Data-Driven Computational Approaches in Pain Medicine

Major Qualifying Project

A Major Qualifying Project submitted to the faculty of
WORCESTER POLYTECHNIC INSTITUTE
in partial fulfillment of the requirements for the Degree of Bachelor of Science

By:
Aidan Burns
Lauren Flanagan
Smera Gora

Report Submitted to:
Professor Carolina Ruiz, Advisor
Department of Computer Science and Data Science Program, WPI

Professor Benjamin Nephew, Co-Advisor
Department of Biology and Biotechnology and Neuroscience Program, WPI

Doctor Lisa Conboy, Co- Advisor
Beth Israel Deaconess Medical Center, Harvard Medical School
April 27, 2023

Abstract

Chronic pain affects a large percentage of the adult population in the United States. The goal of this project was to analyze data from a study about the effects of acupuncture treatment on veterans of the Persian Gulf War (1990-91) afflicted with what is known as Gulf War Illness (GWI) to determine for whom acupuncture effectively reduces chronic pain. A series of machine learning models were developed to gain insight into the key factors that help predict whether a patient's chronic pain improved. Logistic Regression yielded the most accurate predictive models for pain improvement in patients. Drawing from the calculated feature importance and Logistic Regression modeling, the most important factors for the prediction are derived from the McGill Pain Scale, SF-36 questionnaire, Locus of Control questionnaire, Pittsburgh Sleep Quality Index, and Carroll Depression Scale.

Table of Contents

List of Figures

List of Tables

1. Introduction

Chronic pain affects over 20% of adults in the United States and has a large negative impact on many aspects of their health and well-being (Yong et al., 2022). A specific form of chronic pain only observed in veterans of the Persian Gulf War (1990-91) is known as Gulf War Illness (GWI). There has been much debate and uncertainty surrounding the source of this illness since the mid-90s; however, over the years, a consensus has been reached that one of the leading causes was the numerous toxic exposures that veterans faced during their time in the Gulf War (Wessely & Freedman, 2006). Veterans with GWI experience a multitude of chronic symptoms that can be partitioned into three main categories comprising fatigue, mood and cognition impairments, and musculoskeletal issues (Centers for Disease Control and Prevention (CDC), 1995). These categories cover a wide range of different ailments that Gulf War veterans face daily and present the need for analysis on how to improve their health and well-being.

There have been studies that investigate the efficacy of different treatments for GWI patients, including acupuncture and other Traditional Chinese Medicine (TCM). A current study into the effects of acupuncture treatment on GWI patients has yielded data that is important to understanding patient response (Conboy et al., 2012). Further analysis of the patients' data was required to gain insight into whether (and for whom) the acupuncture treatment resulted in an improvement of chronic pain as well as the corresponding degree of significance. The goal of this project was to utilize the data from this study to identify and predict for whom the treatment effectively helped reduce chronic pain, through the means of predictive modeling and other analysis.

We addressed this goal through the development of a series of regression and classification models to predict the reduction of chronic pain in veterans. The specific modeling techniques used were Linear Regression, Logistic Regression, K-Nearest Neighbors, Softmax Regression, and XG Boosting. The Logistic Regression technique produced the best performing model. The key factors

for predicting the pain outcome were found to be in the McGill Pain Scale, the SF-36 questionnaire, the Locus of Control questionnaire, Pittsburgh Sleep Quality Index, and Carroll Depression Scale. These important features for predicting pain outcomes in patients can be further utilized to determine if treatment will be effective for other Gulf War veterans.

2. Background

2.1 Chronic pain

Chronic pain can be defined as a pain that lasts “months or years,” and “interferes with daily life” (Dydyk & Conermann, 2023). It is one of the leading health issues in the US. In fact, prior studies show that “findings indicate that more than one in five adults in America experiences chronic pain” (Yong et al., 2022). This condition has a vast variety of root causes, and impacts people in different ways. Some causes of chronic pain can be due to long-term illnesses that result in persistent pain, major injuries, diseases that can leave you “more sensitive to pain,” or even pain “caused by psychological factors such as stress, anxiety and depression” (Dydyk & Conermann, 2023).

Recent studies have further investigated the science behind chronic pain. These studies show that the nervous system’s neuroinflammation is associated and probably even “mediates the persistence and chronification of human pain conditions” (Ji et al., 2018). Another study also looks into pain that often coexists with chronic pain, referred to as chronic overlapping pain. According to the study, patients with chronic overlapping pain do not show “a compensatory increase in antiinflammatory cytokines” (Ji et al., 2018) to reduce pain compared to patients who are suffering from regular localized pain. In addition to this, people with chronic overlapping pain fail to “augment immune response and proinflammatory cytokine production” (Ji et al., 2018). This means that the pain these patients feel is sustained as their nervous system has suffered changes making it hard for them to fight the pain. Lastly, studies show that issues that cause chronic pain such as tissue and nerve damage further “heighten[s] synaptic transmission” (Ji et al., 2018), or an increase in which neurons communicate which in turn lowers a person’s pain threshold, as well as “amplification of pain responses, and a spread of pain sensitivity to noninjured areas” (Ji et al., 2018). The reason this increased synaptic activity is disadvantageous is because neurons that regulate pain (nociceptive neurons) take part in more of these neuron transmissions.

2.1.1 Impacts on the Individual

As previously mentioned, chronic pain affects around one in five adults in America. Unfortunately, chronic pain frequently comes with other conditions including depression. A study shows that "injury sensory pathways of body pains have been shown to share the same brain regions involved in mood management ... which form a histological structural foundation for the coexistence of pain and depression" (Sheng et al., 2017). Further studies have also shown that of patients who suffer from persistent pain as much as 85% suffer from depression. Another interesting concept this study investigates is that depression and chronic pain act bidirectionally (Bair et al., 2003). Furthermore, people with depression are at risk of having pain complaints. The study concludes that "on average, 65% of patients with depression experience one or more pain complaints (Bair et al., 2003).

In addition, pain patients incur much more costs than those with less pain, and have trouble being productive in the workplace. A study calculated that a person suffering from "severe pain had health care expenditures $3,210 [annually] higher than those of persons with moderate pain" (Gaskin & Richard, 2012). Exacerbating this financial challenge, these people usually make less money in the workplace as they work fewer hours due to their pain. This study observed that a person "with severe pain worked 717 fewer hours [annually]" (Gaskin & Richard, 2012).

2.1.2 Impacts on the Economy and Society

Chronic pain not only impacts the individual but the economy as well. One event that shed light on this topic was the opioid epidemic starting in the late 1990s. Opioids, widely prescribed as a pain-relieving medicine, soon caused addiction and thousands of deaths due to overdosing. Many misused and took advantage of the prescriptions due to addiction, and this epidemic indicated a substantial need for non-addictive inventions for pain(National Academies of Sciences, Engineering, and Medicine (U.S.) et al., 2017).

Another way chronic pain impacts society is due to low worker productivity as mentioned earlier. Patients with chronic pain work less than others who do not suffer from this condition. In

addition, undertreatment is a large issue in today's society. Tracing back to opioid misuse, financially struggling patients would likely opt in to using opioid medication compared to financially stable people, leaving them more susceptible to opioid addiction and drug misuse (Newman et al., 2018). In addition to this, studies show racially underrepresented groups face more severe chronic pain, due to epigenetic mechanisms that increase sensitivity to pain (Aroke et al., 2019).

2.2 Gulf War Illness (GWI)

One specific illness that involves chronic pain is Gulf War Illness (GWI), otherwise known as Gulf War syndrome. GWI refers to the prominent chronic symptoms shared by many veterans from the Persian Gulf War (1990-91) and has been the topic of numerous medical studies involving both the investigation of its causes and subsequent treatments of its symptoms (Kerr, 2015).

Early studies, conducted in the first few years after the Gulf War, did not produce sufficient evidence of the existence of GWI. While there were some reports of veterans with similar symptoms, which they believed were tied to their deployment in the Gulf War, there was very little data at the time to make strong conclusions about the existence of a syndrome (Beale et al., 1997). In addition, a common hypothesis at this time was that the symptoms present in these veterans were "simply the result of war- related stress," which did not consider all that these individuals were exposed to during the war (Kerr, 2015). Another reason for the lack of concrete evidence and definition of GWI in these early years was the poor documentation of medical services (during and before deployment) that were provided to veterans (Mawson & Croft, 2019). In conjunction with this, the vast number of different symptoms reported made accurately defining the root cause an immense task (Joyce & Holton, 2020). These factors all contributed to the unexplained nature of this illness and made the specific identification of its causes difficult. However, over time and with the accumulation of more reports and data on these Gulf War veterans, more trends and detailed ideas of GWI's origins arose.

2.2.1 Causes of GWI

It was concluded that GWI can be traced back to "numerous potentially toxic deployment-related exposures that appeared to vary by country of deployment, by location within the theater, by unit, and by personal job types," encapsulating the broad scope of the potential origins of this illness (Kerr, 2015). While there have been many theories and studies into the origins of this illness, it is known that Gulf War veterans suffered a multitude of exposures during the war that are considered prominent causes of GWI (Kerr, 2015).

As is present in war settings, Wessely and Freedman (2006) reported that there is always the risk "to life and limb from the various munitions and explosives that are part and parcel of modern war," as well as "dangers to the psyche." Other toxic exposures, such as the smoke from oil fires started by the Iraqis, were unanticipated but not unheard of during this time. However, one major difference in the Gulf War was the added threat of more recent advancements in dangerous weapons and technology. Expected presence of "large stocks of chemical and biological weapons" incited countermeasures to protect against these deadly weapons, which included "giving vaccinations against biological agents such as plague and anthrax, [and] taking pyridostigmine bromide tablets to protect against exposure to organophosphate (OP) nerve agents" (Wessely & Freedman, 2006). The exposure to both the vaccinations meant to protect and the chemical agents themselves proved to be very harmful and a potential cause of the "rashes, muscle pain, fatigue, headaches, and other mysterious symptoms" reported ("Update on Gulf War Illness," 1997). This combination of recently developed health threats and protective measures used to combat them, along with the many pre-established dangers of war, resulted in the emergence of the many chronic symptoms experienced by veterans.

2.2.2 GWI Symptoms and Symptom Clusters

GWI symptoms have been sorted into three main categories by the CDC: fatigability, mood and cognition, and musculoskeletal (Centers for Disease Control and Prevention (CDC), 1995). Although commonly seen individually, the aggregation of these three symptom clusters is

helpful in defining the nature of GWI. They facilitate patient diagnoses and studies into potential treatments for this illness. A patient is diagnosed with GWI if experiencing at least one symptom from two of the three clusters defined by the CDC (Nettleman, 2015).

The first of these symptom clusters, fatigability, refers to exhaustion or fatigue that persists for longer than 24 hours subsequent to exertion (Conboy et al., 2012). The second category of GWI symptoms is mood and cognition; contained in this cluster is a myriad of different possible mood-related conditions, such as feeling depressed, irritable, or anxious, as well as symptoms related to difficulty thinking or concentrating (Conboy et al., 2012). Also included in this cluster is difficulty sleeping. Of the three components of GWI, "cognitive problems remain one of the most prevalent and distressing symptoms of GW veterans," making it an important component of GWI screening (Jeffrey et al., 2019). The third and last component of GWI deals with joint and muscle pain (Blanchard et al., 2006). Chronic skeletal pain has been observed in many Gulf War veterans, and even includes suffering from bone fractures and soft tissue injuries (Mawson & Croft, 2019). While these clusters are useful in defining GWI and diagnosing veterans, there are other symptoms related to GWI, including reproductive disorders, gastrointestinal issues, and Amyotrophic Lateral Sclerosis (ALS), among various others, that affect Gulf War veterans (Mawson & Croft, 2019). Overall, defining these symptoms in patients is a practical and attainable way of identifying GWI using both qualitative and quantitative measurements.

2.3 Measurements of Health

The measurement of subjective experiences in health is inherently difficult. To accomplish this task, many studies have been conducted to determine accurate and easy ways to gather information from participants. Early health questionnaires often were criticized for being too long with ambiguous statements, too narrow in scope, and involving the addition of subscores that were unrelated (Hunt et al., 1985). Over time, more specific and concise questionnaires were created to improve the shortcomings of previous questionnaires.

2.3.1 McGill Pain Scale

The short form McGill Pain questionnaire was developed in order to track an individual's perceived pain over time for academic research (Melzack, 1987). It consists of 15 sensory adjectives that participants describe as either 'none', 'mild', 'moderate', or 'severe'. It was constructed with 11 sensory words, and four affective words from the original McGill pain scale. It was thought that there were three dimensions to pain, the experience of the pain, the quality of the pain, and the intensity of the total pain experience (Melzack & Raja, 2005). The McGill pain questionnaire was also found to be more accurate than the Nottingham profile (Melzack & Raja, 2005).

Table 1. Example of a Filled-Out Short Form McGill Pain Questionnaire (Melzack, 1987)

Word	None	Mild	Moderate	Severe
Throbbing	✓			
Shooting			✓	
Stabbing		✓		
Sharp		✓		
Cramping				✓
Gnawing		✓		
Hot-Burning				✓
Aching				✓
Heavy			✓	
Tender	✓			
Splitting	✓			
Tiring-Exhausting				✓
Sickening	✓			
Fearful		✓		
Punishing-Cruel		✓		

2.3.2 SF-36 Scale

In 1992, the RAND Corporation developed a 36-question short form survey (SF-36) in order to easily assess the quality of life of patients (Brazier et al., 1992). Quality of life was meant as the combination of health limiting physical and social activities, bodily pain, mental health, energy and fatigue, the perception of personal health, and the impact health had on one's profession (Ware & Sherbourne, 1992). The SF-36 was an attempt at improving the older Nottingham health profile. At 36 questions long, it was short enough to have a high response rate and also have a high rate of completion. The SF-36 was also able to 'detect low levels of ill patients' who had scored perfect health on the Nottingham questionnaire.

2.3.3 Other Scales

With the focus on chronic pain, there are many scales that standardize how an individual perceives their health. Within mental health, there is the Beck Anxiety Inventory (BAI), the Carroll Depression Scale, and the Whitely Depression Scale. The Pittsburgh Sleep Quality Index (PSQI) measures sleep, Measure Yourself Medical Outcome Profile (MYMOP) measures general health, while the Profile of Mood States (POMS) provides understanding of mood (see Table 2). These questionnaires and surveys can also be tailored to fit a study (i.e., by removing or adding questions).

Table 2. List of Scales Used in Data Modeling

Scale	Description
BAI	• 21 items • Measures of anxiety of the psyche and cognition • Self-reported (Leyfer et al., 2006)
Carroll	• 24 items • Self-rating instrument used for measuring the severity of one's depression (Carroll et al., 1981)
Whitely	• Seven items • Assesses the anxiety experienced due to the fear of having an undiagnosed illness (Chen et al., 2021)
PSQI	• 17 items • Measures the quality of sleep and sleep disturbances • Said to be more reflective of a negative cognitive viewpoint or depressive symptoms than actual sleep parameters (Grandner et al., 2006)
POMS	• 65 items • Measures mood swings of tension and anxiety, anger and hostility, and fatigue and inertia through a 5-point scale (Shahid et al., 2011)
MYMOP	• 10 items • Offers an individualized approach to evaluating general health • Works particularly well for musculoskeletal and respiratory conditions, by focusing on impacted symptoms and activities (Paterson, 1996)
WAI	• 36 items • Working Alliance Inventory is a self-reported instrument used to measure the quality of alliance between a patient and a therapist • Based on bonds, goals, and tasks (Paap et al., 2019)

2.4 Related Work

An article of review that summarizes nine Gulf War Illness studies concluded that Gulf War veterans have been affected by neuropsychological illnesses after being deployed (Jeffrey et al., 2019). This article also summarizes studies comparing veterans who have fallen symptomatic with GWI and Gulf War veterans who have not been affected by it. Conboy et al. have been investigating the effects of acupuncture on GWI patients to determine whether it will decrease chronic pain levels (Conboy et al., 2012). The data from this study has been used in the present analysis.

2.4.1 Previous Studies on GWI

Several investigations have focused on the neuropsychological outcomes of Gulf War veterans, as many of them seemed to have been affected by neuropsychological impairments. Previous studies have also concluded that "visuospatial abilities, attention/executive functioning, and learning/memory" are rated lower in GWI patients than in non GWI patients using a variety of assessments (Jeffrey et al., 2019).

Another attribute that GWI affects is memory, which was assessed and concluded by using the Expanded Health Symptom Checklist. GWI veterans have also suffered from "high levels of neuropsychological symptoms also reported tension, fatigue, confusion, and decreased vigor," as concluded by the Profile of Moods survey. Studies have also shown a decreased ability to retain information (Jeffrey et al., 2019).

In another study, GWI patients took the SCL-90-R questionnaire that screens for a "range of psychological symptoms and psychopathological features on nine subscales" (Jensen et al., 2013). People who were deployed during the Gulf War showed significantly poorer performance over all on this questionnaire when compared to a control group (Jeffrey et al., 2019). The authors were not able to draw conclusive results due to the small sample size (n=103) and other limitations, but there seemed to have been some correlation between deployment and psychological impact. Overall, all studies conducted on Gulf War veterans' cognitive aspects have shown them to be

lower or worse off than other veterans who have not been deployed in the Gulf War and people who did not serve in the military at all.

A few other studies compared symptomatic vs. non-symptomatic GW veterans in terms of neuropsychological performance. Utilizing a validated survey, these studies have shown that veterans who exhibit symptoms perform worse on measures of brain function. In addition to this, symptomatic veterans performed worse on a test that measured their "abstract reasoning and problem solving/flexibility; measures of executive functioning" (Jeffrey et al., 2019).

2.4.2 A Study of the Effectiveness of Acupuncture on GWI Patients

A previous randomized clinical trial evaluated the effects of personalized acupuncture treatment on patients suffering from chronic pain due to Gulf War Illness (GWI). Assessments used to measure the outcome of this study were the McGill Pain Scale and the SF-36 Scale, both validated and reliable scales. Individuals first had to go through a screening to make sure that they had GWI before being admitted into the trial (Melzack & Raja, 2005; Ware & Sherbourne, 1992). During testing, patients were split up into two groups. The first group of patients (Group 1) were getting acupuncture twice a week, and the second group (Group 2) got put on a two-month waitlist and was then able to get weekly acupuncture. These acupuncture treatments were individualized and conducted by experienced acupuncture practitioners. Other methods of Traditional Chinese Medicine supplemented the acupuncture, such as heat therapies, electro-acupuncture, Chinese massage, and press balls (Conboy et al., 2012). These two groups completed the SF-36 survey and the McGill Pain Questionnaire bimonthly to track progress. Here, the latter group can be viewed as a control group, as the first two months they did not receive treatment, and there also seems to have been no change in their pain level during these first two months. This study provided treatment to all participants, allowing the evaluation of the effects of frequency of acupuncture treatment on pain levels.

This study reported a statistically significant improvement in the pain of the patients who received treatment twice a week. An increase in SF-36 score positively correlates with good quality of life, as larger values on the SF-36 scale are mapped to a higher rating of wellbeing. Patients who

were getting bi-weekly acupuncture after month 4 saw around a nine-point increase in their point value in the SF-36 scale (Figure 1). For the waitlisted group, however, during their two months on the waitlist they saw a slight decline in quality of life, but in the end there was a net increase of three points. It can be inferred here that getting acupuncture more often gets one's body more acclimated to the treatment, and therefore people have better results further down the line with the treatment.

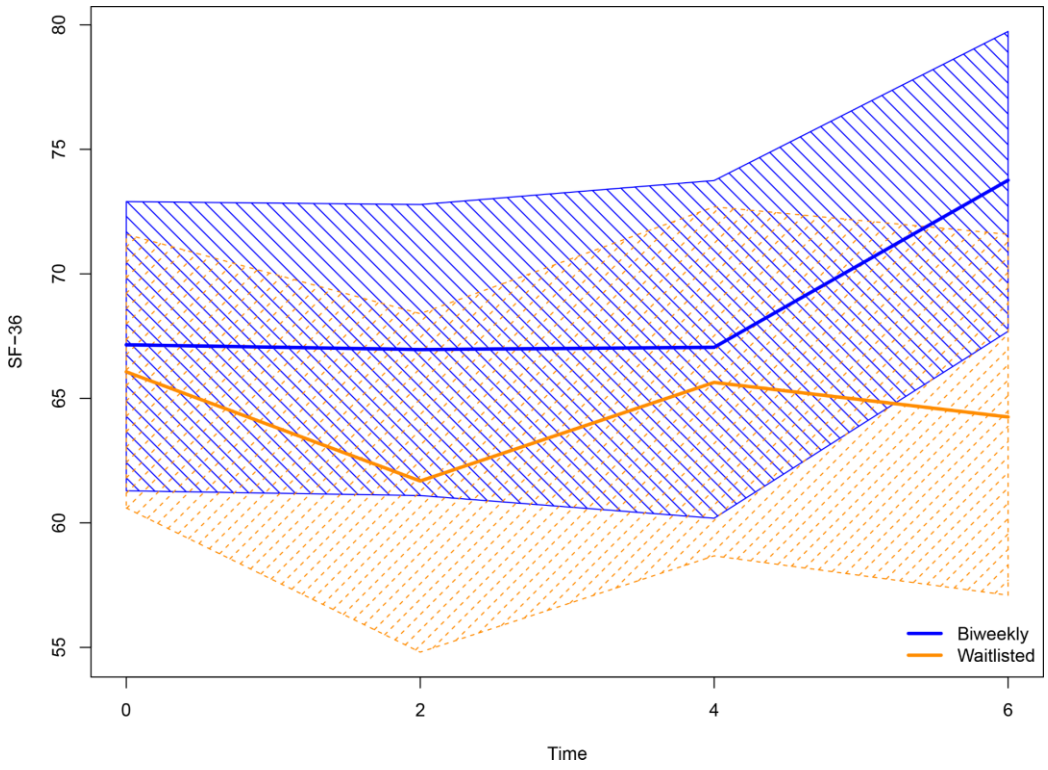

Figure 1. SF-36 Scores of Biweekly and Waitlisted Groups Over Time (Conboy et al., 2012). Scores improve more significantly for people receiving acupuncture more frequently.

The McGill scale (see Figure 2) works in reverse, where a lower score value indicates less pain, so a decrease in score is considered positive. Here again, a steady decrease can be seen in the group that received acupuncture twice a week. For the waitlisted group, during their two months of no treatment, there was an increase in their McGill pain score, which starts decreasing after they receive acupuncture (but not as rapidly as the group that receives acupuncture twice a week).

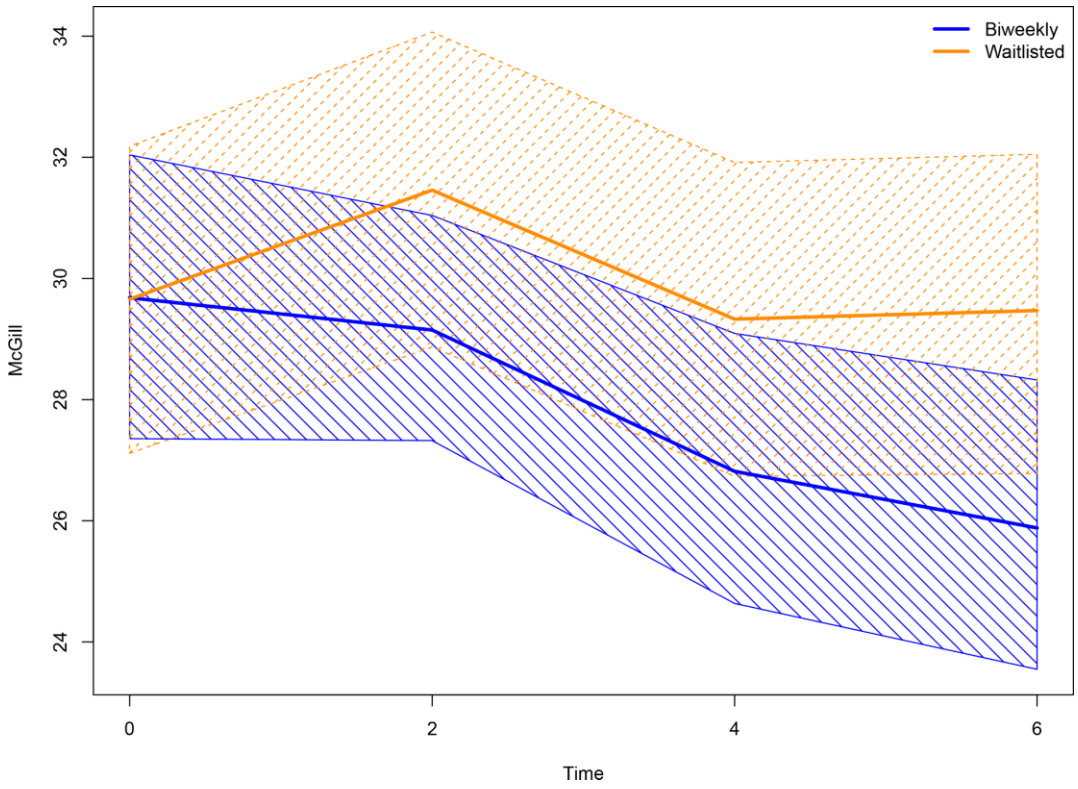

Figure 2. Pain Levels of Biweekly and Waitlisted Groups Over Time (Conboy et al., 2012). Pain levels decrease more significantly for patients receiving acupuncture more often.

3. Methodology

3.1 Data Description

Data was provided to the team courtesy of Dr. Conboy from her previous research into acupuncture and GWI. The data consists of 103 patients split into two treatment groups and observed over the course of the 6-month study (outlined in Section 2.4.2 above). Group 1 patients received acupuncture treatment twice a week, while Group 2 received it once a week after an initial two-month waitlist. The dataset contains over 1900 attributes describing each participant and their answers to various questionnaires over the course of the study. Including data from over 20 standard questionnaires, the surveys were administered four times throughout treatment. The first survey (T1) was first administered before treatment had started. The second (T2) and third (T3) surveys were administered at two and four months into the treatment, respectively. The final survey was administered at the conclusion of treatment (6 months). The scope of the questionnaires includes topics such as mood, pain, anxiety, sleep, depression, and patient-healthcare provider relations. For more information about the individual scales, refer to Section 2.3.

3.1.1 Demographics

There is a high percentage of participants in their forties, with the median age of the group being 47 years old, the mean being 48.2 years old, and the oldest being in their late 60s (Figure 3). Out of the 103 participants, seven identified as Hispanic or Latino. The largest identifying race was white, followed by African American, 2+ races, other, and then American Indian (Figure 4).

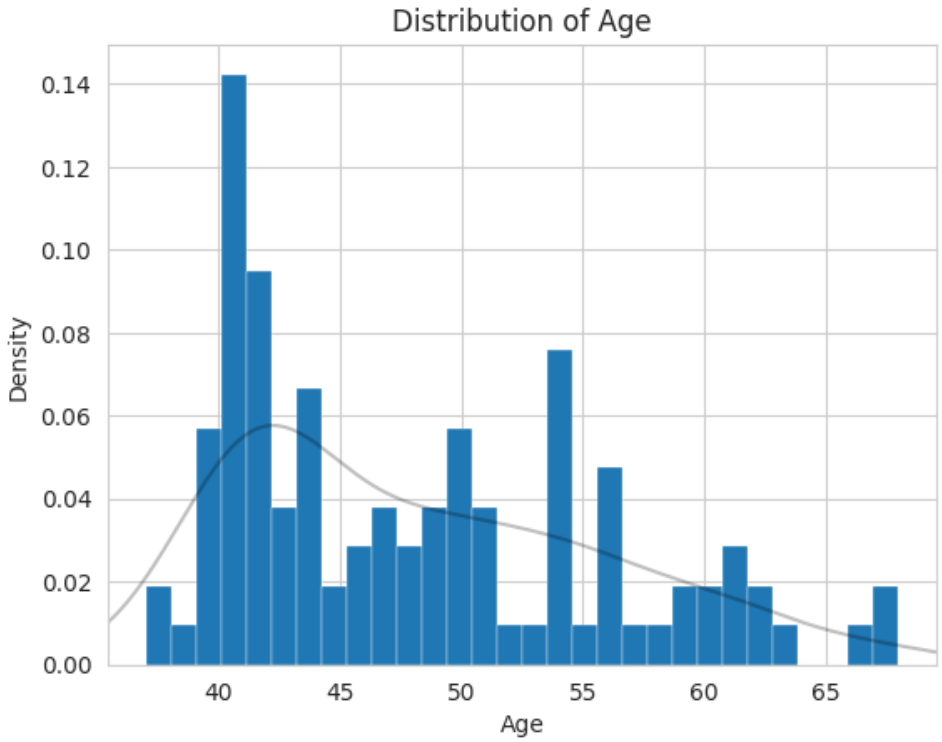

Figure 3. Age Distribution of Study Participants

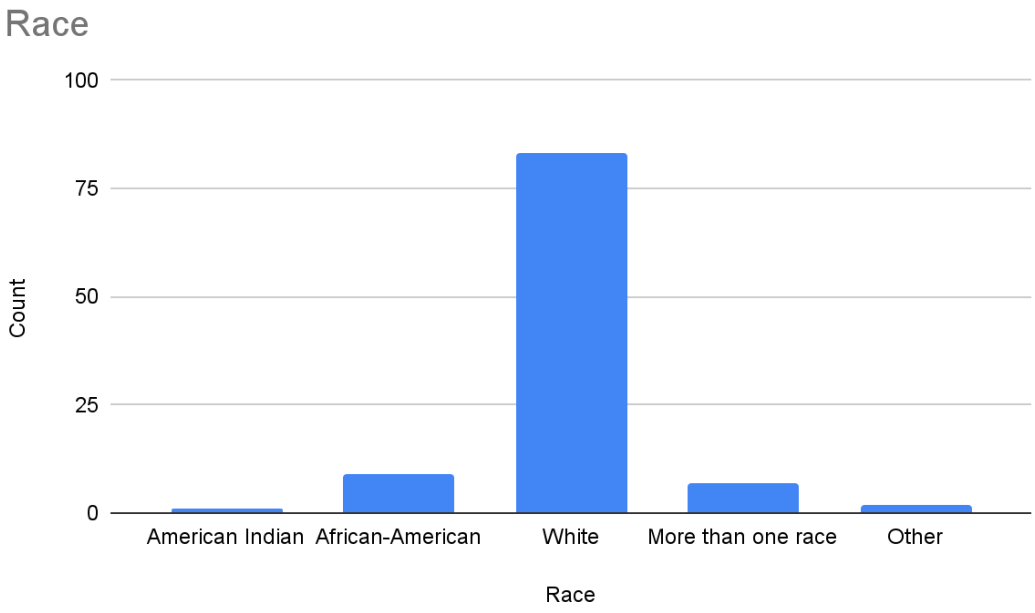

Figure 4. The distribution of identifying races for study participants

3.1.2 Pain Score Distributions

With the inclusion of over 20 questionnaires, there is a great deal of sensitivity about different aspects of how treatment affects the study participants. Many questions have remarkably similar responses between the two participant groups, and there are few that stand out with different distributions of the respondents.

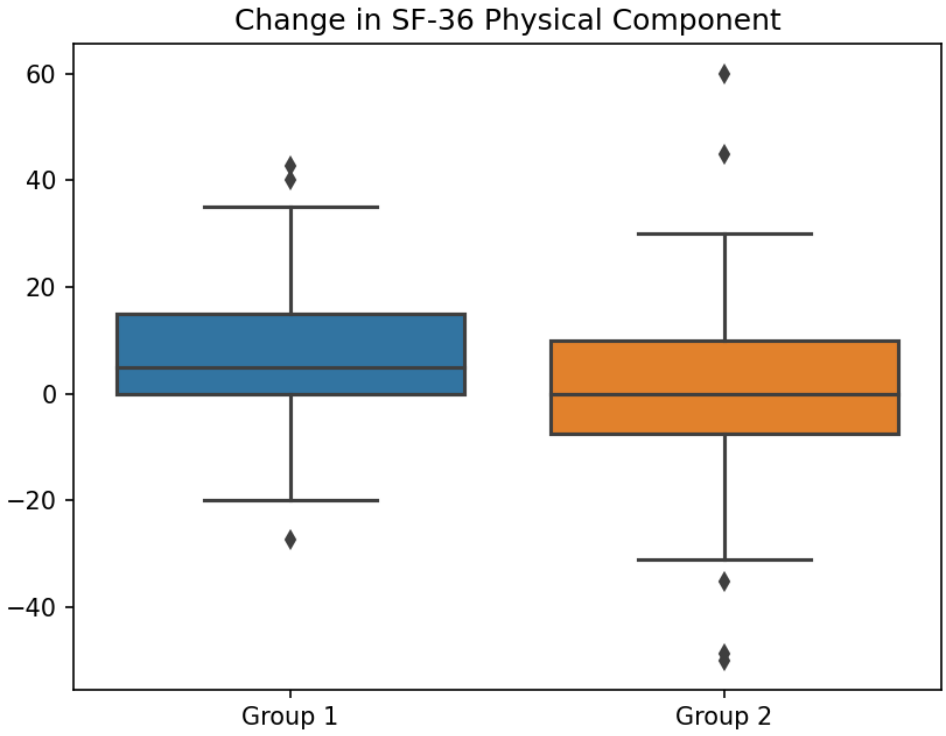

Figure 5. Change in SF-36 Physical Functioning Components. Higher SF-36 Physical Component value is more desirable (scores range from 0-100).

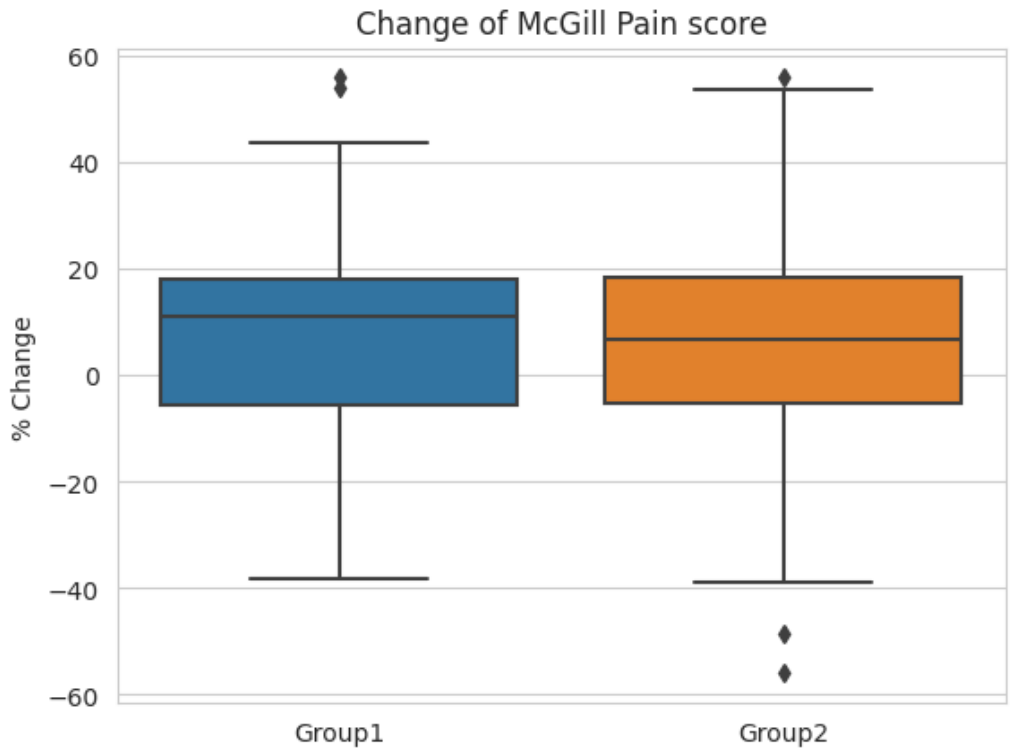

Figure 6. Percent Change in McGill Pain. Higher reduction in McGill pain measure is more desirable.

3.1.3 Traditional Chinese Medicine

Additional data provided to the team by Dr. Conboy included Traditional Chinese Medicine (TCM) Diagnoses data that was directly associated with each patient in the dataset. Attributes in this supplementary dataset specified three different categorical diagnoses for patients (with some overlap), resulting in the following breakdown (Taylor-Swanson et al., 2019):

- Single Diagnosis (Excess, Deficiency, or Channel Imbalance)
- Excess and Deficiency
- Excess, Deficiency, and Channel Imbalance
- Deficiency and Channel Imbalance
- Excess and Channel

3.2 Data Pre-processing

This section outlines the pre-processing performed on the given data, both prior work and that of the team. Prior pre-processing of the data was conducted by Ruofan Hu, a Ph. D. student at WPI and collaborator of Prof. Ruiz and Dr. Conboy, and the version of the data produced from this work was used as the base version for the team's usage.

3.2.1 Prior Pre-processing

Based on previous work on the data by Ruofan Hu, some pre-processing of features had already been performed and was used as the basis for further pre-processing. This work included handling missing values and calculating delta variables for use in analysis, as well as creating summary variables to condense the large number of questionnaire scales and subscales.

3.2.1.1 Missing Values

To handle the large number of attributes in the data, certain patients were removed from the data based on the percentage of missing values observed. Specifically, there were 20 patients who were missing over 50% of the survey questions, so their values were excluded from the dataset.

To further handle the more intricate missing values observed in the data, two methods of imputation were used to cater to the specific variable for which there was a missing value. The first method consisted of calculating and imputing the mean of whatever was the smallest or most specific subscale or sub-questionnaire to which the missing value belonged. This approach was used for most variables and not applied in cases where using a mean value was illogical in the context of the variable. The second method used for imputing values was to impute with zero, which applied to questions under the Social Network Index (SNI) category (such as "How many children do you have?"). Both of these methods were utilized where applicable in the data.

3.2.1.2 Summary Variables

Further pre-processing of the data included calculating summary variables for the many scales and questionnaires observed in the data. For each questionnaire, a sum score was calculated to combine the numerous questions into a single variable that could be used for analysis and modeling. Each sum score that was calculated followed a similar process to that described for the delta variables, in which the positive or negative impact of a higher value on each scale was considered and accounted for in the sum.

3.2.1.3 Delta Variables

For additional use in modeling and other analyses, some transformed variables were formulated based on the calculated summary variables. Specifically, delta variables were created to represent the difference between the first timepoint in the study (T1) and the last timepoint (T4). These delta variables were calculated in two different ways based on the specific variable, as each attribute could either improve or worsen with a positive or negative delta value. Furthermore, the delta values for variables such as McGill Pain were calculated as [Baseline – T4] because the higher the value on this scale meant worse pain. Using this formula for the delta allows reduction in pain (and reduction in other variables for which lower is better) to be observed as a positive change; meanwhile, for variables such as Social Support, the formula [T4 – Baseline] was used to indicate that higher values are better. This method of calculating delta values was applied to all attributes in the data, where each attribute was identified as having either a positive or negative impact with an increase between start and end points.

3.2.2 Further Pre-processing

Before developing predictive models for the data, further pre-processing and transformations were performed. Some extra rows were removed that programs (Excel, Python, R) were identifying as data, although they were actually metadata. These programs also viewed all data as strings, so casting of these attributes to numerical values was necessary. Transformations of the

initial model data involved scaling all the values using the min-max normalization method (scaled between 0 and 1). The dataset was then broken into two different subsets, one containing the data of patients from Group 1 and the other from Group 2. Lastly, we decided to use data from surveys that were filled out prior to a patient's treatment, and survey data that was collected at the end of the study. Attributes that computed the changes between attributes from the beginning to the end of the study were also included in the data. Further feature selection and other data computations took place during modeling, which will be discussed in the following sections.

Data that had low response rates and low variability were removed as they would not be useful to analysis and modeling. The first round of pre-processing involved removing any attributes with response rates less than 60%. Subsequently, we removed all attributes with less than 10% unique data.

Combining identical information into one attribute was important in one case. The attribute “Race” was repeated for all four surveys in the study, and they contained some conflicting and incomplete information. The data was corrected manually, paying attention to each individual’s responses and selecting the one that was most accurate.

3.2.2.1 Feature Selection with Lasso Regression

One modeling technique that was used to obtain values for feature selection was Lasso Regression. Lasso Regression was implemented in R (data programming language) using the glmnet and caret packages. This modeling technique was useful in determining which attributes to keep in the training set because it regularizes input data, which reduces many coefficient values to zero. Running Lasso Regression in R Studio outputs the attributes and their coefficients that were not reduced to zero. The attributes selected as well as their coefficients can be seen in Appendix A.

To build this model, the data was first split up into two separate datasets based on group. This was done because it was hypothesized that each group would have a different set of coefficients that are useful for predicting decreases in pain. Next, all data between the two groups was scaled, and two Lasso Regression models were run for each group. The first Lasso model used

the Timepoint 1 data as inputs and predicted the percent change in a veteran's pain at the last timepoint. This model would be useful in a prediction environment if one was trying to choose who should receive treatment. The second Lasso model used the delta values between the timepoints as its input, and the output was again pain reduction. This was done to see if there were any relationships between changes in other attributes of a person and how they perceive pain. For both of these models, the input data was min-max scaled (so all values were between 0 and 1) to make it easier to read the coefficients and see which factors have the greatest impact on pain.

3.2.2.2 Feature Selection with Random Forest

The Random Forest modeling technique was also implemented, and its output was leveraged for feature selection. The Random Forest method was run on scaled data from Timepoint 1 in both Group 1 and Group 2 separately. The output that was being predicted by this model was the percent change in a veteran's pain at the last timepoint.

This method was implemented in both R Studio and Python. In R Studio, the Random Forest model output includes each attribute and its "increase in mean squared error" value. This value indicates how much the model error would increase if the specific attribute it is attached to was removed from the model. For this reason, values with an increase in mean squared error value above zero were chosen. In later experiments with Python's implementation of Random Forest from the sklearn package, the output listed each attribute and its "feature importance" value (where higher-valued attributes were considered more important to the model). These attributes were graphed and then selected based on their feature importance ratings. Many of the same features were chosen from the R and Python implementations. The features chosen from the implementation of Random Forest in R Studio were used in Linear Regression, Logistic Regression, KNN, and Softmax. The features chosen from the Python Random Forest implementation were used in XG Boosting. Output from Random Forest models in both Python and R can be found in Appendix B.

3.2.2.3 Data Features and Subsets for Modeling

As a result of the further pre-processing conducted, specifically feature selection, the following subsets of the original data were created for modeling purposes. For all subsets of the data created, only features from Timepoint 1 (T1) were included to be utilized as model input data. The outcome variable associated with each of these data subsets was the McGill Pain Reduction percentage delta value (T4 – T1). This feature was used as the target variable for predictions across all regression and classification models. For classification models, the target variable was divided into two classes, split by a threshold to indicate a cutoff for significant improvement in pain vs. insignificant. This threshold was adjusted at values of +12%, +20%, and +30% to test different cutoffs and to see how each model performed as a result.

Table 3 outlines a list of the full features in the dataset, after prior pre-processing by Ruofan Hu, along with their corresponding occurrences in each of the study timepoints. Some features were not present in the first Timepoint and were seen later in the study.

Table 3. List of the full features, after prior pre-processing, and timepoint occurrences

Attribute	T1	T2	T3	T4
WAI: Working Alliance Bond		✓	✓	✓
WAI: Working Alliance Task		✓	✓	✓
WAI: Working Alliance Goal		✓	✓	✓
WDEP Whitely Depression	✓	✓	✓	✓
sumFI Fatigue	✓	✓	✓	✓
SOC Social Support Open	✓	✓	✓	✓
SNI Support Networks Num People	✓	✓	✓	✓
SNI Support Networks Num Embedded	✓	✓	✓	✓
SNI Support Networks High Contact Role	✓	✓	✓	✓
SF_PHYS Physical Functioning	✓	✓	✓	✓
SCHN Stress	✓	✓	✓	✓

PSQI Subjective Sleep Quality	✓	✓	✓	✓
PSQI Sleep Medications	✓	✓	✓	✓
PSQI Sleep Disturbances	✓	✓	✓	✓
PSQI Sleep Daytime Dysfunction	✓	✓	✓	✓
POMS Mood States	✓	✓	✓	✓
OPTI Optimism	✓	✓	✓	✓
MYMOP Self-reported Medical Outcome	✓	✓	✓	✓
McPain Pain score	✓	✓	✓	✓
McPain_s Sensory Pain score	✓	✓	✓	✓
McPain_a Affective Pain score	✓	✓	✓	✓
LCTR Locus of Control Powerful	✓	✓	✓	✓
LCTR Locus of Control Internal	✓	✓	✓	✓
LCTR Locus of Control Chance	✓	✓	✓	✓
ISEL Social Support	✓	✓	✓	✓
CDEP Work Interests	✓	✓	✓	✓
CDEP Retardation	✓	✓	✓	✓
CDEP Psy Anxiety	✓	✓	✓	✓
CDEP Guilt	✓	✓	✓	✓
CDEP Depression	✓	✓	✓	✓
CDEP Agitation	✓	✓	✓	✓
CAT Catastrophizing	✓	✓	✓	✓
BCX Body Consciousness	✓	✓	✓	✓
BANX Anxiety	✓	✓	✓	✓

Table 4 outlines each of the data subsets used as model input. The datasets labeled "Original" illustrate the original data for each group (without any feature selection) and the corresponding information about that dataset. All other data subsets listed specify from which method(s) the features were selected (namely Lasso Regression and Random Forest (RF)), along

with the same information about each dataset. TCM refers to the "Traditional Chinese Medicine" Diagnoses data that was received later in the modeling process and was appended to some of the following data subsets.

Table 4. Standard Data Subsets Utilized for Modeling from Pre-Processing and Feature Selection. The "∩" symbol was used to signify the intersect of attributes chosen from Lasso Regression and Random Forest. Where present, the TCM Diagnoses data was appended.

Features Selected From	Group	# of Features Included	# of Features Removed	# of Patients
Original	1	34	0	44
Lasso	1	16	18	44
RF	1	19	15	44
Lasso ∩ RF	1	10	24	44
Lasso ∩ RF, TCM	1	11	23	41
Original	2	34	0	39
Lasso	2	9	25	39
RF	2	19	15	39
Lasso ∩ RF	2	8	26	39
Lasso ∩ RF, TCM	2	9	25	39

3.3 Modeling Experiments

The following sections outline modeling experiments conducted for the purpose of predicting the McGill Pain Reduction percentage (target attribute). The experiments involved modeling techniques Linear Regression, Logistic Regression, K-nearest neighbors (KNN), and Softmax Regression to view relationships between improvements in chronic pain and other features in the data as well as to predict whether patients experienced an improvement in or worsening of chronic pain based on their baseline survey.

For all modeling experiments, the model validation method used was Leave-One-Out cross validation; this method was used to maximize usage of the given small data. For the purpose of evaluating all binary classification models, the most important metric is precision (Hicks et al., 2022).

3.3.1 Linear Regression

Linear Regression was trained on all datasets outlined in Section 3.2.2.3, with a target variable of the reduction in McGill Pain score. Using Linear Regression as our first technique, we aimed to predict the improvement in patients' chronic pain through regression analysis. Evaluation metrics used to determine prediction performance included R^2, mean average error (MAE), and mean squared error (MSE).

3.3.2 Logistic Regression

Logistic Regression was trained on all datasets outlined in Section 3.2.2.3. The purpose of using Logistic Regression was to serve as a classification method in an attempt to increase accuracy over models similar to a majority class classifier. As previously mentioned, binary classes were created by adjusting a threshold for the McGill Pain score. The pain outcome thresholds used included +12%, +20%, and +30%.

3.3.3 K-Nearest Neighbors

The K-Nearest Neighbors (KNN) classification method was employed to train models on all datasets outlined in Section 3.2.2.3. As previously mentioned, the output was mapped to binary values. To do this, we set different thresholds and considered all patients who had a pain reduction percentage above the given threshold to have gotten better. The thresholds for the classification of the target attribute were the same as in previous models (+12%, +20%, and +30%).

3.3.4 Softmax Regression

The Softmax Regression technique was used to create further models following a similar process to previous models for the input data and outcome variables (i.e., utilizing the feature-reduced datasets from the Lasso Regression and Random Forest models, as well as the TCM data). The outcome variable used for these experiments was the percentage reduction in McGill pain score (T1 to T4) that was used for other models. Models were trained over the two separate groups using the input data subsets outlined in Section 3.2.2.3.

The Softmax Regression method was run to classify the McGill Pain reduction percentage using three classes to differentiate between different levels of pain reduction. Namely, the first class represented that a patient had a significant pain reduction ("Significantly Better"), which was defined by a pain reduction threshold of significance. The second class contained all insignificant pain reduction from below the threshold to above zero ("Insignificantly Better"), while the third class represented anything zero and below that demonstrated that the pain became worse over time ("Worse"). The threshold that was used varied for certain experiments based on what is considered clinically significant pain reduction and the distribution of the given data. Accordingly, experiments were run each with pain outcome thresholds of +12%, +20%, and +30% improvement to observe which would lead to the best model performance (as outlined in Section 3.2.2.3). Following the same process as KNN, these models used LOOCV to fully utilize the data for training and testing purposes. To evaluate the model's predictive performance, test accuracy and confusion matrices were also calculated.

3.3.5 XG Boosting

The XG Boosting models were trained using the input data explained in Table 5. Additional datasets were created to run experiments on XG Boosting, as this modeling technique worked best with our data. The new data that was engineered for these experiments are outlined in Table 5. The output that was being predicted was the McGill Pain Reduction percentage. The

models were trained over the two separate groups, as well as the groups combined. A new data subset that was created for XG Boosting modeling was "shifted baseline data" for Group 2 to account for the two-month period at the beginning of the study that this group did not receive treatment. Specifically, Group 2's Timepoint 2 data was considered the new baseline values of pain. This "shift" included recalculating the delta value for the McGill Pain Reduction score to compute using Timepoint 2 as the initial value (T4 – T2). After recalculating the outcome values, the TCM Diagnoses data was appended with both the Group 1 and Group 2 datasets. Random Forest and Lasso Regression were run on that data to obtain values used for feature selection.

Table 5. Additional Data Subsets from Feature Selection Utilized for XG Boosting

Features Selected From	Group	# of Attributes Included	# of Attributes Removed	# of Patients
Original	1 & 2	34	0	80
Shifted Baseline	2	34	0	39
Shifted Baseline, Lasso	2	9	25	39
Shifted Baseline, RF	2	16	18	39
Shifted Baseline, Lasso ∩ RF	2	7	27	39
Both Groups Combined, Lasso	1 & 2	6	28	80
Both Groups Combined, RF	1 & 2	10	24	80
Both Groups Combined, Lasso ∩ RF	1 & 2	4	30	80

Another dataset that was used for only XG Boosting training was a dataset that consisted of Group 1 Timepoint 1 data with appended Group 2 Timepoint 2 data (Table 5). A new attribute was added to this new dataset to identify which patient came from which group. TCM Diagnoses data was also appended into the dataset. This data was then run through Random Forest and Lasso

Regression to obtain values for feature selection. The output that was being predicted was the McGill Pain Reduction percentage. The pain reduction for all patients in Group 2 was recalculated so that Timepoint 2 would be considered their baseline pain score instead of Timepoint 1. Test accuracy, F1 scores, and confusion matrices were calculated to measure how well the model was performing.

4. Results

4.1 Feature Selection

4.1.1 Feature Selection with Lasso Regression

Tables 6 and 7 show the features in decreasing order of the (absolute value) coefficients produced from the models using Lasso Regression for Group 1 and Group 2 respectively.

Table 6. Lasso Regression coefficients using Timepoint 1 attribute values as input and McGill Pain Reduction Percentage as output on Group 1

Attribute	Coefficient
McPain..Pain.score..T1.	34.305116
SF_PHYS..Physical.Functioning..T1.	27.816196
SNI..Support.Networks.High.Contact.Role..T1.	-19.280454
LCTR..Locus.of.Control...Powerful..T1.	19.10707
PSQI..Sleep.Daytime.Dysfunction.C7..T1.	17.185981
CDEP.Work.Interests..T1.	15.851624
OPTI..Optimism..T1.	-14.28442
PSQI..Subjective.Sleep.Quality.C1..T1.	12.094371
CDEP.Retardation..T1.	-10.389943
CDEP.Depression..T1.	-9.3270251
McPain_a..Affective.Pain.score..T1.	7.692115

PSQI..Sleep.Medications.C6..T1.	6.2181086
MYMOP..Self.reported.Medical.Outcome..T1.	-4.0800111
BCX..Body.Consciousness..T1.	-1.952934
CDEP.Psy.Anxiety..T1.	1.6491213
POMS..Mood.States..T1.	0.2065523

Table 7. Lasso Regression coefficients using Timepoint 1 attribute values as input and McGill Pain Reduction Percentage as output on Group 2

Attribute	Coefficient
CAT..Catastrophizing..T1.	40.031677
PSQI..Sleep.Daytime.Dysfunction.C7..T1.	-29.841435
CDEP.Retardation..T1.	28.531807
PSQI..Sleep.Disturbances.C5..T1.	-26.471346
SCHN..Stress..T1.	-25.687251
CDEP.Depression..T1.	6.6761321
CDEP.Work.Interests..T1.	-4.2347112
CDEP.Psy.Anxiety..T1.	-3.7659624
BCX..Body.Consciousness..T1.	0.2012609

4.1.2 Feature Selection with Random Forest

Tables 8 and 9 show the features for Group 1 and Group 2 respectively, sorted in decreasing order of predictive importance as measured by the increase in the Mean Squared Error value produced when a feature is removed from consideration in the construction of Random Forest models.

Table 8. Random Forest output using Timepoint 1 attribute values as input and McGill Pain Reduction percentage as output on Group 1

Attribute	Increase in Mean Squared Error
POMS..Mood.States..T1.	65.514135
McPain..Pain.score..T1.	34.405286
SNI..Support.Networks.Num.People..T1.	26.676747
McPain_a..Affective.Pain.score..T1.	20.087733
CDEP.Guilt..T1.	17.495748
OPTI..Optimism..T1.	15.879208
CDEP.Psy.Anxiety..T1.	15.483043
CDEP.Retardation..T1.	12.631556
ISEL..Social.Support..T1.	12.212556
sumFI..Fatigue..T1.	12.104012
PSQI..Subjective.Sleep.Quality.C1..T1.	11.646765
LCTR..Locus.of.Control...Chance..T1.	11.534417

SOC..Social.Support.Open..T1.	7.8988537
BANX..Anxiety..T1.	5.4351759
McPain_s..Sensory.Pain.score..T1.	5.4117133
BCX..Body.Consciousness..T1.	4.4306142
MYMOP..Self.reported.Medical.Outcome..T1.	4.0834094
PSQI..Sleep.Medications.C6..T1.	3.9236771
CDEP.Agitation..T1.	3.9071771

Table 9. Random Forest output using Timepoint 1 attribute values as input and McGill Pain Reduction percentage as output on Group 2

Attribute	Increase in Mean Squared Error
CAT..Catastrophizing..T1.	105.8876
LCTR..Locus.of.Control...Internal..T1.	44.154603
CDEP.Retardation..T1.	21.933869
POMS..Mood.States..T1.	19.651913
CDEP.Psy.Anxiety..T1.	18.767761
MYMOP..Self.reported.Medical.Outcome..T1.	15.702524
SCHN..Stress..T1.	12.146609
PSQI..Sleep.Disturbances.C5..T1.	12.063518

CDEP.Agitation..T1.	9.1321974
LCTR..Locus.of.Control...Powerful..T1.	9.0829226
BCX..Body.Consciousness..T1.	8.5886245
PSQI..Sleep.Medications.C6..T1.	8.3178172
SNI..Support.Networks.High.Contact.Role..T1.	7.897154
CDEP.Work.Interests..T1.	3.1674251
LCTR..Locus.of.Control...Chance..T1.	3.0122292
sumFI..Fatigue..T1.	2.7611721
PSQI..Sleep.Daytime.Dysfunction.C7..T1.	2.1993682
CDEP.Guilt..T1.	1.1477933
McPain..Pain.score..T1.	1.0967201

4.1.3 Feature Selection Summary

Based on the feature selection conducted using both Lasso Regression and Random Forest values from the tables above, the features that appeared most frequently in the top-ranked feature lists (see Appendix A and B for full lists) were from the following questionnaires: McGill Pain Scale, Pittsburgh Sleep Quality Index (PSQI), and Carroll Depression Scale (CDEP). Values from these questionnaires are important to predicting the pain outcome for veterans.

4.2 Linear Regression

Linear Regression was found to be powerful for predicting the McGill Pain Reduction percentage. Table 10 shows the results of the Linear Regression experiments.

Table 10. Linear Regression Results

Group	# of patients	Attributes selected by ML method	R^2	MAE	MSE
1	44	Lasso regression	0.44	12.87	254
1	41	Lasso ∩ RF, TCM data	-0.14	16.44	341
2	39	Lasso ∩ RF	0.37	17.32	450
2	39	Lasso ∩ RF, TCM data	0.39	17.52	433

The attributes selected using Lasso Regression for Group 1 proved to be the most predictive, while the inclusion of TCM Diagnoses only seemed to hinder the performance. For Group 2, the most accurate model was trained on attributes at the intersection of Lasso Regression and Random Forest, together with an attribute that includes TCM Diagnoses. The models trained on attributes selected by Random Forest alone resulted in worse performance compared to what would be expected by simply predicting the mean change in pain over the patients in the dataset.

Due to the considerable variability in the regressors for the models using Linear Regression, as demonstrated in Table 10, we could not derive conclusions directly from the models. A full list of regressors can be found in Appendix C.

4.3 Logistic Regression

Figure 7 shows the performance of the most accurate Logistic Regression model for Group 1. Trained on attributes selected by Lasso Regression, this model had an accuracy of 66%. It correctly categorized 22 out of the 26 patients labeled as "No Significant Improvement", and correctly predicted 7 out of the 18 patients that were categorized as "Significant Improvement."

Thresholds above 12% resulted in models that classified all patients as having "No Significant Improvement."

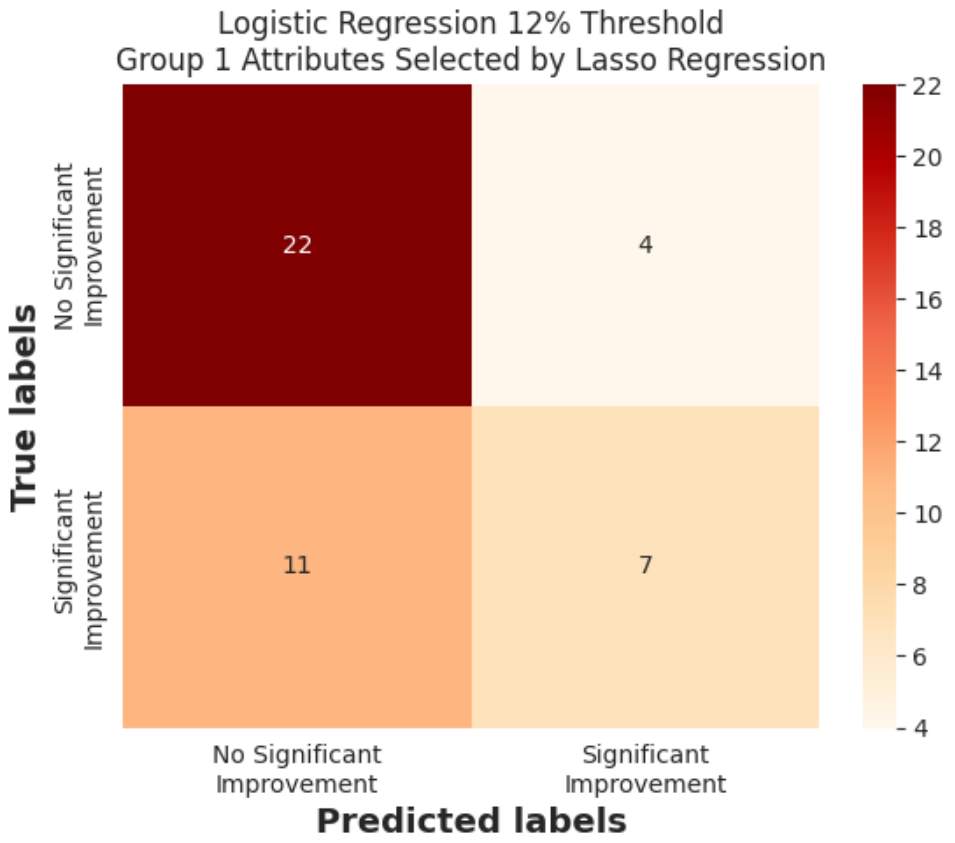

Figure 7. Logistic Regression Confusion Matrix for Group 1 data selected by Lasso Regression

Table 11 presents the values of regressors for the top-performing Logistic Regression model, along with those from a poor performing model. These regressors were selected to highlight the large difference in the values of the coefficients between the models. Features that were deemed important from these regressors for the best performing model include the Carroll Depression Scale (CDEP), the Locus of Control questionnaire, and the SF-36 Pain questionnaire. These features were the highest weighted in the better performing models. A full list of regressors can be found in Appendix D.

Table 11. Select Logistic Regression Regressors for Group 1

	Coefficients From Model Trained on Lasso Attributes	Coefficients From Model Trained on Lasso ∩ RF + TCM data
MYMOP	-0.37	-0.49
McGill Pain	0.50	0.43
PSQI Subjective Sleep Quality	0.63	0.40
SF-36	0.75	
Locus of Control Powerful	0.77	
Carroll Depression	-0.94	

In Figure 8, it shows the confusion matrix of the most accurate Logistic Regression model for Group 2. Trained on data that included TCM Diagnoses one-hot encoded, this model had an accuracy of 79%, correctly categorizing all 26 of the patients labeled as "No Significant Improvement," and correctly predicted 4 out of 12 of the patients who were categorized as "Significant Improvement." Thresholds above 12% resulted in models that classified all patients as having "No Significant Improvement."

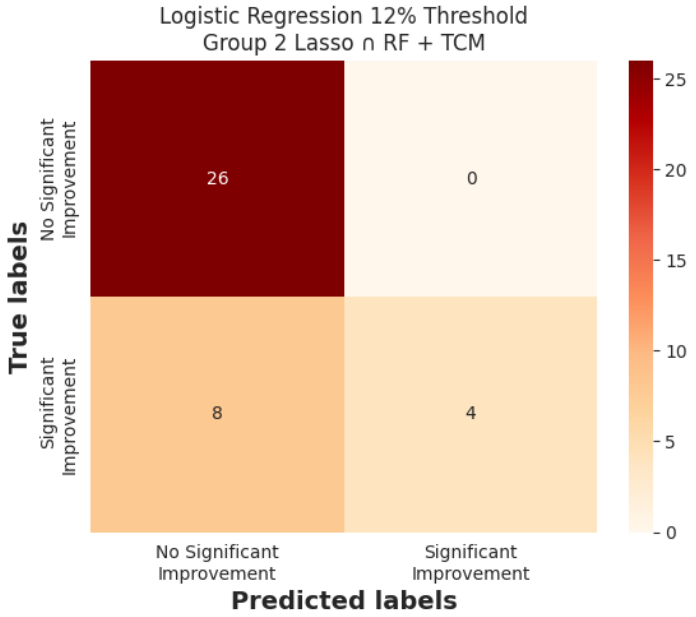

Figure 8. Logistic Regression Confusion Matrix for Group 2 with TCM Data

4.4 K-Nearest Neighbors

Table 12 shows the results obtained with the K-Nearest Neighbors technique.

Table 12. KNN Test Accuracy for Different Pain Outcome Thresholds and Datasets. Each prediction in the LOOCV was aggregated into a confusion matrix, and the accuracies were calculated from here. Test accuracies and thresholds for each dataset are listed.

Group	Data	Test Accuracy (Threshold: 12%, 20%, 30%)
1	Lasso	59.09% 72.73% 90.91%
1	RF	56.82% 75% 90.91%
1	Lasso ∩ RF	52.27% 72.73% 90.91%
1	Lasso ∩ RF, TCM	51.16% 72.09% 90.70%
2	Lasso	74.36% 82.05% 84.62%
2	RF	61.54% 76.92% 79.49%
2	Lasso ∩ RF	61.54% 76.92% 79.49%
2	Lasso ∩ RF, TCM	64.10% 76.92% 79.49%

For Group 1, KNN worked best on attributes selected by Lasso Regression. The model was 59% accurate at a 12% threshold and had an F1 score of 0.41. The confusion matrix for this model can be seen in Figure 9.

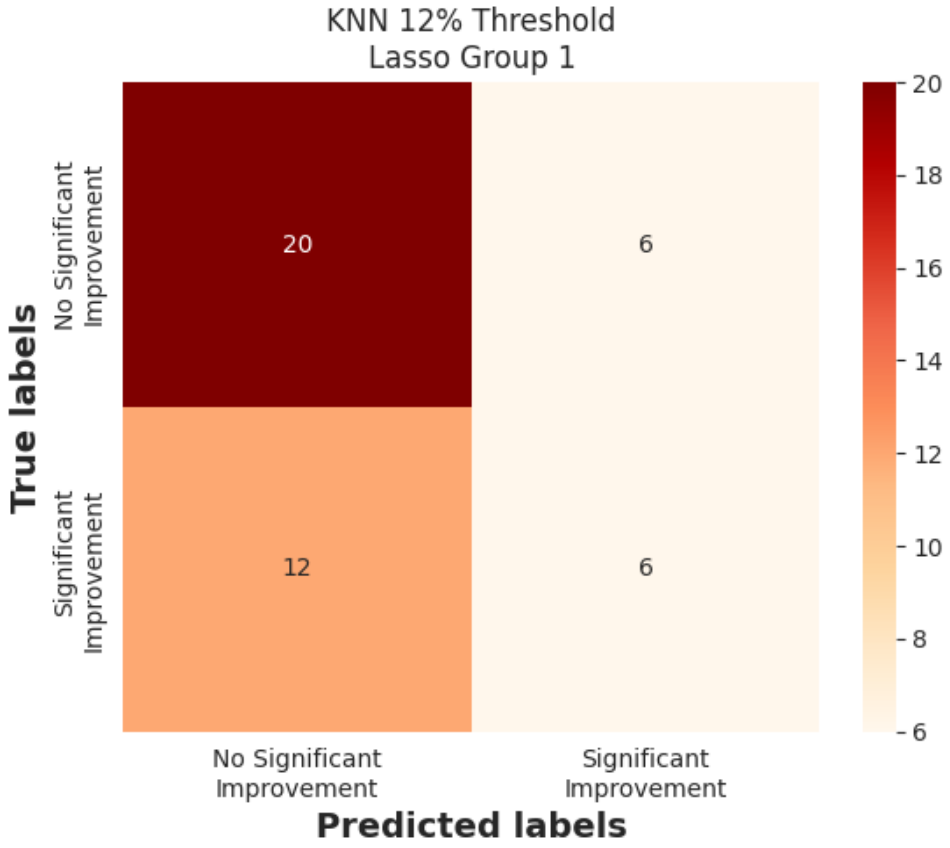

Figure 9. KNN Confusion Matrix on Lasso Group 1, 12% Threshold

At a 20% threshold, the model does not predict that any patients have gotten better and gives an F1 score of zero. The model predicted that every patient had gotten worse except for one, who was classified incorrectly. These predictions were very close to the model that predicted the majority class. The predictive power of the model is low, despite its accuracy of 73%. The confusion matrix for this model can be seen in Figure 10.

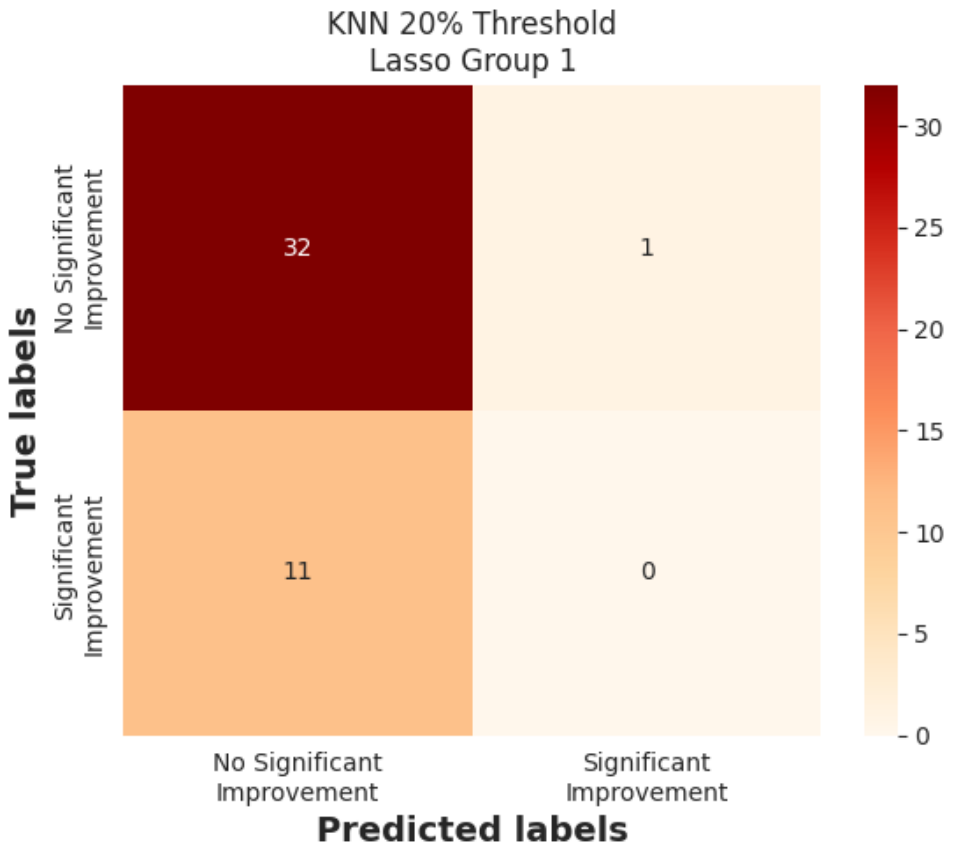

Figure 10. KNN Confusion Matrix on Lasso Group 1, 20% Threshold

The model created using the Random Forest variables performed second-to-best for predicting the pain change of patients in Group 1. The correlation matrix of this model ran at a 12% threshold percent and is shown in Figure 11. The F1 score was 0.42, with an accuracy of 56.8%.

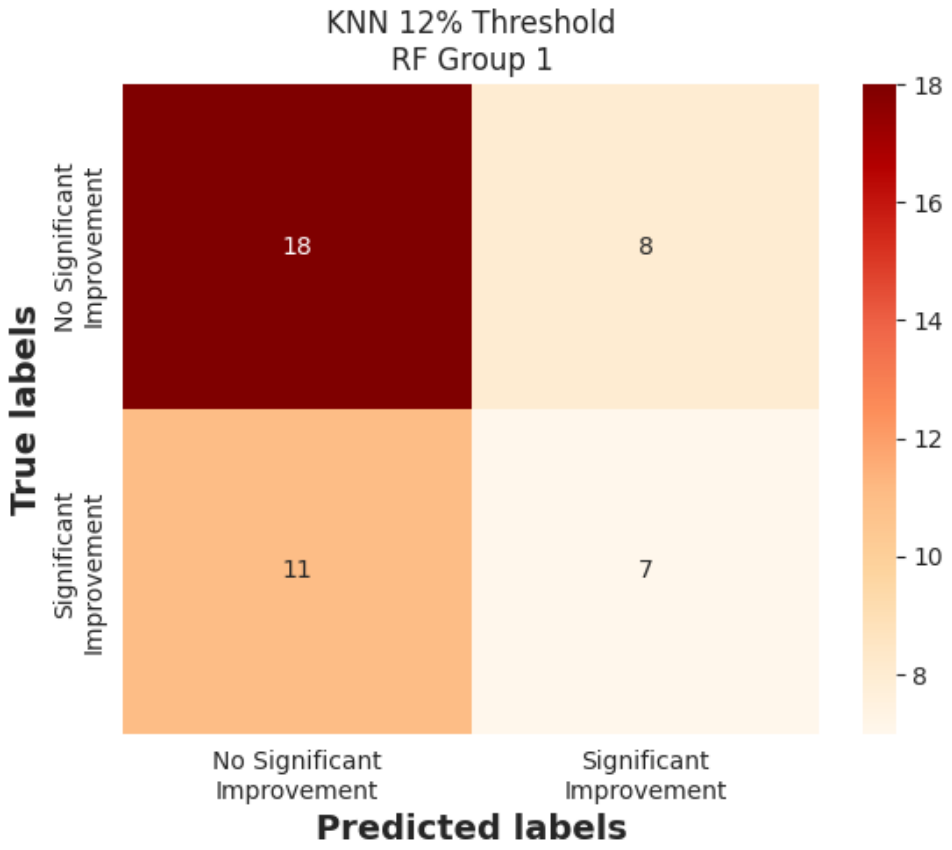

Figure 11. KNN Confusion Matrix on RF Group 1, 12% Threshold

The accuracy of using Random Forest coefficients was 75% at a threshold of 20%. The F1 score in this case, however, is zero. This model has weak predictive power as it predicts that all patients have not gotten better when, in reality, 33% of them have, so although the accuracy of this model is high, it is not a good model because of the incorrect predictions. The confusion matrix for this model can be seen in Figure 12.

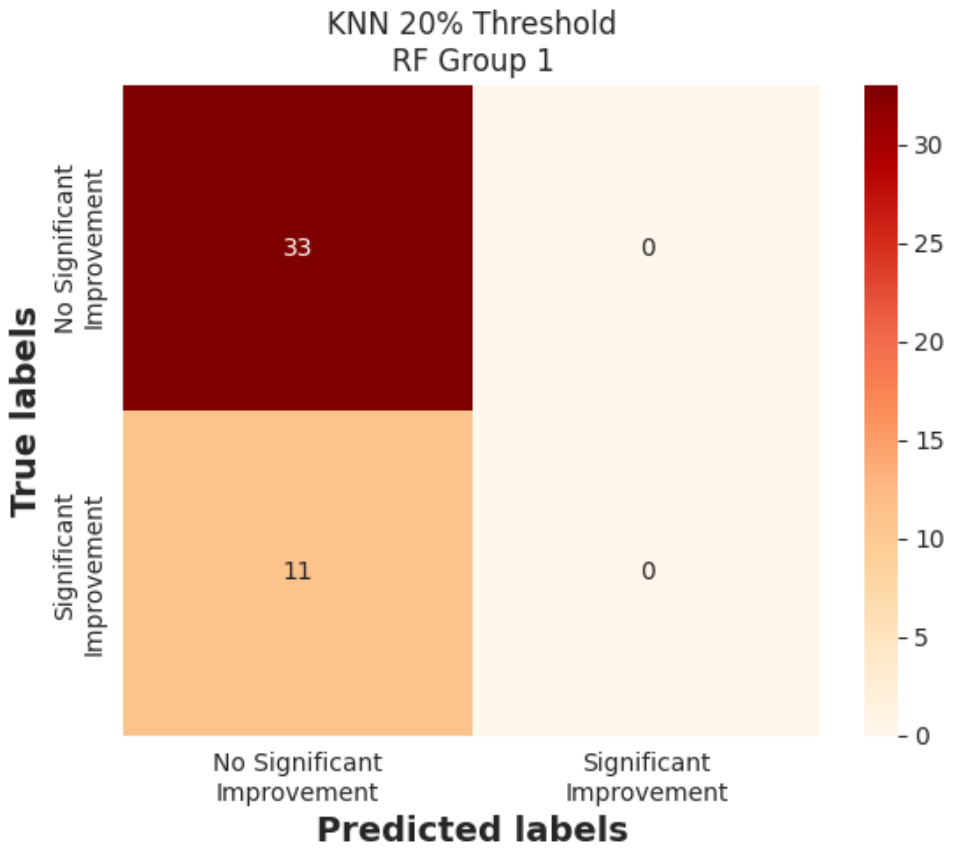

Figure 12. KNN Confusion Matrix on RF Group 1, 20% Threshold

The best KNN model for predicting values for Group 2 used the field values generated by the Random Forest model. This was the only model that did not return a F1 score of 0 for thresholds of 20% and 30%. This model's confusion matrix is shown in Figure 13, and the F1 score for it was 0.375 at the threshold of 12%.

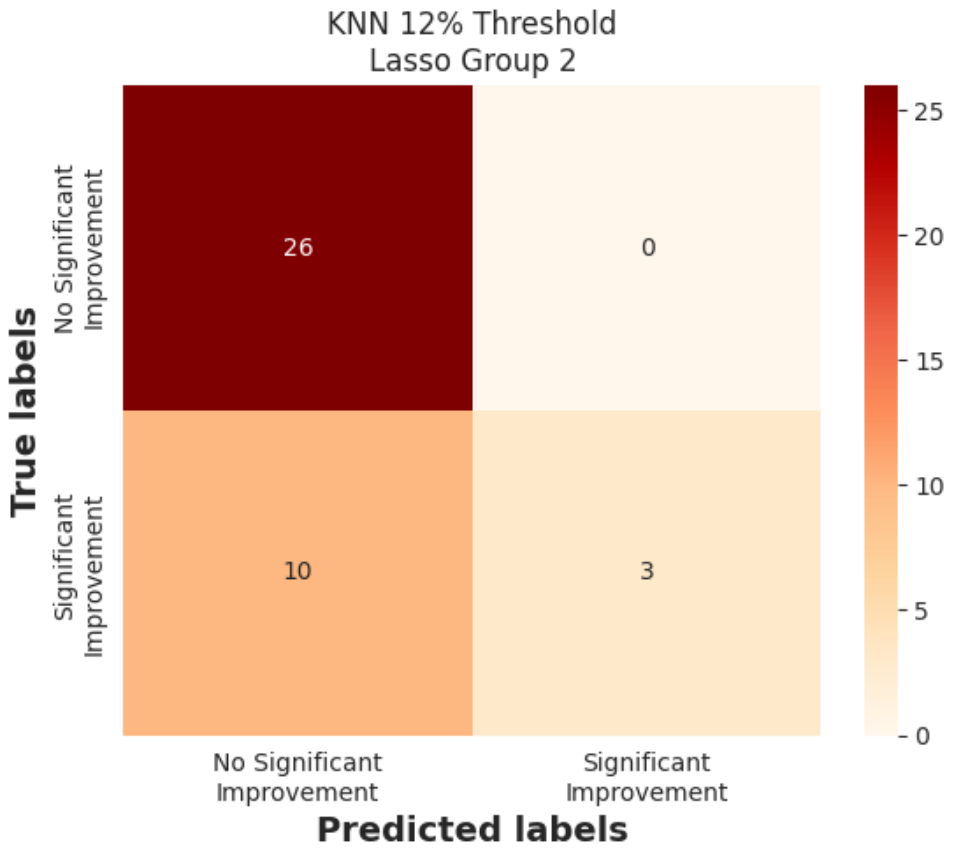

Figure 13. KNN Confusion Matrix on RF Group 2, 12% Threshold

4.5 Softmax Regression

Table 13 presents the results obtained on the Softmax regression experiments.

Table 13. Softmax Regression Test Accuracy for Different Pain Outcome Thresholds and Datasets. Each prediction in the LOOCV was aggregated into a confusion matrix, and the accuracies were calculated from here. Test accuracies and thresholds for each dataset are listed.

Group	Data	Test Accuracy (Threshold: 12%, 20%, 30%)
1	Lasso	34.09%, 31.82% 50.00%
1	RF	27.27% 34.09% 45.45%
1	Lasso ∩ RF	29.55% 38.64% 47.73%
1	Lasso ∩ RF, TCM	32.56% 25.58% 58.14%
2	Lasso	38.46% 33.33% 38.46%
2	RF	28.21% 41.03% 38.46%
2	Lasso ∩ RF	35.90% 30.77% 33.33%
2	Lasso ∩ RF, TCM	30.77% 30.77% 41.03%

For Group 1, the Softmax method worked best on attributes selected by the intersection of Lasso Regression and Random Forest joined with the TCM data. The model was 58% accurate at a 30% threshold, and sufficiently predicted which patients fell into the second category (improvement below the 30% threshold but above 0). However, it is important to note that a large portion of individuals who got worse were incorrectly classified as seeing some improvement, as depicted in Figure 14.

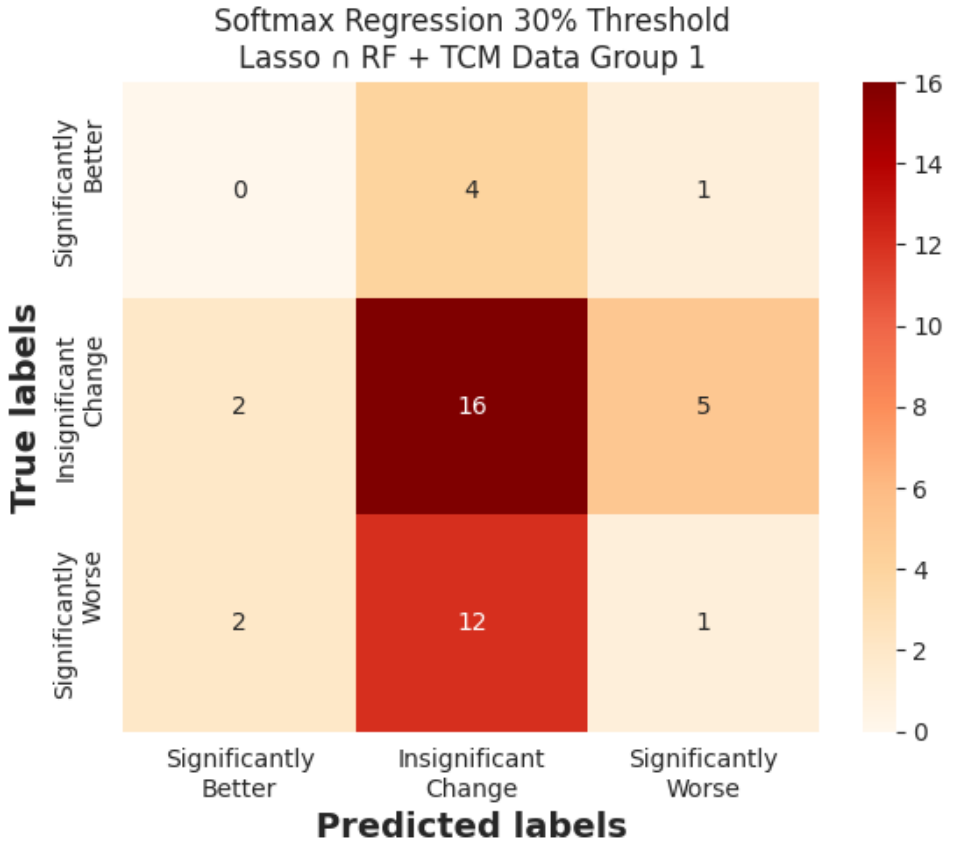

Figure 14. Softmax Regression Confusion Matrix on Lasso ∩ RF + TCM Data Group 1, 30% Threshold

For Group 1, the Softmax method's next-best performance was on attributes selected by only Lasso Regression. The model was 50% accurate and sufficiently predicted which patients got insignificantly better at a 30% threshold. However, the accuracy at predicting which patients have gotten significantly better was not as high as depicted in Figure 15.

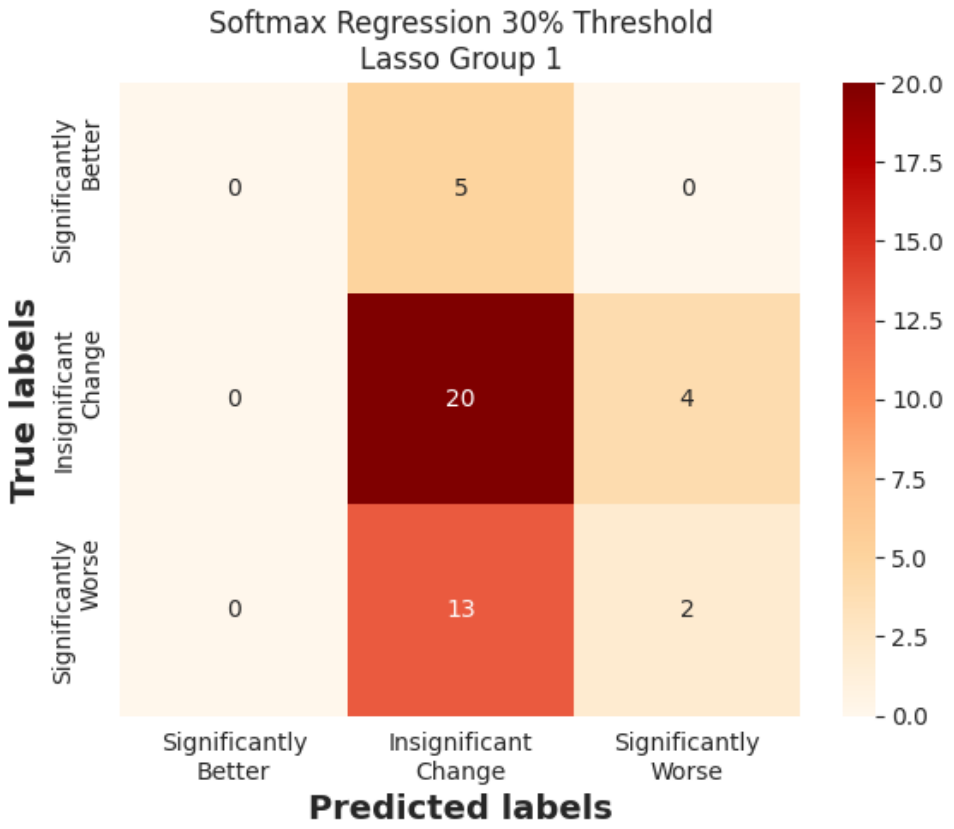

Figure 15. Softmax Regression Confusion Matrix on Lasso Group 1, 30% Threshold

For Group 2, the Softmax method's best performance was on attributes selected by Random Forest at a 20% threshold and on the intersection of Lasso Regression and Random Forest joined with the TCM data at a 30% threshold. Both of these experiments yielded a 41% test accuracy (see Table 13). For Random Forest at the 20% threshold, the model was able to predict correctly for each of the 3 classes (with varying accuracy), including nearly half of the patients who saw significant improvement. It is important to note that for the class "Worse," the model classified a majority of these patients as "Insignificantly Better," as depicted in Figure 16.

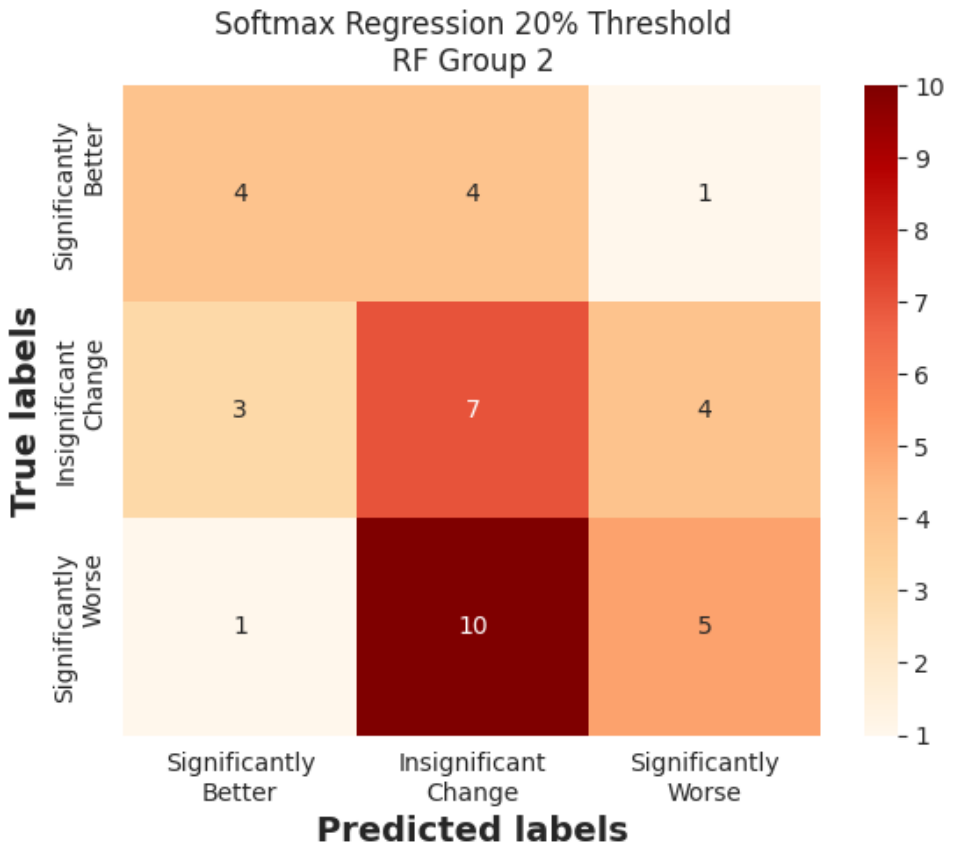

Figure 16. Softmax Regression Confusion Matrix on RF Group 2, 20% Threshold

For the intersection of Lasso Regression and Random Forest joined with the TCM data at a 30% threshold (see Figure 17), the model was also able to accurately predict some from each class, most prominently for the insignificant improvement classification. However, it performed poorly for classifying the third category ("Worse"), a majority of which were incorrectly classified as the second category.

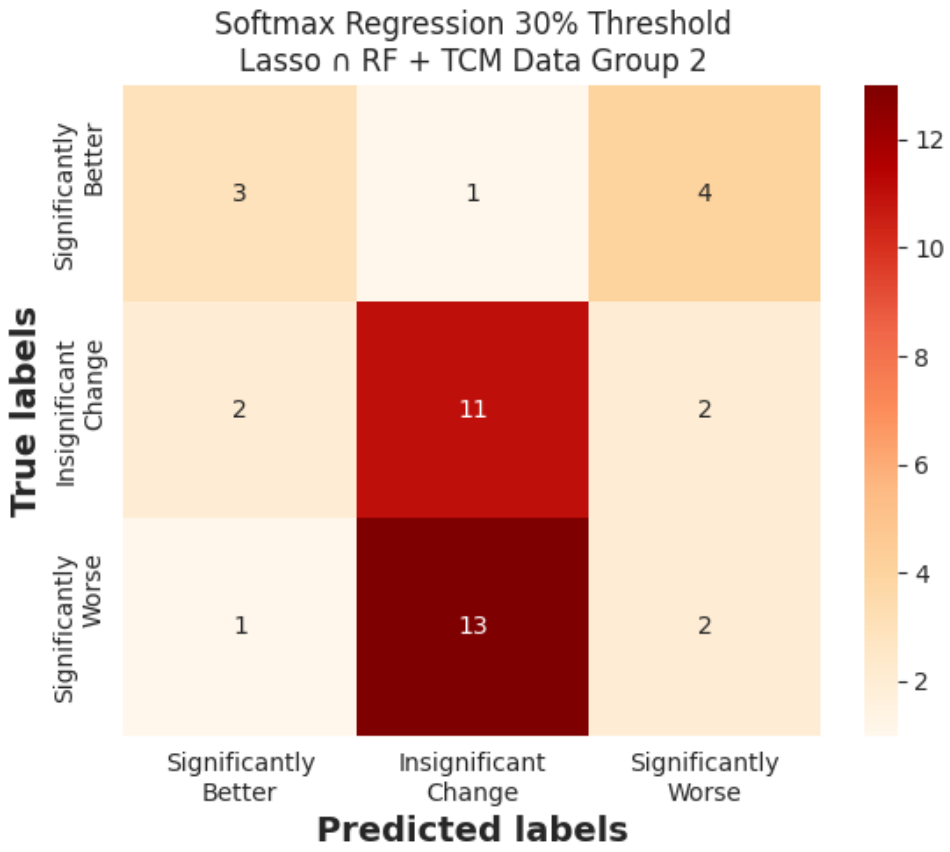

Figure 17. Softmax Regression Confusion Matrix on Lasso ∩ RF Group 2 with TCM Data, 30% Threshold

Overall, for both Groups 1 and 2, with the varying thresholds and data, a common trend that was observed in most confusion matrices was that the "Worse" true labels were classified as "Insignificantly Better" by the model. In addition, out of the 12 experiments for Group 1, there were seven that accurately classified patients with significant improvement for at least one patient. For Group 2, 10 out of 12 were classified correctly for significant improvement. Although the accuracy was not the best compared to other Group 2 experiments, the Softmax method on only Lasso Regression attributes at the 12% threshold yielded the highest number of correct predictions for significant improvement (see Figure 18).

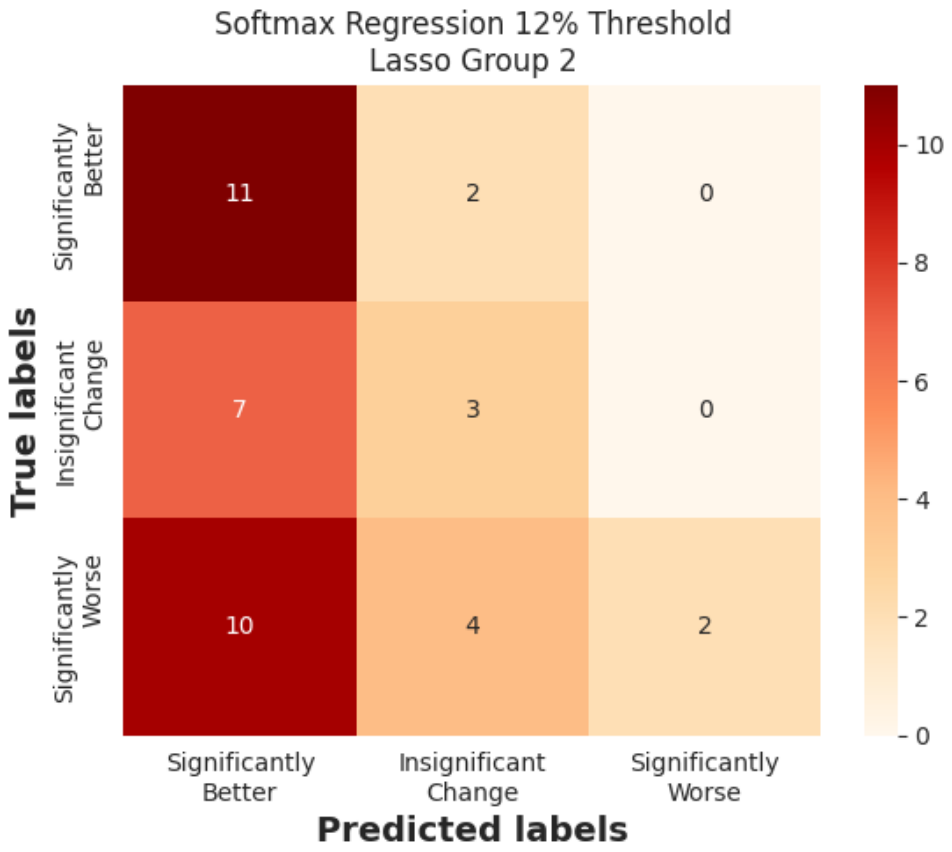

Figure 18. Softmax Regression Confusion Matrix on Lasso Group 2, 12% Threshold

4.6 XG Boosting

The results of the XG Boosting experiments are presented in Table 14.

Table 14. XG Boosting Test Accuracy at Different Pain Outcome Thresholds and Datasets. Each prediction in the LOOCV was aggregated into a confusion matrix, and the accuracies were calculated from here. Test accuracies and thresholds for each dataset are listed.

Group	Data	Test Accuracy (Threshold: 12%, 20%, 30%)	F1 Score (Threshold: 12%, 20%, 30%)
1	Lasso	56.82% 61.82% 84.90%	0.4571 0.3000 0.0000
1	RF	50.00%, 70.45% 86.36%	0.3889 0.2353 0.0000
1	Lasso ∩ RF	54.55%, 70.45% 88.64%	0.4118 0.3810 0.0000
1	Lasso ∩ RF, TCM	60.98%, 68.29% 82.93%	0.5000 0.3810 0.0000
2	Lasso	58.97%, 79.50% 87.18%	0.2727 0.4286 0.6154
2	Lasso (TCM added into Lasso selection), Group 2 shifted to start at T2	56.41%, 66.67% 76.92%	0.2609 0.2353 0.4000
2	RF	53.85%, 76.92% 82.05%	0.1818 0.3077 0.3636

2	RF (TCM added into RF selection), Group 2 shifted to start at T2	53.85%, 61.54% 71.79%	0.2500 0.1176 0.1167
2	Lasso ∩ RF (TCM added into Lasso and RF selection), Group 2 shifted to start at T2	56.41%, 76.92% 82.05%	0.2609 0.4000 0.4615
2	Lasso ∩ RF, TCM	56.41%, 79.49% 87.18%	0.1905 0.4286 0.6154
1 & 2	Lasso ∩ RF (TCM added into Lasso and RF selection), Group 2 shifted to start at T2	60.00%, 61.25% 81.25%	0.4483 0.1143 0.1176
1 & 2	RF (TCM added into RF selection), Group 2 shifted to start at T2	66.25%, 70.00% 77.50%	0.4906 0.2500 0.0000
1 & 2	Lasso (TCM added into Lasso selection), Group 2 shifted to start at T2	61.25%, 62.50% 80.00%	0.4561 0.1667 0.0000
1 & 2	All T1 features, Group 2 shifted to start at T2	61.25%, 68.75% 80.00%	0.3922 0.1935 0.0000

For Group 1, the confusion matrix for the best performing model can be seen in Figure 19. The data this model was run on had features selected by Lasso Regression as well as Random Forest. Additional TCM data was then appended to the data. For Group 1, the confusion matrix for this model (seen in Figure 19) does not have the highest accuracy, but it does have the most values predicted correctly for patients who got better at the given threshold. The accuracy for the

confusion matrix below is calculated to be 60.98% at a 12% threshold, with an F1 score of 0.50.

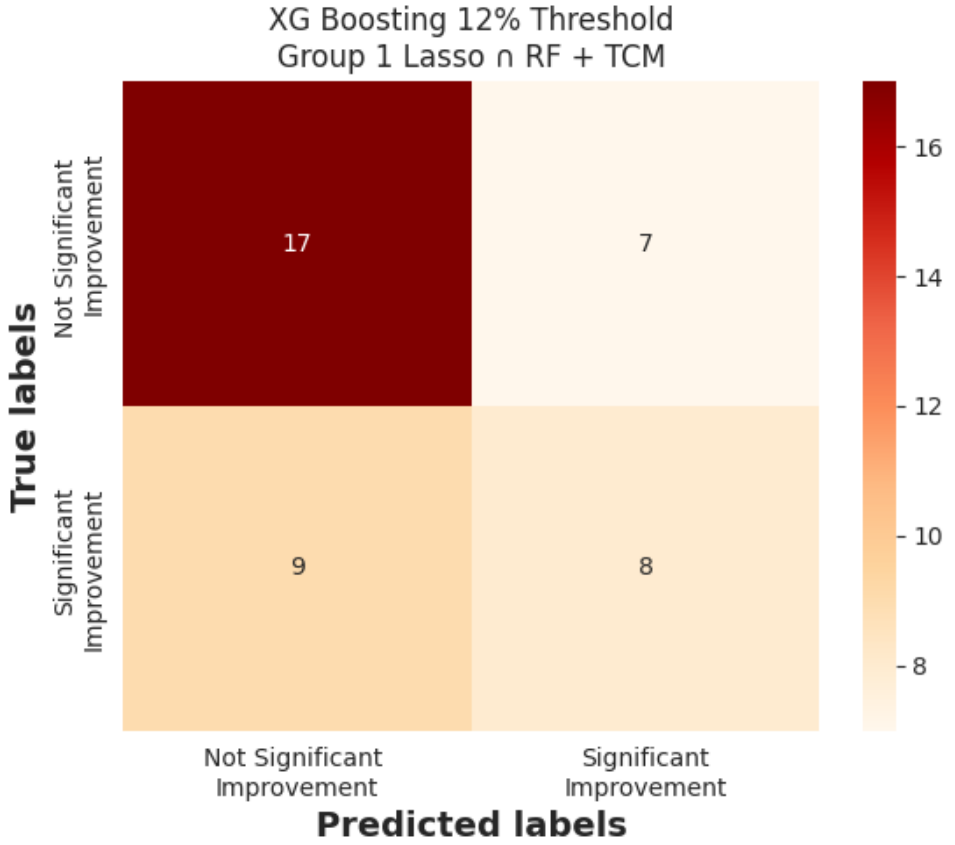

Figure 19. XG Boosting Confusion Matrix on Lasso ∩ RF, TCM Group 1, 12% Threshold

As explained prior, some feature selection for XG Boosting was done with the TCM data included. The TCM Diagnoses feature was considered significant during Lasso feature selection and hence was part of the model data for the XG Boosting iteration, whose confusion matrix is shown in Figure 20. In general, predicting which patients have gotten better in Group 2 was much more successful than it was in Group 1. For the XG Boosting technique, at the 30% threshold level, not all implementations had a True Positive rate of 0, but for Group 2, almost every model had a True Positive Rate greater than 0. A possibility for this could be that 8 people in Group 2 had a pain reduction of 30% or above as opposed to the 4 in Group 1. Therefore, models that ran on Group 2 encountered more training data for patients who had gotten better at the 30% threshold and therefore had more predictive power. The accuracy of the model shown below in Figure 20 is 76.92%, and it has an F1 score of 0.40.

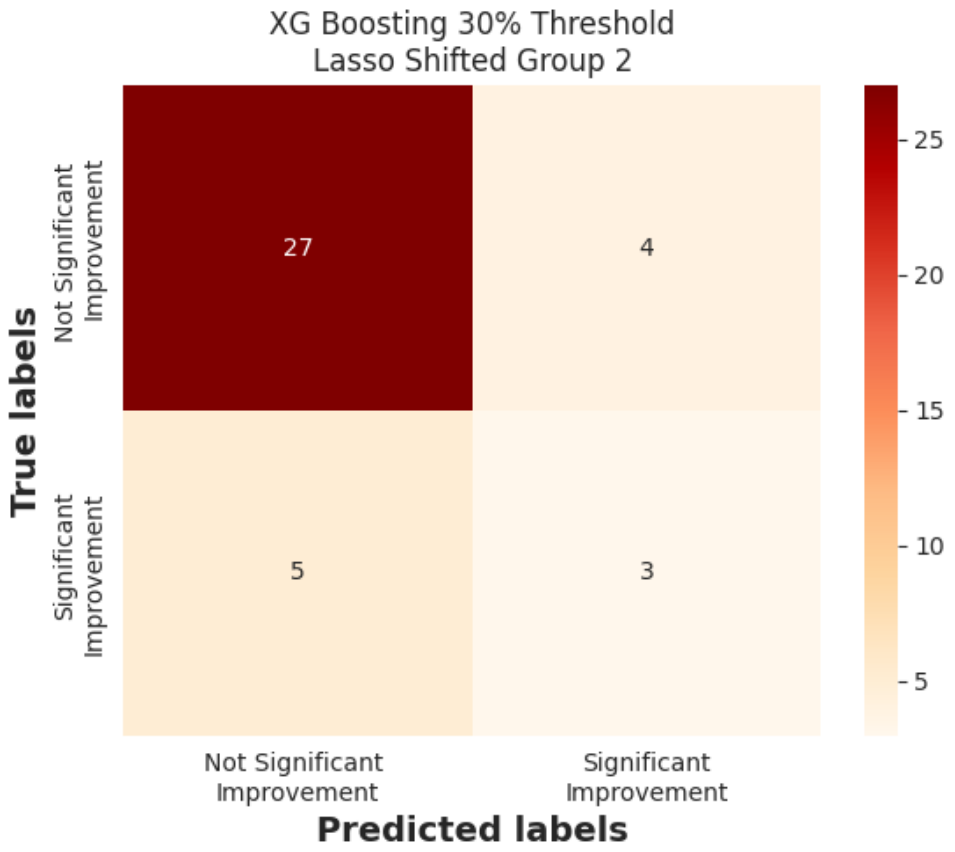

Figure 20. XG Boosting Confusion Matrix on Lasso Group 2 Shifted, 30% Threshold

4.7 Results Summary

To recommend best treatment for veterans, it is important to maintain the highest precision (Hicks et al., 2022). For Group 1, the best performing model was trained using Logistic Regression on the attributes selected by Lasso Regression. As seen in Table 15, the precision of the model was 64%. Similarly, for Group 2, the best performing model was trained using Logistic Regression, but trained on the data that included TCM Diagnoses. As seen in Table 15, the precision was 100%, and F1 score was 0.50. As stated in Sections 4.1.3 and 4.3, some of the most prevalent questionnaires that can be used for accurately recommending acupuncture include Pittsburgh Sleep Quality Index, the Locus of Control questionnaire, the SF-36 questionnaire, the Carroll Depression Scale, and McGill Pain Scale.

Table 15. Best Performing Binary Classification Models from Each Modeling Technique

Technique	Training Data	Group	Threshold	Precision	F1 score	Recall	Accuracy
Logistic Regression	Lasso	1	+12%	64%	0.48	39%	66%
KNN	RF	1	+12%	47%	0.42	39%	65%
XG Boosting	Lasso ∩ RF, TCM	1	+12%	53%	0.50	47%	61%
Logistic Regression	Lasso ∩ RF, TCM	2	+12%	100%	0.50	39%	79%
KNN	RF	2	+12%	100%	0.38	23%	74%
XG Boosting	Lasso	2	+12%	30%	0.27	23%	59%

5. Conclusions

The overall goal of this project was to predict pain response to acupuncture treatment for veterans suffering from Gulf War Illness (GWI). We developed both regression and classification models to predict the pain reduction outcome for patients over the course of the four timepoints in the study. To enhance these predictive models, the most important features were chosen as input by their predictive strength of the pain reduction outcome. Through a number of experiments, we attempted to model how different features of veterans' health contribute to their response to acupuncture treatment (predicting the change in pain). The best performing model produced, in terms of precision and accuracy, used the Logistic Regression method with input data of features selected from Lasso Regression for Group 1, which produced a precision of 64%, and an F1 score of 0.4830. The key factors for predicting the pain outcome were found to be in the McGill Pain Scale, the Pittsburgh Sleep Quality Index, the Locus of Control questionnaire, the SF-36 questionnaire, and Carroll Depression questionnaire.

To improve the effectiveness of modeling how acupuncture impacts chronic pain in veterans with GWI, it could prove helpful to acquire additional data about the patients and train more models. Because the training data utilized in this project consisted of less than 50 patients per group, the addition of more data could also enhance the generalization of our models and reduce the risk of overfitting. Furthermore, other classification techniques such as support vector machines and naïve Bayes classifiers could offer more insight into the effectiveness of acupuncture for treating Gulf War Illness. Overall, more investigation of effective treatments for GWI is necessary to improve the well-being of Gulf War veterans with chronic pain.

Appendix A: Lasso Regression Output Equations

The tables below show the equations from the Lasso Regression method used for feature selection.

Table A1. *Lasso Regression coefficients using Timepoint 1 attribute values as input and McGill Pain Reduction Percentage as output on Group 1.*

Attribute	Coefficient
McPain..Pain.score..T1.	34.305116
SF_PHYS..Physical.Functioning..T1.	27.816196
SNI..Support.Networks.High.Contact.Role..T1.	-19.280454
LCTR..Locus.of.Control...Powerful..T1.	19.10707
PSQI..Sleep.Daytime.Dysfunction.C7..T1.	17.185981
CDEP.Work.Interests..T1.	15.851624
OPTI..Optimism..T1.	-14.28442
PSQI..Subjective.Sleep.Quality.C1..T1.	12.094371
CDEP.Retardation..T1.	-10.389943
CDEP.Depression..T1.	-9.3270251
McPain_a..Affective.Pain.score..T1.	7.692115
PSQI..Sleep.Medications.C6..T1.	6.2181086
MYMOP..Self.reported.Medical.Outcome..T1.	-4.0800111
BCX..Body.Consciousness..T1.	-1.952934

CDEP.Psy.Anxiety..T1.	1.6491213
POMS..Mood.States..T1.	0.2065523

Table A2. *Lasso Regression coefficients using delta attribute values as input and McGill Pain Reduction Percentage as output on Group 1.*

Attribute	Coefficient
SNI..Support.Networks.Num.People..D.	98.392821
SNI..Support.Networks.Num.Embedded..D.	-83.925385
WAI..Working.Alliance.Task..D.	38.356932
CDEP.Retardation..D.	-36.040853
SNI..Support.Networks.High.Contact.Role..D.	34.685764
LCTR..Locus.of.Control...Internal..D.	-26.955414
ISEL..Social.Support..D.	-26.923819
CDEP.Depression..D.	-26.007694
POMS..Mood.States..D.	24.045611
PSQI..Sleep.Disturbances.C5..D.	24.008163
SCHN..Stress..D.	22.140698
PSQI..Sleep.Medications.C6..D.	20.710779
CDEP.Agitation..D.	-17.966602

sumFI..Fatigue..D.	16.389921
CDEP.Guilt..D.	15.730232
WAI..Working.Alliance.Goal..D.	12.992023
PSQI..Subjective.Sleep.Quality.C1..D.	11.889857
LCTR..Locus.of.Control...Chance..D.	-8.631689
LCTR..Locus.of.Control...Powerful..D.	7.6654509
OPTI..Optimism..D.	6.1441566
WDEP..Whitely.Depression..D.	-4.371798
WAI..Working.Alliance.Bond..D.	3.9802694
BCX..Body.Consciousness..D.	3.9602985
BANX..Anxiety..D.	-3.9297184
CAT..Catastrophizing..D.	1.5874542
PSQI..Sleep.Daytime.Dysfunction.C7..D.	1.1883199
SF_PHYS..Physical.Functioning..D.	-0.4178776

Table A3. *Lasso Regression coefficients using Timepoint 1 attribute values as input and McGill Pain Reduction Percentage as output on Group 2.*

Attribute	Coefficient
CAT..Catastrophizing..T1.	40.031677
PSQI..Sleep.Daytime.Dysfunction.C7..T1.	-29.841435
CDEP.Retardation..T1.	28.531807
PSQI..Sleep.Disturbances.C5..T1.	-26.471346
SCHN..Stress..T1.	-25.687251
CDEP.Depression..T1.	6.6761321
CDEP.Work.Interests..T1.	-4.2347112
CDEP.Psy.Anxiety..T1.	-3.7659624
BCX..Body.Consciousness..T1.	0.2012609

Table A4. *Lasso Regression coefficients using delta attributes values as input and McGill Pain Reduction Percentage as output on Group 2.*

Attribute	Coefficient
BCX..Body.Consciousness..D.	14.891542
sumFI..Fatigue..D.	14.261043
CAT..Catastrophizing..D.	13.957125
SF_PHYS..Physical.Functioning..D.	13.644653
PSQI..Subjective.Sleep.Quality.C1..D.	13.25602

LCTR..Locus.of.Control...Powerful..D.	-11.072026
LCTR..Locus.of.Control...Internal..D.	-9.3825041
LCTR..Locus.of.Control...Chance..D.	0.8708093

Table A5. *Lasso Regression coefficients using Timepoint 2 (baseline) attributes values as input and McGill Pain Reduction Percentage as output on Group 2.*

Attribute	Coefficient
McPain_s..Sensory.Pain.score..T2.	24.188121
TCM Diagnoses	-19.286312
LCTR..Locus.of.Control...Chance..T2.	17.562496
CDEP.Work.Interests..T2.	-17.04251
McPain_a..Affective.Pain.score..T2.	7.74638
LCTR..Locus.of.Control...Internal..T2.	-6.8016597
MYMOP..Self.reported.Medical.Outcome..T2.	-6.2946499
CDEP.Guilt..T2.	-1.8986728
SCHN..Stress..T2.	-1.8187312

Table A6. *Lasso Regression coefficients using baseline attribute values as input and McGill Pain Reduction Percentage as output on Group 1 and Group 2 combined.*

Attribute	Coefficient
PSQI..Subjective.Sleep.Quality.C1	0.4908474
TCM Diagnoses	-7.7313663
LCTR..Locus.of.Control...Chance	8.796358
MYMOP..Self.reported.Medical.Outcome	-13.996435
McPain_s..Sensory.Pain.score	17.58949
McPain_a..Affective.Pain.score	17.772838

Appendix B: Random Forest Outputs

The tables below display the output of the Random Forest method used for feature selection.

Table B1. *Random Forest output using Timepoint 1 attribute values as input and McGill Pain Reduction percentage as output on Group 1.*

Attribute	Increase in Mean Squared Error
POMS..Mood.States..T1.	65.514135
McPain..Pain.score..T1.	34.405286
SNI..Support.Networks.Num.People..T1.	26.676747
McPain_a..Affective.Pain.score..T1.	20.087733
CDEP.Guilt..T1.	17.495748
OPTI..Optimism..T1.	15.879208
CDEP.Psy.Anxiety..T1.	15.483043
CDEP.Retardation..T1.	12.631556
ISEL..Social.Support..T1.	12.212556
sumFI..Fatigue..T1.	12.104012
PSQI..Subjective.Sleep.Quality.C1..T1.	11.646765
LCTR..Locus.of.Control...Chance..T1.	11.534417
SOC..Social.Support.Open..T1.	7.8988537
BANX..Anxiety..T1.	5.4351759

McPain_s..Sensory.Pain.score..T1.	5.4117133
BCX..Body.Consciousness..T1.	4.4306142
MYMOP..Self.reported.Medical.Outcome..T1.	4.0834094
PSQI..Sleep.Medications.C6..T1.	3.9236771
CDEP.Agitation..T1.	3.9071771

Table B2. *Random Forest output using Timepoint 1 attribute values as input and McGill Pain Reduction percentage as output on Group 2.*

Attribute	Increase in Mean Squared Error
CAT..Catastrophizing..T1.	105.8876
LCTR..Locus.of.Control...Internal..T1.	44.154603
CDEP.Retardation..T1.	21.933869
POMS..Mood.States..T1.	19.651913
CDEP.Psy.Anxiety..T1.	18.767761
MYMOP..Self.reported.Medical.Outcome..T1.	15.702524
SCHN..Stress..T1.	12.146609
PSQI..Sleep.Disturbances.C5..T1.	12.063518
CDEP.Agitation..T1.	9.1321974
LCTR..Locus.of.Control...Powerful..T1.	9.0829226

BCX..Body.Consciousness..T1.	8.5886245
PSQI..Sleep.Medications.C6..T1.	8.3178172
SNI..Support.Networks.High.Contact.Role..T1.	7.897154
CDEP.Work.Interests..T1.	3.1674251
LCTR..Locus.of.Control...Chance..T1.	3.0122292
sumFI..Fatigue..T1.	2.7611721
PSQI..Sleep.Daytime.Dysfunction.C7..T1.	2.1993682
CDEP.Guilt..T1.	1.1477933
McPain..Pain.score..T1.	1.0967201

Table B3. *Random Forest output using Timepoint 2 (baseline) attribute values as input and McGill Pain Reduction percentage as output on Group 2.*

Attribute	Feature Importance
McPain..Pain.score..T2.	0.112978893
sumFI..Fatigue..T2.	0.106164661
McPain_a..Affective.Pain.score..T2.	0.075239175
LCTR..Locus.of.Control...Chance..T2.	0.070870446
LCTR..Locus.of.Control...Internal..T2.	0.065342829
CAT..Catastrophizing..T2.	0.037889037

SCHN..Stress..T2.	0.037665217
ISEL..Social.Support..T2.	0.034023221
SOC..Social.Support.Open..T2.	0.032924948
MYMOP..Self.reported.Medical.Outcome..T2.	0.031844426
PSQI..Subjective.Sleep.Quality.C1..T2.	0.030703327
WDEP..Whitely.Depression..T2.	0.02843588
CDEP.Work.Interests..T2.	0.028190882
McPain_s..Sensory.Pain.score..T2.	0.026521947
SNI..Support.Networks.High.Contact.Role..T2.	0.025440301
PSQI..Sleep.Disturbances.C5..T2.	0.025341946
BANX..Anxiety..T2.	0.022482692
SNI..Support.Networks.Num.Embedded..T2.	0.018652001
WAI..Working.Alliance.Goal..T2.	0.017991554

Table B4. *Random Forest output using baseline attribute values as input and McGill Pain Reduction percentage as output on Group 1 and Group 2 combined.*

Attribute	Feature Importance
McPain..Pain.score	0.150639602
McPain_a..Affective.Pain.score	0.100955726

MYMOP..Self.reported.Medical.Outcome	0.062663902
LCTR..Locus.of.Control...Chance	0.059170522
sumFI..Fatigue	0.055079062
SNI..Support.Networks.Num.People	0.052864793
ISEL..Social.Support	0.044250739
McPain_s..Sensory.Pain.score	0.040819915
LCTR..Locus.of.Control...Powerful	0.036234279
POMS..Mood.States	0.033789155
LCTR..Locus.of.Control...Internal	0.032560676
BCX..Body.Consciousness	0.031144256
SOC..Social.Support.Open	0.029228507
CAT..Catastrophizing	0.027543248
PSQI..Subjective.Sleep.Quality.C1	0.027202287
SF_PHYS..Physical.Functioning	0.026475576
SNI..Support.Networks.High.Contact.Role	0.026398574
OPTI..Optimism	0.024635537
PSQI..Sleep.Disturbances.C5	0.016775461

Appendix C: Linear Regression Models

Table C1: *A complete list of regressors from the best performing data with Group 1.*

	Lasso R^2: 0.44	**Lasso ∩ RF, TCM** R^2: -0.14
BCX Body Consciousness	-255.8	-1079.8
CDEP Psy Anxiety	590.8	1264.7
CDEP Retardation	-908.2	-1079.9
MYMOP Self reported Medical Outcome	-719.6	-1397.1
McPain Pain score	2246.4	1802.9
McPain_a Affective Pain score	186.7	461.9
OPTI Optimism	-815.0	-820.3
POMS Mood States	-544.1	-178.4
PSQI Sleep Medications C6	653.3	793.9
PSQI Subjective Sleep Quality C1	586.8	264.5
CDEP Depression	-648.7	
CDEP Work Interests	1001.9	
LCTR Locus of Control Powerful	981.5	
PSQI Sleep Daytime Dysfunction C7	987.7	
SF_PHYS Physical Functioning	1499.3	
SNI Support Networks High Contact Role	-1006.7	
Deficiency and Channel (TCM diagnosis)		617.8
Excess and Channel (TCM diagnosis)		774.7
Excess and Deficiency (TCM diagnosis)		646.5
Excess and Deficiency and Channel (TCM diagnosis)		598.4
Single Category (TCM diagnosis)		777.8

Appendix D: Logistic Regression Models

Table D1: *A complete list of regressors from the best performing data with Group 1.*

	Lasso Accuracy: 66%	**Lasso ∩ RF, TCM** Accuracy: 51%
BCXBody Consciousness	-0.82	-0.86
CDEP Depression	-0.94	
CDEP Psy Anxiety	-0.38	-0.21
CDEP Retardation	-0.69	-0.52
CDEP Work Interests	0.20	
LCTR Locus of Control Powerful	0.77	
MYMOP Self reported Medical Outcome	-0.37	-0.49
McGill Pain score	0.50	0.43
McGill Affective Pain score	0.42	0.60
OPTI Optimism	-0.64	-0.62
POMS Mood States	0.04	0.10
PSQI Sleep Daytime Dysfunction	0.50	
PSQI Sleep Medications	0.04	-0.19
PSQI Subjective Sleep Quality	0.63	0.40
SF_PHYS Physical Functioning	0.75	
SNI Support Networks High Contact Role	-0.36	
TCM Diagnoses		-0.19

Appendix E: Star Glyphs

The star glyphs below (generated using the seaborn Python package) represent how patients' subscores changed over time, allowing for easy deciphering of how patients were affected by the treatment. The attributes shown in the glyphs include PSQI Sleep Daytime Dysfunction, Support Networks High Contact Role, Locus of Control (Powerful), SF-36 Physical Functioning, and McGill Pain Score. The blue highlighted region represents the patient's scores at the beginning of the study (at Timepoint 1), while the red highlighted region indicates the scores after treatment was completed (at Timepoint 4). Furthemore, each point for the blue and red regions corresponds with an attribute value (as described above). By examining the differences between the blue and red areas, one can visually observe the impact or changes resulting from the treatment.

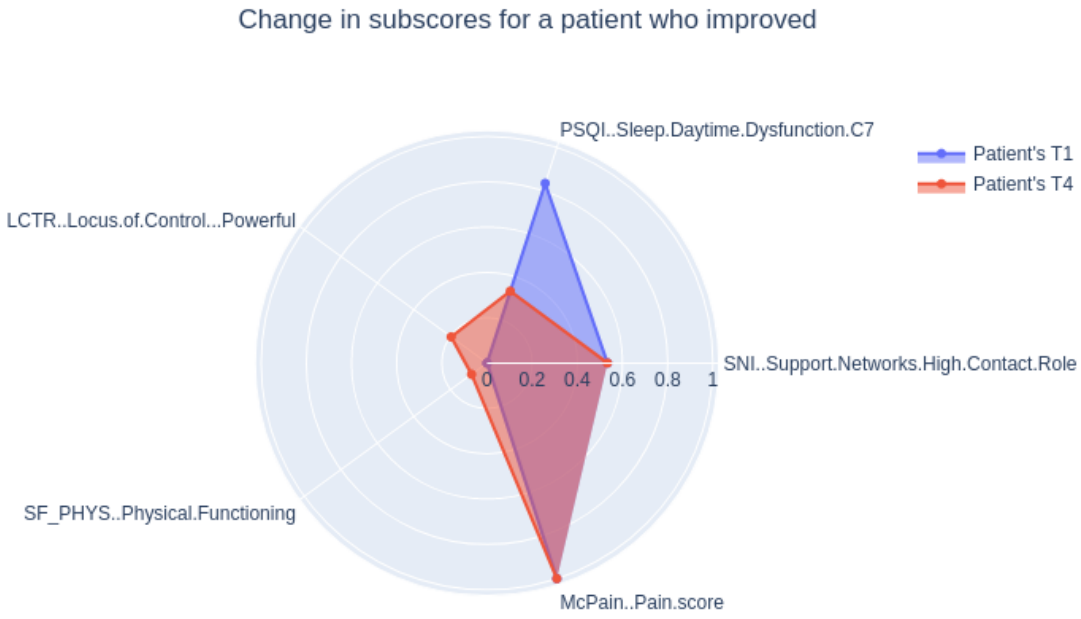

Figure E1. *A star glyph depicting a patient whose McGill Pain Score improved from Timepoint 1 to Timepoint 4.*

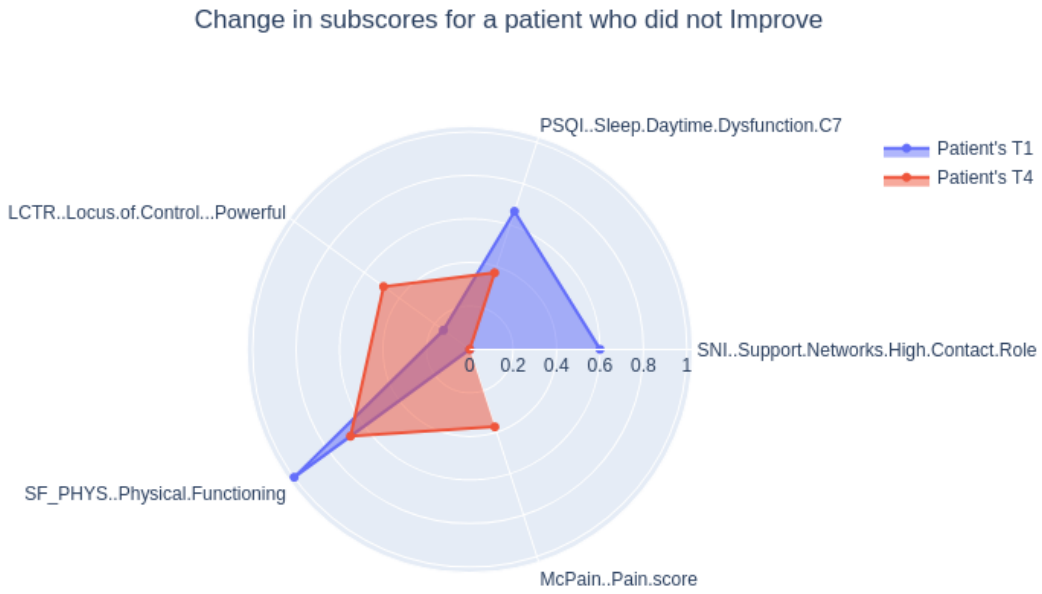

Figure E2. *A star glyph depicting a patient whose McGill Pain Score did not improve from Timepoint 1 to Timepoint 4.*

Appendix F: Parallel Coordinates

The following parallel coordinate graphs indicate certain relationships among key features in the given data over time. These graphs were generated using the plotly package in Python. Each line in the graph represents a patient in the specified group of the study. Each vertical axis depicts the delta score of an attribute calculated from the first to last timepoint. The color scale is a blue-orange gradient that is based on the values of the McGill Pain Reduction (blue indicating a reduction in pain and orange indicating an increase in pain).

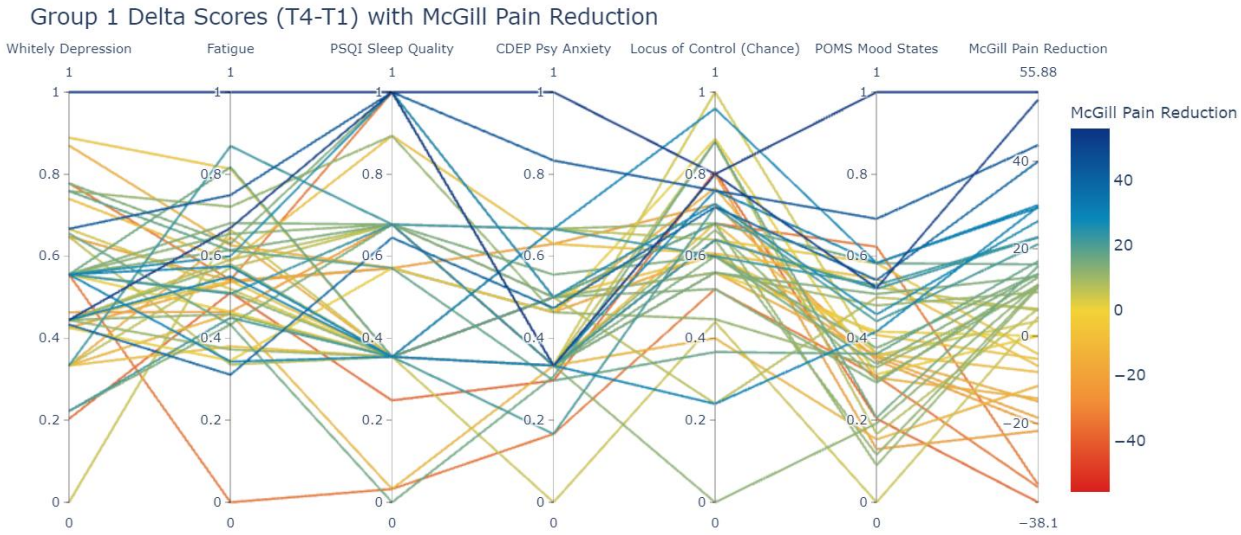

Figure F1. *A parallel coordinate graph of patients in Group 1 with delta scores of Depression, Fatigue, Sleep Quality, Anxiety, Locus of Control (Chance), Mood, and McGill Pain Reduction.*

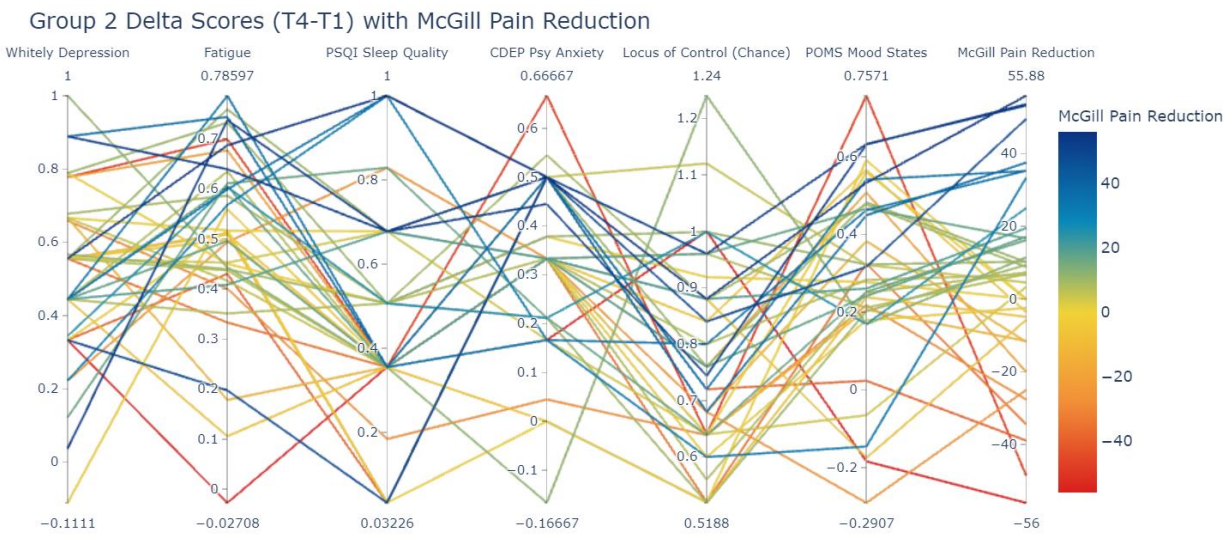

Figure F2. *A parallel coordinate graph of patients in Group 2 with delta scores of Depression, Fatigue, Sleep Quality, Anxiety, Locus of Control (Chance), Mood, and McGill Pain Reduction.*

References

Aroke, E. N., P. V. Joseph, A. Roy, D. S. Overstreet, T. O. Tollefsbol, D. E. Vance, and B. R. Goodin. "Could Epigenetics Help Explain Racial Disparities in Chronic Pain?" *J Pain Res* 12 (2019): 701–10. https://doi.org/10.2147/jpr.s191848.

Bair, Matthew J., Rebecca L. Robinson, Wayne Katon, and Kurt Kroenke. "Depression and Pain Comorbidity: A Literature Review." *Archives of Internal Medicine* 163, no. 20 (2003): 2433–45. https://doi.org/10.1001/archinte.163.20.2433.

Beale, Peter, Peter Heaf, and Norman Jones. "Gulf War Illness." *BMJ: British Medical Journal* 314, no. 7086 (1997): 1041–1041.

Blanchard, Melvin S., Seth A. Eisen, Renee Alpern, Joel Karlinsky, Rosemary Toomey, Domenic J. Reda, Frances M. Murphy, Leila W. Jackson, and Han K. Kang. "Chronic Multisymptom Illness Complex in Gulf War I Veterans 10 Years Later." *American Journal of Epidemiology* 163, no. 1 (January 1, 2006): 66–75. https://doi.org/10.1093/aje/kwj008.

Brazier, J. E., R. Harper, N. M. Jones, A. O'Cathain, K. J. Thomas, T. Usherwood, and L. Westlake. "Validating the SF-36 Health Survey Questionnaire: New Outcome Measure for Primary Care." *British Medical Journal* 305, no. 6846 (1992): 160–64. https://doi.org/10.1136/bmj.305.6846.160.

Carroll, Bernard J., Michael Feinberg, Peter E. Smouse, Sarah G. Rawson, and John F. Greden. "The Carroll Rating Scale for Depression I. Development, Reliability and Validation." *British Journal of Psychiatry* 138, no. 3 (March 1981): 194–200. https://doi.org/10.1192/bjp.138.3.194.

Centers for Disease Control and Prevention (CDC). "Unexplained Illness among Persian Gulf War Veterans in an Air National Guard Unit: Preliminary Report--August 1990-March 1995." *MMWR. Morbidity and Mortality Weekly Report* 44, no. 23 (June 16, 1995): 443–47.

Chen, Yixiao, Per Fink, Jing Wei, Anne-Kristin Toussaint, Lan Zhang, Yaoyin Zhang, Hua Chen, et al. "Psychometric Evaluation of the Whiteley Index-8 in Chinese Outpatients in General Hospitals." *Frontiers in Psychology* 12 (July 1, 2021): 557662. https://doi.org/10.3389/fpsyg.2021.557662.

Conboy, Lisa, Meredith St John, and Rosa Schnyer. "The Effectiveness of Acupuncture in the Treatment of Gulf War Illness." *Contemporary Clinical Trials* 33, no. 3 (May 1, 2012): 557–62. https://doi.org/10.1016/j.cct.2012.02.006.

Dydyk, Alexander M., and Till Conermann. "Chronic Pain." In *StatPearls*. Treasure Island (FL): StatPearls Publishing, 2023. http://www.ncbi.nlm.nih.gov/books/NBK553030/.

Gaskin, Darrell J., and Patrick Richard. "The Economic Costs of Pain in the United States." *The Journal of Pain* 13, no. 8 (August 1, 2012): 715–24. https://doi.org/10.1016/j.jpain.2012.03.009.

Grandner, Michael A, Daniel F Kripke, In-Young Yoon, and Shawn D Youngstedt. "Criterion Validity of the Pittsburgh Sleep Quality Index: Investigation in a Non-Clinical Sample." *Sleep and Biological Rhythms* 4, no. 2 (June 2006): 129–36. https://doi.org/10.1111/j.1479-8425.2006.00207.x.

Hicks, S. A., Strümke, I., Thambawita, V., Hammou, M., Riegler, M. A., Halvorsen, P., & Parasa, S. (2022). On evaluation metrics for medical applications of artificial intelligence. Scientific Reports, 12(1), 5979. https://doi.org/10.1038/s41598-022-09954-8

Hunt, S. M., J. McEwen, and S. P. McKenna. "Measuring Health Status: A New Tool for Clinicians and Epidemiologists." *The Journal of the Royal College of General Practitioners* 35, no. 273 (April 1985): 185–88.

Jeffrey, Mary G., Maxine Krengel, Jeffrey L. Kibler, Clara Zundel, Nancy G. Klimas, Kimberly Sullivan, and Travis J. A. Craddock. "Neuropsychological Findings in Gulf War Illness: A Review." *Frontiers in Psychology* 10 (September 2019). https://doi.org/10.3389/fpsyg.2019.02088.

Jensen, H. H., E. L. Mortensen, and M. Lotz. "Scl-90-R Symptom Profiles and Outcome of Short-Term Psychodynamic Group Therapy." *ISRN Psychiatry* 2013 (2013): 540134. https://doi.org/10.1155/2013/540134.

Ji, Ru-Rong, Andrea Nackley, Yul Huh, Niccolò Terrando, and William Maixner. "Neuroinflammation and Central Sensitization in Chronic and Widespread Pain." *Anesthesiology* 129, no. 2 (2018): 343–66. https://doi.org/10.1097/aln.0000000000002130.

Joyce, Michelle R., and Kathleen F. Holton. "Neurotoxicity in Gulf War Illness and the Potential Role of Glutamate." *NeuroToxicology* 80 (September 1, 2020): 60–70. https://doi.org/10.1016/j.neuro.2020.06.008.

Kerr, Kathleen J. "Gulf War Illness: An Overview of Events, Most Prevalent Health Outcomes, Exposures, and Clues as to Pathogenesis." *Reviews on Environmental Health* 30, no. 4 (January 1, 2015). https://doi.org/10.1515/reveh-2015-0032.

Leyfer, Ovsanna T., Joshua L. Ruberg, and Janet Woodruff-Borden. "Examination of the Utility of the Beck Anxiety Inventory and Its Factors as a Screener for Anxiety Disorders." *Journal of Anxiety Disorders* 20, no. 4 (January 2006): 444–58. https://doi.org/10.1016/j.janxdis.2005.05.004.

Mawson, Anthony R., and Ashley M. Croft. "Gulf War Illness: Unifying Hypothesis for a Continuing Health Problem." *International Journal of Environmental Research and Public Health* 16, no. 1 (2019): 111.

Melzack, Ronald. "The Short-Form McGill Pain Questionnaire." *PAIN* 30, no. 2 (1987): 191–97. https://doi.org/10.1016/0304-3959(87)91074-8.

Melzack, Ronald, and Srinivasa N. Raja. "The McGill Pain Questionnaire: From Description to Measurement." *Anesthesiology* 103, no. 1 (2005): 199–202. https://doi.org/10.1097/00000542-200507000-00028.

National Academies of Sciences, Engineering, and Medicine (U.S.), Richard J. Bonnie, Morgan A. Ford, and Jonathan Phillips, eds. *Pain Management and the Opioid Epidemic: Balancing Societal and Individual Benefits and Risks of Prescription Opioid Use*. Washington, DC: The National Academies Press, 2017.

Nephew, B. C., A. C. Incollingo Rodriguez, V. Melican, J. J. Polcari, K. E. Nippert, M. Rashkovskii, L. B. Linnell, et al. "Depression Predicts Chronic Pain Interference in Racially Diverse, Income-Disadvantaged Patients." *Pain Med* 23, no. 7 (July 1, 2022): 1239–48. https://doi.org/10.1093/pm/pnab342.

Nettleman, M. "Gulf War Illness: Challenges Persist." *Trans Am Clin Climatol Assoc* 126 (2015): 237–47.

Newman, Andrea K, Shweta Kapoor, and Beverly E Thorn. "Health Care Utilization for Chronic Pain in Low-Income Settings." *Pain Medicine* 19, no. 12 (2018): 2387–97. https://doi.org/10.1093/pm/pny119.

Paap, Davy, Ernst Schrier, and Pieter U. Dijkstra. "Development and Validation of the Working Alliance Inventory Dutch Version for Use in Rehabilitation Setting." *Physiotherapy Theory and Practice* 35, no. 12 (December 2, 2019): 1292–1303. https://doi.org/10.1080/09593985.2018.1471112.

Paterson, C. "Measuring Outcomes in Primary Care: A Patient Generated Measure, MYMOP, Compared with the SF-36 Health Survey." *BMJ* 312, no. 7037 (April 20, 1996): 1016–20. https://doi.org/10.1136/bmj.312.7037.1016.

Shahid, Azmeh, Kate Wilkinson, Shai Marcu, and Colin M. Shapiro. "Profile of Mood States (POMS)." In *STOP, THAT and One Hundred Other Sleep Scales*, edited by Azmeh Shahid, Kate Wilkinson, Shai Marcu, and Colin M Shapiro, 285–86. New York, NY: Springer New York, 2011. https://doi.org/10.1007/978-1-4419-9893-4_68.

Sheng, Jiyao, Shui Liu, Yicun Wang, Ranji Cui, and Xuewen Zhang. "The Link between Depression and Chronic Pain: Neural Mechanisms in the Brain." *Neural Plasticity* 2017 (June 19, 2017): 9724371. https://doi.org/10.1155/2017/9724371.

Taylor-Swanson, Lisa, Joe Chang, Rosa Schnyer, Kai-Yin Hsu, Beth Ann Schmitt, and Lisa A. Conboy. "Matrix Analysis of Traditional Chinese Medicine Differential Diagnoses in Gulf War Illness." *The Journal of Alternative and Complementary Medicine* 25, no. 11 (November 1, 2019): 1097–1102. https://doi.org/10.1089/acm.2017.0299.

"Update on Gulf War Illness." *Environmental Health Perspectives* 105, no. 5 (1997): 474–76. https://doi.org/10.2307/3433570.

Ware, J. E., Jr., and C. D. Sherbourne. "The MOS 36-Item Short-Form Health Survey (SF-36). I. Conceptual Framework and Item Selection." *Med Care* 30, no. 6 (June 1992): 473–83.

Wessely, S., and L. Freedman. "Reflections on Gulf War Illness." *Philos Trans R Soc Lond B Biol Sci* 361, no. 1468 (April 29, 2006): 721–30. https://doi.org/10.1098/rstb.2006.1830.

Yong, R. Jason, Peter M. Mullins, and Neil Bhattacharyya. “Prevalence of Chronic Pain among Adults in the United States.” *PAIN* 163, no. 2 (2022): e328–32. https://doi.org/10.1097/j.pain.0000000000002291.