chronic pain patients with lower extremity amputation [85] Vlaeyen et al. found a significant covariation between lumbar muscular activity and the pain report, suggesting that the presence of pain results in tensing of the muscles in patients [86]. Lackner found that self-efficacy was a better predictor of lifting ability than measures of perceived control over pain or psychological distress [87]. Verbunt et al found an association between decreased quadriceps muscle strength of CLBP subjects and increased self-reported pain intensity and psychological distress [88].

The significant differences in the self-reported measures of pain and self-efficacy, number of lifts and lifting patterns between the two CLBP groups suggest that these measures are related. Specifically, higher pain intensity possible translates to guarded lifting style and lower endurance for some unknown reason. All subjects were asked for their best performance during the lifting task, but it is possible that some CLBP subjects choose to perform at sub-maximal level because they may have remembered a past incident that caused increased pain and were fearful of reoccurrence of that pain. This type of behavior has been labeled as fear-avoidance.

A fear-avoidance model has been constructed that attempts to describe why chronic pain develops. The model based on multiple research studies that were performed over several years [89-92], describes two pathways when an individual has an injury: fear-avoidance or confrontation. In the fear-avoidance pathway, patients experience pain that leads to catastrophizing thoughts about their pain which evolves into pain-related fear. The fear leads to increased attention to a pain threat or hyperviligance to body sensations which leads to disability, depression and deconditioning due to disuse or avoidance of activities of daily life. In the confrontation pathway, patients experience pain with the injury but do not develop pain-related fear and continue to confront daily activities, which lead to faster recovery.

Although no significant differences were found for the cognitive, disability and psychosocial domains, the effect sizes of measures within these domains were moderate suggesting that at larger sample sizes, significant differences may be found. Assuming this trend in the data, the guarded CLBP lifters would report higher disability, greater MPI Dysfunctional composite scores and more catastrophizing statements than the high performing CLBP lifters. Combining these characteristics with the higher report of pain, lower self-efficacy and constrained motion lifting pattern, the results suggest that guarded CLBP lifters demonstrate greater fear-avoidance behavior than high performing CLBP lifters and this behavior may explain why the two CLBP groups lift differently.

The self-reported measures at pre-treatment and post-treatment assessments were compared between the high performing CLBP lifters and the guarded CLBP lifters to determine whether the HMM classifications can predict which CLBP patients that will benefit most from treatment. The results showed that the group-by-treatment interaction was not significant for any of the domains or measures in the domains, indicating that it is not possible to predict the patients that will benefit most from treatment based on HMM classification.

To assess the treatment effect on the HMM classifications, the CLBP subjects that completed the lifting task after treatment were classified to either the control or CLBP HMM. The guarded CLBP lifters were labeled as changers or non-changers based on the post-treatment HMM classification. Half of the guarded lifters were labeled changers suggesting that changers alter in lifting patterns after treatment and that the HMM classification procedure can evaluate treatment effectiveness. The high performing CLBP lifters were also classified to a HMM at post-treatment and all except two of the high performing CLBP lifters were classified to the

control HMM, suggesting that high performing CLBP subject did not perform poorer lifting patterns (i.e. classified to the CLBP HMM) at the post-treatment assessment.

The non-changers used token 1 lifting pattern more frequently than any other token in the post-treatment lifting task and this is the same lifting pattern non-changers used frequently in the pre-treatment lifting task. Since no significant group difference was found between the changers and non-changers for pain intensity ratings during the functional capacity evaluation, pain is probably not the reason for lifting pattern differences. During treatment, CLBP patients received instructions about body mechanics, cognitive behavior and pain management. It appears that the changers may have developed skills to decrease fear-avoidance, increased coping and increased pain management techniques. Further evidence to this idea is the significant difference between the changers and non-changers in the spinal mobility domain. Changers were found to have a greater range of flexion/extension motion which may have translated to greater flexibility and less stress on the lumbar back when performing the lifting task, possibly enabling the changers to perform lifting patterns that are more similar to controls than to the lifting patterns of non-changers.

The effect sizes were large to medium for coping strategies, MPI Dysfunctional composite score, pain intensity and pain severity when comparing the post-treatment assessment of the changers and non-changers. In addition, in post-treatment, the non-changers used the guarded, more constrained lifting style associated with token 1. These results indicate that the non-changers are still demonstrating fear-avoidance behavior after treatment, suggesting that treatment was not as effective for these patients. Some reasons for the differential treatment outcomes may be the short length of the treatment program, the unwillingness of the subjects to change pain behavior or the learning behavior of the individuals. The treatment program was 3.5

weeks and involved verbal instructions about body mechanics and cognitive behavior. It is possible that the non-changers may need a longer treatment programs, learn through visual stimulation or are not willing to alter their perception of pain. Future studies that evaluate the effectiveness of longer treatment programs and different treatment modalities such as biofeedback [93] may lead to better treatment options for these patients.

CLBP subjects were assigned to a HMM based on their lifting patterns perform during a repetitive lifting task. Five lifting patterns were found with a data reduction procedure and of these lifting patterns, two were related to jerk. In this project, jerk was calculated with a hyperbolic tangent model to obtain smooth estimates and the parameters of maximum jerk, rms jerk and time at maximum jerk were determined from the estimates. Previously, jerk was calculated with a hepatic spline and a rms measure of jerk was compared between a control group and a CLBP group [15, 25]. The results showed that control subjects and CLBP subjects perform lifts with significantly different patterns of jerk. The hyperbolic tangent model was used as the smoothing method in this project because we were interested in characterizing maximum jerk. It was not possible to characterize maximum jerk with the spline estimates of jerk due to the variability in the waveforms. A limitation of using jerk to describe motion in this project is that it is not clear whether the estimates from the either the hyperbolic tangent model or spline smoothing method are truly measuring jerk. Since jerk was calculated as the third derivative of position, it was a noisy measure that could only be estimated using smoothing methods. Future studies may focus on validating the estimates of jerk from the smoothing methods possibly by comparing jerk calculated with accelerometers and jerk calculated from the displacement markers using the different smoothing methods.

A limitation of the HMM classification procedure is that the reliability of the HMMs must be estimated with simulation studies. The objective of the study is to reliably identify the group of CLBP subjects who performed lifting patterns that more similar to the lifting pattern of control subjects than other CLBP subject within a sample using HMMs. A simulation study was designed to assess how reliably the HMMs can classify the simulated sequences to the correct HMM when a percentage of the sequences are intentionally mislabeled to the wrong group and the HMMs are trained with the mislabeled sequences. The reliability results of the simulation study are only estimates of HMM reliability because the high performing CLBP lifters are represented as control sequences in the simulation, while in the clinical data the CLBP subjects are not control subjects. The control sequences were used as approximations because at the start of the project, it was impossible to know which of the CLBP subjects were high performing CLBP lifters and which were the guarded CLBP lifters.

Another limitation of the study is the sample size especially in the analysis of the treatment data. Several of the self-reported measures had moderate effect sizes indicating potential for significant differences in larger samples sizes. A power analysis was performed and showed that approximately 218 subjects (95 high performing CLBP lifters and 123 guarded CLBP lifters) would be needed to identify significant differences between guarded CLBP lifters and high performing CLBP lifter on the measures of coping strategies, MPI Dysfunctional composite score, functional status from Jan van Breemen and walking speed (Appendix C). Future studies may focus on using the HMM classification procedure to determine whether these variables are different between the two CLBP groups at larger sample sizes.

Other systems to identify groups of CLBP subjects that were based on clinical examination and self-reported measures have been described. O'Sullivan identified groups

within a CLBP patient population based on clinical examinations and found differences in posture and trunk muscle activation during a seated task [5]. Turk and Rudy identified groups based on responses to the MPI [3], and Dunn et al. identified CLBP groups based on self-reported pain intensity, disability and psychosocial measures [6]. The advantage of using the HMMs to identify the CLBP groups is that the classifications are not dependent on the experience of the clinician or the subjectivity of the CLBP patient's self-report as is found with the other classification systems. The HMMs also provide a method to incorporate temporal patterns into the classification strategy.

One of the motivating factors of this project was from the observations of an experienced occupational therapist, who described differences in the lifting motion of CLBP subjects during a lifting task that were not discernable to the untrained or inexperienced observer. Specifically, the clinician observed during the clinical research study that some of the CLBP subjects lifted like control subjects and others lifted differently from controls. The HMM classification procedure was able to identify these CLBP groups. Since few tools are available to quantify the observations of experienced clinicians, the results of this study suggest that the HMM classification procedure is a useful research tool for validating the clinician observations, identifying CLBP groups and evaluating the effectiveness of treatment.

This study provides a classification method for identifying groups within a CLBP population based on time series data and can be easily adapted to classify patients from other clinical populations besides CLBP. One reason for developing a classification method is that groups of chronic pain patients have shown differential response to standardize treatment protocol when separated into groups based on responses to the MPI. In identifying difference between sub-groups, more diverse treatment options can become available to patients that could

possibly increase recovery time and treatment effectiveness. The classification method used HMM to classify subjects based on their time series data. An advantage of the HMMs is that information over duration of the performance or measures of the disease over time can be used to classify patient instead of discrete or short-term clinical parameters

7.0 CONCLUSION

Patients diagnosed with CLBP are not a homogenous group and different treatments may be beneficial to different groups of patients. The HMM classification procedure described in this thesis provides a technique to identify sub-groups within a sample of CLBP subjects based on time series data. Simulation studies demonstrated the reliability of the classification method. In this study, a sub-group of CLBP patients that performed a repetitive lifting task more like control subjects than like other CLBP subjects was identified. The sub-group performed more lifts than the other CLBP subjects, reported lower pain intensity at the before and after the repetitive lifting task, lower levels of pain intensity during activities of daily life and higher levels of self-efficacy. Treatment outcomes could not be predicted based on the pre-treatment HMM classifications of the CLBP subjects. However, half of the CLBP subjects that were classified to the CLBP HMM at pre-treatment assessment were classified to the control HMM at post-treatment and the other half were assigned to the CLBP HMM. Almost all of the CLBP subjects classified to the control HMM at pre-treatment were classified to the control HMM at post-treatment, indicating that the HMM classification procedure is a useful research tool to measure treatment effectiveness.

APPENDIX A

3-STATE CLBP HMM

To determine whether the 3-state CLBP HMM or 2-state CLBP HMM was the appropriate HMM to apply to the clinical data, the reliability of both models was assessed in the second simulation. The 2-state CLBP HMM and 3-state CLBP HMM were compared to the 3-state control HMM to determine whether these models could reliably identify mislabeled simulated sequences to the correct HMM. The results of the second simulation comparing the 2-state CLBP HMM and 3-state control HMM found that the models could reliable identify 41% or fewer mislabeled sequences to the correct model, as shown in Figure 24 (Chapter 4). The results of the second simulation study assessing the reliability of the 3-state control HMM and the 3-state CLBP HMM showed that the HMMs could reliably identify 22% or fewer of the mislabeled simulated sequences to the correct model (Figure 36). Based on these results, the 2-state CLBP HMM was chosen as the appropriate HMM to describe the CLBP time series data.

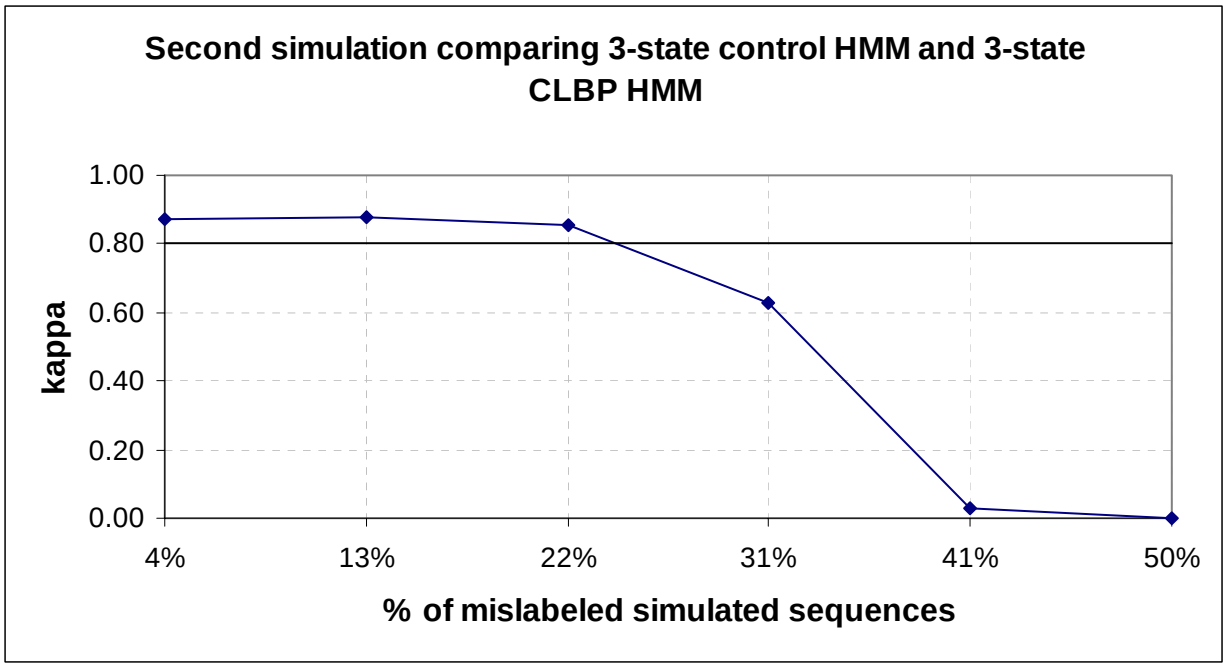


Figure 36: Results of the second simulation study assessing the reliability of the 3-state CLBP HMM and the 3-state control HMM. HMMs were considered reliable if kappa was > 0.8.

APPENDIX B

DESIGN OF TWO CLBP HMMS

Since two groups of CLBP subjects (guarded CLBP lifters and high performing CLBP lifters) were found, HMMs were designed to determine whether it is more appropriate to classify sub-groups of the CLBP subjects with two CLBP HMMs or with a control HMM and a CLBP HMM. This section describes the design of the guarded CLBP lifters (GL) HMM and the high performing CLBP lifters (HP) HMM, simulation studies to determine reliability, and results of the HMMs when applied to the clinical data.

B.1 DESIGN

The two CLBP HMMs were designed according to the methods described in Chapter 3. Briefly, a fully-connected 4-state, 3-state and 2-state HMMs were trained for the GL and the HP CLBP groups. A pruning procedure was applied to the each of the HMMs to reduce the topology of the models. The likelihood probability, K-L measure and the entropy of the token distribution were used to identify the simplest topology of the 4-state, 3-state and 2-states GL HMMs, and the simplest topology of the 4-state, 3-state and 2-states HP HMMs. The Viterbi algorithm was applied to the resulting HMMs, and the frequency that each state was occupied was calculated.

The HMM that contained states that were frequently occupied by the subject's sequences was chosen.

B.2 SIMULATION STUDIES

In order to determine how reliably the GL HMM and HP HMM can identify mislabeled sequences to the correct model, an experiment similar to the second simulation study was performed. In this simulation, simulated sequences were switched between the GL HMM and the HP HMM and between the HP HMM and the control HMM. A modified jackknife method was used to train HMMs and the logarithm of likelihood probability determined the classification of the simulated sequence. The sample size of the simulated sequences was chosen to match the clinical data with 35 simulated sequences generated from the HP HMM, 46 simulated sequences generated from the GL HMM and 51 simulated sequences generated from the control HMM. The methods of the second simulation described in Chapter 4 were used.

.

B.3 RESULTS

The pruning algorithm indicated that the simplest topologies of the 4-state, 3-state and 2-state LP and HP HMMs were temporal HMMs. The frequency calculation, based on the results of the Viterbi algorithm, showed that the 2-state HMM was the appropriate HMM to describe the GL time series data as shown in Table 23. For the HP data, it is not clear whether the 2-state HMM or the 3-state HMM was the appropriate model since 14% of the subjects occupied the last state

in the 3-state HMM. Both models were tested in the simulation studies and the HMM with the higher reliability was chosen to describe the HP time series data. The parameters of the 2-state CLBP guarded lifters HMM and the 2-state CLBP high performing lifters HMM are shown in Table 24.

Table 23: Frequency that the states were occupied for the 4-state, 3-state and 2-state GL HMMs and the 4-state, 3-state and 2-state HP HMMs

	4-state HMM				3-state HMM			2-state HMM	
# of states in most likely state path	1	2	3	4	1	2	3	1	2
GL group	63%	23%	14%	0%	70%	24%	6%	68%	32%
HP group	37%	46%	17%	0%	54%	31%	14%	60%	40%

Table 24: Parameters of the guarded CLBP lifters HMM and the high performing CLBP lifters HMM

Parameters of the trained Guarded CLBP lifters HMM					
Guarded HMM Transition Probability					
	Transition to state 1			Transition to state 2	
In state 1	0.9619			0.0381	
In state 2	0.0000			1.0000	
Guarded HMM Token Probability					
	Token 1	Token 2	Token 3	Token 4	Token 5
State 1	0.9670	0.0000	0.0000	0.0293	0.0037
State 2	0.4306	0.0839	0.0989	0.2694	0.1172
Parameters of the trained High Performing CLBP lifters HMM					
High Performing HMM Transition Probability					
	Transition to state 1			Transition to state 2	
In state 1	0.9590			0.0410	
In state 2	0.0000			1.0000	
High Performing HMM Token Probability					
	Token 1	Token 2	Token 3	Token 4	Token 5
State 1	0.0869	0.0000	0.0338	0.8378	0.0415
State 2	0.0119	0.3225	0.2485	0.0050	0.4121

In the simulation, simulated sequences were equally mislabeled between the CLBP groups (i.e. one sequence from the HP group was labeled as a GL sequence and one sequence from the HP group was labeled as a GL sequence) from 4 to 36 (5% to 44% of the total sample) in increments of 2. The maximum percentage of mislabeled sequences was 44% of the total sample (n = 81) because at this percentage, 50% of the HP data (n = 35) was incorrectly labeled to the GL group. For the comparison between the control HMM and the HP HMM, the maximum percentage of mislabeled sequences was 40% of the total sample.

The second simulation results comparing the GL HMM and the 2-state HP HMM showed that these models were reliable when 34% or less of the total sample was mislabeled (Figure 37). This same result was found when comparing the 3-state HP HMM and the GL HMM as shown in Figure 38. The 2-state HP HMM and control HMM were reliable when 22% or less of the total sample was mislabeled as shown in Figure 39. The results of the second simulation comparing 3-state HP HMM and the control HMM showed that the models were reliable when 9% or less of the total sample was mislabeled (Figure 40). Based on these results, a 2-state HP HMM was chosen since this model was more reliable. Both the 2-state GL HMM and the 2-state HP HMM were applied to the clinical data.

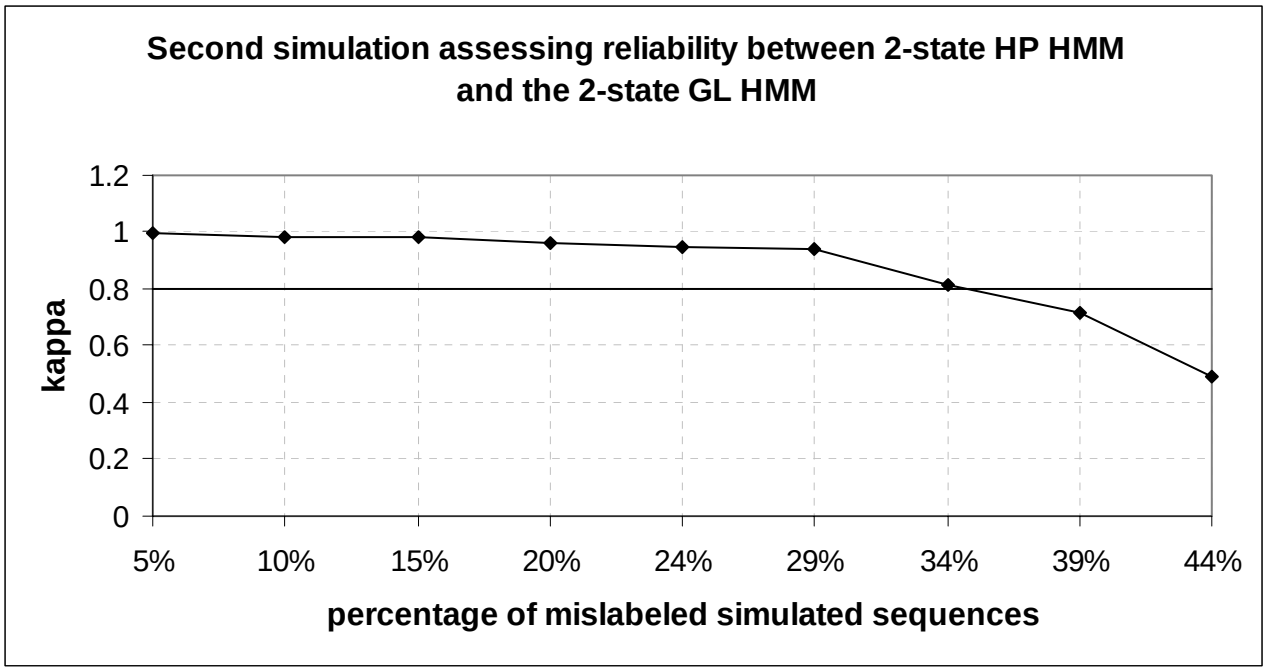


Figure 37: Kappa plotted for the percentage of mislabeled simulated sequences comparing the 2-state HP HMM and the 2-state GL HMM. HMMs were considered reliable if kappa was > 0.8.

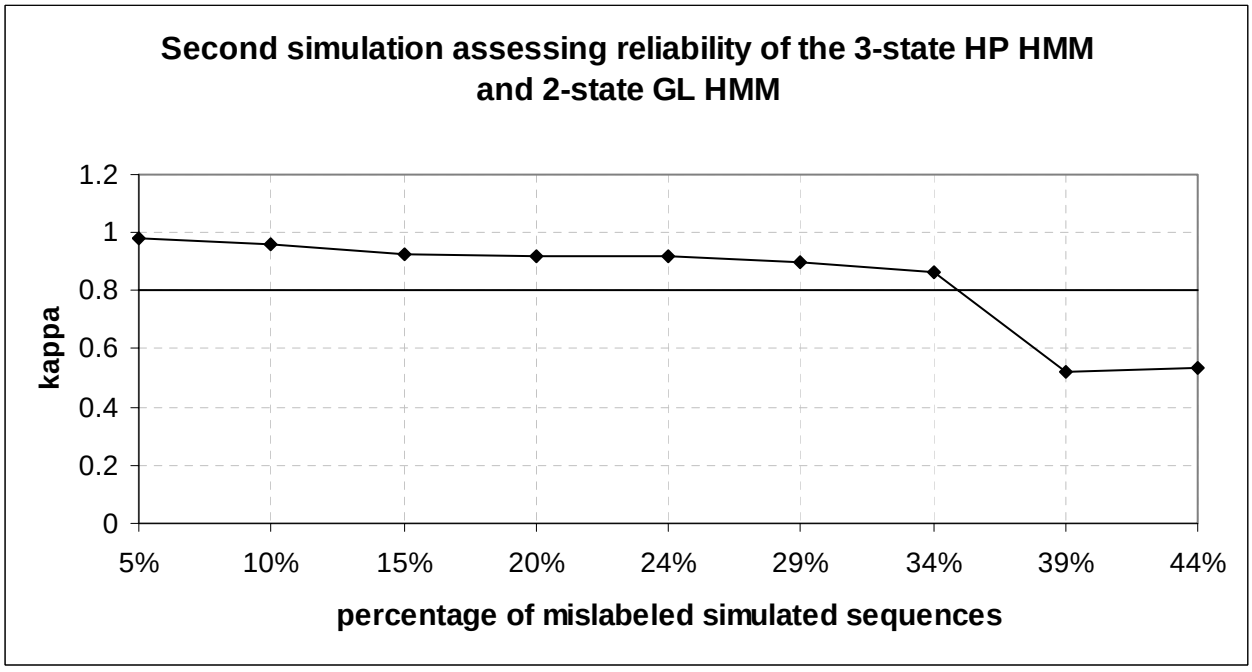


Figure 38: Kappa plotted for the percentage of mislabeled simulated sequences comparing the 3-state HP HMM and the 2-state GL HMM. HMMs were considered reliable if kappa was > 0.8.

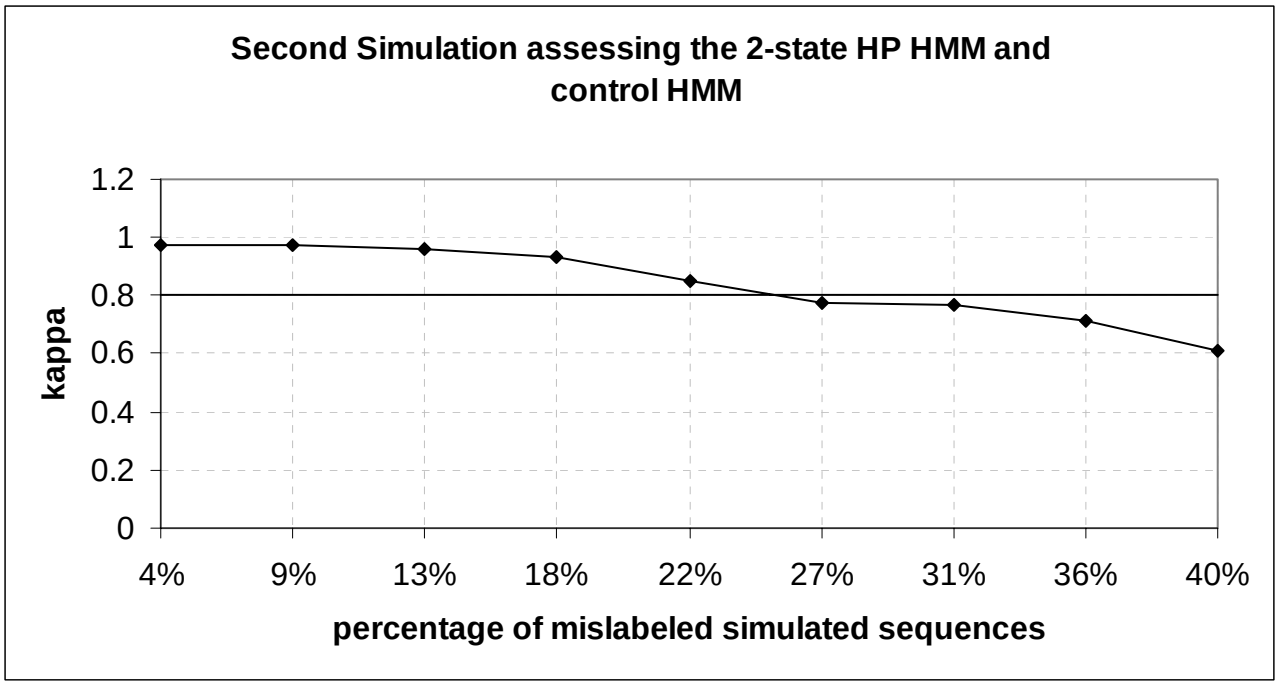


Figure 39: Kappa plotted for the percentage of mislabeled simulated sequences comparing the 2-state HP HMM and the control HMM. HMMs were considered reliable if kappa was > 0.8.

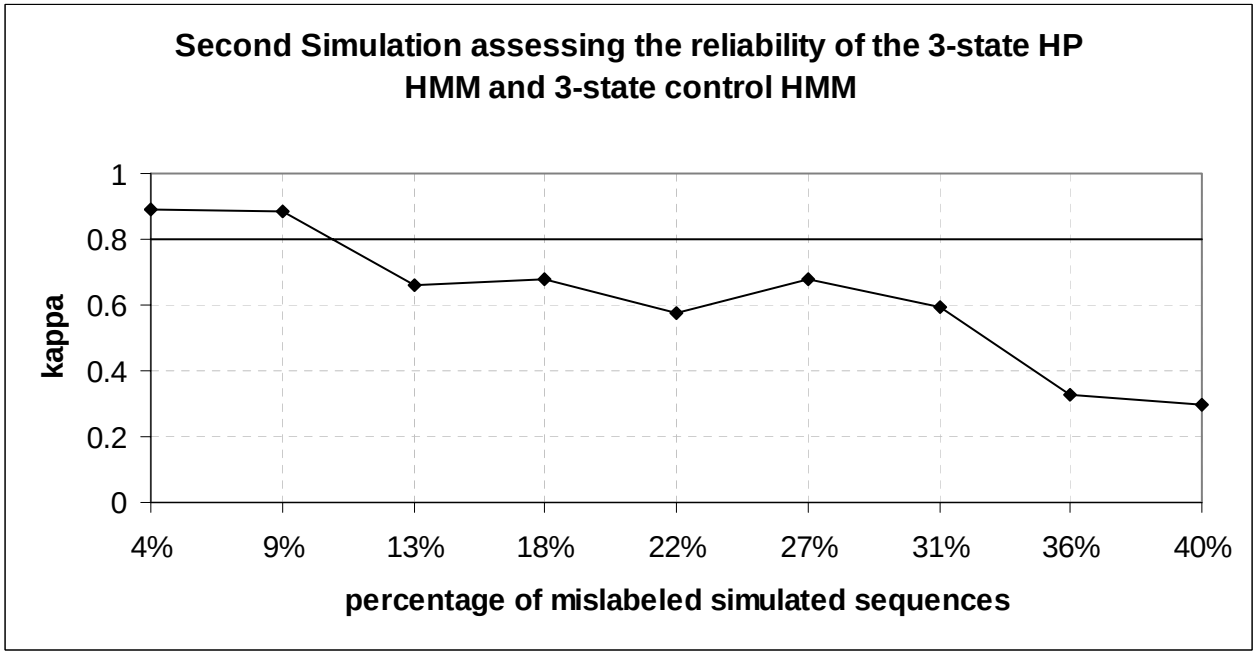


Figure 40: Kappa plotted for the percentage of mislabeled simulated sequences comparing the 3-state HP HMM and the control HMM. HMMs were considered reliable if kappa was > 0.8.

B.4 APPLICATION TO THE CLINICAL DATA

The subjects classified to the GL group were used to train the GL HMM and the subjects in the HP group were used to train the HP HMM. The CLBP subject's lifting sequences were tested against the GL HMM, HP HMM and control HMM using a modified jackknife method and classified to one of the three HMM based on the likelihood probability. The control HMM was trained with all of the control data and the control data were not tested.

When the subjects in the GL group were classified to a HMM, 40 subjects were classified to the GL HMM, 3 subjects were classified to the HP HMM and 3 subjects were classified to the control HMM. In the HP group, 10 subjects were classified to the HP HMM, 2 were classified to the GL HMM and 23 subjects were classified to the control HMM.

Given the small number of subjects in the HP group that were classified to the HP HMM, this model was considered unreliable to classify subjects to a group. The results suggest that lifting pattern distribution of HP subjects is more similar to the lifting patterns distribution of the control subjects. Based on these results, the control HMM and CLBP HMM were used to classify the lifting sequences of the CLBP subjects at pre-treatment and post-treatment.

APPENDIX C

POWER ANALYSIS

Power analysis was performed on the measures that showed moderate effects when comparing the high performing CLBP lifters and the guarded CLBP lifters to determine the sample sizes that could possibly produce significant p-values. Moderate effects were found for the measures of coping strategies, MPI Dysfunctional composite scores, walking speed, walking speed and functional status from the Jan van Breemen. Since more CLBP subjects were classified to the CLBP HMM than to the control HMM, the total sample size needed when comparing the means of two normally distributed samples of unequal sample size was calculated [94]. Based on the number of subjects that were found in each CLBP group when sample size was 81, the power analysis was calculated with the assumption that CLBP subjects classified to the CLBP HMM would be 1.3 times larger than the CLBP subjects classified to the control HMM. The sample size of the group of CLBP subjects classified to the CLBP HMM was calculated using the equation [94] on the left and was denoted as n_1. The sample size of the CLBP subjects classified to the control HMM was calculated using the equation [94] on the right and denoted as n_2.

$$n_1 = \frac{\left(\sigma_1^2 + \frac{\sigma_2^2}{k}\right)\left(Z_{1-\alpha/2} + Z_{1-\beta}\right)^2}{|\mu_2 - \mu_1|^2} \qquad n_2 = \frac{\left(k\sigma_1^2 + \sigma_2^2\right)\left(Z_{1-\alpha/2} + Z_{1-\beta}\right)^2}{|\mu_2 - \mu_1|^2}$$

In both equations, k is 1.3 or the projected ratio of the two samples (k = 46/35), σ represented standard deviation, μ represented the mean values and the values of $Z_{1-\alpha/2}$ was 1.96 and $Z_{1-\beta}$ was 0.84. These Z score values represent a p-value of 0.05 as a significant level and a power level of 80%.

The results of the sample size calculations of the two CLBP groups for the measures of coping strategies, MPI Dysfunctional composite scores, walking speed, and functional status are shown in Table 25. The maximum sample size to detect significant differences ranges from 106 (60 guarded CLBP lifters and 46 high performing CLBP lifters) of the Dysfunctional composite score to 217 (123 guarded CLBP lifters and 94 high performing CLBP lifters) for the coping strategies (anxiety).

Table 25: The projected sample sizes of the high performing CLBP lifter and guarded CLBP lifters and the average (standard deviation) of the variables with moderate effect sizes are listed.

	Guarded CLBP lifters N = 46	High performing CLBP lifters N = 35	Projected Sample size of guarded CLBP lifters	Projected Sample size of high performing CLBP lifters	Total sample size
Coping strategies: emotional	3.85 (1.31)	3.21 (1.32)	76	57	135
Coping strategies: anxiety	4.62 (1.18)	4.13 (1.35)	123	94	217
MPI Dysfunctional composite score	62.60 (10.30)	57.27 (9.24)	60	46	106
Jan van Breemen: functional status	3.91 (1.15)	4.50 (1.62)	107	82	189
Jan van Breemen: walking speed	41.63 (15.89)	35.29 (7.95)	65	50	115

BIBLIOGRAPHY

1. Delitto A., Erhard R.E., Bowling R. W., DeRosa C. P., and Greathouse D.G. (1994). A treatment-based classification approach to low back syndrome: identifying and staging patients for conservative treatment. *Physical Therapy*, 75 (6), 470-489.

2. Fritz J.M., Delitto A., and Erhard R.E. (2003) Comparison of classification-based physical therapy with therapy based on clinical practice guidelines for patients with acute low back pain. *Spine,* 28(13),1363-1372.

3. Turk D.C. and Rudy T.E. (1987). Towards a comprehensive assessment of chronic pain patients. *Behavioral Research Therapy*, 25 (4), 237-249.

4. Turk D.C. and Rudy T.E. (1988). Toward an empirically derived taxonomy of chronic pain patients: Integration of psychological assessment of data. *Journal of Consulting and Clinical Psychology,* 56, 233-238.

5. O'Sullivan P.B. (2004) Clinical instability of the lumbar spine: its pathological basis, diagnosis and conservative management. In: Boyling JD, Jull G (eds.) *Grieve's Modern Manual Therapy* (pp.311-331) 3rd edition. Philadelphia, PA: Elsevier.

6. Dunn K.M., Jordan K. and Croft P.R. (2006). Characterizing the course of low back pain: A latent class analysis. *American Journal of Epidemiology,*163 (8), 754-761.

7. Sparto P.J., Parnianpour M., Reinsel T.E., Simon S. (1997). The effect of fatigue on multijoint kinematics coordination, and postural stability during a repetitive lifting test. *Journal of Orthopaedic and Sports Physical Therapy*, 25(1), 3-12.

8. Bonato P, Boissy P, Corce U.D, and Roy S.H. (2002) Changes in the Surface EMG signal and the biomechanics of motion during a repetitive lifting task. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 10 (1), 38-47.

9. Scholtz J.P. (1992) Low back injury and manual lifting: Review and new perspective. *Physical Therapy Practice,*1 (3), 20-31.

10. Dolan P.A. and Adams M. A. (1998). Repetitive lifting tasks fatigue the back muscles and increase the bending moment acting on the lumbar spine. *Journal of Biomechanics,* 31, 713-721.

11. McIntyre D.R., Glover L.H., Conino M.C., Seeds R.H., and Levene J.A. (1991). A comparison of the characteristics of preferred low-back motion of normal subjects and low-back pain patients. *Journal of Spinal Disorders*, 4(1), 90-95.

12. Bush-Joseph C., Schipplein O., Andersson G.B.J, and Andriacchi T.P. (1998). Influences of dynamic factors on the lumbar spine moment in lifting. *Ergonomics*, 31(2), 211-216.

13. Oddsson L.I.E and DeLuca C. J. (2003). Activation imbalances in lumbar spine muscles in the presence of chronic low back pain. *Journal of Applied Physiology*, 94, 1410-1420.

14. Boston J.R., Rudy T.E., Mercer S.R., and Kubinski J.A. (1993). A measure of body coordination during repetitive dynamic lifting. *IEEE Transactions on Rehabilitation Engineering*, 1(3), 137-144.

15. Slaboda J.C., Boston J.R., Rudy T.E., Lieber S.J. and Rasetshwane D.M. (2005). The use of splines to calculate jerk for a repetitive lifting task involving chronic lower back patients. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 13(3), 406-414.

16. Rudy T.E, Boston J.R., Lieber S.J., Kubinski J.A. and Stacey B.R. (2003). Body motion during repetitive isodynamic lifting: a comparative study of normal subjects and low-back pain patients. *Pain*, 105, 319-326.

17. Larivière C., Gagnon D. and Loisel P. (2000). The effect of load on coordination of the trunk for subjects with and without chronic low back pain during flexion-extension and lateral bending. *Clinical Biomechanics*, 15, 407-416.

18. Marras W.S., Davis K.G, Ferguson S.A, Lucas B. R., and Gupta P. (2001). Spine loading characteristics of patients with low back pain compared with asymptomatic individuals. *Spine,* 26(23), 2566-2574.

19. Kankaanpaa M, Taimela S, Laaksonen D, Hanninen O, and Airaksinen O. (1998). Back and Hip Extensor fatigability in chronic low back pain patients and controls. *Archives of Physical medicine and Rehabilitation*, 79, 412-417.

20. Vasko R.C., El-Jaroudi A. and Boston J.R. (1997). Application of hidden Markov model topology estimation to repetitive lifting data. *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 5, 4073-4076.

21. Vasko, R. Hidden (1997). *Markov model topology inference: The DISSOLVE algorithm*. Doctoral Thesis, University of Pittsburgh, PA.

22. Simmons D.G. and Statham-Simons L. (1989). Chronic mysofascial pain syndrome. In: Tollison C.D. (ed) *Handbook of chronic pain management* (pp. 509-529), Baltimore, MD: Williams and Wilkens.

23. Boston J.R., Rudy T.E., Lieber S.J. and Stacey B.R. (1995). Measuring treatment effects on repetitive lifting for patients with chronic low back pain: Speed, style, and coordination. *Journal of Spinal Disorders*, 8(5), 342-351.

24. Vasko R.C., El-Jaroudi A., and Boston J.R. (1996). An algorithm to determine hidden Markov model topology. *IEEE Proceedings of the International Conference of Acoustic, Speech and Signal Processing*, 6, 3578-3582.

25. Slaboda J.C. (2004). Application of Jerk Analysis to a Repetitive lifting Tasks in Patients with Chronic Low Back Pain, Master's thesis, University of Pittsburgh, PA

26. Flash T. and Hogan N. (1985). The coordination of arm movements: An experimental confirmed mathematical model. *Journal of Neuroscience*, 5(7), 1688-1703.

27. Woltring H.J. (1985). On optimal smoothing and derivative estimation from noisy displacement data in biomechanics. *Human Movement Science*, 4, 229-245.

28. Wells R.P. and Winter D.A. (1980). Assessment of signal and noise in the kinematics of normal, pathological and sporting gaits. *London: Human Locomotion I*, 92-93.

29. Winter D.A. (1990). *Biomechanics and motor control of human movement*. Toronto, Ontario: Wiley-Interscience Inc., 2nd edition.

30. Turk D.C. (2005). The potential of treatment matching for subgroups of patients with chronic pain. *Clinical Journal of Pain,* 21(1), 44-55.

31. Rudy T.E., Turk D.C., Kubinski J.A., and Zaki H.S. (1995). Differential treatment response of TMD patients as a function of psychological characteristics. *Pain*, 61,103-112.

32. Denison, E., Asenlof P., Sandborgh M., and Lindberg P. (2007). Musculoskeletal pain in primary health care: Sub-groups based on pain intensity, disability, self-efficacy, and fear-avoidance variables. *The Journal of Pain*, 8(1), 67-74.

33. Dankaerts W., O'Sullivan, Burnett A., and Straker L. (2006). Differences in sitting postures are associated with nonspecific chronic low back pain disorders when patients are subclassified. *Spine,* 31(6), 698-704.

34. Dankaerts W., O'Sullivan, Burnett A., and Straker L. (2006). Altered patterns of superficial trunk muscle activation during sitting in nonspecific chronic low back pain patients. *Spine,* 31(17), 2017-2023.

35. Brennan G.P., Fritz J.M., Hunter S.J., Thackeray A., Delitto A., and Erhard R.E. (2006). Identifying subgroups of patients with acute/subacute nonspecific low back pain. *Spine*, 31(6), 623-631.

36. Fairbank JCT, Mbaot JC, Davis JB, O'Brien JP. (1980). The Oswestry low back pain disability questionnaire. *Psychotherapy*, 66, 271-273.

37. Wrigley A. T., Albert W. J., Deluzio K.J, and Stevenson J.M. (2005). Differentiating lifting techniques between those who develop low back pain and those who do not. *Clinical Biomechanics,* 20, 254-263.

38. Bishop J.B., Szpalski M., Ananthraman S.K., McIntrye S.K., and Pope M. (1997). Classification of low back pain from dynamic motion characteristics using an artificial neural network. *Spine*, 22(24), 2991-2998.

39. Spitzer W.O., Leblanc F.E., and Dupuis M. (1987). Scientific approaches to the assessment and management of activity related spinal disorders. *Spine*, 12, S1-S59.

40. Rabiner L.R. (1973) A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77 (2), 257-286.

41. Rabiner L.R. and Juang B.H. (1986). An introduction to hidden Markov models. IEEE ASSP Magazine, 4-16.

42. Ephraim Y and Merhav N (2002). Hidden Markov processes. *IEEE Transactions on Information Theory*, 48, 1518-1569.

43. Krogh A., Brown M., Mian I.S., Sjölander K, and Haussler D. (1994). Hidden Markov Models in Computational Biology: Applications to protein modeling. *Journal of Molecular Biology*, 235, 1501-1531.

44. Durbin R., Eddy S., Krogh A., and Mitchison G. (1998) *Biological sequence analysis: probabilistic model of protein and nucleic acids*. Cambridge, UK: Cambridge University Press.

45. Franke A., Caelli T, and Hudson R.J. (2004). Analysis of movements and behavior of caribou (Rangifer tarandus) using hidden Markov models. *Ecological Modeling*, 173, 259-270.

46. Visser I, Raijmakers EJ, and Molenaar (2002). Fitting hidden Markov models to psychological data. *Scientific Programming*, 10, 185-199.

47. Kale A., Sundaresan A., Rajagopalan N., Cuntoor N.P., Roy-Chowdhury A.K., Kruger V., and Chellappa R. (2004). Identification of humans using gait. *IEEE Transactions on image processing*, 13 (9), 1163-1173.

48. Scheffer C., Engelbrecht H., and Heyne P.S. (2005). A comparative evaluation of neural networks and hidden Markov models for monitoring turning tool wear. *Neural Computing and Applications*, 14, 325-336.

49. Yu F, Morgenstern H., Hurwitz E. and Berlin T.R. (2003). Use of a Markov transition model to analyse longitudinal low-back pain data. *Statistical Methods in Medical Research,* 12, 321-331.

50. Pober D M, Staudenmayer J., Raphael C, and Freedson, P.S. (2006). Development of Novel Techniques to classify physical activity mode using accelerometers. *Medicine and Science in Sports and Exercise,* 38(9), 1626-1634.

51. Baum L.E., Petrie T, Soules G., and Weiss N. (1970) A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. *Annuals of Mathematical Statistics*, 41, 164-171.

52. Wong S., Gardner A.B., Kreiger A.M., and Litt B. (2006). A stochastic framework for evaluating seizure algorithms using hidden Markov models. *Journal of Neurophysiology*, doi:10.1152/jn.00190.2006

53. Bair V., Baumert M., Caminal P., Vallverdu M., Faber R., and Voss A. (2006). Hidden Markov models based on symbolic dynamics for statistical modeling of cardiovascular control in hypertensive pregnancy disorders. *IEEE Transactions on Biomedical Engineering*, 35 (1), 140-143.

54. Cooper B. and Lipsitch M. (2004). The analysis of hospital infection data using hidden Markov models. *Biostatistics*, 5(2), 223-237.

55. Milligan G.W. and Cooper M.C., (1980). An examination of procedures for determining the number of clusters in a data set. Psychometrika, 50 (2), 159-179.

56. Milligan G.W. and Cooper M.C. (1987). Methodology Review: Clustering Methods Applied Psychological Measurement, 11(4), 329-354.

57. Aldenderfer M.S. and Blashfield R.K. (1984). *Cluster Analysis.* Newbury Park, CA: Sage Publications.

58. Morey L. and Agresti A., (1984). The measurement of classification agreement: An adjustment to the Rand statistic for chance agreement. Educational and Psychological Measurement, 44, 33-37.

59. Milligan G. (1981b). A review of the Monte Carlo tests of cluster analysis. *Multivariate Behavioral Research,* 16, 379-407.

60. Calinski R. B. and Harabasz J. (1974). A dendrite method for cluster analysis. *Communications in Statistics,* 3, 1-27.

61. Duda R.O. and Hart P.E. (1973) *Pattern classification and scene analysis.* Wiley: New York.

62. McIntyre R.M. and Blashfield R.K., (1980) A nearest-centroid technique for evaluating the minimum-variance clustering procedure. *Multivariate Behavioral Research*, 2, 225-238.

63. McGinn T., Wyer P.C., Newman T.B., Keitz S., Leipzig R. and Guyatt G. (2004) Tips for learners of evidence-based medicine: Measures of observer variability (kappa statistic). *Canadian Medical Association Journal*, 171 (11), 1369-1373.

64. Rudy, T. E.; Turk, D. C.; Brena, S. F.; Stieg, R. L., and Brody, M. C. (1990) Quantification of biomedical findings of chronic pain patients: Development of an index of pathology. *Pain*, 42, 167-82.

65. Turk, D. C.; Wack, J. T., and Kerns, R. D. (1985). An empirical examination of the "pain behavior" construct. *Journal of Behavioral Medicine*, 8, 119-130.

66. Lorig K., Chastain RL, Ung E, Shoor S, Holman HR. (1989). Development and evaluation of a scale to measure perceived self-efficacy in people with arthritis. *Arthritis Rheumatology*, 32, 37-44.

67. Rosenstiel AK and Keefe FJ. (1983). The use of coping strategies in chronic low back pain. *Pain*, 17, 33-44.

68. Rudy T.E. (2005) Multidimensional Pain Inventory, Version 3.0 User's guide, University of Pittsburgh, PA, website: www.pain.pitt.edu/mpi.

69. Bandura A. (1977). Self-efficacy: towards a unifying theory of behavior change. *Psychological Review*, 84, 191-215.

70. Lankhorst GJ, Van de Stadt RJ, Vodeelaar TW, Van der Korst JK and Prevo AJH. (1982) Objectivity and repeatability of measurements in low back pain. *Scandinavian Journal of Rehabilitation Medicine*, 14, 21-26.

71. Roland M and Morris R., (1983). A study of the natural history of back pain-Part I: Development of a reliable and sensitive measure of disability in low-back pain. *Spine*, 8, 141-144.

72. Kuzkaya, C. (1998). *Multi segmental human body modeling using robotic techniques to characterize lifting styles and dynamics*. Master thesis, University of Pittsburgh, PA

73. SYSTAT. SYSTAT for Windows: Statistics, version 11, Chicago, IL: SPSS, 2004.

74. SAS OnlineDoc®, Version 8, Cary, NC: SAS Institute Inc., 1999.

75. Turk D.C. and Rudy T.E. (1994). A cognitive–behavioral perspective on chronic pain: beyond the scalpel and syringe In: Tollison C.D. (eds) *Handbook of pain management* (pp. 136-151), Baltimore, MD: Williams and Wilkens.

76. Cattell R.B. The meaning and strategic use of factor analysis in Lawrence Erlbaum Associates (2nd edition), *A first course in factor* analysis (pp. 131-202), New Jersey: Hillsdale, 1992.

77. Gorsuch R.L. (1992) Exploratory factor analysis In: Lawrence Erlbaum Associates (2nd edition), *A first course in factor analysis* (pp. 231-258), New Jersey: Hillsdale, 1992.

78. Mathworks Inc., MatLab Manual Version 7 Release 14, Natick, Massachusetts: Mathworks Inc. 2005.

79. Viterbi A.J. (1967) Error bounds for convolutional codes and an asymptotically optimal decoding algorithm. *IEEE Transactions on Information Theory*, IT-13, 260-269.

80. Kullback S. (1958) Information Theory and Statistics, New York: Wiley.

81. Liang H, Anderson-Sprecher R.C., Kubichek R.F. and Talwar G. (2005). A novel approach to approximate Kullback-Leibler distance rate for hidden Markov models. *IEEE Conference on Signals, Systems and Computers*, 869-873.

82. Fleiss J. (1981). *Statistical methods for rates and proportions,* New York, NY: Wiley.

83. Eye A. and Schuster C. (1998). *Regression analysis for social sciences.* San Diego, CA: Academic Press.

84. Lipsey MW. (1990), *Design Sensitivity: Statistical Power for Experimental Research*. Newbury Park, CA: Sage Publication, Inc.

85. Rudy T.E., Lieber S.J., Boston J.R., Gourley L. M., and Baysal E. (2003). Psychosocial predictor of physical performance in disabled individuals with chronic pain. *The Clinical Journal of Pain,* 19, 18-30.

86. Vlaeyen JWS, Kole-Snijders AMJ, Boeren RGB, and van Eek H. (1995) Fear of movement/(re)injury in chronic pain and its relation to behavioral performance. *Pain*, 62, 363-372.

87. Lackner J.M, and Carosella A.M. (1999). The relative influence of perceived pain control, anxiety, and functional sefl-efficacy on spinal function among patients with chronic lower back pain. *Spine*, 24(2), 2254-2261.

88. Verbunt J.A., Seelen H.A., Vlaeyen J.W., Bousema E.J., van der Heijden G.J., Heuts P.H., and Knottnerus J.A. (2005). Pain-related factors contributing to muscle inhibition in patients with chronic lower back pain. *Clinical Journal of Pain*, 21(3), 232-240.

89. Lethem J., Slade PD, Troup JDG, and Bentley G. (1983). Outline of a fear-avoidance model of exaggerated pain perceptions. *Behavioral Research Therapy*, 21, 401-408.

90. Phillips HC. (1987). Avoidance behavior and its role in sustaining chronic pain. *Behavioral Research Therapy*, 25, 273-279.

91. Waddell G., Newton M, Henderson I, Somerville D and Main C. (1993). A Fear-Avoidance Beliefs Questionnaire (FABQ) and the role of fear-avoidance beliefs in chronic low back pain and disability. *Pain*, 52, 157-168.

92. Vlaeyen JWS and Linton SJ. (2000). Fear-avoidance and its consequences in chronic musculoskeletal pain: a state of the art. *Pain*, 85, 317-332.

93. Neblett R., Gatchel R. J., and Mayer T. G. (2003) Clincal guide to surface–EMG assisted stretching as an adjunct to chronic musculoskeletal pain rehabilitation. Applied psychophysiology and Biofeedback, 28 (2), 147-160

94. Rosner B. (2000). *Fundamentals of Biostatistics.* Pacific Grove CA: Duxbury.