3-2023

Uncertainty Quantification in Federated Learning for Persistent Post-traumatic Headache

Byungmoo Brian Kim

UNCERTAINTY QUANTIFICATION IN FEDERATED LEARNING FOR PERSISTENT POST-TRAUMATIC HEADACHE

THESIS

Byungmoo Brian Kim, Second Lieutenant, USAF

AFIT-ENS-MS-23-M-132

DEPARTMENT OF THE AIR FORCE
AIR UNIVERSITY

AIR FORCE INSTITUTE OF TECHNOLOGY

Wright-Patterson Air Force Base, Ohio

AFIT-ENS-MS-23-M-132

UNCERTAINTY QUANTIFICATION IN FEDERATED LEARNING FOR PERSISTENT POST-TRAUMATIC HEADACHE

THESIS

Presented to the Faculty

Department of Operational Sciences

Graduate School of Engineering and Management

Air Force Institute of Technology

Air University

Air Education and Training Command

in Partial Fulfillment of the Requirements for the

Degree of Master of Science in Data Science

Byungmoo Brian Kim, B.S.C.S.

Second Lieutenant, USAF

March 23, 2023

AFIT-ENS-MS-23-M-132

UNCERTAINTY QUANTIFICATION IN FEDERATED LEARNING FOR PERSISTENT POST-TRAUMATIC HEADACHE

THESIS

Byungmoo Brian Kim, B.S.C.S.
Second Lieutenant, USAF

Committee Membership:

Nathan B. Gaw, Ph.D
Chair

Major Chancellor A. Johnstone, Ph.D
Reader

AFIT-ENS-MS-23-M-132

Abstract

Each year, millions of college athletes play sports that put them at risk of a variety of physical injuries. Notable among these injuries is mild traumatic brain injury (mTBI). Following a mild traumatic brain injury (mTBI), many athletes suffer from a post-traumatic headache (PTH), which will eventually resolve or develop into persistent post-traumatic headache (PPTH). PPTH can lead to debilitating pain and detrimentally affect an athlete's lifestyle and future professional sports prospects. Although no known cure for PPTH exists, research has shown that receiving treatment at earlier stages of mTBI and PTH lowers the risk of patients developing PPTH. Previous studies have shown machine learning (ML) models capable of predicting a patient's PTH progression (whether resolution to a healthy condition or conversion to PPTH). However, none of them have considered the issue of retaining patient privacy within each institution. When respecting patient privacy, there is typically a lack of data available to train ML models since model training can only be performed within a single institution. Federated learning (FL) has demonstrated the potential of harnessing data from separate institutions without sacrificing patient privacy. Within an FL framework, local institutions can run ML models on their own private dataset and share the trained model parameters without sharing the data between institutions. Additionally, quantifying uncertainty of model parameters associated with key features of interest in predicting PTH progression has not been explored in the context of FL. The proposed data analysis framework, Uncertainty Quantification in Federated Learning (UQFL), combines FL and uncertainty quantification (UQ) to (1) protect patient privacy and (2) provide a measure of uncertainty for each model parameter. UQFL was applied to the Concussion Assessment, Research

and Education (CARE) dataset, which contains clinical measurements that track the condition of college athletes following an mTBI. UQFL identified a variety of clinical measurements that significantly contributed to model prediction of PTH progression ($p < 0.05$); namely, Satisfaction with Life Scale (SWLS) Score, SCAT3 Total Number of Symptoms, Standardized Assessment Concussion (SAC) score, Vestibular Ocular Motor Screening (VOMS) Score, Brief Symptom Inventory (BSI) 18 Score, SCAT3 Total Score, and Clinical Reaction Time (CRT). UQFL demonstrates the capability to significantly capture the same parameter values calculated from the same ML model trained on a centralized dataset. It was even found that some of the FL models outperformed traditional ML models trained a centralized database (likely due to incorporating the heterogeneity of institutions directly into the model framework). Future work will entail making the UQFL model robust to missing data and capable of integrating different types of clinical measures shared from each institution.

Table of Contents

Page

Page

Page

List of Figures

List of Tables

UNCERTAINTY QUANTIFICATION IN FEDERATED LEARNING FOR PERSISTENT POST-TRAUMATIC HEADACHE

I. Introduction

Approximately 1.7 million people suffer from traumatic brain injury (TBI) annually. Among these, post traumatic headache (PTH) is the most common symptom following TBI, which can either resolve or continue into persistent PTH (PPTH), and can eventually lead to lasting brain damage or even death. However, recognizing PTH in the early stages and getting treatment quickly dramatically increases a patient's chance of PTH being cured. Different hospitals and organizations store their own medical data but centralizing those data into one database is impossible due to legal and privacy issues. This project proposes a solution to this issue using Uncertainty Quantification Federated Learning (UQFL). UQFL allows hospitals to learn from data from other hospitals without having to share data. Using data from NCAA athletes, this project builds a UQFL model of clinical measures among student athletes to predict PTH persistence and quantify the uncertainties of model parameters.

1.1 Background

1.1.1 Concussion and Neurodegenerative Diseases

A concussion or a mild traumatic brain injury (mTBI) is a brain injury caused by a bump to the head that results in the brain moving back and forth rapidly [6]. This movement causes the brain to either bump into the inner skull or twist the brain stem,

injuring the brain cells. The most common symptom after a concussion is a PTH. Primary symptoms of PTH include nausea, headaches, vomiting and sluggishness, among many others. Most patients who suffered from a concussion get their PTH resolved within 3-6 months. However, if these symptoms continue after 3-6 months, this type of PTH is called PPTH. This persistent headache results from a continuous degeneration and breakdown of the brain cells [7]. Because of the longer time period that the brain cells were damaged, PPTH is thought to develop into secondary symptoms from the continuously damaged brain cells which may or may not correlate with long-term neurological damage such as neurodegenerative diseases [8].

A neurodegenerative disease is a type of disease where neurons in the nervous system stop functioning over time [9]. A nervous system is a network of nerves throughout the body, including the brain and the spinal cord [10]. It is responsible for relaying signals to bodily functions such as senses, thoughts, movements and heartbeats. The breakdown or degeneration of the nerves, and ultimately the nervous system, from a neurodegenerative disease leads to the loss of such bodily functions and can lead to death. Some examples of neurodegenerative disease include Alzheimers Disease (AD), Parkinson's Disease (PD), or Amyotrophic Lateral Schlerosis (ALS or Lou Gehrig's Disease) [11]. Currently, there are no known cures to neurodegenerative diseases and patients can only get treatment to prolong one's life or ease the symptoms of the disease. Because of the time-related nature of the disease, the probability of developing a neurodegenerative disease increase as one ages. The symptoms worsen as time passes as well.

No studies have conclusively proved that a concussion leads to a neurodegenerative disease. However, many studies have shown that cognitive impairment coming from the damage to the brain cells from an mTBI and the breakdown of neurons in the brain and the nervous system that cause impairment of bodily activities and functions

from neurodegenerative diseases are similar [8]. This similarity has led researchers to presume that an untreated mTBI, such as a PPTH or repetitive concussions could lead to a neurodegenerative disease.

1.1.2 Effects of Concussions and Neurodegenerative Diseases

Concussion is the most common type of mTBI. According to the University of Michigan Health, every year, about 5-10% of athletes are concussed from sports-related activities, which amounts to 3.8 million athletes annually. [12]. Concussions are not contained to just professional athletic settings and can be caused in a variety of ways, such as falling or vehicle crashes. Studies have shown that people who have been concussed are more susceptible of having another concussion even with less force to the head [13]. This means that the person is more at risk of brain damage even with smaller threats. Even though a concussion might seem like a minor injury at the moment, it can have permanent dangerous effects on the human brain, which may lead to more serious neurodegenerative diseases such as Alzheimer's Disease (AD) or Parkinson's Disease (PD).

AD affects more than 6 million Americans, most of whom are over 65 years old, and may cause dementia. The National Institutes of Health cites that AD is the 7th leading cause of death in the United States and is the most common cause of dementia [14]. The loss of cognitive function impedes most patients from performing daily activities even if they are alive. PD causes unintended or uncontrollable movements and patients diagnosed with PD have difficulty walking and talking. Around 60,000 Americans are diagnosed with PD annually, most of whom are also older [15]. Neurodegenerative diseases are thought to be most prevalent in people who have repetitively experienced traumatic brain injuries throughout their lifetime. The two biggest examples are contact athletes and members of the military. This disease also

progresses with age as it starts off mildly with symptoms such as depression, aggression or irritability. However, as it progresses over decades, it can lead to major cognitive impairments and dementia [16]. Finally, ALS is a disease that destroys the motor neurons in the brain that lead to loss of control of skeletal muscles. This means that the patient loses control of activities such as walking, breathing and speaking [17].

According to the National Institute of Health, there are no known cures for all four of these diseases [9]. Another common factor among the four diseases is that they are all progressive neuro-degenerative diseases. This means that the symptoms over time get worse and can eventually lead to death. However, the earlier that these diseases are discovered by doctors in patients, the patients will suffer fewer damaging effects of the diseases through treatment and medicine. This process is analogous to the need of finding cancer at stage 0 compared to finding cancer at stage 4; it is more manageable to address cancer at stage 0 and extremely difficult (sometimes impossible) to cure cancer at stage 4. This means that the earlier doctors can perceive cognitive impairment, the patients have a higher chance of living longer and healthier lives or even be cured.

1.1.3 NCAA/Collecting Data

Even though concussions can happen to anyone, collecting data for every instance of a concussion nationwide is impractical and inefficient. As a result, athletes, who are the most susceptible to concussions, are a natural source for data. Athletes daily endure extreme physical contact that is far beyond the stress levels of an average person. A 400-pound football player tackling another for a sack or 7-feet tall basketball players elbowing each other in the face for a rebound are two examples. As a result, professional athletes are the ideal source to data to study the effects of concussion on.

The assumption is that if there is a treatment that can help the quarterback recover from a concussion caused by a 400-pound man, then the treatment will work well on an average person who received a concussion from a fall.

However, because of the hyper-competitive environment of the professional sports industry, there is a very small sample size to gather data from. In 2021-2022, there were 526 players in the National Basketball Association (NBA), 1123 players in the National Hockey League (NHL) and 1696 players in the National Football League (NFL) [18, 19, 20]. Even if all three leagues were combined, there would only be a total of 3345 players. From an estimated 5-10% of athletes who are concussed every year, it would only be 167 to 334 players every year to gather data from. As a result, many studies study the effects of concussions and traumatic brain injuries on Division I athletes.

In the National Collegiate Athletic Association (NCAA), there are 187,375 athletes that compete at the Division I level, which is about 50 times the number of athletes in the NBA, NFL and NHL combined [21]. College athletic programs are divided into three divisions with each division separated in a hierarchy by the level of competition they play. Division I schools play in games that have a larger fanbase, nationally broadcasted and generate more revenue. This allows Division I athletic programs to recruit among the best high school players in the nation. Consequently, these school have the best college athletes playing in their programs who train and perform almost at the professional level. To demonstrate the high level of competition Division I athletes play, in 2021-2022, 1414 out of 1696 or 83% of the NFL players were drafted from Division I schools [20]. By having a larger pool of data to source from without sacrificing the competitiveness, and the susceptibility of a concussion, of professional athletes, Division I athletes are the ideal source to conduct research on the effects of concussion.

However, data on concussion based on athletes can potentially biased. First, the assessment tools to diagnosis of concussion are based on a series of self-reported clinical scores [22]. These tools make assessing a patient's concussion difficult because it is near impossible to accurately detect the effects of concussion with commonly used assessment tools [23]. For example, a SCAT3 score is one of the commonly used assessment tools to diagnose a patient with concussion [24]. The patient answers a series of questions with seven ordinal scaled responses about their physical and mental well-being (i.e. "Nervous or Anxious" or as simple as "Headache"). The neurological damage and the effects that damage will have on a person in the future cannot be simply categorized into seven ordinal variables. The effects of the concussion are often too complex to be fully comprehended by the assessment tools used today.

Second, athletes under report their symptoms [23]. As mentioned above, it is extremely difficult to play in Division 1 sports, let alone get drafted to play professionally. Health is one of the most important aspects scouts look for in drafting athletes, whether it be at the collegiate or professional level. NFL Hall of Fame Executive Bill Polian states that the reason NFL hosts the NFL Combine every year is to get a "complete assessment of (athlete's) health and their ability to withstand the rigors of pro football, the likes of which they have never had" [25]. As a result, players are incentivized to remain healthy or maintain the appearance of being healthy in order to be scouted and have a higher change of becoming a professional athlete. This is one of the reasons why players under report their symptoms. They also under report their symptoms in order to return to play as much as possible. For similar reasons to play professionally or at the collegiate level, scouts make decisions on players based off of their performances [26]. However, if they are not playing due to injury, then the scouts cannot make positive reports on those players, which would potentially hurt the athletes' futures. In order to prevent this problem, many athletes do anything

they can to keep playing, even by under reporting their concussion symptoms.

These inaccurate responses pose two major problems. First, the injured athletes cannot be properly taken care of and the effects of concussion will potentially have long-lasting effects on the athletes' lives. Second, the effectiveness of any model or studies can be considered ineffective because they were based on inaccurate data. These problems add to the reasons as to why diagnosing and treating concussions have been so difficult.

1.1.4 Military Applications

Not only are TBIs a problem of concern for the professional and amateur athlete community, they are also a problem for military personnel as well. Military TBIs are different from civilian TBI because of the environment, as well as external and internal stress related to war [1]. One of the most common causes of TBI among military personnel were combat related explosions [27]. Additionally, military personnel are deployed to situations, even when not in combat, that require extreme physical and mental demands, forced to inhabit harsh living conditions and are often expected to operate under sustained sleep deprivation. These so-called non-combat situations were shown to induce TBI in military personnel as well [27].

Figure 1 shows the number of incidents of TBI by military branch from 2000 to 2011 and how almost 30,000 personnel from all four branches of the mlitary were diagnosed with TBI in 2011 alone. Figure 1 also shows a sharp increase of TBI incidents from the United States (US) Army since 2005. In 2015, the Congressional Research Service reported that 253,330 TBI incidents occurred from 2000 to 2012, of which 194,561 cases were mTBI [28]. This timeline correlates with US military involvement in Operation Iraqi Freedom and Operation Enduring Freedom, when soldiers were deployed the most and subject to the harsh environments of war where they would

be most likely to sustain TBIs in both combat and non-combat environments [27].

However, there are more conflicts awaiting the United States in the future [29]. These possible future conflicts that could affect US military personnel. If engaged in war with these two countries, the occurrences of TBI in US military personnel will be far greater than the numbers seen in Iraq and Afghanistan. Finding a method to help mitigate the TBIs suffered by US military personnel will be of great importance not only for our current veterans but also future veterans.

1.1.5 Patient Privacy

The most important aspect of gaining insights from a machine learning model is to have good data. In contemporary times, data can be collected from everywhere at any time due to the internet and handheld devices [30]. Companies can track how much time their users spend on a particular web page or what parts of their user interface are most used by a certain demographic in order to target specific ads towards those demographic [31]. This is not only true in commercial industries, but also in the medical industry. With medical tools and technology advancing, doctors are

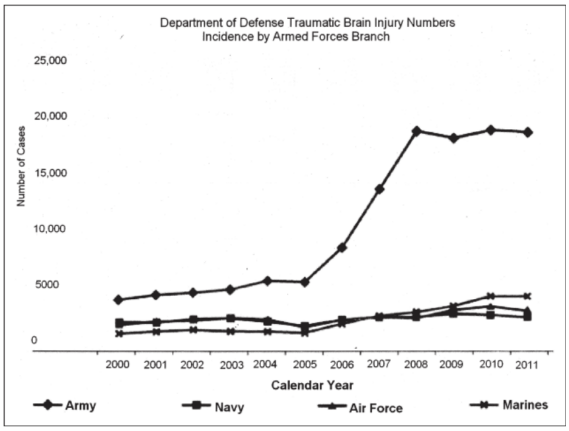


Fig 1.—The number of traumatic brain injury cases reported by each branch of the US military.The data are from the Defense and Veterans Brain Injury Center website.[24] The original source of the data was the US Department of Defense, Theatre Medical Data Store within the Defense Medical Surveillance System.

Figure 1: DOD TBI Occurrences by Branch (2000-2011) [1]

collecting more data than ever on patients in order to help diagnose their treatments with less time and more accuracy [30]. Despite the deluge of data being tracked and stored, accessing data can be challenging, if not impossible, for researchers in medical fields to properly run machine learning models. The Privacy Rule of the Health Insurance Portability and Accountability Act (HIPAA) protects patients from hospitals releasing their personally identifiable information (PII) [32]. HIPAA prevents researchers or doctors from sharing patient data for medical purposes, which severely impedes researchers from making valuable insights on diagnosis of patient conditions such as TBI. One way this research proposes is using Federated Learning (FL) to help mitigate that impediment and allow researchers to make valuable machine learning models for patients based on limited data.

1.1.6 Uncertainty Quantification of Federated Learning

Federated Learning (FL) is a machine learning model that is capable of gaining as much insight from a regular machine learning model with restricted data. FL is able to learn and make predictions without a need for a centralized database, unlike other machine learning models [33]. This means that FL models allow learning from decentralized databases where individual organizations or hospitals do not have to share data and researchers and doctors can still gain meaning insights just by looking at model parameters generated by those individual organizations or hospitals. This partially circumvents HIPAA, with which many research projects would have been previously deemed illegal; with FL no data is shared, only the model parameters.

However, models that learn from decentralized databases might have more uncertainties. This is because there are more steps involved to generate the model output only using model parameters. Unfortunately, there is very little research out there that quantifies uncertainty for a FL model. Having uncertainty quantification for

FL models will allow more credible and reliable research to be done on previously infeasible medical databases all around the world.

II. Literature Review

2.1 Application of Machine Learning in PTH

2.1.1 Overview

Most work performed on diagnosing and studying concussion and its symptoms are done on the actual brain [34]. They study the brain and the physiological and neurological damage that the brain sustains in an mTBI. However, there are few studies that emply machine learning to study the effects of concussion in terms of clinical tests (e.g., human aptitude and symptom tests) [35].

Most studies have used basic machine learning techniques such as logistic regression and random forest. One paper used random forests to impute nominal and ordinal data of PTH clinical tests [36]. Another used logistic regression to place an arbitrary threshold of whether someone had mTBI based on certain features [37]. However, these classifications were based on arbitrary thresholds determined by the researchers of the study.

Other machine learning studies done on TBI and concussion were performed on neurological images or had response variables that were separated into classes. As a result, most studies have methodologies with convolutional neural networks (CNNs) or regressions with binary or multi-class responses. In contrast, this paper uses linear regression to predict continuous data (i.e., concussion severity clinical scores) as output. However, there are still some things to take away from different machine learning techniques done on concussion studies.

2.2 Supervised Machine Learning

Supervised machine learning processes are machine learning techniques that predict future output based on past labeled inputs. Supervised machine learning tech-

niques are often used in PTH and TBI research due to their predictive capabilities and historical data from the long-term design of most PTH and TBI studies [38]. The two techniques discussed in this thesis are regression and classification techniques.

2.2.1 Regression

2.2.1.1 Overview

Regression analysis is a mathematical technique that estimates a dependent value (outcome) based on independent values (features). It determines the relationship between the dependent and independent variables by finding the effect the independent variable has on the dependent variable mathematically. In other words, it finds how much of the independent values has an affect on the outcome of the dependent value. Despite its simplicity, regression is a powerful tool used in almost every aspect of the professional world, such as business, law, medicine, engineering and science [39]. There are numerous different types of regression models, but the two that will be discussed in this paper are Linear Regression and Logistic Regression.

2.2.1.2 Assumptions

For all machine learning models, the data used to train and test the models have to be a good representation of the population as a whole [40]. This is to to ensure that the results from the machine learning models will have insights on the real world problems it was modeled after. Additionally, there are different assumptions for different techniques that need to be met in order for the model to be insightful. In all linear regression models, there are five assumptions that have to be met in order for the model to have statistical significance [40].

1. Linearity

2. Normality

3. No Multicollinearity Among Features

4. No Auto-Correlation of Residuals

5. Homoscedasticity

Linearity is the assumption that the response variables and its feature variables have a linear relationship. If a linear relationship does not exist in a linear regression, the predicted estimate would not be useful as it would be no different than guessing what the next estimate would be. This would severely decrease the accuracy of the model to predict values. Normality assumes that the error terms of the model are normally distributed with zero mean. This is important because we calculate confidence intervals to quantify uncertainty and confidence intervals rely on normally distributed error terms. No multicollinearity among features means that the feature variables are not correlated with one another. This is important because if they are correlated with one another, then we do not know which features impacted the output in what way. Having auto-correlation among the residuals means that the model is wrong. It usually means it is missing features from the regression required to accurately model the data. Finally, homoscedasticity means that the variance associated with each observation is independent of the independent variable. If they are not, it means that the model is incorrectly putting too much weight on certain features or portions of the data [40]. These assumptions are important to be checked if not made in order to have validity of the linear regression model.

2.2.1.3 Linear Regression

A linear regression is a mathematical model that models the relationship between a continuous dependent variable and one or more independent variables. Linear

regression assumes there exists a relationship between the features X and the response variable Y. The model that results is called a regression function and it maps the linear combination of the features to a independent variable [41]. A basic linear model is in the form of $(y = \beta_0 + \beta_k x_k + \epsilon)$, where y is the response variable, k is the number of features, β_0 is the constant coefficient, β_k is the coefficient for the k^{th} feature, x_k is the k^{th} feature, and ϵ is the error. The coefficients are called the model parameters and determine how much of a feature is calculated for the independent variable in the regression function. This linear model has two applications: prediction and variation.

For prediction, a linear model allows the capability of fitting the observed data to predict what the independent variable will be given a new set of dependent features. There are generally two types of predictions available with a linear model: predicting a mean response (general estimate) and predicting a specific response value (point estimation) [41]. This paper will be focused on point estimation because the data are all numerical values. The first reason is that there are no images required for classification. The second reason is point estimates allow us to construct pseudo-distributions and confidence intervals.

For variation, an analysis on the model can calculate how consistent the dependent variables are in relationship to the independent variable. We quantify this through uncertainty quantification. This will be discussed more in depth in Section 2.5.

Because of the predictive power of the linear regression model, it is often used in healthcare research where clinical tests are quantified as numerical values. Researchers in New York University (NYU) studied the severity of the SCAT3 score using PTH presence (headache free), intensity and frequency as independent features [22]. The study was conducted on a heterogeneous group with a varying range in age, gender and severity of the PTH. This was to get as accurate of a representation of the population as possible. The study shows that the intensity of the SCAT3 scores

correlates with the intensity of the headaches and the frequency of the headaches. It also shows that patients who did not have post traumatic headaches had lower SCAT3 scores. The study is able to show a relationship between the SCAT3 score and the independent feature variables. Although this study does not use the exact same features as the ones used by the researchers at NYU, it was still able to define a relationship between the SCAT3 scores and a patient's brain health. The researchers, among many other literature reviewed, conclude that the SCAT3 score is a good assessment to evaluate post-traumatic headache. This is why we use the SCAT3 score and the SCAT3 symptom score as our two main response variables for the study.

2.2.1.4 Logistic Regression

One method of supervised machine learning techniques used often for PTH studies is logistic regression. Logistic regression estimates the binary outcome of a linear function by calculating what the probability of the outcome will be using the parameters of the model [42]. Whereas linear regression has continuous response variables, logistic regression has binary response variables. Logistic regression tries to solve the probability $\Pr(Y = 1 \mid X = x)$ for a set number of feature variables X given $Y = 1$ or $Y = 0$. The parameters of X are solved through maximum likelihood and the probability of Y is solved using a sigmoid function.

Logistic regression, similar to linear regression, also assumes that there exists a relationship between the features X and the response variable Y. The difference between logistic regression and linear regression is that in linear regression, the response variable is a continuous variable while the response variable in logistic regression is a binary variable. However, because the response variable is not continuous, logistic regression requires a link function, called a logit function, to solve for the expected value of the response variable. The logit function is $\ln \frac{p}{1-p}$ where $p \in (0, 1)$ is the prob-

ability that the outcome of a specific combination of features will be 1 ($Y = 1|X = x$). $1 - p$ in turn, is the probability that the outcome will be 0 ($Y = 0|X = x$). The logit function then can be modeled as a linear function of the features just like linear regression, allowing the assumption of linear relationship between the features and response variable to hold for a logistic regression.

The logistic regression model using the logit function then turns into

$$\log \frac{p}{1-p} = \beta_0 + x_k{}'\beta \tag{1}$$

for k number of features.

Solving for p gives,

$$p = \frac{1}{1 + e^{-(\beta_0 + x_k{}'\beta)}} \tag{2}$$

which is a sigmoid function

$$p = \frac{1}{1 + e^{-z}} \tag{3}$$

using a logistic regression model, replacing z with $(\beta_0 + x_k{}'\beta)$, shown in figure below. The sigmoid function allows a probability odds threshold, usually set a 0.5, whether the probability output of the combination of features is closer to 1 or 0.

One of the tests done to predict the effect of concussions on neurocognitive activities was used for logistic regression using the binary variable of whether the athlete had a previous concussion history or not based on neuromechanical performance tests. The authors test whether neuromechanical performance deficiencies are a result of previous sport related concussions. The study is conducted on 35 Olympic athletes were tested for both ordinal scale tests that were self-answered and continuous timed reflex tests [43]. These sets of features made it similar to our study as our dataset contained both ordinal and time data as well.

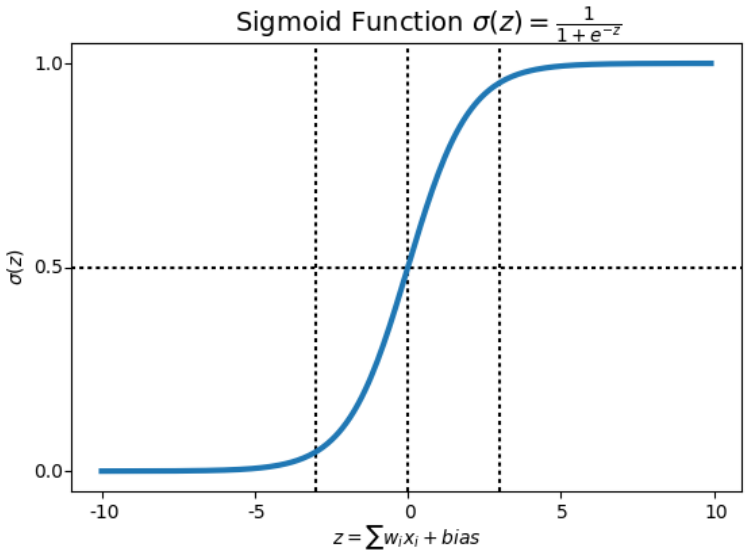


Figure 2: Sigmoid Function Visual Example [2]

Wilkerson et. al studied the affect of athletes' previous concussion history by studying their sensorimotor skills. They had a group of 75 athletes with 12 features and a binary label output of whether that particular athlete had a concussion history or not. The 12 features were sub-tests of visual-motor reaction time (VMRT) and whole-body reactive agility (WBRA) tests. VMRT is similar to the Vestibular Ocular Motor Screening (VOMS) and WBRA is similar to the ImPACT scores, which include visual motor speed scores.

Using similar features to our dataset, the authors calculate the probability of whether performance deficiencies of the athletes are a result of previous concussions.

2.2.2 Classification

One study showed that patients with concussions showed changes in brain electrical activity and used response variables from an electroencephalogram (EEG) to use classification techniques to find relationships with brain activity and patients with

concussions [44]. The study used the Concussion Index, which is an ordinal score of 0 to 100 where a higher index means more cognitive impairment, and arbitrarily used a threshold of 70 as whether patients had a concussion or not.

2.3 Unsupervised Machine Learning

One of the studies performed was employing unsupervised convolutional auto-encoders on brain magnetic resonance imaging (MRI) [3]. Due to the cost of MRIs, there is a lack of data to fully train a regular convolutional neural network. Under the assumption that mTBI only affects certain parts of the brain, the researchers study only parts of the MRI images they called "patches". However, they cannot get the labels of the patches from the subject MRI images because the mTBI does not impact all patches. This is why supervised learning is not viable and they must use unsupervised learning techniques. The researchers use a convolutional auto-encoder, which is an unsupervised learning technique, to uncover latent features. It uncovers latent features by taking a patch image and pushing the image through multiple convolution and pooling layers through the encoding phase. Then it attempts to reconstruct the original image from the convoluted image using the learned features through the decoder phase.

The features that were uncovered through the convolutional auto-encoders are put into a histogram of all the patches in a specific region in visual words. Finally, the histogram of the patients is compared with the histogram of the control to compare where there are differences in the number of visual words represented for different patches of the image. Figure 3 shows the differences of histograms of the number of features constructed from the MRI between the the patients with mTBI (red dotted line) and the control group (blue dotted line). The green dotted line shows the differences between the two groups. The major oscillations seen in the green line

show that there are clear differences in the features extracted from the MRIs of patients with mTBI and the control group. This means that the MRIs between the two groups are different, so patients with mTBI have an altered image of the brain.

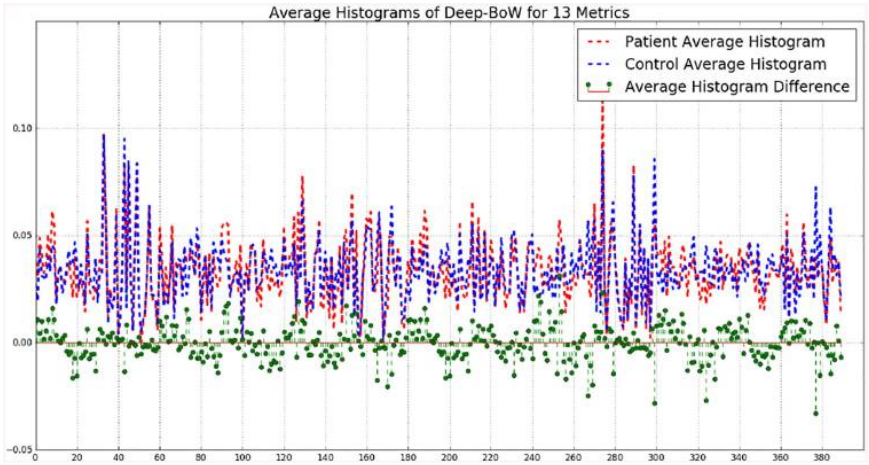


Figure 3: Histogram of Number of Features of TBI Group and Control Group [3]

2.4 Federated Learning

2.4.1 Overview

Currently, technology in the 21^{st} century can be described by a framework called Internet of Things (IoT). IoT is a system of technologies that are connected over a shared network system (typically the Internet) and share each other's data on a single database [45]. The type of technologies can range from handheld devices, such as phones or smartwatches, to vehicles and home appliances that all have connection to the Internet. For example, General Motors (GM) has a program called OnStar. The OnStar program gathers data from all GM vehicles, called edge devices, that are part of this service send monitoring data to a OnStar cloud database. GM then aggregates these data, builds models and sends the current state of the health of the vehicle from the cloud database back to the individual vehicle [46]. By aggregating data from all possible devices a single organization owns, that organization can work

with a larger dataset to have a more accurate representation of the population. This allows organizations to save money by spending time on a few models on a centralized database rather than creating and running models on thousands of individual devices. It also saves time as a few models are applicable to thousands of different edge devices.

Most IoT systems have similar processes where all the data from edge devices are collected on a centralized database. Then, different types of statistical models, such as predictive or maintenance, are then run on the collected data. Finally, those models are sent back to the edge devices so that the users have an improved experience. This process is called a centralized model [46]. This provides many advantages as it improves efficiency, saves on cost and models provide a more accurate representation of the population. However, there are two major issues. The two major issues are (1) security and (2) privacy.

Centralized databases collect, store and train models on all of the data collected from the organization's edge devices [45]. For GM's example, the data of hundreds of thousands of cars that are subscribed to the OnStar program are all on one database in one location. This can potentially be a big security concern. In an era where cybersecurity is a major concern, if a few individuals with malevolent intentions either hack or expose the database, then the data of every single edge device is leaked and exposed to the world. Instead of hacking hundreds of thousands of different cars, hackers can attack just one database, which would take less time due to the difference in quantity alone. Also, no hardware or software is perfect–there are reliability concerns and technology will malfunction or break down eventually. If a mistake happens, software has a bug or hardware breaks down, then the database can be broken and the organization would lose all the data instantly. Worse, the data can be leaked and the organization would be legally liable to every single user.

2.4.2 Concept

Federated Learning (FL) trains models on a central server, just like IoT, without having to collect individual data from its edge devices [47]. Around 20 years ago, mobile phones were just starting to emerge, laptops were beginning to become mainstream and cars were beginning to have more complicated sensors. The computing power in these edge devices were weak or nonexistent. The lack of computing power in these edge devices made FL impossible. It was a necessity to send all the data into a central database because big organizations were the only ones who could afford a big, powerful computer to run machine learning models.

However, in modern times, phones have computer chips that are more powerful than the computers that sent Apollo 11 to the moon [48]. All modern vehicles have computer chips and even home appliances have connection to the Internet. This development of computing power in edge devices allows those devices to run machine learning models and calculate model parameters. It also decreases the necessity of edge devices to send their data to a central database just to run machine learning models. Instead of sending the actual data, federated learning only sends model parameters. It stores all the data of the edge devices locally and each runs an individual machine learning algorithm on its own device. After the algorithm is run locally, it sends the local parameters to a global server and the global server collects all of the individual model parameters from the edge devices. It then calculates an optimization and gradient descent with those collected model parameters, outputs new parameters and sends those updated parameters down to the edge devices. The edge devices then run a new model with the new model parameters from the global server on its local data and repeats the process [47]. This process of sending updated model parameters back and forth between the global server and edge devices prevents local data from being shared with anyone else.

2.4.3 Patient Privacy

In the United States, HIPAA is a law that prevents healthcare companies and organizations from sharing patients' personal health information (PHI). It protects three areas of patients and their privacy [49].

1. "Portability of insurance or the ability of a patient/worker to move to another place of work and be certain that insurance coverage is not denied" [49]
2. "Detection and enforcement of fraud and accountability" [49]
3. "Simplify administrative procedures in health care and other professions" [49]

The first two directly apply to the need of patient privacy for the security of the individuals. The first is the denial of insurance during insurance swapping process. Employers have different insurance companies that they sign contracts with so that all of their employers that are eligible can have healthcare. Due to anti-competitive laws, there are numerous options for insurance companies that employers can choose from. The employees, however, do not have such options and would most likely have to take the healthcare provided by the employers unless the individuals want to pay more out of pocket to get a separate healthcare plan. This means that when an employee changes companies, they would most likely have to change insurance companies as well. However, the accepting insurance company cannot reject the healthcare of the incoming individual based on personally identifiable information. The first part of HIPAA protects this scenario by making it illegal to share personally identifiable healthcare information. The second part is more simple–it is to protect patients from fraud and fake accountability so that people cannot use other people's healthcare plans or incur costs that the individual did not use. Finally, the third part is to allow a smooth and safe process for patients. It is to ensure that patients can move from

one hospital to another or one doctor to another and their medical data will not be lost during transition of files for any reason [49].

HIPAA applies to all healthcare provides and administrative staff as well as researchers who study medical data. PHI can only be shared in specific circumstances. Some include discussing diagnosis and treatment of the patient with other healthcare providers, disclosing laboratory tests with other healthcare providers, or when calling a pharmacists for medication required by the patient. Generally, PHI can only be shared when offering treatment or advise that is directly involved with the patient [49]. This means that collecting data for a research that does not directly involve the health and safety of the patient, such as PTH, is not a reasonable grounds to disclose PHI.

There are some environments where sharing data is not possible due to privacy concerns. One example is the medical field. In the United States, under the Health Insurance Portability and Accountability Act (HIPAA), it is illegal for all hospitals and health insurance companies to reveal personally identifiable information of their patients or customers [32]. Legal jurisdictions, such as HIPAA, prevent implementation of IoT in some industries. However, it also prevents the advantages of efficiency, cost saving and an accurate representation that IoT systems bring. A new method that not only protects the privacy and security of individuals but also brings the advantages of IoT is called Federated Learning.

The lack of capability to share data impedes the collection and gathering of important or rare medical data, forcing research to be infeasible or expensive. Less data also means that it is less representative of the population. For example, a hospital in Los Angeles might have different representation of data than a hospital in Boston on the same clinical problem. The lack of representation and segmented subsets in currently available data can lead to biases in the research or results, making a pro-

posed medical theory or practice ineffective for some patients at best or dangerous at worst [50]. It is impossible to centralize the data and working on subsets of data are problems for researchers and patients. Federated learning allows hospitals to share information without sharing the data [51, 52, 50].

2.4.4 Federated Averaging

Federated Averaging (FedAvg) is one of the parameter aggregation techniques for federated learning. There are two steps to FedAvg. In step one, after each client would run its machine learning model on its local data, the parameters would be sent to the server. In step two, the server then averages all of the clients' parameters and sends the new averaged parameters back to each client. In a typical gradient descent model, these two steps would be repeated until an optimization is reached [53].

2.5 Uncertainty Quantification Problem

The late statistician George Box once said, "All models are wrong, but some are useful" [54]. The quote is often used to show that statistical models cannot entirely replicate the reality of the world. In order to grasp the difference, there needs to be some measure to quantify how much difference there is from the model created and the true reality of the world. Calculating the difference between model and reality and giving it a mathematical value of some sort is called uncertainty quantification [55]. The reason models do not represent the reality exactly is because of the error of the data or erroneous assumptions when constructing the model. That is why models without uncertainty quantification cannot be trusted.

In uncertainty quantification, there are two types of uncertainties: aleatoric and epistemic uncertainties. Aleatoric uncertainties come from the data and how random it is or how it is not completely representative of the population. This randomness is

part of the data and cannot be reduced. Epistemic uncertainties are created when a lack of data creates uncertainties of the model. This is captured through confidence intervals and be one of the main aspect of this paper [56].

The data uncertainty, called aleatoric uncertainty, stems from randomness of the data [57]. This can range from people collecting the wrong data, mislabeling data or just missing data. Because the dataset is wrong to begin with, any model that tries to do prediction using this faulty dataset would not provide accurate or trustworthy results. Unfortunately, nothing can be done to reduce data error and uncertainty is quantified by adding an error term. On the other hand, uncertainty on model assumptions and errors stem from incorrect model assumptions and incorrect parameters, called epistemic uncertainty [57]. This can be the result of a small dataset that is not a good representation of the population or a large dataset that does not have proper training data [58]. These uncertainties can be quantified using variance of the model parameters. Correctly quantifying the uncertainties of a model will present a more accurate representation of the real world.

2.5.1 Bootstrapping

Bootstrapping is a resampling technique to estimate the true variation of population parameters when there is insufficient data [59]. For example, one model using one dataset would output only one mean. Without boostrapping, there would only be one mean and it would be very difficult to find the distribution of that data's mean. Bootstrapping allows us to find that distribution by picking each sample, with replacement, from that dataset. This would create multiple new datasets based on one dataset and allow the model to run multiple times to calculate multiple means. This generation of multiple means creates a distribution of means that would be an esti-

mate of the true population mean. The distribution can generate standard deviations and confidence intervals that were previously unable to be calculated.

III. Methodology

Preamble

This chapter discusses The CARE Consortium dataset and particular features of interest that were chosen for subsequent analysis. On the final two feature subsets chosen, a basic linear regression is performed, that simulates a centralized database, and a FL using linear regression, that simulates a potential methodology for running better models without sharing data. Finally, the uncertainties of both methodologies are quantified to determine the variation of each model to show that uncertainty quantification (UQ) in FL is a viable model.

3.1 NCAA-DoD Grand Alliance CARE Consortium Dataset and Feature Selection

The dataset was collected from student athletes in Division 1 programs from 30 different universities across the United States from 2014 to 2018. There are three main timelines for each patient used in this research: baseline, 24-48 hours and 6 months post injury. Before the start of every season, each athlete is required to do a diagnostic concussion test, which is labeled as baseline. Throughout the season, if the athlete is concussed, they are required to receive the same concussion test within 24-48 hours and after 6 months to check whether the score changed to gauge whether the athlete's concussion was treated or not [23]. However, there is a lot of missing data in the dataset because the data was collected from a variety of universities, each with their own practices and protocols on concussion treatment. There is very limited standardization on the diagnosis and treatment of concussed individuals in the medical industry, let alone in the NCAA. Even the most common tests have been proven to be unreliable at times but necessary due to a lack of standardization in

how to diagnose a concussion [60]. There are a total of 37 different clinical tests that can possibly be used to diagnose a patient for a concussion. In reality, doctors use a mixture of those different tests. This lack of standardization allows doctors and physicians at each university to administer their own concussion procedures and tests, which leads to different athletes taking one diagnosis test over the other. The different procedures affect the dataset because not everyone is taking the same tests or the same number of tests. Some universities have students that take only one or two tests while other universities might take 20-30 tests. This imbalance creates a lot of features but a lot of the data in those features are missing, creating missing data in the aggregated dataset.

In order to mitigate the negative impacts of the missing data as much as possible, the focus was on the features that were heavily used in other studies, such as the SCAT3 score or standardized assessment of concussion (SAC) score, and had a complete dataset in accordance with other features. This means that patients that had the same combination of features were studied only if those features all contained data. In the end, 12 different scores were obtained, each with all three of the timelines, and obtained the best possible combination by comparing each combination based on the highest R^2 and R^2-adjusted scores.

Sport Concussion Assessment Test (SCAT) is a collected of 22 reported symptoms that are self-rated on 7-point scale (0: no symptoms, 6: severe symptoms) [24]. Figure 4 shows the SCAT3 scorecard that scores individual symptoms ranging from 0, no symptoms, to 132, the highest score with worst symptom [4]. As shown in the figure, the SCAT3 tests 22 different sub-tests that is self-reported by the athlete, who grades each question based on a score of 0 to 6, with 0 feeling no symptom at all and 6 feeling severe symptom. The maximum score that is available for the SCAT3 test is

132 and the lowest score is 0. The maximum number of symptoms the patient can mark is 22; the lowest is 0.

How do you feel?

"You should score yourself on the following symptoms, based on how you feel now".

	none	mild		moderate		severe	
Headache	0	1	2	3	4	5	6
"Pressure in head"	0	1	2	3	4	5	6
Neck Pain	0	1	2	3	4	5	6
Nausea or vomiting	0	1	2	3	4	5	6
Dizziness	0	1	2	3	4	5	6
Blurred vision	0	1	2	3	4	5	6
Balance problems	0	1	2	3	4	5	6
Sensitivity to light	0	1	2	3	4	5	6
Sensitivity to noise	0	1	2	3	4	5	6
Feeling slowed down	0	1	2	3	4	5	6
Feeling like "in a fog"	0	1	2	3	4	5	6
"Don't feel right"	0	1	2	3	4	5	6
Difficulty concentrating	0	1	2	3	4	5	6
Difficulty remembering	0	1	2	3	4	5	6
Fatigue or low energy	0	1	2	3	4	5	6
Confusion	0	1	2	3	4	5	6
Drowsiness	0	1	2	3	4	5	6
Trouble falling asleep	0	1	2	3	4	5	6
More emotional	0	1	2	3	4	5	6
Irritability	0	1	2	3	4	5	6
Sadness	0	1	2	3	4	5	6
Nervous or Anxious	0	1	2	3	4	5	6

Total number of symptoms (Maximum possible 22)

Symptom severity score (Maximum possible 132)

Figure 4: SCAT3 Scorecard [4]

3.2 Centralized Model

The centralized model simulates the concussed athletes all being in one hospital. Although this is not realistic, as these athletes were gathered from 30 different universities, the centralized model is the standard in which the FL model will be compared.

The linear model used:

$$y = x_k{}'\beta + \epsilon \tag{4}$$

where $\beta = (\beta_0, \beta_1, ..., \beta_k)$ for k number of features.

y is the response variable, k is the number of features, β_0 is the constant coefficient, β_k is the coefficient for the k^{th} feature, x_k is the k^{th} feature, and ϵ is the error. The two response variables for y are SCAT3 score and SCAT3 number of symptoms. The SCAT3 score is the score of the diagnosis test of the SCAT test. A Sport Concussion Assessment Test (SCAT) 3 test is the most widely used diagnosis test for concussion in individuals who are believed to have suffered from a concussion [24].

The two main response variables each has its own set of features. The total number of combination of features from the 12 for each response variable was 559. Using the 559 total combinations, the best possible combination of features for each response variable recorded is calculated on four different time periods based on R^2 and R^2-adjusted scores. In the end, eight different possible combinations are compared.

3.3 Federated Learning Model

Compared to the central model, the FL model simulates more realistically how hospitals work and take HIPAA into account. This situation is simulated by creating four clients, representing four hospitals, along with a central server. This is modeled below.

There are four clients and one server. After each client runs a certain machine learning model, the end parameters end up in these four equations.

$$y_{client1} = x_{k,1}{}'\beta_1 + \epsilon_1 \tag{5}$$

$$y_{client2} = x_{k,2}{}'\beta_2 + \epsilon_2 \tag{6}$$

$$y_{client3} = x_{k,3}{}'\beta_3 + \epsilon_3 \tag{7}$$

$$y_{client4} = x_{k,4}{}'\beta_4 + \epsilon_4 \tag{8}$$

After these parameters are calculated in the client side, the parameters from each equation are sent to the server and are then averaged. Thus, the server model, with n being the number of clients, ends up looking like this:

$$y_{server} = \frac{\beta_1 + \beta_2 + \beta_3 + \beta_4}{n} * x_{server} + \epsilon_{server} \tag{9}$$

Simplifying,

$$y_{server} = avg(\beta) * x_{server} + \epsilon_{server} \tag{10}$$

Finally, these new model parameters are sent back to the clients to restart the federated process. This equation is assuming that each client has the same number of samples. This means that each client is weighted equally in the averaging process. However, if the number of samples are not equally distributed in each client, the clients are weighted differently based on the number of samples.

For example, if K is the total number of samples, there are four clients and μ is the percentage of samples in each client based on K, then:

$$K + clienti = \mu_i \tag{11}$$

and

$$1 = \mu_1 + \mu_2 + \mu_3 + \mu_4 \tag{12}$$

Using μ into the original federated averaging equation, the new equation for federated averaging not assuming each client has the same number of data samples would look like this:

$$y_{server} = (\mu_1\beta_1 + \mu_2\beta_2 + \mu_3\beta_3 + \mu_4\beta_4)x_{server} + \epsilon_{server} \tag{13}$$

The reason four clients is used is because of the limitation of the size of the dataset. Most of the datasets tested were between 100 and 200 samples. The reason is that linear regression does not work with incomplete data in the features. In a dataset filled with missing data, the size of the datasets tested had to be trimmed significantly to create a complete dataset with all the data in the features for the model to work.

Using the same dataset from the central model, the data is randomly shuffled and split into four clients evenly, meaning each client would get the same number of samples. Then, each client would run its own linear regression model and calculate model parameters and R^2 and R^2-adjusted scores. These parameters and scores are then sent to a simulated central server that averages the values calculated from each client. This simulates the FedAvg modeling. In a more traditional FL model using a neural network, the model parameters and averages calculated from the server are sent back to the clients to run a new iteration on the same data and model but with the new received parameters. This process would continue iteratively until an optimal is reached. However, because linear regression is a closed-form solution, there is no need for the iterative process like a neural network. This means that the values calculated from the server do not go back to the client.

These R^2 and R^2-adjusted scores are then compared to the corresponding central

model values in order to see if there is a loss in information by only sharing model parameters instead of the actual data. If the R^2 and R^2-adjusted scores for the FL model are lower than corresponding values from the centralized model, then there is a loss in information during the process in the federated model and may not be a viable option for healthcare purposes. The reason is that every piece of information regarding patients and their health records can be paramount in the health and safety of the individual. If the values are the same or higher than the central model, then this would mean that not only is the FL model not losing patient information but also learning more by sharing model parameters rather than collecting data in a central database. This would show that the federated model can be seen as a viable solution to the problem of HIPAA, where hospitals cannot share data.

3.4 Uncertainty Quantification Federated Learning (UQFL)

The biggest reason a linear regression model is used is to be able to calculate the uncertainty in a FL model setting. This allows us to see with a $1 - \alpha$ confidence how accurate the parameters of the individual features are to the real world. UQFL calculates the individual variances of the features for each client. Then, it calculates a global model variance for each feature by using FedAvg and averaging the variances from individual clients. Finally, using the global variance, it calculates the confidence intervals of each feature with a $1-\alpha$ confidence. Figure 5 visualizes the UQFL system, with four hospitals acting as local clients, sharing variances to generate confidence intervals.

The formulation is modeled below.

Client: The variances of each model parameters are calculated for each client.

$$y_{i,k} = {x_{i,k}}'\beta_k + \epsilon_{i,k} \text{ for } i^{th} \text{ value in } K \text{ client, assuming } \epsilon_{i,k}\ N(0, \sigma_k^2) \tag{14}$$

$$\hat{\beta}_k = (\vec{x_k'}\vec{x_k})^{-1}\vec{x_k'}\vec{y_k} \tag{15}$$

$$Var(\hat{\beta}_k) = Var((\vec{x_k'}\vec{x_k})^{-1}\vec{x_k'}\vec{y_k}) \tag{16}$$

$$Var(\hat{\beta}_k) = (\vec{x_k'}\vec{x_k})^{-1}\vec{x_k'}\vec{x_k}(\vec{x_k'}\vec{x_k})^{-1}Var(\vec{y_k}) \tag{17}$$

$$Var(\hat{\beta}_k) = (\vec{x_k'}\vec{x_k})^{-1}\sigma_k \tag{18}$$

$$\hat{\sigma_k^2} = \frac{1}{n-p}\sum_{i=1}^{n}(y_{k,1} - \hat{y_{k,i}})^2 \tag{19}$$

$$Var(\hat{\beta}_k) = (\vec{x_k'}\vec{x_k})^{-1}\hat{\sigma_k} \tag{20}$$

Server: The client parameter variances are aggregated and averaged in the server.

$$Var(\hat{\beta}) = (\sum_{K=1}^{K} Var(\hat{\beta_K})) \tag{21}$$

Confidence Intervals: Confidence intervals are calculated using the final server parameter variances with a $1-\alpha$ confidence.

$$[lower, upper] = \hat{\beta_K} \pm z_{\frac{\alpha}{2}}\hat{\sigma}_{\beta_K} \tag{22}$$

where:

$\hat{\beta}_k$ = Estimate of Coefficient for Client k

$\hat{\beta}$ = Estimate of Coefficient for Server

$\hat{\sigma_k}^2$ = Variance estimate for Client k

$\hat{\sigma_\beta}$ = Std Dev for Server Coefficient

K = Number of Clients

$z_{\frac{\alpha}{2}}$ = z-score

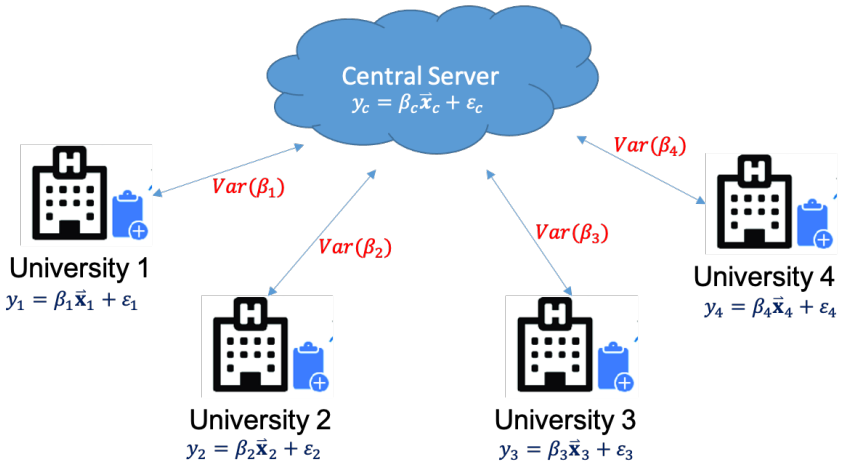


Figure 5: Example of UQFL with Four Hospital Clients

IV. Results and Analysis

4.1 Choosing Subsets with the Dataset

Shown in the Appendix 13, the best R^2 and R^2-adjusted scores come from the time periods in which the response variable is the 6 months post injury time scores and the features are the baseline time scores. For this time period, the two response variables tested, SCAT3 Total Number of Symptoms and SCAT3 Total Score, had two distinct set of features that gave the highest R^2 and R^2-adjusted scores. For SCAT3 Total Number of Symptoms, there were five features used in the model: Satisfaction with Life Scale (SWLS) Score, SCAT3 Total Number of Symptoms, Standardized Assessment Concussion (SAC) score, Vestibular Ocular Motor Screening (VOMS) score and Brief Symptom Inventory (BSI) 18 Score. For SCAT Total Score, there were four features used in the model: SCAT3 Total Score, Clinical Reaction Time (CRT) score, SAC score and BSI 18 score. Because the two response variables have their own data, the datasets will be called subsets.

4.1.1 Features Explained

BSI 18 is a self-reported checklist that has 18 symptoms listed. It measures a patient's brain health. The patient records each symptom listed in the test on a five-point scale based on how much that specific symptom bothers the patient (0: no symptoms, 5: felt worst pain). The total scores can range from 0 to 90 (0: no symptoms, 90: all symptoms, felt worst pain on all symptoms). The patient is considered to have a healthier brain if he has a lower BSI 18 score [61].

SAC measures a patient's brain health for concussion. It is separated into four parts: Orientation Score, Immediate Memory Score, Concentration Score and Delayed Recall Score. The Orientation score has five questions that asks the patient the

month, date, day of week, year and time. The patient receives one point from each correctly answered question for a maximum of five points. The Immediate Memory Score has five words that the patient must recall in order immediately after the test administrator reads them aloud. This test is repeated three times and the patient receives one point from each word remembered correctly for a maximum of 15 points. The Concentration Score has five questions. Four of the questions require the patient to repeat a series of numbers read by the test administrator backwards. For example, the test administrator reads 5-8-3, the patient must answer 3-8-5. The fifth question requires the patient to recall the months of the year backwards starting from December. The patient receives one point from each correctly answered question for a maximum of five points. Finally, the Delayed Recall requires the patient to recall the five words from the Immediate Memory Score in any order. The patient receives one point from each correctly answered question for a maximum of five points [62]. All the points are summed up for a maximum score of 30 and a minimum score of 0. The patient is considered to have a healthier brain if they has a higher SAC score.

SWLS measures the patient's current life satisfaction. The questions consists of five statements: "In most ways my life is close to my ideal.", "The conditions of my life are excellent.", "I am satisfied with my life.", "So far I have gotten the important things I want in life." and "If I could live my life over, I would change almost nothing.". The patient places a value on each question based on a seven point scale (1: "strongly disagree", 7: "strongly agree"). The responses are summed up with a maximum score of 35 and a minimum score of 5. A patient is considered to have a higher life satisfaction if they has a higher SWLS score [63].

VOMS measures a patient's visual and balance control. It has eight categories: a baseline and seven sub-tests. The sub-tests include smooth pursuit, saccades (Horizontal) saccades (Vertical) convergence (Near Point), Vestibular-Ocular Reflex (VOR)

Horizontal, VOR Vertical Visual Motion and sensitivity. These tests include checking if the patient gets double vision looking at an object in close proximity (convergence) or if the patient can maintain eye contact with an object while the patients head moves side to side (VOR Horizontal). Each category measures four symptoms: headache, dizziness, nausea, fogginess. These four symptoms are measured at baseline and once again after conducing each sub-test. Each symptom for each category is measured on a ten-point scale (0: feels no symptom, 10: feels most symptom). The responses are summed up with a maximum score of 320 and a minimum score of 0. A patient is considered to have a better visual and balance control, thus better brain health, if he has a lower VOMS score [64].

CRT measures the hand eye coordination of a patient through reaction time. The test administrator holds a stick in the air and the patient places his hand at the bottom end of the stick. The test administrator drops the stick and the patient has to catch the falling stick as fast as possible. An example of how this test is conducted is shown in Figure 6. The time to catch the stick is measured as a reaction time. The patient is considered to have a healthier brain if he has a faster reaction time [65].

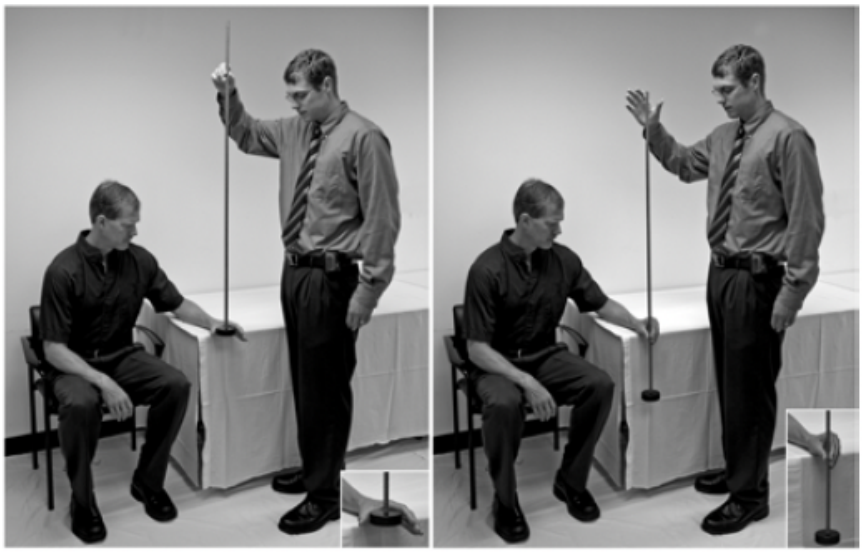


Figure 6: CRT Test Example [4]

4.1.2 Transformation

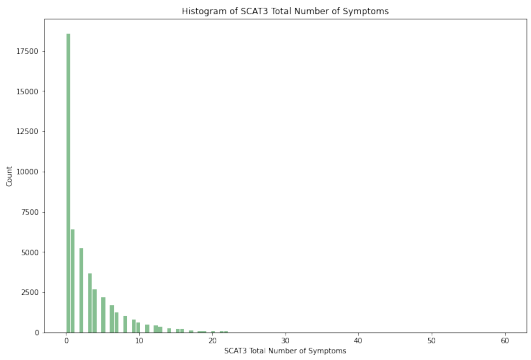
Figure 7 shows that the response variables of interest (SCAT3 Total Score and SCAT3 Total Symptoms) take on a Poisson distribution. As a result, I transformed the y-variable by square rooting its values, shown in Figure 8, ($y' = \sqrt{y}$, because the variance of the square root of a Poisson random variable is independent of the mean [5]. The transformation of the response variable allows the variance to be more stable, predict better and meet the constant variance assumption. However, the model did not perform better compared to the original, untransformed data. The average values of the R^2 and R^2-adjusted scores for the transformed y-values for 1000 runs were about the same as the average values of the R^2 and R^2-adjusted scores for the original y-values for both central and FL models shown by Table 1. Given that the data is mostly ordinal data and there was little improvement of the results for the transformed data, the untransformed data was chosen in order to maximize results [66].

Table 1: Transformed y-values for SCAT3 Total Number of Symptoms

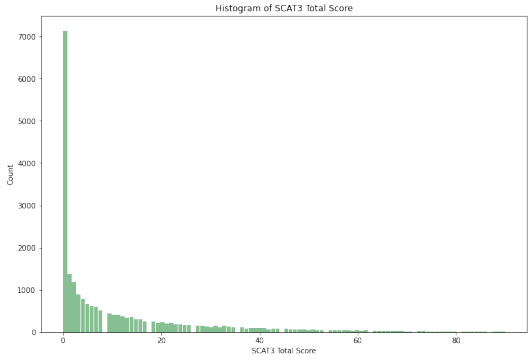
Response Variable	**Features**	**95% Confidence Interval for Central Model**	**95% Confidence Interval for FL Model**	**Number of DataPoints**
SCAT3TotalSymptoms Transformed	Intercept	(-5.5634, 2.7096)	(-6.0847, 3.2308)	102
	Satisfaction with Life Scale.SWLSTotalScore	(0.0012, 0.0032)	(0.0012, 0.0032)	
	SCAT3TotalSymptoms	(0.1115, 0.1185)	(0.1099, 0.1202)	
	SACScore	(0.0492, 0.0573)	(0.0483, 0.0581)	
	VOMS Scoring.VOMSTotalScore	(0.0047, 0.0051)	(0.0046, 0.0052)	
	BSI18Score	(0.0089, 0.0134)	(0.0064, 0.0159)	
R^2 Central (Mean±Std Dev)	**R^2 FL (Mean±Std Dev)**	**R^2-adjusted Central (Mean±Std Dev)**	**R^2-adjusted FL (Mean±Std Dev)**	**Number of Trials**
(0.1048±0.2842)	(0.2481±0.2095)	(-0.1677±0.3707)	(0.0192±0.4535)	1000
R^2 (25^{th} Percentile, Max) Central	**R^2 (25^{th} Percentile, Max) FL**	**R^2-adjusted (25^{th} Percentile, Max) Central**	**R^2-adjusted (25^{th} Percentile, Max) FL**	
(-0.0084, 0.6236)	(0.1234, 0.7649)	(-0.3152, 0.509)	(-0.1433, 0.6933)	
MSE (Mean±Std Dev) Central	**MSE (Mean±Std Dev) Empirical Total**	**MSE (25^{th} Percentile, Max) Central**	**MSE (25^{th} Percentile, Max) Empirical Total**	
(0.6056±0.184)	(0.5113±0.1573)	(0.1648, 1.4166)	(0.3416, 1.0427)	

4.1.3 Normalization

SWLS, SAC, BSI 18, SCAT3 and VOMS are all ordinal variables. They have a clear ordering to their values and a lower score has one meaning for a particular test



(a) Histogram of SCAT3 Total Number of Symptom Response Variables



(b) Histogram of SCAT3 Total Score Response variables

Figure 7: Distribution of Potential Response Variables

Relationship of σ^2 to E(y)	Transformation
$\sigma^2 \propto$ constant	$y' = y$ (no transformation)
$\sigma^2 \propto E(y)$	$y' = \sqrt{y}$ (square root; Poisson data)
$\sigma^2 \propto E(y)[1 - E(y)]$	$y' = \sin^{-1}(\sqrt{y})$ (arcsin; binomial proportions $0 \le y_i \le 1$)
$\sigma^2 \propto [E(y)]^2$	$y' = \ln(y)$ (log)
$\sigma^2 \propto [E(y)]^3$	$y' = y^{-1/2}$ (reciprocal square root)
$\sigma^2 \propto [E(y)]^4$	$y' = y^{-1}$ (reciprocal)

Figure 8: Variance-Stabilizing Transformations based on Expected y-values [5]

and higher score means another for a different test. CRT is a continuous variable as it is the only test that is measured with time. As a result, the CRT scores are normalized at the very beginning. However, the ordinal variables are not. This section shows that the normalization of the ordinal data does not improve the R^2 and R^2-adjusted scores. Using the SCAT3 Total Number of Symptoms at 6 months post injury as a response and SCAT3 Total Number of Symptoms, SWLS, SAC, VOMS, BSI scores at baseline as features, figures below show the comparison of the R^2 and R^2-adjusted scores for the original subset and the normalized subset. As shown in Table 2, the normalized subset is only marginally better than the original subset in terms of R^2 and R^2-adjusted scores. This pattern is also true for SCAT3 Total Score at 6 months and SCAT3 Total Score, CRT, SAC, and BSI 18 scores at baseline. Given that the data

is mostly ordinal data and there are many challenges involving normalizing medical data, I chose the original subset as the final subsets to perform tests on [66].

Table 2: Normalized y-values for SCAT3 Total Number of Symptoms

Response Variable	Features	95% Confidence Interval for Central Model	95% Confidence Interval for FL Model	Number of DataPoints
SCAT3TotalSymptoms Normalized	Intercept	(-0.0218, 0.0195)	(-0.0266, 0.0243)	102
	Satisfaction with Life Scale.SWLSTotalScore	(-0.0156, 0.0348)	(-0.0162, 0.0354)	
	SCAT3TotalSymptoms	(0.3961, 0.5022)	(0.3737, 0.5245)	
	SACScore	(0.1002, 0.1447)	(0.0958, 0.1491)	
	VOMS Scoring.VOMSTotalScore	(0.0246, 0.0834)	(0.0158, 0.0922)	
	BSI18Score	(-0.0039, 0.1230)	(-0.0690, 0.1881)	
R^2 Central (Mean±Std Dev)	R^2 FL (Mean±Std Dev)	R^2-adjusted Central (Mean±Std Dev)	R^2-adjusted FL (Mean±Std Dev)	Number of Trials
(0.1079±0.2668)	(0.2467±0.2142)	(-0.1636±0.3479)	(0.0174±0.2795)	1000
R^2 (25^{th} Percentile, Max) Central	R^2 (25^{th} Percentile, Max) FL	R^2-adjusted (25^{th} Percentile, Max) Central	R^2-adjusted (25^{th} Percentile, Max) FL	
(-0.0181, 0.6247)	(0.1134, 0.7612)	(0.1134, 0.7612)	(-0.1565, 0.6885)	
MSE (Mean±Std Dev) Central	MSE (Mean±Std Dev) Empirical Total	MSE (25^{th} Percentile, Max) Central	MSE (25^{th} Percentile, Max) Empirical Total	
(0.8292±0.2544)	(0.7015±0.217)	(0.279, 2.1588)	(0.1912, 1.4717)	

4.1.4 Data Imputation

In order to combat the missing data, I implemented a K-near neighbor (kNN) data imputation technique. kNN imputes data using data proximity [67]. The major concept of kNN data imputation is to group data that are close to one another so that if a new or missing data is introduced, kNN estimates the value of the missing data based on how close it is to other groups. The imputed value is the average of the closest group of data [68]. However, increasing the subset size did not improve the model scores. Using the SCAT3 Total Score data as the original subset, the original and the imputed subset are compared to see which subset performs better with higher R^2 and R^2-adjusted scores for both the central and federated models, based on 1000 iterations. The original subset had 103 data points and the imputed data had 1579 data points. Despite having more data to work with, shown in Appendix A, the imputed subset performed worse than the original subset in almost every score that was measured. In the original subset, the mean R^2 and R^2-adjusted scores for the central model were 0.14408 and -0.0699 respectively. For the imputed subset, the mean R^2 and

R^2-adjusted scores for the central model were 0.0679 and 0.0457 respectively. In the original subset, the mean R^2 and R^2-adjusted scores for the FL model is 0.3909 and -0.2387 respectively. For the imputed subset, the mean R^2 and R^2-adjusted scores for the FL model is 0.0878 and 0.0661 respectively. In both central and FL models, the original subset scores much higher on the R^2 but the imputed subset scores marginally higher on the R^2-adjusted scores. Also, the max R^2 and R^2-adjusted scores were much higher for the original subset, scoring in the 80 and 90 percentile, while the imputed subset scores maxed out in the 10 percentiles. Due to the results, I chose not to impute the data and use the original subset to perform tests on.

4.1.5 Data Assumptions

4.1.5.1 SCAT3 Total Number of Symptoms

1. Linearity
    - Figure 9 shows a fairly linear relationship between the the SCAT3 Total Number of Symptoms and its residuals.
2. Normality
    - Figure 10 shows a pretty normal distribution of SCAT3 Total Number of Symptoms Residuals.
3. No Multicollinearity Among Features
    - Figure 11 shows little multicollinearity among SCAT3 Total Number of Symptoms features.
4. No Auto-Correlation of Residuals
    - The Durbin-Watson value = 1.9159

- Little to no autocorrelation of SCAT3 Total Number of Symptoms residuals.

5. Homoscedasticity

- Figure 12 shows no apparent pattern in SCAT3 Total Number of Symptoms residuals.

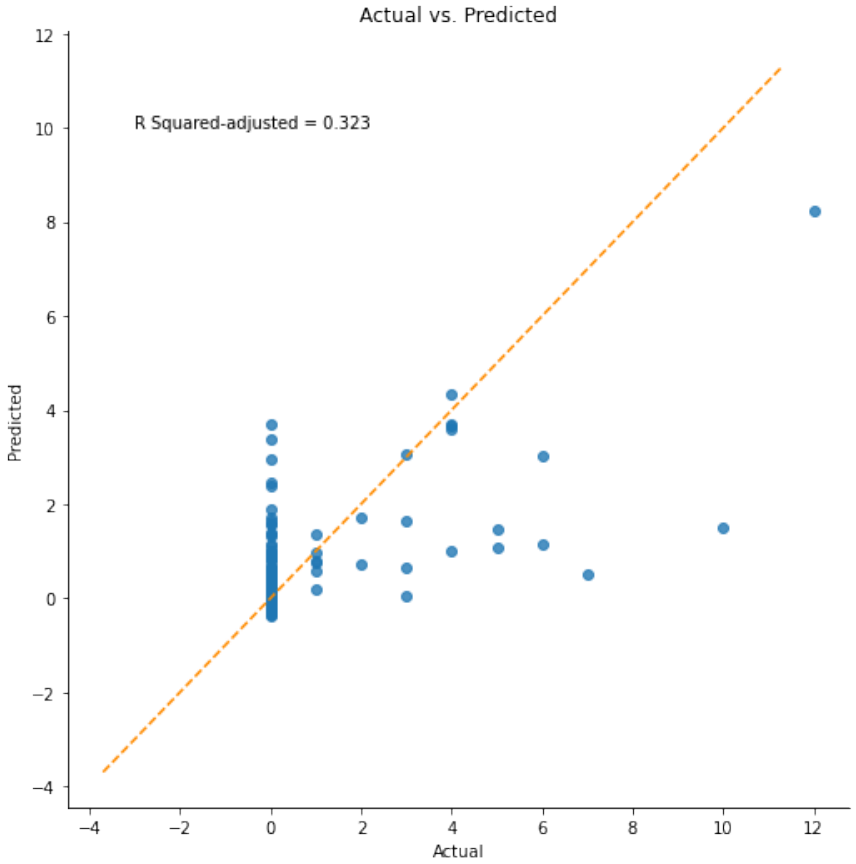


Figure 9: Linearity Test for SCAT3 Total Number of Symptoms

4.1.5.2 SCAT3 Total Score

1. Linearity

- Figure 13 shows a fairly linear relationship between the the SCAT3 Total Score and its residuals.

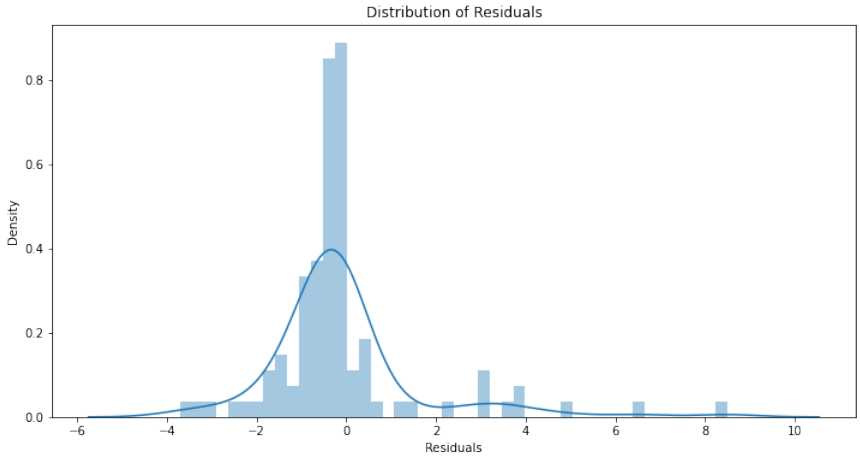


Figure 10: Normality Test for SCAT3 Total Number of Symptoms

2. Normality

    - Figure 14 shows a pretty normal distribution of SCAT3 Total Score Residuals.

3. No Multicollinearity Among Features

    - Figure 15 shows little multicollinearity among SCAT3 Total Score features.

4. No Auto-Correlation of Residuals

    - The Durbin-Watson value = 2.4058
    - Little to no autocorrelation of SCAT3 Total Score residuals.

5. Homoscedasticity

    - Figure 16 shows no apparent pattern in SCAT3 Total Score residuals.

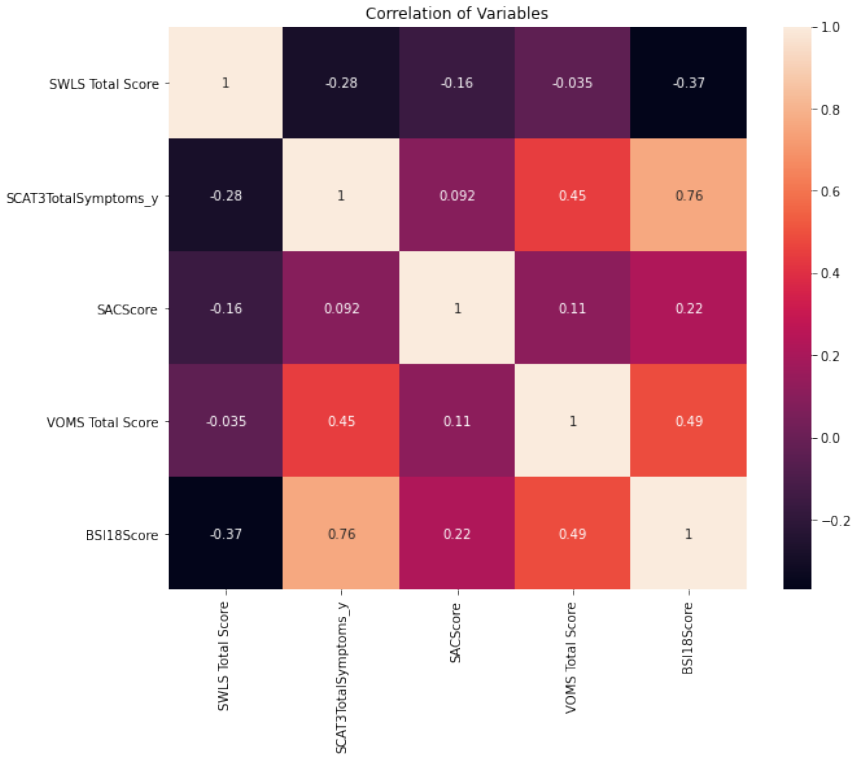


Figure 11: Multicollinearity Test for SCAT3 Total Number of Symptoms

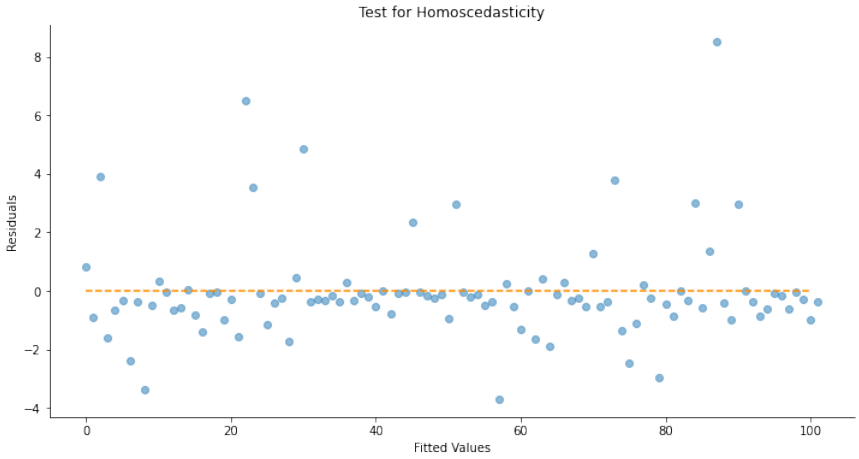


Figure 12: Homoscedasticity Test for SCAT3 Total Number of Symptoms

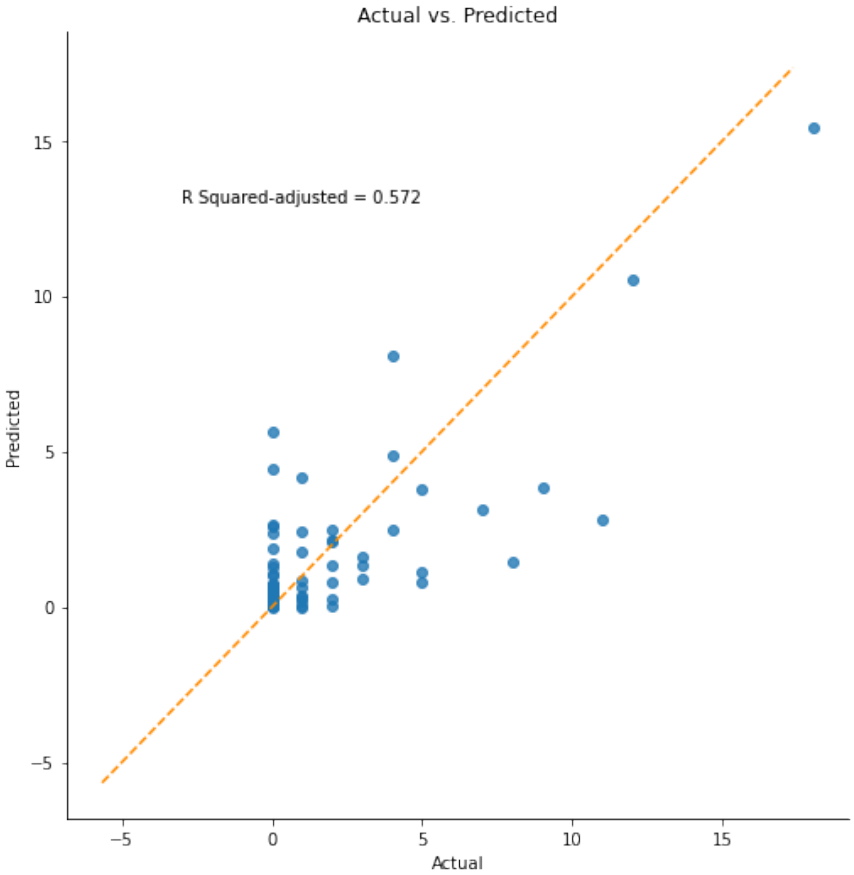


Figure 13: Linearity Test for SCAT3 Total Score

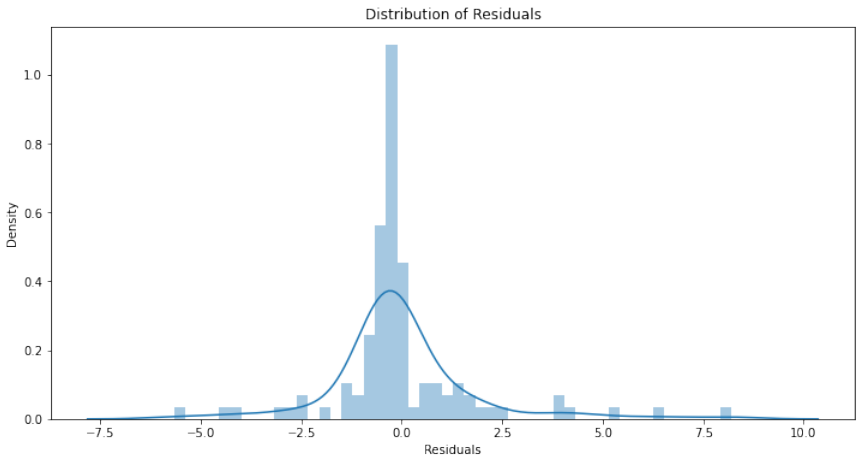


Figure 14: Normality Test for SCAT3 Total Score

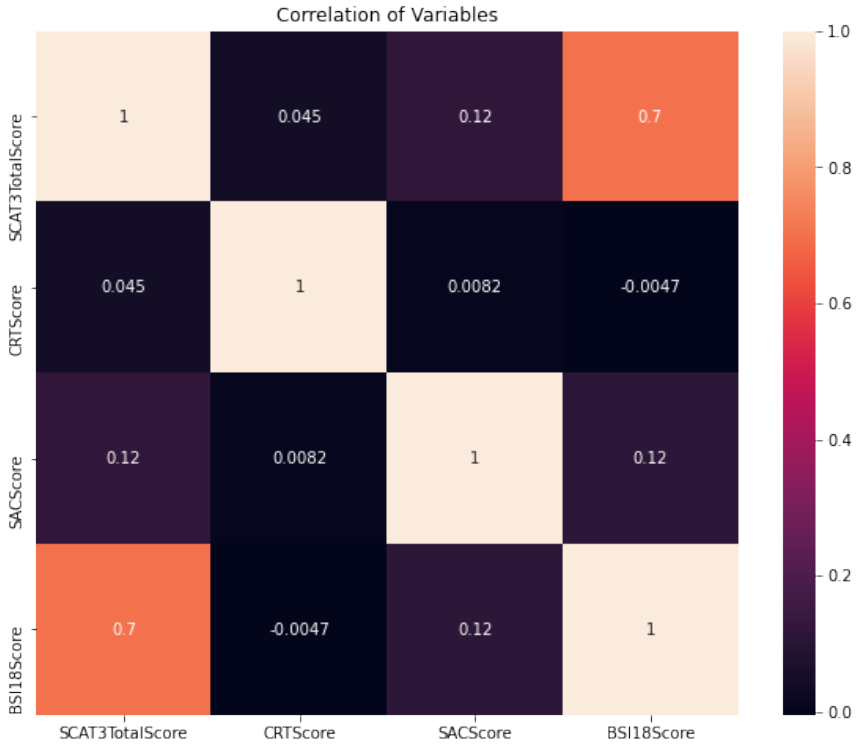


Figure 15: Multicollinearity Test for SCAT3 Total Score

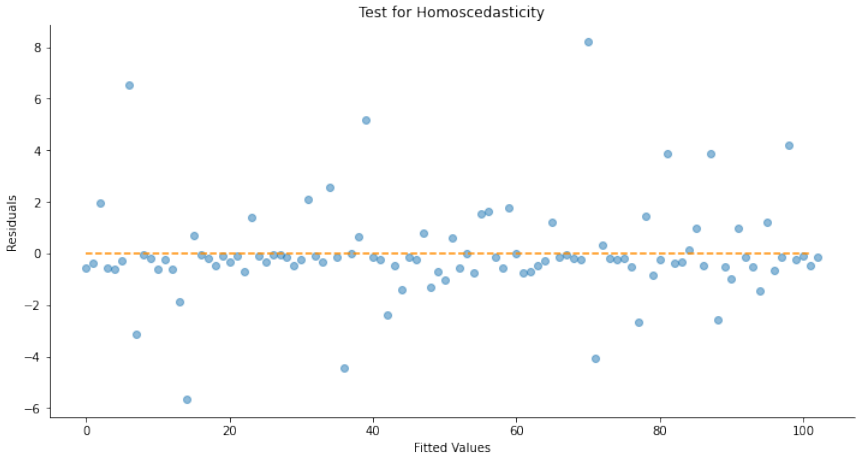


Figure 16: Homoscedasticity Test for SCAT3 Total Score

4.2 Final Tests

4.2.1 FL vs Central

The final tests done on both the SCAT3 Total number of Symptoms and SCAT3 Total Score each has three sections: FL vs Central comparisons, Client vs Total Bootstrapping comparisons, and Percentile vs Empirical Bootstrapping comparisons. First, the tests compare the differences based on three main scores: R^2 scores, R^2-adjusted scores, and Mean Squared Error (MSE) scores. For each of the three scores, there is a mean, along with a standard deviation, and a minimum (min) and maximum (max) value for each score. The min values are actually the 25^{th} percentile of each test. The reason is that there are anomalies with the size, structure and randomness of the test subsets that causes the R^2 scores and R^2-adjusted scores to output values that are wildy different. For example, Figure 17 shows a histogram of 1000 R^2 scores for the FL model of the SCAT3 Total Number of Symptoms subset. The histogram shows values of -3 and -1.5 for the R^2 scores, but there are only 1 to 5 of those values out of 1000. Coupled with the fact that R^2 scores can only range from -1 to 1, these values are seen as outliers, which is why the 25^{th} percentile is used as the min values. In addition to these three tests, there is a confidence interval associated with each model parameter and the intercept for both the central and FL models. The number of data points used for each of the two subsets is also provided.

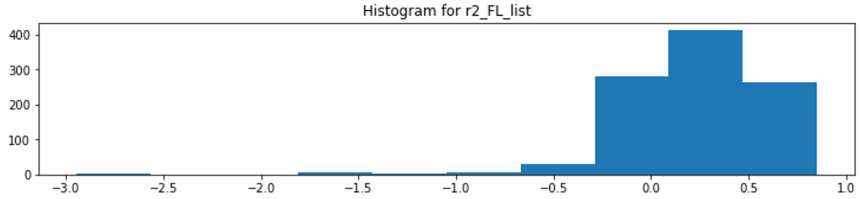


Figure 17: Histogram of R^2 for FL for SCAT3 Total Number of Symptoms (1000 Runs)

Furthermore, for the FL vs Central comparisons, the results display a histogram of the R^2 and R^2-adjusted scores for each model. For all the tests, with the exception of Empirical Total Bootstrapping at 300 iterations for SCAT3 Total Number of Symptoms, the results exhibit confidence intervals (CI) for all of the subset features based on the number trials. Even in the exceptional case, the only parts that the central model scored better than the FL model was the 25^{th} percentile and max MSE, where the central model had lower max and higher 25^{th} percentile scores compared to the FL model. For the client bootstrapping, there are four tests, each with different number of trials: 1000, 3000, 5000, and 10000 trials. For the total bootstrapping, there are also four tests with different numbers of trials. However, due to a lack of computational power and limited time, the numbers of trials are 300, 500, 1000 and 2000. Finally, the times it took to complete the tests are shown for the corresponding number of trials.

4.2.2 Client vs Total Bootstrapping

The client and total bootstrapping technique comparisons show the CI of the model parameters for both the SCAT3 Total Number of Symptoms and SCAT3 Total Score subsets for the FL models only. The difference between the two techniques is that the client bootstrapping technique re-samples data after the subset is split evenly into the four clients. The total bootstrapping technique re-samples data before the is split into the clients. The client technique is more representative of how hospitals and research organizations would operate FL models and the total technique is a simulation to show a comparison. The reason is that in the real world, hospitals do not have access to a centralized database nor does a centralized database of patient information exist. As a result, the total bootstrapping technique would not be possible in the present application.

4.2.3 Percentile vs Empirical Bootstrapping

The original central and FL models made distributional assumptions; both models assume a normal distribution of the data in order to calculate model parameters and its variances. The client and total bootstrapping techniques explained in the previous section uses a percentile bootstrapping technique. Percentile bootstrapping is a parametric technique that assumes that the subset can be modeled by a distribution. However, parametric bootstrapping is sensitive to the data and its procedures. It only works well if the data is a true representation of the real world. However, the model cannot make this assumption for the subset due to myriad of different issues such as missing data and faulty data collection procedures. Because the data used for both the client and total bootstrapping techniques assumes a normal distribution, the two techniques would also have similar faults regarded with parametric bootstrapping. This means that the client and total bootstrapping used in these tests are parametric bootstrapping techniques.

In order to combat the sensitivity of parametric models, an empirical bootstrapping technique is tested. Empirical bootstrapping technique is a non-parametric procedure, meaning the distribution of the data is not assumed. Empirical bootstrap re-samples the data just like the percentile distribution. However, for each iteration of sampling, it subtracts the model parameter of the FL model from percentile bootstrap parameters. This is done for a certain number of iterations and creates a CI of those differences. Finally, at the end, I add the true model parameter to the CI in order to center around the original data and get an estimated true distribution. Equation 23 shows the logic that is used to estimate the true distribution of the data. The distribution of F_{FL}, FL model, is assumed in order to estimate distribution of F_B, percentile bootstrap. Then, the newly formed relationship is used to assume that relationship would be similar between the distribution of F_{FL} and the distribu-

tion of F, an estimate of the true population distribution. The formulation of the distribution is created using Equation 24.

$$F_B \rightarrow F_{FL} \approx F_{FL} \rightarrow F \tag{23}$$

where:

F_B = Distribution of Percentile Bootstrap

F_{FL} = Distribution of the FL Model

F = Estimate of True Population Distribution

$\approx$ = Mimics

$$(\hat{\beta_k^B} - \hat{\beta_k^{FL}})_{quantile_{\frac{\alpha}{2}}} + \beta_k^{FL} \tag{24}$$

where:

$\hat{\beta_k^b}$ = Estimated Model Parameter at feature k for b$^{\text{th}}$ Bootstrap

$\hat{\beta_k^{FL}}$ = Estimated Model Parameter at feature k for FL Model

β_k^{FL} = True Model Parameter at feature k for FL Model

By relaying the relationship between $F_B \rightarrow F_{FL}$ to $F_{FL} \rightarrow F$, the true distribution of the data can be estimated. The CI are compared between the parametric percentile bootstrapping and the non-parametric empirical bootstrapping.

4.3 SCAT3 Total Number of Symptoms

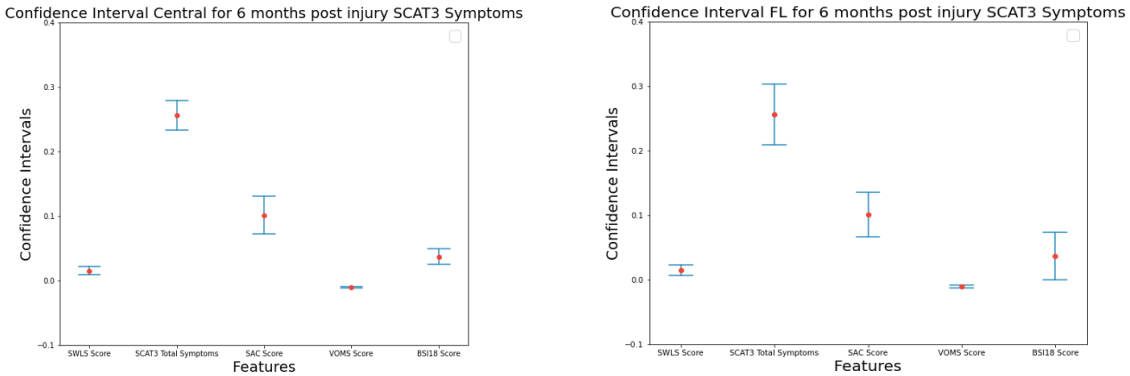
4.3.1 FL vs Central

Table 4 depicts the results of the FL and the central models for SCAT3 Total Number of Symptoms under the assumption of normality. Figure 18 shows that the CI for the central model is narrower than the CI for the FL model. The difference in CI is also shown numerically in Table 3, where the average difference of the CI

for the central model is 0.0275 and the average difference of the CI for the FL model is 0.0397. The central model is 0.0122 narrower than the FL model. The smaller difference in the CI for the central model means that the central model has lower variance than the FL model. This is to be expected because FL models have less information to work with. FL models have more uncertainty due to the fact that they gather the model parameters of the client models instead of running models on a centralized database. This means that FL models have higher variance and, thus, larger confidence intervals.

4.3.2 Client vs Total Bootstrapping

Figure 19 shows the comparisons of the CI between the client and total bootstrap techniques for SCAT3 Total Number of Symptoms subset for 1000 iterations. It shows that the client bootstrap has a larger CI on average for the model parameters than the total bootstrap technique. This is shown numerically by Table 5. Table 5 shows the CI differences between client and total Bootstrapping for the FL model of the SCAT3 Total Number of Symptoms. The total bootstrapping technique for the FL model has an average CI difference of 0.0242. The client bootstrapping technique has



(a) Central CI for SCAT3 Total Number of Symptoms Features

(b) FL CI for SCAT3 Total Number of Symptoms Features

Figure 18: Comparison of CI for SCAT3 Total Number of Symptoms

Table 3: FL vs Central CI Average Length Differences for SCAT3 Total Number of Symptoms Features

	Features	**Lower CI**	**Upper CI**	**Difference**	**Central CI Difference Average**
Central Model	SWLS Total Score	0.0162	0.0278	0.0116	0.02752
	SCAT3TotalSymptoms	0.2228	0.263	0.0402	
	SACScore	0.0909	0.138	0.0471	
	VOMS Total Score	0.0133	0.0161	0.0028	
	BSI18Score	0.0742	0.1101	0.0359	
	Features	**Lower CI**	**Upper CI**	**Difference**	**Fl CI Difference Average**
FL Model	SWLS Total Score	0.0156	0.0284	0.0128	0.0397
	SCAT3TotalSymptoms	0.2109	0.2749	0.064	
	SACScore	0.0831	0.1459	0.0628	
	VOMS Total Score	0.0128	0.0166	0.0038	
	BSI18Score	0.0596	0.1147	0.0551	

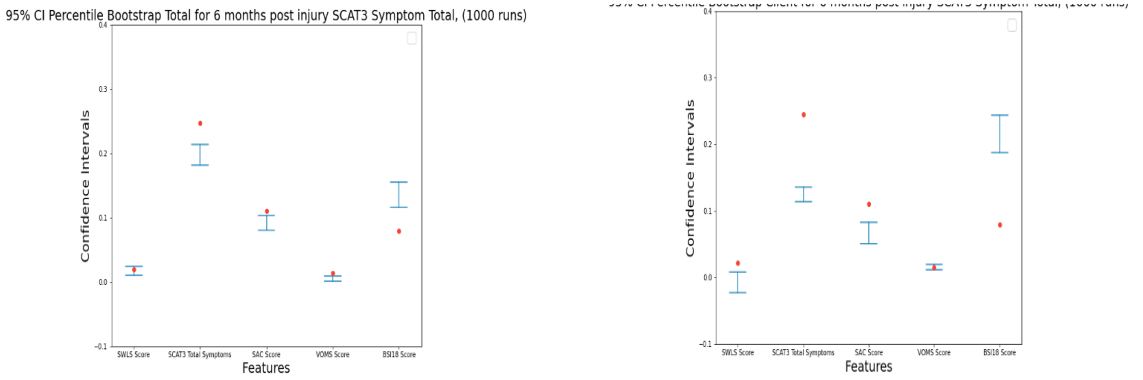
Table 4: FL vs Central for SCAT3 Total Number of Symptoms

Response Variable	**Features**	**95% Confidence Interval for Central Model**	**95% Confidence Interval for FL Model**	**Number of DataPoints**
SCAT3TotalSymptoms	Intercept	(-27.8990, 20.4172)	(-32.6229, 25.1411)	102
	SWLS Total Score	(0.0162, 0.0278)	(0.0156, 0.0284)	
	SCAT3TotalSymptoms	(0.2228, 0.2630)	(0.2109, 0.2749)	
	SACScore	(0.0909, 0.1380)	(0.0831, 0.1459)	
	VOMS Total Score	(0.0133, 0.0161)	(0.0128, 0.0166)	
	BSI18Score	(0.0742, 0.1101)	(0.0596, 0.1147)	
R Squared Central (Mean±Std Dev)	**R Squared FL (Mean±Std Dev)**	**R Squared-adjusted Central (Mean±Std Dev)**	**R Squared-adjusted FL (Mean±Std Dev)**	**Number of Trials**
(0.0079±0.5896)	(0.2237±0.3477)	(-0.294±0.7691)	(-0.0125±0.4535)	1000
R Squared (25^{th} percentile, Max) Central	**R Squared (25^{th} percentile, Max) FL**	**R Squared-adjusted (25^{th} percentile, Max) Central**	**R Squared-adjusted (25^{th} percentile, Max) FL**	
(-0.0994, 0.7035)	(0.0747, 0.7850)	(-0.4339, 0.6132)	(-0.2069, 0.7196)	
MSE Central (Mean±Std Dev)	**MSE FL (Mean±Std Dev)**	**MSE (25^{th} percentile, Max) Central**	**MSE (25^{th} percentile, Max) FL**	
(3.7156±1.5914)	(3.0052±1.3774)	(0.4314, 9.4947)	(0.3416, 6.9325)	

an average CI difference of 0.0303. The total bootstrapping technique has a smaller average CI difference by 0.0062 compared to the client bootstrapping technique.

4.3.3 Percentile vs Empirical Bootstrapping

Figures 20 and 21 show comparison of the CI for the percentile and empirical bootstraps for both the client and total techniques. Figure 20 shows that the empirical bootstrap CIs have larger lengths than the percentile bootstrap CIs for the client bootstrap technique. Table 6 also shows that the average difference of the CI for the empirical bootstrap is 0.0403 and the average difference of the CI for the percentile bootstrap is 0.0303. The percentile bootstrap CI have a smaller average difference than the empirical CI by 0.01. This means that there is more variance in the empirical bootstrap technique using the the client bootstrap technique. Figure 21 also shows that the empirical bootstrap CIs have smaller lengths than the percentile bootstrap CIs for the client bootstrap technique. Table 7 backs this visual as the average difference of the CI for the empirical bootstrap is 0.0231 and the average difference of the CI for the percentile bootstrap is 0.0255. The empirical bootstrap CI have a smaller



(a) Total Bootstrap SCAT3 Total Number of Symptoms

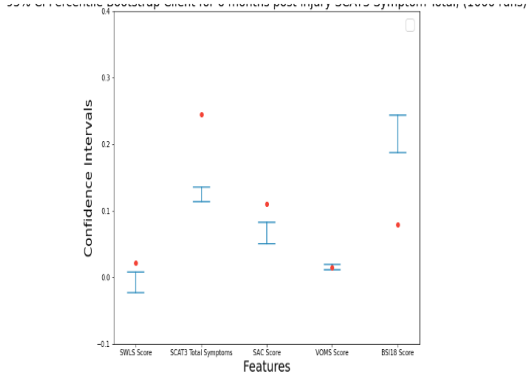
(b) Client Bootstrap SCAT3 Total Number of Symptoms

Figure 19: Comparison of Client and Total Bootstrap CI for SCAT3 Total Number of Symptoms (1000 Runs)

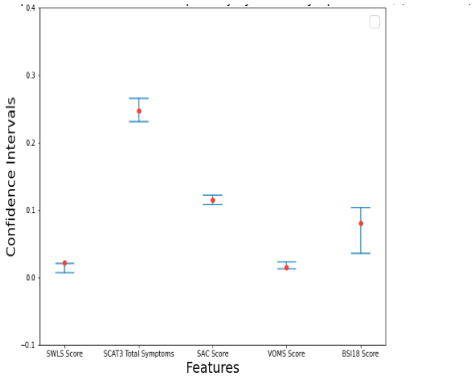
Table 5: FL CI Differences between Client and Total Bootstrapping SCAT3 Total Number of Symptoms

	Features	FL Lower CI	FL Upper CI	FL CI Difference	FL CI Difference Average Total
FL Total Bootstrapping (1000 Runs)	SWLS Total Score	0.0148	0.0286	0.0138	0.02416
	SCAT3TotalSymptoms	0.1833	0.2152	0.0319	
	SACScore	0.0847	0.1028	0.0181	
	VOMS Total Score	-0.0008	0.0067	0.0075	
	BSI18Score	0.1199	0.1694	0.0495	
	Features	**FL Lower CI**	**FL Upper CI**	**FL CI Difference**	**FL CI Difference Average Client**
FL Client Bootstrapping (1000 Runs)	SWLS Total Score	-0.0057	0.0121	0.0178	0.030322
	SCAT3TotalSymptoms	0.08379	0.1289	0.04511	
	SACScore	0.0469	0.0679	0.021	
	VOMS Total Score	0.0147	0.0245	0.0098	
	BSI18Score	0.1925	0.2504	0.0579	

average difference than the empirical CI by 0.0024. This means that there is more variance in the percentile bootstrap technique using the total bootstrap technique.

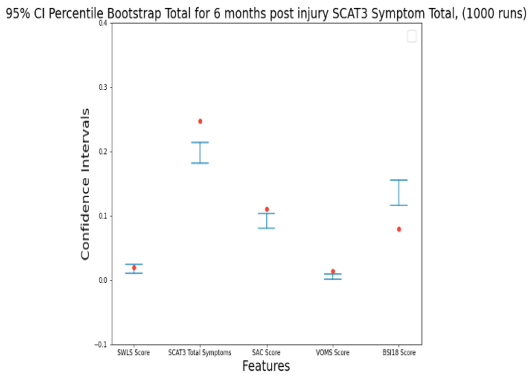


(a) Client Percentile Bootstrap SCAT3 Total Number of Symptoms

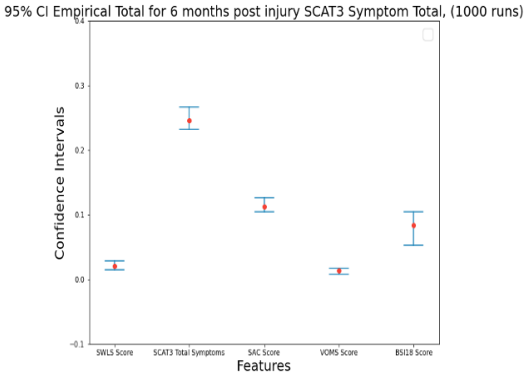


(b) Client Empirical Bootstrap SCAT3 Total Number of Symptoms

Figure 20: Comparison of CI for Client Percentile and Empirical Bootstrap SCAT3 Total Number of Symptoms (1000 Runs)



(a) Total Percentile Bootstrap SCAT3 Total Number of Symptoms



(b) Total Empirical Bootstrap SCAT3 Total Number of Symptoms

Figure 21: Comparison of CI for Total Percentile and Empirical Bootstrap SCAT3 Total Number of Symptoms (1000 Runs)

Table 6: Comparison of Client Percentile and Empirical Techniques for SCAT3 Total Number of Symptoms (1000 Runs)

	Features	**Lower CI**	**Upper CI**	**CI Difference**	**CI Difference Average Total**
Client Percentile Bootstrapping (1000 Runs)	Satisfaction with Life Scale.SWLSTotalScore	-0.0057	0.0121	0.0178	0.030322
	SCAT3TotalSymptoms	0.08379	0.1289	0.04511	
	SACScore	0.0469	0.0679	0.021	
	VOMS Scoring.VOMSTotalScore	0.0147	0.0245	0.0098	
	BSI18Score	0.1925	0.2504	0.0579	
	Features	**Lower CI**	**Upper CI**	**CI Difference**	**CI Difference Average Total**
Client Empirical Bootstrapping (1000 Runs)	Satisfaction with Life Scale.SWLSTotalScore	0.0151	0.0268	0.0117	0.04034
	SCAT3TotalSymptoms	0.2359	0.3075	0.0716	
	SACScore	0.0962	0.1116	0.0154	
	VOMS Scoring.VOMSTotalScore	0.0016	0.0203	0.0187	
	BSI18Score	0.0083	0.0926	0.0843	

Table 7: Comparison of Total Percentile and Empirical Techniques for SCAT3 Total Number of Symptoms (1000 Runs)

	Features	**Lower CI**	**Upper CI**	**CI Difference**	**CI Difference Average Total**
Total Percentile Bootstrapping (1000 Runs)	Satisfaction with Life Scale.SWLSTotalScore	0.0172	0.0288	0.0116	0.0255
	SCAT3TotalSymptoms	0.2277	0.2677	0.04	
	SACScore	0.0851	0.1325	0.0474	
	VOMS Scoring.VOMSTotalScore	0.0124	0.0151	0.0027	
	BSI18Score	0.0716	0.0972	0.0256	
	Features	**Lower CI**	**Upper CI**	**CI Difference**	**CI Difference Average Total**
Total Empirical Bootstrapping (1000 Runs)	Satisfaction with Life Scale.SWLSTotalScore	0.0129	0.0265	0.0136	0.0231
	SCAT3TotalSymptoms	0.2392	0.2679	0.0287	
	SACScore	0.1006	0.1214	0.0208	
	VOMS Scoring.VOMSTotalScore	0.0085	0.0161	0.0076	
	BSI18Score	0.0537	0.0984	0.0447	

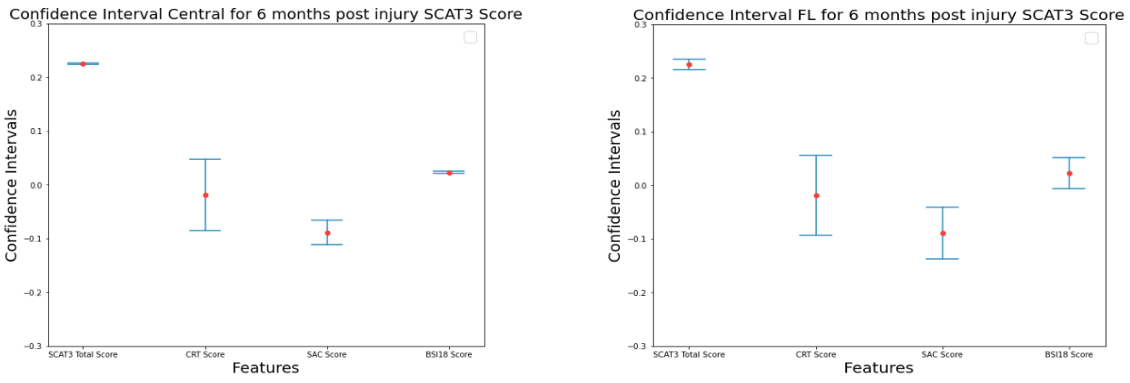
4.4 SCAT3 Total Score

4.4.1 FL vs Central

Table 9 depicts the results of the FL and the central models for SCAT3 Total Score under the assumption of normality. Figure 22, similar to the SCAT3 Total Number of Symptoms, shows that the CI for the central model is also narrower than the CI for the FL model. The difference in CI is also shown again numerically in Table 8. The average difference of the CI for the central model is 0.0743 and the average difference of the CI for the FL model is 0.0835. The central model is 0.0092 narrower than the CI model, showing that the central model has indeed less variance than the FL model.

4.4.2 Client vs Total Bootstrapping

Figure 23 shows the two different types of bootstrapping techniques compared side by side. Visually, the client bootstrap has a larger CI than the total bootstrap. This difference in the width of the CI is shown in Table 10. Table 10 shows the comparisons of the CI for the Total and Client Bootstrapping models for 1000 iterations for the FL model. The FL model client bootstrap has a larger average interval by 0.0099 than



(a) Central CI SCAT3 Total Score

(b) FL CI SCAT3 Total Score

Figure 22: Comparison of CI for SCAT3 Total Score

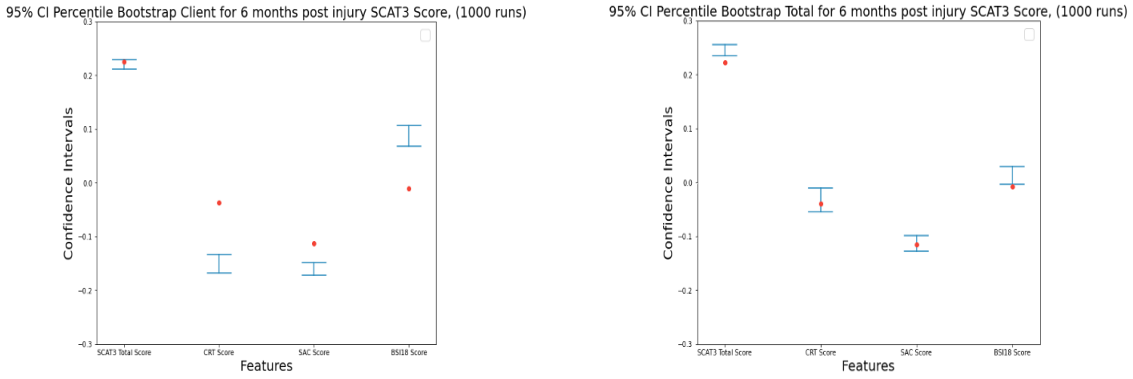
Table 8: FL vs Central CI Differences Total Score

	Features	Lower CI	Upper CI	Difference	Central CI Difference Average
Central Model	SCAT3TotalScore	0.2232	0.2278	0.0046	0.074325
	CRTScore	-0.1443	0.0754	0.2197	
	SACScore	-0.1443	-0.0808	0.0635	
	BSI18Score	-0.0129	-0.0034	0.0095	
	Features	**Lower CI**	**Upper CI**	**Difference**	**Fl CI Difference Average**
FL Model	SCAT3TotalScore	0.2199	0.231	0.0111	0.0835
	CRTScore	-0.143	0.0742	0.2172	
	SACScore	-0.1502	-0.0749	0.0753	
	BSI18Score	-0.0234	0.007	0.0304	

Table 9: FL vs Central for SCAT3 Total Score

Response Variable	Features	95% Confidence Interval for Central Model	95% Confidence Interval for FL Model	Number of DataPoints
SCAT3TotalScore	Intercept	(-21.6516, 28.4347)	(-26.8669, 33.6499)	103
	SCAT3TotalScore_baseline	(0.2232, 0.2278)	(0.2199, 0.2310)	
	CRTScore	(-0.1443, 0.0754)	(-0.1430, 0.0742)	
	SACScore	(-0.1443, -0.0808)	(-0.1502, -0.0749)	
	BSI18Score	(-0.0129, -0.0034)	(-0.0234, 0.0070)	
R Squared Central (Mean±Std Dev)	**R Squared FL (Mean±Std Dev)**	**R Squared-adjusted Central (Mean±Std Dev)**	**R Squared-adjusted FL (Mean±Std Dev)**	**Number of Trials**
(0.1430±0.7613)	(0.3783±0.5083)	(-0.0713±0.9516)	(0.2229±0.6354)	1000
R Squared (25^{th} percentile, Max) Central	**R Squared (25^{th} percentile, Max) FL**	**R Squared-adjusted (25^{th} percentile, Max) Central**	**R Squared-adjusted (25^{th} percentile, Max) FL**	
(-0.0198, 0.8659)	(0.1860, 0.9572)	(-0.2748, 0.8324)	(-0.0175, 0.9465)	
MSE Central (Mean±Std Dev)	**MSE FL (Mean±Std Dev)**	**MSE (25^{th} percentile, Max) Central**	**MSE (25^{th} percentile, Max) FL**	
(4.5366±2.1148)	(3.2882±1.4147)	(0.5154,16.773)	(0.4599, 7.531)	

the total bootstrap. Both Figure 23 and Table 10 show that the client bootstrap has a larger CI, thus more variance, than the total bootstrap technique.



(a) FL Client Bootstrap SCAT3 Total Score

(b) FL Total Bootstrap SCAT3 Total Score

Figure 23: Comparison of Total and Client CI for FL model of SCAT3 Total Score (1000 Runs)

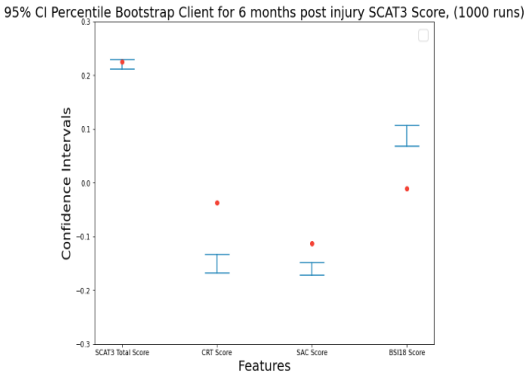
Table 10: FL CI Comparison for Total and Client Bootstrapping for SCAT3 Total Score (1000 Runs)

	Features	FL Lower CI	FL Upper CI	FL CI Difference	FL CI Difference Average Total
Total Bootstrapping (1000 Runs)	SCAT3TotalScore	0.23	0.2539	0.0239	0.03225
	CRTScore	-0.0437	-0.0042	0.0395	
	SACScore	-0.1342	-0.1011	0.0331	
	BSI18Score	-0.0061	0.0264	0.0325	
	Features	**FL Lower CI**	**FL Upper CI**	**FL CI Difference**	**FL CI Difference Average Client**
Client Bootstrapping (1000 Runs)	SCAT3TotalScore	0.1981	0.2202	0.0221	0.0422
	CRTScore	-0.1695	-0.0929	0.0766	
	SACScore	-0.1588	-0.1356	0.0232	
	BSI18Score	0.0864	0.1333	0.0469	

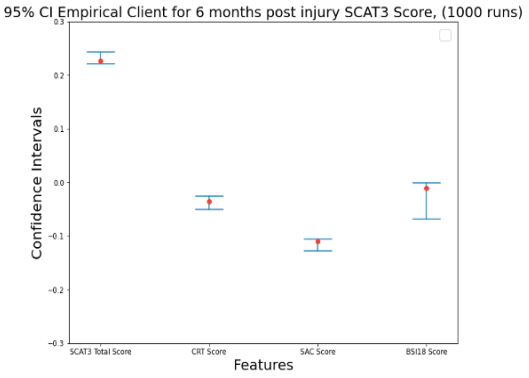
4.4.3 Percentile vs Empirical Bootstrapping

Figure 24 shows the comparison of CI for the percentile and empirical techniques for the client bootstrapping SCAT3 Total Score. Figure 25 shows the comparison of CI for the percentile and empirical techniques for the total bootstrapping SCAT3 Total Score. Both figures show that the percentile bootstrap technique has a larger CI on average than the empirical bootstrap techniques.

The visualization is translated numerically with Tables 11 and reftab:CI Comparison for Percentile and Empirical Total Bootstrapping SCAT3 Total Score. Table 11 shows the comparison of the client percentile and empirical bootstrapping techniques. Comparing the two client bootstrapping techniques, the percentile technique

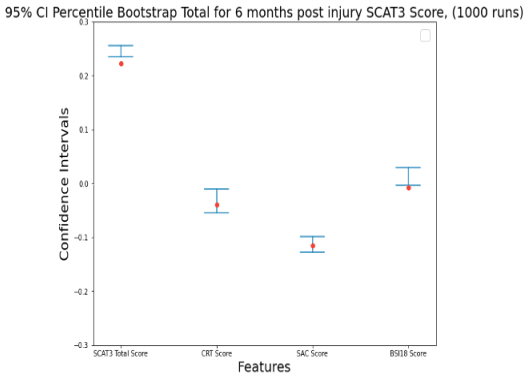


(a) Client Percentile Bootstrap SCAT3 Total Score

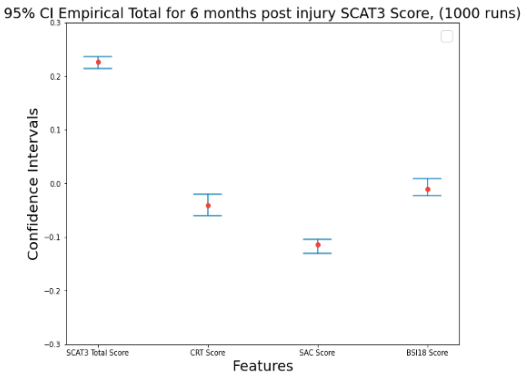


(b) Client Empirical Bootstrap SCAT3 Total Score

Figure 24: Comparison Client CI for for Percentile and Empirical Bootstrapping of SCAT3 Total Score (1000 Runs)



(a) Total Percentile Bootstrap SCAT3 Total Score



(b) Total Empirical Bootstrap SCAT3 Total Score

Figure 25: Comparison Total CI for for Percentile and Empirical Bootstrapping of SCAT3 Total Score (1000 Runs)

has a wider average CI than the empirical technique by 0.0243. Table 12 shows the comparison of the total percentile and empirical bootstrapping techniques. The percentile technique also has a wider average CI than the empirical technique, which is wider by 0.0257. These two tables agree with Figures 24 and 25 in that the empirical bootstrapping technique has a smaller CI than the percentile techniques. This means that the empirical techniques have a lower variance.

Table 11: CI Comparison for Percentile and Empirical Client Bootstrapping for SCAT3 Total Score (1000 Runs)

	Features	**Lower CI**	**Upper CI**	**CI Difference**	**CI Difference Average Total**
Client Percentile Bootstrapping (1000 Runs)	SCAT3TotalScore	0.1981	0.2202	0.0221	0.0422
	CRTScore	-0.1695	-0.0929	0.0766	
	SACScore	-0.1588	-0.1356	0.0232	
	BSI18Score	0.0864	0.1333	0.0469	
	Features	**Lower CI**	**Upper CI**	**CI Difference**	**CI Difference Average Client**
Client Empirical Bootstrapping (1000 Runs)	SCAT3TotalScore	0.229	0.2375	0.0085	0.017925
	CRTScore	-0.0461	-0.0155	0.0306	
	SACScore	-0.1137	-0.1035	0.0102	
	BSI18Score	-0.0296	-0.0072	0.0224	

Table 12: CI Comparison for Percentile and Empirical Total Bootstrapping for SCAT3 Total Score (1000 Runs)

	Features	Lower CI	Upper CI	CI Difference	CI Difference Average Total
Total Percentile Bootstrapping (1000 Runs)	SCAT3TotalScore	0.23	0.2539	0.0239	0.05725
	CRTScore	-0.0437	-0.0042	0.0395	
	SACScore	-0.2342	-0.1011	0.1331	
	BSI18Score	-0.0061	0.0264	0.0325	
	Features	**Lower CI**	**Upper CI**	**CI Difference**	**CI Difference Average Total**
Total Empirical Bootstrapping (1000 Runs)	SCAT3TotalScore	0.2111	0.2316	0.0205	0.0316
	CRTScore	-0.0488	-0.0095	0.0393	
	SACScore	-0.1342	-0.1079	0.0263	
	BSI18Score	-0.0142	0.0261	0.0403	

4.5 Final Results

For both subsets, the results were consistent. In all of the bootstrap types and the different number of trials, the FL models consistently produced higher scores but had more variance for its model parameters. The CI for the FL model features were consistently larger. The FL models produced higher mean R^2 scores, R^2-adjusted scores, lower standard deviations for each score, lower 25^{th} percentiles and higher maximums. The FL models also outputted lower average MSE scores, lower standard deviations, higher 25^{th} percentile and lower maximums. These results show the trade-offs of using a FL model versus a central model. Although the FL models can produce higher scores, there is more variability within the model parameters when compared to its central model counterpart. The empirical bootstrapping technique was also able to capture the estimate of the true distribution while the percentile bootstrapping technique could not. The difference of the results of the two bootstrapping techniques shows it was incorrect to assume the residuals are normally distributed. However, the empirical bootstrapping technique can accurately estimate the true distribution of the data, unlike the percentile bootstrapping technique.

V. Conclusions

5.1 Key Findings and Contributions

One of the key findings was that the FL models produced better results but had a larger variance of the model parameters than the central models. This is to be expected because the FL model learns information by sharing model parameters of its clients while the central model learns information by having the actual data itself. This means that the central model learns from a more true representation of the population than the FL models. This will cause the FL models to have a higher variance of its model parameters compared to those of the central model. The unexpected part was that the FL models consistently produced better results than the central model. Because the central model has a more true representation of the data, I expected better results. However, in most of the tests, the FL models score higher average R^2 and R^2 adjusted scores with higher minima and maxima and lower MSE scores with lower maxima and minima.

Another key finding was the the client bootstrapping technique had larger average CI lengths compared to total bootstrapping technique. This means that the client bootstrapping has a larger variance compared to the total bootstrapping technique. Client bootstrapping re-samples data within its own clients so there is no sharing of data with the other clients. Unlike the client bootstrapping technique, the total bootstrapping technique re-samples the data centrally before splitting up into clients. This means that, similar to the comparison of the FL and central models, the total bootstrapping techniques have more data to work with, which means a more true representation of the underlying population to work with, compared to the client bootstrapping techniques. As a result, there is limited information for the client

bootstrapping technique so there is more variance and uncertainty compared to the total bootstrapping technique.

The final key finding was that the the non-parametric empirical bootstrapping technique calculated lower average CI lengths, thus smaller variance, compared to the parametric percentile bootstrapping technique for both the client and total bootstrapping methods. This means that, as expected, the empirical bootstrapping technique was able to capture an estimate closer to the true distribution of the data than the percentile bootstrapping technique.

5.2 Limitations

There were two major limitations. The first was the computational power. The lack of computational power is shown in the total bootstrapping techniques where 2000 iterations of the SCAT3 Total Score total bootstrapping takes 1305 minutes or roughly 22 hours to run. 10,000 iterations of the client bootstrapping for the same data takes 11 minutes to run. As a result, a one-to-one comparison for each iteration was infeasible. The total bootstrapping technique tests had iterations of 300, 500, 1000 and 2000 due to limited computational power while the the client bootstrapping technique tests had iterations of 1000, 3000, 5000 and 10,000.

The second major limitation was the dataset itself. The dataset was a compilation of individual clinical scores from 30 different universities with 30 or more doctors and test administrators providing their own input and interpretations of the clinical tests. Because some universities did not do certain tests that others did and vice versa, not everybody was following the same procedures and hurt the data collecting process. The lack of standardization in which tests to administer to collect data caused a major missing data problem for the end user of the dataset. The missing data forced

major assumptions of linearity to be broken and large chunks of the dataset rendered useless.

5.3 Future Work

For future work, there are three things that I would like to test. First, I would like to apply bootstrapping techniques on more complex models such as neural networks. This research is limited to closed form solutions in order to find the variance and CI of the model parameters. This limitation is why I strictly used linear regression for this thesis. However, bootstrapping techniques allow open form solutions to re-sample data and calculate the CI of the model parameters.

Second, I would like to try different imputation techniques to enlarge the dataset size. Using a kNN imputation techniques on subsets of the entire dataset proved to be not useful and produced poor results. With more time, applying different imputation techniques not only on subsets of the data but also the entire dataset itself might produce better results. Because of the poor collection process, there was no use for large parts of the dataset. By correctly testing different imputation techniques, the entirety of the dataset would be more useful. This would give us more samples to run tests on, be more representative of the true population, and offer more insight into the problem.

Finally, the last part would be to test larger iterations of each bootstrapping techniques, which was limited by time and computational power. Either with more time or computational resources, trying the 3000, 5000 and 10000 iterations for the total bootstrapping techniques might produce different results. Testing these iterations would provide a more direct comparison between the client and total bootstrapping techniques.

Appendix A. Comparison of Original vs Imputed Datasets for SCAT3 Total Score

1.1 Original Data

Table 13: Original Data SCAT3 Total Score

Response Variable	**Features**	**Number of DataPoints**	**Number of DataPoints**
SCAT3TotalScore (Original)	Intercept	103	1000
	SCAT3TotalScore		
	CRTScore		
	SACScore		
	BSI18Score		
R^2 **Central (Mean±Std Dev)**	R^2 **FL (Mean±Std Dev)**	R^2**-adjusted Central (Mean±Std Dev)**	R^2**-adjusted FL (Mean±Std Dev)**
(0.1430±0.7613)	(0.3783±0.5083)	(-0.0713±0.9516)	(0.2229±0.6354)
R^2 **(Min, Max) Central**	R^2 **(Min, Max) FL**	R^2**-adjusted (Min, Max) Central**	R^2**-adjusted (Min, Max) FL**
(-0.0198, 0.8659)	(0.1860, 0.9572)	(-0.2748, 0.8324)	(-0.0175, 0.9465)
MSE Central (Mean±Std Dev)	**MSE FL (Mean±Std Dev)**	**MSE (Min, Max) Central**	**MSE (Min, Max) FL**
(4.5366±2.1148)	(3.2882±1.4147)	(0.5154,16.773)	(0.4599, 7.531)

1.2 Imputed Data

Table 14: Imputed Data SCAT3 Total Score

Response Variable	**Features**	**Number of DataPoints**	**Number of DataPoints**
SCAT3TotalScore (Imputed)	Intercept	1579	1000
	SCAT3TotalScore		
	CRTScore		
	SACScore		
	BSI18Score		
R^2 **Central (Mean±Std Dev)**	R^2 **FL (Mean±Std Dev)**	R^2**-adjusted Central (Mean±Std Dev)**	R^2**-adjusted FL (Mean±Std Dev)**
(0.0679±0.0408)	(0.0878±0.0353)	(0.0457±0.0418)	(0.0661±0.0362)
R^2 **(Min, Max) Central**	R^2 **(Min, Max) FL**	R^2**-adjusted (Min, Max) Central**	R^2**-adjusted (Min, Max) FL**
(0.0394, 0.1302)	(0.0596, 0.149)	(0.0165, 0.1095)	(0.0372, 0.1287)
MSE Central (Mean±Std Dev)	**MSE FL (Mean±Std Dev)**	**MSE (Min, Max) Central**	**MSE (Min, Max) FL**
(10.4705±3.8071)	(10.2584±3.7328)	(5.3276, 16.4912)	(5.1456, 16.1147)

Appendix B. Comparison of FL vs Central Model

2.1 SCAT3 Total Number of Symptoms

Table 15: SCAT3 Total Number of Symptoms

Response Variable	**Features**	**95% Confidence Interval for Central Model**	**95% Confidence Interval for FL Model**	**Number of DataPoints**
SCAT3TotalSymptoms	Intercept	(-27.8990, 20.4172)	(-32.6229, 25.1411)	102
	SWLS Total Score	(0.0162, 0.0278)	(0.0156, 0.0284)	
	SCAT3TotalSymptoms	(0.2228, 0.2630)	(0.2109, 0.2749)	
	SACScore	(0.0909, 0.1380)	(0.0831, 0.1459)	
	VOMS Total Score	(0.0133, 0.0161)	(0.0128, 0.0166)	
	BSI18Score	(0.0742, 0.1101)	(0.0596, 0.1147)	
R^2 Central (Mean±Std Dev)	**R^2 FL (Mean±Std Dev)**	**R^2-adjusted Central (Mean±Std Dev)**	**R^2-adjusted FL (Mean±Std Dev)**	**Number of Trials**
(0.0079±0.5896)	(0.2237±0.3477)	(-0.294±0.7691)	(-0.0125±0.4535)	1000
R^2 (25^{th} Percentile, Max) Central	**R^2 (25^{th} Percentile, Max) FL**	**R^2-adjusted (25^{th} Percentile, Max) Central**	**R^2-adjusted (25^{th} Percentile, Max) FL**	
(-0.0994, 0.7035)	(0.0747, 0.7850)	(-0.4339, 0.6132)	(-0.2069, 0.7196)	
MSE (Mean±Std Dev) Central	**MSE (Mean±Std Dev) Empirical Total**	**MSE (25^{th} Percentile, Max) Central**	**MSE (25^{th} Percentile, Max) Empirical Total**	
(3.7156±1.5914)	(3.0052±1.3774)	(0.4314, 9.4947)	(0.3416, 6.9325)	

2.2 SCAT3 Total Score

Table 16: SCAT3 Total Score

Response Variable	**Features**	**95% Confidence Interval for Central Model**	**95% Confidence Interval for FL Model**	**Number of DataPoints**
SCAT3TotalScore	Intercept	(-21.6516, 28.4347)	(-26.8669, 33.6499)	103
	SCAT3TotalScore	(0.2232, 0.2278)	(0.2199, 0.2310)	
	CRTScore	(-0.1443, 0.0754)	(-0.1430, 0.0742)	
	SACScore	(-0.1443, -0.0808)	(-0.1502, -0.0749)	
	BSI18Score	(-0.0129, -0.0034)	(-0.0234, 0.0070)	
R^2 **Central (Mean±Std Dev)**	R^2 **FL (Mean±Std Dev)**	R^2**-adjusted Central (Mean±Std Dev)**	R^2**-adjusted FL (Mean±Std Dev)**	**Number of Trials**
(0.1430±0.7613)	(0.3783±0.5083)	(-0.0713±0.9516)	(0.2229±0.6354)	1000
R^2 **(25th Percentile, Max) Central**	R^2 **(25th Percentile, Max) FL**	R^2**-adjusted (25th Percentile, Max) Central**	R^2**-adjusted (25th Percentile, Max) FL**	
(-0.0198, 0.8659)	(0.1860, 0.9572)	(-0.2748, 0.8324)	(-0.0175, 0.9465)	
MSE (Mean±Std Dev) Central	**MSE (Mean±Std Dev) Empirical Total**	**MSE (25th Percentile, Max) Central**	**MSE (25th Percentile, Max) Empirical Total**	
(4.5366±2.1148)	(3.2882±1.4147)	(0.5154,16.773)	(0.4599, 7.531)	

Appendix C. Transformed Data Results

3.1 SCAT3 Total Number of Symptoms

Table 17: Transformed SCAT3 Total Number of Symptoms

Response Variable	**Features**	**95% Confidence Interval for Central Model**	**95% Confidence Interval for FL Model**	**Number of DataPoints**
SCAT3TotalSymptoms Transformed	Intercept	(-5.5634, 2.7096)	(-6.0847, 3.2308)	102
	Satisfaction with Life Scale.SWLSTotalScore	(0.0012, 0.0032)	(0.0012, 0.0032)	
	SCAT3TotalSymptoms	(0.1115, 0.1185)	(0.1099, 0.1202)	
	SACScore	(0.0492, 0.0573)	(0.0483, 0.0581)	
	VOMS Scoring.VOMSTotalScore	(0.0047, 0.0051)	(0.0046, 0.0052)	
	BSI18Score	(0.0089, 0.0134)	(0.0064, 0.0159)	
R^2 **Central (Mean±Std Dev)**	R^2 **FL (Mean±Std Dev)**	R^2**-adjusted Central (Mean±Std Dev)**	R^2**-adjusted FL (Mean±Std Dev)**	**Number of Trials**
(0.1048±0.2842)	(0.2481±0.2095)	(-0.1677±0.3707)	(0.0192±0.4535)	1000
R^2 **(25^{th} Percentile, Max) Central**	R^2 **(25^{th} Percentile, Max) FL**	R^2**-adjusted (25^{th} Percentile, Max) Central**	R^2**-adjusted (25^{th} Percentile, Max) FL**	
(-0.0084, 0.6236)	(0.1234, 0.7649)	(-0.3152, 0.509)	(-0.1433, 0.6933)	
MSE (Mean±Std Dev) Central	**MSE (Mean±Std Dev) Empirical Total**	**MSE (25^{th} Percentile, Max) Central**	**MSE (25^{th} Percentile, Max) Empirical Total**	
(0.6056±0.184)	(0.5113±0.1573)	(0.1648, 1.4166)	(0.3416, 1.0427)	

3.2 SCAT3 Total Score

Table 18: Transformed SCAT3 Total Score

Response Variable	**Features**	**95% Confidence Interval for Central Model**	**95% Confidence Interval for FL Model**	**Number of DataPoints**
SCAT3TotalScore Transformed	Intercept	(-1.9624, 5.8610)	(-3.0164, 6.9149)	103
	SCAT3TotalScore	(0.0571, 0.0578)	(0.0565, 0.0584)	
	CRTScore	(-0.0222, 0.0124)	(-0.0242, 0.0143)	
	SACScore	(-0.0653, -0.0553)	(-0.0665, -0.0541)	
	BSI18Score	(0.0165, 0.0181)	(0.0149, 0.0197)	
R^2 **Central (Mean±Std Dev)**	R^2 **FL (Mean±Std Dev)**	R^2**-adjusted Central (Mean±Std Dev)**	R^2**-adjusted FL (Mean±Std Dev)**	**Number of Trials**
(0.2569±0.2639)	(0.358±0.2170)	(0.0711±0.3298)	(0.1975±0.2713)	1000
R^2 **(25^{th} Percentile, Max) Central**	R^2 **(25^{th} Percentile, Max) FL**	R^2**-adjusted (25^{th} Percentile, Max) Central**	R^2**-adjusted (25^{th} Percentile, Max) FL**	
(0.1231, 0.7461)	(0.2497, 0.8219)	(-0.0962, 0.6827)	(0.0621, 0.7773)	
MSE (Mean±Std Dev) Central	**MSE (Mean±Std Dev) Empirical Total**	**MSE (25^{th} Percentile, Max) Central**	**MSE (25^{th} Percentile, Max) Empirical Total**	
(0.6079±0.1988)	(0.5241±0.147)	(0.2191, 2.1019)	(0.1717, 1.138)	

Appendix D. Normalized Data Results

4.1 SCAT3 Total Number of Symptoms

Table 19: Normalized SCAT3 Total Number of Symptoms

Response Variable	**Features**	**95% Confidence Interval for Central Model**	**95% Confidence Interval for FL Model**	**Number of DataPoints**
SCAT3TotalSymptoms Normalized	Intercept	(-0.0218, 0.0195)	(-0.0266, 0.0243)	102
	Satisfaction with Life Scale.SWLSTotalScore	(-0.0156, 0.0348)	(-0.0162, 0.0354)	
	SCAT3TotalSymptoms	(0.3961, 0.5022)	(0.3737, 0.5245)	
	SACScore	(0.1002, 0.1447)	(0.0958, 0.1491)	
	VOMS Scoring.VOMSTotalScore	(0.0246, 0.0834)	(0.0158, 0.0922)	
	BSI18Score	(-0.0039, 0.1230)	(-0.0690, 0.1881)	
R^2 **Central (Mean±Std Dev)**	R^2 **FL (Mean±Std Dev)**	R^2**-adjusted Central (Mean±Std Dev)**	R^2**-adjusted FL (Mean±Std Dev)**	**Number of Trials**
(0.1079±0.2668)	(0.2467±0.2142)	(-0.1636±0.3479)	(0.0174±0.2795)	1000
R^2 **(25th Percentile, Max) Central**	R^2 **(25th Percentile, Max) FL**	R^2**-adjusted (25th Percentile, Max) Central**	R^2**-adjusted (25th Percentile, Max) FL**	
(-0.0181, 0.6247)	(0.1134, 0.7612)	(0.1134, 0.7612)	(-0.1565, 0.6885)	
MSE (Mean±Std Dev) Central	**MSE (Mean±Std Dev) Empirical Total**	**MSE (25th Percentile, Max) Central**	**MSE (25th Percentile, Max) Empirical Total**	
(0.8292±0.2544)	(0.7015±0.217)	(0.279, 2.1588)	(0.1912, 1.4717)	

4.2 SCAT3 Total Score

Table 20: Normalized SCAT3 Total Score

Response Variable	**Features**	**95% Confidence Interval for Central Model**	**95% Confidence Interval for FL Model**	**Number of DataPoints**
SCAT3TotalScore Normalized	Intercept	(-0.0156, 0.0166)	(-0.0208, 0.0218)	103
	SCAT3TotalScore	(0.5537, 0.6335)	(0.4947, 0.6925)	
	CRTScore	(-0.0207, 0.0121)	(-0.0224, 0.0138)	
	SACScore	(-0.1263, -0.0934)	(-0.1299, -0.0897)	
	BSI18Score	(0.0812, 0.1603)	(-0.0085, 0.2501)	
R^2 **Central (Mean±Std Dev)**	R^2 **FL (Mean±Std Dev)**	R^2**-adjusted Central (Mean±Std Dev)**	R^2**-adjusted FL (Mean±Std Dev)**	**Number of Trials**
(0.2779±0.2724)	(0.3658±0.2421)	(0.0973±0.3405)	(0.2073±0.3027)	1000
R^2 **(25th Percentile, Max) Central**	R^2 **(25th Percentile, Max) FL**	R^2**-adjusted (25th Percentile, Max) Central**	R^2**-adjusted (25th Percentile, Max) FL**	
(0.1446, 0.7818)	(0.2648, 0.7729)	(-0.0692, 0.7272)	(0.081, 0.7162)	
MSE (Mean±Std Dev) Central	**MSE (Mean±Std Dev) Empirical Total**	**MSE (25th Percentile, Max) Central**	**MSE (25th Percentile, Max) Empirical Total**	
(0.6691±0.2468)	(0.5828±0.1706)	(0.2191,3.9023)	(0.1925, 1.1335)	

Appendix E. Percentile Client Bootstrap

5.1 SCAT3 Total Number of Symptoms

5.1.1 1000 Runs

Table 21: Percentile Client SCAT3 Total Number of Symptoms 1000 Runs

Response Variable	**Features**	**95% Confidence Interval for Central Model**	**95% Confidence Interval for Bootstrap Client Model**	**Number of DataPoints**
SCAT3TotalSymptoms				102
	SWLS Total Score	(0.0172, 0.0288)	(-0.0057, 0.0121)	
	SCAT3TotalSymptoms	(0.2277, 0.2677)	(0.08379, 0.1289)	
	SACScore	(0.0851, 0.1325)	(0.0469, 0.0679)	
	VOMS Total Score	(0.0124, 0.0151)	(0.0147, 0.0245)	
	BSI18Score	(0.0716, 0.0972)	(0.1925, 0.2504)	
R^2 (Mean±Std Dev) Central	**R^2 (Mean±Std Dev) Bootstrap Client**	**R^2-adjusted (Mean±Std Dev) Central**	**R^2-adjusted (Mean±Std Dev) Bootstrap Client**	**Number of Trials**
(0.0142±0.6036)	(0.1565±0.4839)	(-0.2859±0.7874)	(-0.1002±0.6311)	1000
R^2 (25^{th} Percentile, Max) Central	**R^2 (25^{th} Percentile, Max) FL**	**R^2-adjusted (25^{th} Percentile, Max) Central**	**R^2-adjusted (25^{th} Percentile, Max) FL**	**Time**
(-0.1029,0.6482)	(-0.0088,0.8737)	(-0.4386, 0.5412)	(-0.3158, 0.8352)	31.8s
MSE (Mean±Std Dev) Central	**MSE (Mean±Std Dev) Empirical Total**	**MSE (25^{th} Percentile, Max) Central**	**MSE (25^{th} Percentile, Max) Empirical Total**	
(3.6432±1.5968)	(3.0612±1.3182)	(0.6233, 10.871)	(0.4776, 7.3475)	

5.1.2 3000 Runs

Table 22: Percentile Client SCAT3 Total Number of Symptoms 3000 Runs

Response Variable	**Features**	**95% Confidence Interval for Central Model**	**95% Confidence Interval for Bootstrap Client Model**	**Number of DataPoints**
SCAT3TotalSymptoms				102
	SWLS Total Score	(0.0165, 0.0282)	(0.0041, 0.01304)	
	SCAT3TotalSymptoms	(0.2261,0.2666)	(0.0723, 0.1143)	
	SACScore	(0.0881, 0.1359)	(0.0378, 0.0579)	
	VOMS Total Score	(0.0123, 0.0151)	(0.0168, 0.0293)	
	BSI18Score	(0.0675, 0.0933)	(0.2007, 0.2246)	
R^2 (Mean±Std Dev) Central	**R^2 (Mean±Std Dev) Bootstrap Client**	**R^2-adjusted (Mean±Std Dev) Central**	**R^2-adjusted (Mean±Std Dev) Bootstrap Client**	**Number of Trials**
(0.0126±1.2155)	(0.1562±0.6911)	(-0.2913±1.5854)	(-0.1006±0.9014)	3000
R^2 (25^{th} Percentile, Max) Central	**R^2 (25^{th} Percentile, Max) FL**	**R^2-adjusted (25^{th} Percentile, Max) Central**	**R^2-adjusted (25^{th} Percentile, Max) FL**	**Time**
(-0.1087, 0.6812)	(-0.0213, 0.8606)	(-0.4461, 0.5842)	(-0.3321, 0.8182)	3m 49s
MSE (Mean±Std Dev) Central	**MSE (Mean±Std Dev) Empirical Total**	**MSE (25^{th} Percentile, Max) Central**	**MSE (25^{th} Percentile, Max) Empirical Total**	
(3.7184±1.6169)	(3.1356±1.3379)	(0.5422, 11.4692)	(0.4598, 7.511)	

5.1.3 5000 Runs

Table 23: Percentile Client SCAT3 Total Number of Symptoms 5000 Runs

Response Variable	**Features**	**95% Confidence Interval for Central Model**	**95% Confidence Interval for Bootstrap Client Model**	**Number of DataPoints**
SCAT3TotalSymptoms				102
	SWLS Total Score	(0.0153, 0.0269)	(0.0009, 0.0071)	
	SCAT3TotalSymptoms	(0.2260, 0.2662)	(0.0931, 0.1449)	
	SACScore	(0.0887, 0.1362)	(0.0399, 0.0621)	
	VOMS Total Score	(0.0127, 0.0154)	(0.0156, 0.0222)	
	BSI18Score	(0.0687, 0.0943)	(0.2053, 0.2272)	
R^2 (Mean±Std Dev) Central	**R^2 (Mean±Std Dev) Bootstrap Client**	**R^2-adjusted (Mean±Std Dev) Central**	**R^2-adjusted (Mean±Std Dev) Bootstrap Client**	**Number of Trials**
(0.0105±0.7972)	(0.1415±0.6271)	(-0.2907±1.0398)	(-0.1198±0.818)	5000
R^2 (25^{th} Percentile, Max) Central	**R^2 (25^{th} Percentile, Max) FL**	**R^2-adjusted (25^{th} Percentile, Max) Central**	**R^2-adjusted (25^{th} Percentile, Max) FL**	**Time**
(-0.1053, 0.6714)	(-0.0220, 0.8572)	(-0.4417, 0.5713)	(-0.3331, 0.08138)	10m 25s
MSE (Mean±Std Dev) Central	**MSE (Mean±Std Dev) Empirical Total**	**MSE (25^{th} Percentile, Max) Central**	**MSE (25^{th} Percentile, Max) Empirical Total**	
(3.6756±1.5861)	(3.1206±1.3097)	(0.5134, 10.8711)	(0.3505, 7.1678)	

5.1.4 10,000 Runs

Table 24: Percentile Client SCAT3 Total Number of Symptoms 10,000 Runs

Response Variable	**Features**	**95% Confidence Interval for Central Model**	**95% Confidence Interval for Bootstrap Client Model**	**Number of DataPoints**
SCAT3TotalSymptoms				102
	SWLS Total Score	(0.0158, 0.0275)	(0.0083, 0.0169)	
	SCAT3TotalSymptoms	(0.2258, 0.2661)	(0.0897, 0.1155)	
	SACScore	(0.0892, 0.1367)	(0.0409, 0.0538)	
	VOMS Total Score	(0.0129, 0.0156)	(0.0156, 0.0183)	
	BSI18Score	(0.0683, 0.0940)	(0.2086, 0.2250)	
R^2 **(Mean±Std Dev) Central**	R^2 **(Mean±Std Dev) Bootstrap Client**	R^2**-adjusted (Mean±Std Dev) Central**	R^2**-adjusted (Mean±Std Dev) Bootstrap Client**	**Number of Trials**
(-0.023±1.3356)	(0.1205±1.0314)	(-0.3344±1.742)	(-0.1472±1.3454)	10000
R^2 **(25th Percentile, Max) Central**	R^2 **(25th Percentile, Max) FL**	R^2**-adjusted (25th Percentile, Max) Central**	R^2**-adjusted (25th Percentile, Max) FL**	**Time**
(-0.1165, 0.7288)	(-0.0242, 0.9006)	(-0.4564, 0.6463)	(-0.3359, 0.8703)	42m 58s
MSE (Mean±Std Dev) Central	**MSE (Mean±Std Dev) Empirical Total**	**MSE (25th Percentile, Max) Central**	**MSE (25th Percentile, Max) Empirical Total**	
(3.6773±1.575)	(3.1137±1.3149)	(0.3828, 11.6019)	(0.2990, 8.1362)	

5.2 SCAT3 Total Score

5.2.1 1000 Runs

Table 25: Percentile Client SCAT3 Total Score 1000 Runs

Response Variable	**Features**	**95% Confidence Interval for Central Model**	**95% Confidence Interval for Bootstrap Client Model**	**Number of DataPoints**
SCAT3TotalScore				103
	SCAT3TotalScore	(0.2251, 0.2299)	(0.1981 ,0.2202)	
	CRTScore	(-0.1512, 0.0654)	(-0.1695, -0.0929)	
	SACScore	(-0.1419, -0.0797)	(-0.1588, -0.1356)	
	BSI18Score	(-0.0178, 0.0080)	(0.0864, 0.1333)	
R^2 **(Mean±Std Dev) Central**	R^2 **(Mean±Std Dev) Bootstrap Client**	R^2**-adjusted (Mean±Std Dev) Central**	R^2**-adjusted (Mean±Std Dev) Bootstrap Client**	**Number of Trials**
(0.1407±0.8929)	(0.2321±0.8696)	(-0.0741±1.1161)	(0.0402±1.087)	1000
R^2 **(25th Percentile, Max) Central**	R^2 **(25th Percentile, Max) FL**	R^2**-adjusted (25th Percentile, Max) Central**	R^2**-adjusted (25th Percentile, Max) FL**	**Time**
(0.0266, 0.8915)	(0.0516, 0.9675)	(-0.2168, 0.8644)	(-0.1856, 0.9594)	20s
MSE (Mean±Std Dev) Central	**MSE (Mean±Std Dev) Empirical Total**	**MSE (25th Percentile, Max) Central**	**MSE (25th Percentile, Max) Empirical Total**	
(4.4923±2.0396)	(3.7899±1.6589)	(0.4108, 14.9036)	(0.3461, 9.5259)	

5.2.2 3000 Runs

Table 26: Percentile Client SCAT3 Total Score 3000 Runs

Response Variable	**Features**	**95% Confidence Interval for Central Model**	**95% Confidence Interval for Bootstrap Client Model**	**Number of DataPoints**
SCAT3TotalScore				103
	SCAT3TotalScore	(0.2237, 0.2285)	(0.2125, 0.2215)	
	CRTScore	(-0.1490, 0.0726)	(-0.1840, -0.1545)	
	SACScore	(-0.1447, -0.0810)	(-0.1693, -0.1408)	
	BSI18Score	(-0.0151, -0.0051)	(0.0820, 0.1172)	
R^2 **(Mean±Std Dev) Central**	R^2 **(Mean±Std Dev) Bootstrap Client**	R^2**-adjusted (Mean±Std Dev) Central**	R^2**-adjusted (Mean±Std Dev) Bootstrap Client**	**Number of Trials**
(0.146±0.8543)	(0.2437±0.8486)	(-0.0675±1.0679)	(0.0546±1.0608)	3000
R^2 **(25th Percentile, Max) Central**	R^2 **(25th Percentile, Max) FL**	R^2**-adjusted (25th Percentile, Max) Central**	R^2**-adjusted (25th Percentile, Max) FL**	**Time**
(0.0032, 0.8807)	(0.0752,0.9656)	(-0.246, 0.8508)	(-0.1560, 0.957)	2m 51s
MSE (Mean±Std Dev) Central	**MSE (Mean±Std Dev) Empirical Total**	**MSE (25th Percentile, Max) Central**	**MSE (25th Percentile, Max) Empirical Total**	
(4.5618±2.0505)	(3.7995±1.6997)	(0.4127, 19.0703)	(0.3696, 9.6066)	

5.2.3 5000 Runs

Table 27: Percentile Client SCAT3 Total Score 5000 Runs

Response Variable	**Features**	**95% Confidence Interval for Central Model**	**95% Confidence Interval for Bootstrap Client Model**	**Number of DataPoints**
SCAT3TotalScore				103
	SCAT3TotalScore	(0.2224, 0.2272)	(0.2068, 0.2124)	
	CRTScore	(-0.1503, 0.0720)	(-0.1724, -0.1583)	
	SACScore	(-0.1452, -0.0815)	(-0.1567, -0.1462)	
	BSI18Score	(-0.0116, -0.0020)	(0.1088 , 0.1199)	
R^2 **(Mean±Std Dev) Central**	R^2 **(Mean±Std Dev) Bootstrap Client**	R^2**-adjusted (Mean±Std Dev) Central**	R^2**-adjusted (Mean±Std Dev) Bootstrap Client**	**Number of Trials**
(0.1666±0.7749)	(0.2604±0.7633)	(-0.0418±0.9687)	(0.0755±0.9541)	5000
R^2 **(25th Percentile, Max) Central**	R^2 **(25th Percentile, Max) FL**	R^2**-adjusted (25th Percentile, Max) Central**	R^2**-adjusted (25th Percentile, Max) FL**	**Time**
(-0.0013, 0.899)	(0.0761, 0.9642)	(-0.2516, 0.8738)	(-0.1548, 0.9552)	8m 33s
MSE (Mean±Std Dev) Central	**MSE (Mean±Std Dev) Empirical Total**	**MSE (25th Percentile, Max) Central**	**MSE (25th Percentile, Max) Empirical Total**	
(4.4626±2.0294)	(3.775±1.7489)	(0.4239, 15.7181)	(0.2741, 10.5301)	

5.2.4 10,000 Runs

Table 28: Percentile Client SCAT3 Total Score 10,000 Runs

Response Variable	**Features**	**95% Confidence Interval for Central Model**	**95% Confidence Interval for Bootstrap Client Model**	**Number of DataPoints**
SCAT3TotalScore				103
	SCAT3TotalScore	(0.2221, 0.2268)	(0.2158, 0.2244)	
	CRTScore	(-0.1479, 0.0722)	(-0.1690, -0.1599)	
	SACScore	(-0.1459, -0.0825)	(–0.1464, -0.1403)	
	BSI18Score	(-0.0117, -0.0019)	(0.0940, 0.1087)	
R^2 (Mean±Std Dev) Central	**R^2 (Mean±Std Dev) Bootstrap Client**	**R^2-adjusted (Mean±Std Dev) Central**	**R^2-adjusted (Mean±Std Dev) Bootstrap Client**	**Number of Trials**
(0.1418±0.7929)	(0.2383±0.7677)	(-0.0728±0.9911)	(0.0478±0.9596)	10000
R^2 (25^{th} Percentile, Max) Central	**R^2 (25^{th} Percentile, Max) FL**	**R^2-adjusted (25^{th} Percentile, Max) Central**	**R^2-adjusted (25^{th} Percentile, Max) FL**	**Time**
(-0.0159, 0.9132)	(0.0518, 0.9734)	(-0.2699, .8915)	(-0.1853, 0.9668)	48m 28s
MSE (Mean±Std Dev) Central	**MSE (Mean±Std Dev) Empirical Total**	**MSE (25^{th} Percentile, Max) Central**	**MSE (25^{th} Percentile, Max) Empirical Total**	
(4.5132±2.0595)	(3.7855±1.6929)	(0.3535, 17.0.110)	(0.3526, 9.9625)	

Appendix F. Empirical Client Bootstrap

6.1 SCAT3 Total Number of Symptoms

6.1.1 1000 Runs

Table 29: Empirical Client SCAT3 Total Number of Symptoms 1000 Runs

Response Variable	**Features**	**95% Confidence Interval for Central Model**	**95% Confidence Interval for Empirical Client Model**	**Number of DataPoints**
SCAT3TotalSymptoms				102
	SWLS Total Score	(0.0172, 0.0288)	(0.0151, 0.0268)	
	SCAT3TotalSymptoms	(0.2277, 0.2677)	(0.2359, 0.3075)	
	SACScore	(0.0851, 0.1325)	(0.0962, 0.1116)	
	VOMS Total Score	(0.0124, 0.0151)	(0.0016, 0.0203)	
	BSI18Score	(0.0716, 0.0972)	(0.0083, 0.0926)	
R^2 (Mean±Std Dev) Central	**R^2 (Mean±Std Dev) Empirical Client**	**R^2-adjusted (Mean±Std Dev) Central**	**R^2-adjusted (Mean±Std Dev) Empirical Client**	**Number of Trials**
(-0.0365±1.3381)	(0.0963±1.1444)	(-0.352±1.7453)	(-0.1787±1.4927)	1000
R^2 (25^{th} Percentile, Max) Central	**R^2 (25^{th} Percentile, Max) FL**	**R^2-adjusted (25^{th} Percentile, Max) Central**	**R^2-adjusted (25^{th} Percentile, Max) FL**	**Time**
(-0.1361, 0.652)	(-0.0420, 0.8959)	(-0.4818, 0.5461)	(-0.3591, 0.8643)	34.9s
MSE (Mean±Std Dev) Central	**MSE (Mean±Std Dev) Empirical Total**	**MSE (25^{th} Percentile, Max) Central**	**MSE (25^{th} Percentile, Max) Empirical Total**	
(3.6747±1.5971)	(3.1260±1.3382)	(0.4199, 9.9705)	(0.3622, 7.7736)	

6.1.2 3000 Runs

Table 30: Empirical Client SCAT3 Total Number of Symptoms 3000 Runs

Response Variable	**Features**	**95% Confidence Interval for Central Model**	**95% Confidence Interval for Empirical Client Model**	**Number of DataPoints**
SCAT3TotalSymptoms				102
	SWLS Total Score	(0.0165, 0.0282)	(0.0213, 0.0389)	
	SCAT3TotalSymptoms	(0.2261,0.2666)	(0.2144, 0.2514)	
	SACScore	(0.0881, 0.1359)	(0.0919, 0.1126)	
	VOMS Total Score	(0.0123, 0.0151)	(0.0104, 0.0157)	
	BSI18Score	(0.0675, 0.0933)	(0.0487, 0.0887)	
R^2 (Mean±Std Dev) Central	**R^2 (Mean±Std Dev) Empirical Client**	**R^2-adjusted (Mean±Std Dev) Central**	**R^2-adjusted (Mean±Std Dev) Empirical Client**	**Number of Trials**
(0.0144±0.9311)	(0.1703±0.5681)	(-0.2856±1.2145)	(-0.0823±0.7411)	3000
R^2 (25^{th} Percentile, Max) Central	**R^2 (25^{th} Percentile, Max) FL**	**R^2-adjusted (25^{th} Percentile, Max) Central**	**R^2-adjusted (25^{th} Percentile, Max) FL**	**Time**
(-0.1001, 0.7129)	(0.0054, 0.8906)	(-0.4349, 0.6256)	(-0.2973, 0.8573)	6m 39s
MSE (Mean±Std Dev) Central	**MSE (Mean±Std Dev) Empirical Total**	**MSE (25^{th} Percentile, Max) Central**	**MSE (25^{th} Percentile, Max) Empirical Total**	
(3.6562±1.6055)	(3.0584±1.312)	(0.4755, 10.4407)	(0.4674, 7.4325)	

6.1.3 5000 Runs

Table 31: Empirical Client SCAT3 Total Number of Symptoms 5000 Runs

Response Variable	**Features**	**95% Confidence Interval for Central Model**	**95% Confidence Interval for Empirical Client Model**	**Number of DataPoints**
SCAT3TotalSymptoms				102
	SWLS Total Score	(0.0153, 0.0269)	(0.0185, 0.0252)	
	SCAT3TotalSymptoms	(0.2260, 0.2662)	(0.2389, 0.2694)	
	SACScore	(0.0887, 0.1362)	(0.1084, 0.1189)	
	VOMS Total Score	(0.0127, 0.0154)	(0.0085, 0.0154)	
	BSI18Score	(0.0687, 0.0943)	(0.0732, 0.0903)	
R^2 (Mean±Std Dev) Central	**R^2 (Mean±Std Dev) Empirical Client**	**R^2-adjusted (Mean±Std Dev) Central**	**R^2-adjusted (Mean±Std Dev) Empirical Client**	**Number of Trials**
(-0.0068±0.9411)	(0.1392±0.7011)	(-0.3132±1.2275)	(-0.1228±0.9145)	5000
R^2 (25^{th} Percentile, Max) Central	**R^2 (25^{th} Percentile, Max) FL**	**R^2-adjusted (25^{th} Percentile, Max) Central**	**R^2-adjusted (25^{th} Percentile, Max) FL**	**Time**
(-0.1126, 0.7191)	(-0.0124, 0.8636)	(-0.4513, 0.6336)	(-0.3205, 0.8221)	20m 54s
MSE (Mean±Std Dev) Central	**MSE (Mean±Std Dev) Empirical Total**	**MSE (25^{th} Percentile, Max) Central**	**MSE (25^{th} Percentile, Max) Empirical Total**	
(3.7022±1.6096)	(3.1171±1.3344)	(0.4630, 10.6612)	(0.3300, 7.4443)	

6.1.4 10,000 Runs

Table 32: Empirical Client SCAT3 Total Number of Symptoms 10,000 Runs

Response Variable	**Features**	**95% Confidence Interval for Central Model**	**95% Confidence Interval for Empirical Client Model**	**Number of DataPoints**
SCAT3TotalSymptoms				102
	SWLS Total Score	(0.0158, 0.0275)	(0.0215, 0.0305)	
	SCAT3TotalSymptoms	(0.2258, 0.2661)	(0.2202, 0.2473)	
	SACScore	(0.0892, 0.1367)	(0.1105, 0.1229)	
	VOMS Total Score	(0.0129, 0.0156)	(0.0123, 0.0152)	
	BSI18Score	(0.0683, 0.0940)	(0.0787, 0.1163)	
R^2 (Mean±Std Dev) Central	**R^2 (Mean±Std Dev) Empirical Client**	**R^2-adjusted (Mean±Std Dev) Central**	**R^2-adjusted (Mean±Std Dev) Empirical Client**	**Number of Trials**
(0.0189±0.6450)	(0.1584±0.5146)	(-0.2796±0.8413)	(-0.0978±0.6713)	10000
R^2 (25^{th} Percentile, Max) Central	**R^2 (25^{th} Percentile, Max) FL**	**R^2-adjusted (25^{th} Percentile, Max) Central**	**R^2-adjusted (25^{th} Percentile, Max) FL**	**Time**
(-0.1020, 0.7262)	(-0.0114, 0.9053)	(-0.4374, 0.6429)	(-0.3192, 0.8765)	81m 21s
MSE (Mean±Std Dev) Central	**MSE (Mean±Std Dev) Empirical Total**	**MSE (25^{th} Percentile, Max) Central**	**MSE (25^{th} Percentile, Max) Empirical Total**	
(3.681±1.5911)	(3.1055±1.3242)	(0.4025, 12.7403)	(0.2939, 8.0476)	

6.2 SCAT3 Total Score

6.2.1 1000 Runs

Table 33: Empirical Client SCAT3 Total Score 1000 Runs

Response Variable	**Features**	**95% Confidence Interval for Central Model**	**95% Confidence Interval for Empirical Client Model**	**Number of DataPoints**
SCAT3TotalScore				103
	SCAT3TotalScore	(0.2251, 0.2299)	(0.229, 0.2375)	
	CRTScore	(-0.1512, 0.0654)	(-0.0461, -0.0155)	
	SACScore	(-0.1419, -0.0797)	(-0.1137, -0.1035)	
	BSI18Score	(-0.0178, 0.0080)	(-0.0296, -0.0072)	
R^2 (Mean±Std Dev) Central	**R^2 (Mean±Std Dev) Empirical Client**	**R^2-adjusted (Mean±Std Dev) Central**	**R^2-adjusted (Mean±Std Dev) Empirical Client**	**Number of Trials**
(0.1364±0.7790)	(0.3179±0.6006)	(-0.0795±0.9737)	(0.1474±0.7508)	1000
R^2 (25^{th} Percentile, Max) Central	**R^2 (25^{th} Percentile, Max) FL**	**R^2-adjusted (25^{th} Percentile, Max) Central**	**R^2-adjusted (25^{th} Percentile, Max) FL**	**Time**
(-0.0423, 0.9009)	(0.1361, 0.9498)	(-0.3029, 0.8761)	(-0.0799, 0.9373)	24.4s
MSE (Mean±Std Dev) Central	**MSE (Mean±Std Dev) Empirical Total**	**MSE (25^{th} Percentile, Max) Central**	**MSE (25^{th} Percentile, Max) Empirical Total**	
(4.6821±2.1133)	(3.6248±1.5226)	(0.4823, 18.8282)	(0.4237, 8.5066)	

6.2.2 3000 Runs

Table 34: Empirical Client SCAT3 Total Score 3000 Runs

Response Variable	**Features**	**95% Confidence Interval for Central Model**	**95% Confidence Interval for Empirical Client Model**	**Number of DataPoints**
SCAT3TotalScore				103
	SCAT3TotalScore	(0.2237, 0.2285)	(0.2223, 0.2268)	
	CRTScore	(-0.1490, 0.0726)	(-0.0382, -0.0197)	
	SACScore	(-0.1447, -0.0810)	(-0.1230, -0.1129)	
	BSI18Score	(-0.0151, -0.0051)	(-0.0103, -0.0012)	
R^2 **(Mean±Std Dev) Central**	R^2 **(Mean±Std Dev) Empirical Client**	R^2**-adjusted (Mean±Std Dev) Central**	R^2**-adjusted (Mean±Std Dev) Empirical Client**	**Number of Trials**
(0.1503±0.8031)	(0.3204±0.6324532107275694)	(0.1505±0.7906)	(0.1474±0.7508)	3000
R^2 **(25th Percentile, Max) Central**	R^2 **(25th Percentile, Max) FL**	R^2**-adjusted (25th Percentile, Max) Central**	R^2**-adjusted (25th Percentile, Max) FL**	**Time**
(-0.0132, 0.8775)	(0.1426, 0.9558)	(-0.2665, 0.8469)	(-0.0718, 0.9448)	9m 34s
MSE (Mean±Std Dev) Central	**MSE (Mean±Std Dev) Empirical Total**	**MSE (25th Percentile, Max) Central**	**MSE (25th Percentile, Max) Empirical Total**	
(4.4938±1.9977)	(3.5714±1.5745)	(0.388, 14.9746)	(0.3411, 9.4782)	

6.2.3 5000 Runs

Table 35: Empirical Client SCAT3 Total Score 5000 Runs

Response Variable	**Features**	**95% Confidence Interval for Central Model**	**95% Confidence Interval for Empirical Client Model**	**Number of DataPoints**
SCAT3TotalScore				103
	SCAT3TotalScore	(0.2224, 0.2272)	(0.2220, 0.2256)	
	CRTScore	(-0.1503, 0.0720)	(-0.0391, -0.0288)	
	SACScore	(-0.1452, -0.0815)	(-0.1135, -0.1078)	
	BSI18Score	(-0.0116, -0.0020)	(-0.0114, 0.0054)	
R^2 **(Mean±Std Dev) Central**	R^2 **(Mean±Std Dev) Empirical Client**	R^2**-adjusted (Mean±Std Dev) Central**	R^2**-adjusted (Mean±Std Dev) Empirical Client**	**Number of Trials**
(0.1450±0.7859)	(0.3111±0.6219)	(-0.0687±0.9823)	(0.1389±0.7774)	5000
R^2 **(25th Percentile, Max) Central**	R^2 **(25th Percentile, Max) FL**	R^2**-adjusted (25th Percentile, Max) Central**	R^2**-adjusted (25th Percentile, Max) FL**	**Time**
(-0.0158, 0.9023)	(0.1290, 0.9684)	(-0.2697, 0.8779)	(-0.0888, 0.9605)	36m 12s
MSE (Mean±Std Dev) Central	**MSE (Mean±Std Dev) Empirical Total**	**MSE (25th Percentile, Max) Central**	**MSE (25th Percentile, Max) Empirical Total**	
(4.4696±2.0336)	(3.5512±1.5475)	(0.4151, 19.1195)	(0.3889, 9.5458)	

6.2.4 10,000 Runs

Table 36: Empirical Client SCAT3 Total Score 10,000 Runs

Response Variable	**Features**	**95% Confidence Interval for Central Model**	**95% Confidence Interval for Empirical Client Model**	**Number of DataPoints**
SCAT3TotalScore				103
	SCAT3TotalScore	(0.2221, 0.2268)	(0.2234, 0.2280)	
	CRTScore	(-0.1479, 0.0722)	(-0.0393, -0.3092)	
	SACScore	(-0.1459, -0.0825)	(-0.1249, -0.11367)	
	BSI18Score	(-0.0117, -0.0019)	(−0.0099, -0.0054)	
R^2 (Mean±Std Dev) Central	**R^2 (Mean±Std Dev) Empirical Client**	**R^2-adjusted (Mean±Std Dev) Central**	**R^2-adjusted (Mean±Std Dev) Empirical Client**	**Number of Trials**
(0.1454±0.7827)	(0.3111±0.6141)	(-0.0683±0.9784)	(0.1389±0.7676)	10000
R^2 (25^{th} Percentile, Max) Central	**R^2 (25^{th} Percentile, Max) FL**	**R^2-adjusted (25^{th} Percentile, Max) Central**	**R^2-adjusted (25^{th} Percentile, Max) FL**	**Time**
(-0.0171, 0.9003)	(0.1200, 0.9572)	(-0.2713, 0.8753)	(-0.0999, 0.9465)	92m 33s
MSE (Mean±Std Dev) Central	**MSE (Mean±Std Dev) Empirical Total**	**MSE (25^{th} Percentile, Max) Central**	**MSE (25^{th} Percentile, Max) Empirical Total**	
(4.5323±2.0379)	(3.601±1.5855)	(0.449, 16.3193)	(0.2822, 9.1031)	

Appendix G. Percentile Total Bootstrap

7.1 SCAT3 Total Number of Symptoms

7.1.1 300 Runs

Table 37: Percentile Total SCAT3 Total Number of Symptoms 300 Runs

Response Variable	**Features**	**95% Confidence Interval for Central Model**	**95% Confidence Interval for Bootstrap Total Model**	**Number of DataPoints**
SCAT3TotalSymptoms				102
	SWLS Total Score	(0.01566, 0.0270)	(0.0109, 0.0316)	
	SCAT3TotalSymptoms	(0.2266,0.2664)	(0.1768, 0.2377)	
	SACScore	(0.0872, 0.1332)	(0.0775, 0.1134)	
	VOMS Total Score	(0.0112, 0.0138)	(-0.0049, 0.0084)	
	BSI18Score	(0.0656, 0.0911)	(0.1046, 0.1699)	
R^2 (Mean±Std Dev) Central	**R^2 (Mean±Std Dev) Bootstrap Total**	**R^2-adjusted (Mean±Std Dev) Central**	**R^2-adjusted (Mean±Std Dev) Bootstrap Total**	**Number of Trials**
(0.0540±0.4481)	(0.2450±0.3162)	(-0.2339±0.5845)	(0.0152±0.4125)	300
R^2 (25^{th} Percentile, Max) Central	**R^2 (25^{th} Percentile, Max) FL**	**R^2-adjusted (25^{th} Percentile, Max) Central**	**R^2-adjusted (25^{th} Percentile, Max) FL**	**Time**
(-0.0692, 0.6711)	(0.0751, 0.8158)	(-0.3946, 0.571)	(-0.2064, 0.7597)	12m 45s
MSE (Mean±Std Dev) Central	**MSE (Mean±Std Dev) Empirical Total**	**MSE (25^{th} Percentile, Max) Central**	**MSE (25^{th} Percentile, Max) Empirical Total**	
(3.7279±1.5705)	(2.9882±1.2650)	(0.6775, 8.8768)	(0.4883, 6.2516)	

7.1.2 500 Runs

Table 38: Percentile Total SCAT3 Total Number of Symptoms 500 Runs

Response Variable	**Features**	**95% Confidence Interval for Central Model**	**95% Confidence Interval for Bootstrap Total Model**	**Number of DataPoints**
SCAT3TotalSymptoms				102
	SWLS Total Score	(0.0147, 0.0263)	(0.0124, 0.0308)	
	SCAT3TotalSymptoms	(0.2252, 0.2659)	(0.1671, 0.2243)	
	SACScore	(0.0886, 0.1361)	(0.0827, 0.1122)	
	VOMS Total Score	(0.0120, 0.0147)	(-0.0044, 0.0061)	
	BSI18Score	(0.0722, 0.0981)	(0.1108, 0.1617)	
R^2 **(Mean±Std Dev) Central**	R^2 **(Mean±Std Dev) Bootstrap Total**	R^2**-adjusted (Mean±Std Dev) Central**	R^2**-adjusted (Mean±Std Dev) Bootstrap Total**	**Number of Trials**
(0.0427±0.4707)	(0.2415±0.3173)	(-0.2486±0.6140)	(0.0106±0.4139)	500
R^2 **(25th Percentile, Max) Central**	R^2 **(25th Percentile, Max) FL**	R^2**-adjusted (25th Percentile, Max) Central**	R^2**-adjusted (25th Percentile, Max) FL**	**Time**
(-0.0969, 0.6376)	(0.0567, 0.8246)	(-0.4307, 0.5273)	(-0.2304, 0.7713)	40m 6s
MSE (Mean±Std Dev) Central	**MSE (Mean±Std Dev) Empirical Total**	**MSE (25th Percentile, Max) Central**	**MSE (25th Percentile, Max) Empirical Total**	
(3.6782±1.5959)	(2.9646±1.3640)	(0.7921, 8.8637)	(0.3148, 6.9409)	

7.1.3 1000 Runs

Table 39: Percentile Total SCAT3 Total Number of Symptoms 1000 Runs

Response Variable	**Features**	**95% Confidence Interval for Central Model**	**95% Confidence Interval for Bootstrap Total Model**	**Number of DataPoints**
SCAT3TotalSymptoms				102
	SWLS Total Score	(0.0155, 0.0270)	(0.0148, 0.0286)	
	SCAT3TotalSymptoms	(0.2239, 0.2642)	(0.1833, 0.2152)	
	SACScore	(0.0886, 0.1356)	(0.0847, 0.1028)	
	VOMS Total Score	(0.0124, 0.0152)	(-0.0008, 0.0067)	
	BSI18Score	(0.0707, 0.0961)	(0.1199, 0.1694)	
R^2 **(Mean±Std Dev) Central**	R^2 **(Mean±Std Dev) Bootstrap Total**	R^2**-adjusted (Mean±Std Dev) Central**	R^2**-adjusted (Mean±Std Dev) Bootstrap Total**	**Number of Trials**
(-0.0091±1.0217)	(0.2001±0.6808)	(-0.3162±1.3326)	(-0.0434±0.8880)	1000
R^2 **(25th Percentile, Max) Central**	R^2 **(25th Percentile, Max) FL**	R^2**-adjusted (25th Percentile, Max) Central**	R^2**-adjusted (25th Percentile, Max) FL**	**Time**
(-0.1101, 0.6557)	(0.0532, 0.8319)	(-0.4479, 0.551)	(-0.2350, 0.7807)	191m 21s
MSE (Mean±Std Dev) Central	**MSE (Mean±Std Dev) Empirical Total**	**MSE (25th Percentile, Max) Central**	**MSE (25th Percentile, Max) Empirical Total**	
(3.7506±1.6433)	(3.0208±1.375)	(0.4406, 10.2498)	(0.2151, 7.1407)	

7.1.4 2000 Runs

Table 40: Percentile Total SCAT3 Total Number of Symptoms 2000 Runs

Response Variable	**Features**	**95% Confidence Interval for Central Model**	**95% Confidence Interval for Bootstrap Total Model**	**Number of DataPoints**
SCAT3TotalSymptoms				102
	SWLS Total Score	(0.0152, 0.02679)	(0.0147, 0.02421)	
	SCAT3TotalSymptoms	(0.2280, 0.2681)	(0.1926, 0.2152)	
	SACScore	(0.0897, 0.1368)	(0.0861, 0.1013)	
	VOMS Total Score	(0.0138 , 0.0165)	(0.0003, 0.0057)	
	BSI18Score	(0.0656, 0.0911)	(0.1232, 0.1548)	
R^2 (Mean±Std Dev) Central	**R^2 (Mean±Std Dev) Bootstrap Total**	**R^2-adjusted (Mean±Std Dev) Central**	**R^2-adjusted (Mean±Std Dev) Bootstrap Total**	**Number of Trials**
(0.0203±0.5642)	(0.2158±0.3937)	(-0.2779±0.7359)	(-0.0229±0.5136)	2000
R^2 (25^{th} Percentile, Max) Central	**R^2 (25^{th} Percentile, Max) FL**	**R^2-adjusted (25^{th} Percentile, Max) Central**	**R^2-adjusted (25^{th} Percentile, Max) FL**	**Time**
(-0.1006, 0.6449)	(0.0584, 0.8631)	(-0.4356, 0.5369)	(-0.2282, 0.8214)	1347m 24s
MSE (Mean±Std Dev) Central	**MSE (Mean±Std Dev) Empirical Total**	**MSE (25^{th} Percentile, Max) Central**	**MSE (25^{th} Percentile, Max) Empirical Total**	
(3.671±1.5844)	(2.975±1.3481)	(0.4643, 9.9618)	(0.3228, 7.3726)	

7.2 SCAT3 Total Score

7.2.1 300 Runs

Table 41: Percentile Total SCAT3 Total Score 300 Runs

Response Variable	**Features**	**95% Confidence Interval for Central Model**	**95% Confidence Interval for Bootstrap Total Model**	**Number of DataPoints**
SCAT3TotalScore				103
	SCAT3TotalScore	(0.2271, 0.2319)	(0.2303, 0.2612)	
	CRTScore	(-0.1525, 0.0720)	(-0.0791, 0.0229)	
	SACScore	(-0.1490, -0.0849)	(-0.1281, -0.0812)	
	BSI18Score	(-0.0163, -0.0063)	(-0.0215, 0.0451)	
R^2 (Mean±Std Dev) Central	**R^2 (Mean±Std Dev) Bootstrap Total**	**R^2-adjusted (Mean±Std Dev) Central**	**R^2-adjusted (Mean±Std Dev) Bootstrap Total**	**Number of Trials**
(0.1426±0.7390)	(0.3415±0.5093)	(-0.0718±0.9238)	(0.1768±0.6366)	300
R^2 (25^{th} Percentile, Max) Central	**R^2 (25^{th} Percentile, Max) FL**	**R^2-adjusted (25^{th} Percentile, Max) Central**	**R^2-adjusted (25^{th} Percentile, Max) FL**	**Time**
(-0.0644, 0.8849)	(0.1300, 0.9630)	(-0.3305, 0.8561)	(-0.0875, 0.8537)	11m 1s
MSE (Mean±Std Dev) Central	**MSE (Mean±Std Dev) Empirical Total**	**MSE (25^{th} Percentile, Max) Central**	**MSE (25^{th} Percentile, Max) Empirical Total**	
(4.3774±1.9747)	(3.3071±1.4478)	(0.5509, 13.8465)	(0.3893, 8.2739)	

7.2.2 500 Runs

Table 42: Percentile Total SCAT3 Total Score 500 Runs

Response Variable	**Features**	**95% Confidence Interval for Central Model**	**95% Confidence Interval for Bootstrap Total Model**	**Number of DataPoints**
SCAT3TotalScore				103
	SCAT3TotalScore	(0.2213, 0.2261)	(0.2279, 0.2614)	
	CRTScore	(-0.1466, 0.0754)	(-0.0697, -0.0057)	
	SACScore	(-0.1435, -0.0802)	(-0.1231, 0.0845)	
	BSI18Score	(-0.0106, -0.0010)	(-0.0096, 0.0359)	
R^2 **(Mean±Std Dev) Central**	R^2 **(Mean±Std Dev) Bootstrap Total**	R^2**-adjusted (Mean±Std Dev) Central**	R^2**-adjusted (Mean±Std Dev) Bootstrap Total**	**Number of Trials**
(0.1862±0.6647)	(0.385±0.4694)	(-0.0173±0.8309)	(0.2313±0.5868)	500
R^2 **(25^{th} Percentile, Max) Central**	R^2 **(25^{th} Percentile, Max) FL**	R^2**-adjusted (25^{th} Percentile, Max) Central**	R^2**-adjusted (25^{th} Percentile, Max) FL**	**Time**
(0.0494, 0.8662)	(0.2024, 0.9433)	(-0.1883, 0.8327)	(0.0030, 0.9291)	36m 19s
MSE (Mean±Std Dev) Central	**MSE (Mean±Std Dev) Empirical Total**	**MSE (25^{th} Percentile, Max) Central**	**MSE (25^{th} Percentile, Max) Empirical Total**	
(4.5752±1.9825)	(3.4283±1.3988)	(0.5338, 13.7961)	(0.5751, 8.8899)	

7.2.3 1000 Runs

Table 43: Percentile Total SCAT3 Total Score 1000 Runs

Response Variable	**Features**	**95% Confidence Interval for Central Model**	**95% Confidence Interval for Bootstrap Total Model**	**Number of DataPoints**
SCAT3TotalScore				103
	SCAT3TotalScore	(0.2227, 0.2274)	(0.2300, 0.2539)	
	CRTScore	(-0.1439, 0.0752)	(-0.0437, -0.0042)	
	SACScore	(-0.1454, -0.0822)	(-0.1342, -0.1011)	
	BSI18Score	(-0.0139, 0.0039)	(-0.0061, 0.0264)	
R^2 **(Mean±Std Dev) Central**	R^2 **(Mean±Std Dev) Bootstrap Total**	R^2**-adjusted (Mean±Std Dev) Central**	R^2**-adjusted (Mean±Std Dev) Bootstrap Total**	**Number of Trials**
(0.1320±0.9364)	(0.3363±0.7223)	(-0.0850±1.1705)	(0.1703±0.9029)	1000
R^2 **(25^{th} Percentile, Max) Central**	R^2 **(25^{th} Percentile, Max) FL**	R^2**-adjusted (25^{th} Percentile, Max) Central**	R^2**-adjusted (25^{th} Percentile, Max) FL**	**Time**
(-0.0216, 0.8841)	(-0.1647, 0.9430)	(-0.2770, 0.8552)	(-0.0441, 0.9287)	175m 37s
MSE (Mean±Std Dev) Central	**MSE (Mean±Std Dev) Empirical Total**	**MSE (25^{th} Percentile, Max) Central**	**MSE (25^{th} Percentile, Max) Empirical Total**	
(4.498±2.0371)	(3.3501±1.3674)	(0.5381, 18.0542)	(0.4969, 8.2852)	

7.2.4 2000 Runs

Table 44: Percentile Total SCAT3 Total Score 2000 Runs

Response Variable	**Features**	**95% Confidence Interval for Central Model**	**95% Confidence Interval for Bootstrap Total Model**	**Number of DataPoints**
SCAT3TotalScore				103
	SCAT3TotalScore	(0.2218, 0.2265)	(0.2387, 0.2529)	
	CRTScore	(-0.1492, 0.0711)	(-0.0397, -0.0111)	
	SACScore	(-0.1445, -0.0817)	(-0.1188, -0.1009)	
	BSI18Score	(-0.0124, -0.0027)	(-0.0032, 0.0205)	
R^2 **(Mean±Std Dev) Central**	R^2 **(Mean±Std Dev) Bootstrap Total**	R^2**-adjusted (Mean±Std Dev) Central**	R^2**-adjusted (Mean±Std Dev) Bootstrap Total**	**Number of Trials**
(0.1392±0.8291)	(0.3437±0.6361)	(-0.0759±1.0364)	(0.1797±0.7952)	2000
R^2 **(25th Percentile, Max) Central**	R^2 **(25th Percentile, Max) FL**	R^2**-adjusted (25th Percentile, Max) Central**	R^2**-adjusted (25th Percentile, Max) FL**	**Time**
(-0.0137, 0.8833)	(0.1876, 0.9592)	(-0.2671, 0.8542)	(-0.0155, 0.9490)	1305m 49s
MSE (Mean±Std Dev) Central	**MSE (Mean±Std Dev) Empirical Total**	**MSE (25th Percentile, Max) Central**	**MSE (25th Percentile, Max) Empirical Total**	
(4.5184±2.0537)	(3.35±1.4147)	(0.5072, 14.9126)	(0.4183, 8.0642)	

Appendix H. Empirical Total Bootstrap

8.1 SCAT3 Total Number of Symptoms

8.1.1 300 Runs

Table 45: Empirical Total SCAT3 Total Number of Symptoms 300 Runs

Response Variable	**Features**	**95% Confidence Interval for Central Model**	**95% Confidence Interval for Empirical Total Model**	**Number of DataPoints**
SCAT3TotalSymptoms				102
	SWLS Total Score	(0.01566, 0.0270)	(0.0115, 0.0324)	
	SCAT3TotalSymptoms	(0.2266,0.2664)	(0.2154, 0.2655)	
	SACScore	(0.0872, 0.1332)	(0.0846, 0.1302)	
	VOMS Total Score	(0.0112, 0.0138)	(0.0083, 0.0248)	
	BSI18Score	(0.0656, 0.0911)	(0.0566, 0.1308)	
R^2 **(Mean±Std Dev) Central**	R^2 **(Mean±Std Dev) Empirical Total**	R^2**-adjusted (Mean±Std Dev) Central**	R^2**-adjusted (Mean±Std Dev) Empirical Total**	**Number of Trials**
(0.0733±0.3986)	(0.2333±0.4503)	(-0.2087±0.5199)	(0±0.5873)	300
R^2 **(25^{th} Percentile, Max) Central**	R^2 **(25^{th} Percentile, Max) FL**	R^2**-adjusted (25^{th} Percentile, Max) Central**	R^2**-adjusted (25^{th} Percentile, Max) FL**	**Time**
(-0.0757, 0.6548)	(0.0882, 0.8593)	(-0.4031, 0.5498)	(-0.1894, 0.8165)	11m 33s
MSE (Mean±Std Dev) Central	**MSE (Mean±Std Dev) Empirical Total**	**MSE (25^{th} Percentile, Max) Central**	**MSE (25^{th} Percentile, Max) Empirical Total**	
(3.8667±1.6569)	(3.1504±1.49)	(0.5539, 10.6878)	(0.3646, 14.1696)	

8.1.2 500 Runs

Table 46: Empirical Total SCAT3 Total Number of Symptoms 500 Runs

Response Variable	**Features**	**95% Confidence Interval for Central Model**	**95% Confidence Interval for Empirical Total Model**	**Number of DataPoints**
SCAT3TotalSymptoms				102
	SWLS Total Score	(0.0147, 0.0263)	(0.0176, 0.0380)	
	SCAT3TotalSymptoms	(0.2252, 0.2659)	(0.2174, 0.2609)	
	SACScore	(0.0886, 0.1361)	(0.0939, 0.1219)	
	VOMS Total Score	(0.0120, 0.0147)	(0.0107, 0.0220)	
	BSI18Score	(0.0722, 0.0981)	(0.0632, 0.1149)	
R^2 **(Mean±Std Dev) Central**	R^2 **(Mean±Std Dev) Empirical Total**	R^2**-adjusted (Mean±Std Dev) Central**	R^2**-adjusted (Mean±Std Dev) Empirical Total**	**Number of Trials**
(-0.0006±0.5606)	(0.2082±0.3749)	(-0.3051±0.7312)	(-0.0328±0.4891)	500
R^2 **(25^{th} Percentile, Max) Central**	R^2 **(25^{th} Percentile, Max) FL**	R^2**-adjusted (25^{th} Percentile, Max) Central**	R^2**-adjusted (25^{th} Percentile, Max) FL**	**Time**
(-0.1342, 0.641)	(0.0417, 0.8233)	(-0.4793, 0.5317)	(-0.2499, 0.7695)	39m 29s
MSE (Mean±Std Dev) Central	**MSE (Mean±Std Dev) Empirical Total**	**MSE (25^{th} Percentile, Max) Central**	**MSE (25^{th} Percentile, Max) Empirical Total**	
(3.7158±1.5671)	(2.9856±1.3175)	(0.5275,9.9532)	(0.4175, 6.753)	

8.1.3 1000 Runs

Table 47: Empirical Total SCAT3 Total Number of Symptoms 1000 Runs

Response Variable	**Features**	**95% Confidence Interval for Central Model**	**95% Confidence Interval for Empirical Total Model**	**Number of DataPoints**
SCAT3TotalSymptoms				102
	SWLS Total Score	(0.0155, 0.0270)	(0.0129, 0.0265)	
	SCAT3TotalSymptoms	(0.2239, 0.2642)	(0.2392, 0.2679)	
	SACScore	(0.0886, 0.1356)	(0.1006, 0.1214)	
	VOMS Total Score	(0.0124, 0.0152)	(0.0085, 0.0161)	
	BSI18Score	(0.0707, 0.0961)	(0.0537, 0.0984)	
R^2 **(Mean±Std Dev) Central**	R^2 **(Mean±Std Dev) Empirical Total**	R^2**-adjusted (Mean±Std Dev) Central**	R^2**-adjusted (Mean±Std Dev) Empirical Total**	**Number of Trials**
(0.015±0.5987)	(0.2152±0.3956)	(-0.2847±0.7809)	(-0.0236±0.516)	1000
R^2 **(25^{th} Percentile, Max) Central**	R^2 **(25^{th} Percentile, Max) FL**	R^2**-adjusted (25^{th} Percentile, Max) Central**	R^2**-adjusted (25^{th} Percentile, Max) FL**	**Time**
(-0.1161, 0.6531)	(0.0584, 0.854)	(-0.4558, 0.5475)	(-0.2281, 0.8096)	204m 9s
MSE (Mean±Std Dev) Central	**MSE (Mean±Std Dev) Empirical Total**	**MSE (25^{th} Percentile, Max) Central**	**MSE (25^{th} Percentile, Max) Empirical Total**	
(3.7533±1.6028)	(3.0271±1.3425)	(0.5467, 8.7968)	(0.3964, 7.1324)	

8.1.4 2000 Runs

Table 48: Empirical Total SCAT3 Total Number of Symptoms 2000 Runs

Response Variable	**Features**	**95% Confidence Interval for Central Model**	**95% Confidence Interval for Empirical Total Model**	**Number of DataPoints**
SCAT3TotalSymptoms				102
	SWLS Total Score	(0.0152, 0.02679)	(0.0134, 0.0239)	
	SCAT3TotalSymptoms	(0.2280, 0.2681)	(0.2381, 0.2623)	
	SACScore	(0.0897, 0.1368)	(0.1019, 0.1175)	
	VOMS Total Score	(0.0138 , 0.0165)	(0.0100, 0.0167)	
	BSI18Score	(0.0656, 0.0911)	(0.0668, 0.0972)	
R^2 **(Mean±Std Dev) Central**	R^2 **(Mean±Std Dev) Empirical Total**	R^2**-adjusted (Mean±Std Dev) Central**	R^2**-adjusted (Mean±Std Dev) Empirical Total**	**Number of Trials**
(0.0094±0.6436)	(0.2167±0.4169)	(-0.2921±0.8395)	(-0.0217±0.5438)	2000
R^2 **(25th Percentile, Max) Central**	R^2 **(25th Percentile, Max) FL**	R^2**-adjusted (25th Percentile, Max) Central**	R^2**-adjusted (25th Percentile, Max) FL**	**Time**
(-0.1039, 0.6685)	(0.0578, 0.8502)	(-0.4398, 0.5676)	(-0.2290, 08046)	1316m 23s
MSE (Mean±Std Dev) Central	**MSE (Mean±Std Dev) Empirical Total**	**MSE (25th Percentile, Max) Central**	**MSE (25th Percentile, Max) Empirical Total**	
(3.7298±1.6022)	(3.0008±1.3281)	(0.5458, 10.6987)	(0.27, 7.1546)	

8.2 SCAT3 Total Score

8.2.1 300 Runs

Table 49: Empirical Total SCAT3 Total Score 300 Runs

Response Variable	**Features**	**95% Confidence Interval for Central Model**	**95% Confidence Interval for Empirical Total Model**	**Number of DataPoints**
SCAT3TotalScore				103
	SCAT3TotalScore	(0.2271, 0.2319)	(0.2044, 0.2391)	
	CRTScore	(-0.1525, 0.0720)	(-0.0639, -0.0079)	
	SACScore	(-0.1490, -0.0849)	(-0.1303, -0.0879)	
	BSI18Score	(-0.0163, -0.0063)	(-0.0340, 0.0215)	
R^2 **(Mean±Std Dev) Central**	R^2 **(Mean±Std Dev) Empirical Total**	R^2**-adjusted (Mean±Std Dev) Central**	R^2**-adjusted (Mean±Std Dev) Empirical Total**	**Number of Trials**
(0.0913±0.8926)	(0.3032±0.7107)	(-0.1358±1.1158)	(0.1290±0.8884)	300
R^2 **(25th Percentile, Max) Central**	R^2 **(25th Percentile, Max) FL**	R^2**-adjusted (25th Percentile, Max) Central**	R^2**-adjusted (25th Percentile, Max) FL**	**Time**
(-0.0498, 0.8886)	(0.1547, 0.9472)	(-0.3123, 0.8608)	(-0.0566, 0.9340)	14m 59s
MSE (Mean±Std Dev) Central	**MSE (Mean±Std Dev) Empirical Total**	**MSE (25th Percentile, Max) Central**	**MSE (25th Percentile, Max) Empirical Total**	
(4.4518±1.9832)	(3.2989±1.3669)	(0.5595, 12.3045)	(0.4475, 7.1673)	

8.2.2 500 Runs

Table 50: Empirical Total SCAT3 Total Score 500 Runs

Response Variable	Features	95% Confidence Interval for Central Model	95% Confidence Interval for Empirical Total Model	Number of DataPoints
SCAT3TotalScore				103
	SCAT3TotalScore	(0.2213, 0.2261)	(0.2058, 0.2383)	
	CRTScore	(-0.1466, 0.0754)	(-0.0533, -0.0006)	
	SACScore	(-0.1435, -0.0802)	(-0.1198, -0.0821)	
	BSI18Score	(-0.0106, -0.0010)	(-0.0365, 0.0106)	
R^2 (Mean±Std Dev) Central	R^2 (Mean±Std Dev) Empirical Total	R^2-adjusted (Mean±Std Dev) Central	R^2-adjusted (Mean±Std Dev) Empirical Total	Number of Trials
(0.1724±0.7023)	(0.3655±0.5121)	(-0.0345±0.8779)	(0.2069±0.6401)	500
R^2 (25^{th} Percentile, Max) Central	R^2 (25^{th} Percentile, Max) FL	R^2-adjusted (25^{th} Percentile, Max) Central	R^2-adjusted (25^{th} Percentile, Max) FL	Time
(0.0457, 0.8784)	(0.1993, 0.9541)	(-0.1929, 0.8479)	(-0.0009, 0.9427)	47m 29s
MSE (Mean±Std Dev) Central	MSE (Mean±Std Dev) Empirical Total	MSE (25^{th} Percentile, Max) Central	MSE (25^{th} Percentile, Max) Empirical Total	
(4.4257±1.9363)	(3.3625±1.4107)	(0.6223, 11.4299)	(0.4982, 7.5234)	

8.2.3 1000 Runs

Table 51: Empirical Total SCAT3 Total Score 1000 Runs

Response Variable	Features	95% Confidence Interval for Central Model	95% Confidence Interval for Empirical Total Model	Number of DataPoints
SCAT3TotalScore				103
	SCAT3TotalScore	(0.2227, 0.2274)	(0.2111, 0.2316)	
	CRTScore	(-0.1439, 0.0752)	(-0.0488, -0.0095)	
	SACScore	(-0.1454, -0.0822)	(-0.1342, -0.1079)	
	BSI18Score	(-0.0139, 0.0039)	(-0.0142, 0.0261)	
R^2 (Mean±Std Dev) Central	R^2 (Mean±Std Dev) Empirical Total	R^2-adjusted (Mean±Std Dev) Central	R^2-adjusted (Mean±Std Dev) Empirical Total	Number of Trials
(0.1303±0.9299)	(0.3207±0.7162)	(-0.0871±1.1624)	(0.1509±0.8953)	1000
R^2 (25^{th} Percentile, Max) Central	R^2 (25^{th} Percentile, Max) FL	R^2-adjusted (25^{th} Percentile, Max) Central	R^2-adjusted (25^{th} Percentile, Max) FL	Time
(-0.0319, 0.8941)	(0.1314, 0.9635)	(-0.2898, 0.8676)	(-0.0858, 0.9544)	210m 16s
MSE (Mean±Std Dev) Central	MSE (Mean±Std Dev) Empirical Total	MSE (25^{th} Percentile, Max) Central	MSE (25^{th} Percentile, Max) Empirical Total	
(4.6306±2.1642)	(4.6306±2.1642)	(0.3118, 13.7749)	(0.3118, 13.7749)	

8.2.4 2000 Runs

Table 52: Empirical Total SCAT3 Total Score 2000 Runs

Response Variable	**Features**	**95% Confidence Interval for Central Model**	**95% Confidence Interval for Empirical Total Model**	**Number of DataPoints**
SCAT3TotalScore				103
	SCAT3TotalScore	(0.2218, 0.2265)	(0.2209, 0.2352)	
	CRTScore	(-0.1492, 0.0711)	(-0.0624, -0.0324)	
	SACScore	(-0.1445, -0.0817)	(-0.1189, -0.0979)	
	BSI18Score	(-0.0124, -0.0027)	(-0.0265, -0.0023)	
R^2 **(Mean±Std Dev) Central**	R^2 **(Mean±Std Dev) Empirical Total**	R^2**-adjusted (Mean±Std Dev) Central**	R^2**-adjusted (Mean±Std Dev) Empirical Total**	**Number of Trials**
(0.1675±0.7424)	(0.3616±0.5362)	(-0.0406±0.928)	(0.202±0.6703)	2000
R^2 **(25^{th} Percentile, Max) Central**	R^2 **(25^{th} Percentile, Max) FL**	R^2**-adjusted (25^{th} Percentile, Max) Central**	R^2**-adjusted (25^{th} Percentile, Max) FL**	**Time**
(0.0137, 0.8874)	(0.1854, 0.9696)	(-0.2328, 0.8593)	(-0.0183, 0.962)	1360m 15s
MSE (Mean±Std Dev) Central	**MSE (Mean±Std Dev) Empirical Total**	**MSE (25^{th} Percentile, Max) Central**	**MSE (25^{th} Percentile, Max) Empirical Total**	
(4.5536±2.0399)	(3.4111±1.4061)	(0.39912, 14.0953)	(0.2515, 9.2929)	

Bibliography

1. Brett Theeler, Sylvia Lucas, Ronald G Riechers, and Robert L Ruff. Post-traumatic headaches in civilians and military personnel: A comparative, clinical review. *Headache: The Journal of Head and Face Pain*, 53(6):881–900, May 2013.

2. Kang Nahua. Multi-layer neural networks with sigmoid function- deep learning for rookies (2), Jun 2017.

3. Shervin Minaee, Yao Wang, Anna Choromanska, Sohae Chung, Xiuyuan Wang, Els Fieremans, Steven Flanagan, Joseph Rath, and Yvonne W. Lui. A deep unsupervised learning approach toward mtbi identification using diffusion mri. In *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 1267–1270, 2018.

4. Thomas W. McAllister, Michael McCrea, and Steven Broglio. Concussion assessment, research and education (care) consortium - longitudinal clinical study core assessment manual, Mar 2016.

5. Douglas C. Montgomery, Elizabeth A. Peck, and G. Geoffrey Vining. *Chapter 5/Transformations and Weighting to Correct Model Inadequacies*, volume 821 of *Wiley Series in Probability and Statistics*, page 294–409. Wiley, fifth edition, 2021.

6. CDC. What is a concussion?, Feb 2019.

7. Randolph W. Evans. Persistent post-traumatic headache, postconcussion syndrome, and whiplash injuries: The evidence for a non-traumatic basis with an historical review. *Headache: The Journal of Head and Face Pain*, 50(4):716–724, Apr 2010.

8. Marcela Cruz-Haces, Jonathan Tang, Glen Acosta, Joseph Fernandez, and Riyi Shi. Pathological correlations between traumatic brain injury and chronic neurodegenerative diseases. *Translational Neurodegeneration*, 6(1), 2017.

9. Jonathan A. Hollander and Cindy Lawler. Neurodegenerative diseases, Jun 2022.

10. Cleveland Clinic. Nervous system: What it is, types, symptoms, May 2020.

11. Daniel H. Daneshvar, David O. Riley, Christopher J. Nowinski, Ann C. McKee, Robert A. Stern, and Robert C. Cantu. Long-term consequences: Effects on normal development profile after concussion. *Physical Medicine and Rehabilitation Clinics of North America*, 22(4):683–700, Sep 2011.

12. Andrea Almeida, Andrea Aagesen, and Kristi Waters Ray. Concussion in athletes, Aug 2014.

13. Charles H. Tator. Concussions and their consequences: Current diagnosis, management and prevention. *Canadian Medical Association Journal*, 185(11):975–979, Aug 2013.

14. National Institute on Aging. Alzheimer's disease fact sheet, Jul 2021.

15. National Institutes of Health. Parkinson's disease, Feb 2022.

16. Thor D Stein, Victor E Alvarez, and Ann C McKee. Chronic traumatic encephalopathy: A spectrum of neuropathological changes following repetitive brain trauma in athletes and military personnel. *Alzheimer's Research Therapy*, 6(1):4, Jan 2014.

17. National Institutes of Health. Amyotrophic lateral sclerosis (als) fact sheet, Jun 2013.

18. Sportradar. 2021-22 nba season summary — basketball-reference.com, Nov 2022.

19. Sandro Azerrad. How many players are in the nhl? (updated 2022-23), Nov 2022.

20. Gary Putnik. Nfl players by college on 2021 rosters, Sep 2021.

21. NCAA. Our division i members, 2022.

22. Olivia Begasse de Dhaem, William B. Barr, Laura J. Balcer, Steven L. Galetta, and Mia T. Minen. Post-traumatic headache: The use of the sport concussion assessment tool (scat-3) as a predictor of post-concussion recovery. *The Journal of Headache and Pain*, 18(1), May 2017.

23. NCAA-DOD. Ncaa - dod grand alliance, May 2022.

24. Johanna M Hurtubise, Cindy E Hughes, Lauren E Sergio, and Alison K Macpherson. Comparison of baseline and postconcussion scat3 scores and symptoms in varsity athletes: An investigation into differences by sex and history of concussion. *BMJ Open Sport amp; Exercise Medicine*, 4(1), Mar 2018.

25. Bill Polian. History of nfl scouting combine and why it's important to teams, Feb 2023.

26. Edwin Weathersby. A scout's take on how college football recruits are evaluated, Oct 2017.

27. Deborah Warden. Military tbi during the iraq and afghanistan wars. *Journal of Head Trauma Rehabilitation*, 21(5):398–402, Oct 2006.

28. Hannah Fischer. A guide to u.s. military casualty statistics: Operation freedom's sentinel, operation inherent resolve, operation new dawn, operation iraqi freedom, and operation enduring freedom, Aug 2015.

29. Joseph R Biden. National security strategy of the united states of america. Technical report, Executive Office of The President Washington DC Washington United States, 2022.

30. Mehmet Kayaalp. Patient privacy in the era of big data. *Balkan medical journal*, 35(1):8–17, 2018.

31. Avi Goldfarb and Catherine E. Tucker. Online advertising, behavioral targeting, and privacy. *Commun. ACM*, 54(5):25–27, may 2011.

32. Office for Civil Rights. Your rights under hipaa, Jan 2022.

33. H. Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Agüera y Arcas. Communication-efficient learning of deep networks from decentralized data. 2016.

34. Yunliang Cai, Shaoju Wu, Wei Zhao, Zhigang Li, Zheyang Wu, and Songbai Ji. Concussion classification via deep learning using whole-brain white matter fiber strains. *PLOS ONE*, 13(5), May 2018.

35. Spencer W Liebel, Caroline G Turner, Anna Camille Svirsko, Gian-Gabriel P Garcia, Paul F Pasquina, Thomas McAllister, Michael McCrea, and Steven P Broglio. Temporal network architectures of neurocognitive functioning and psychological symptoms in collegiate athletes following concussion. *Journal of Neurotrauma*, (ja).

36. Lou Grangeon, Emer O'Connor, Chun-Kong Chan, Layan Akijian, Thanh Mai Pham Ngoc, and Manjit Singh Matharu. New insights in post-traumatic headache with cluster headache phenotype: a cohort study. *Journal of Neurology, Neurosurgery & Psychiatry*, 91(6):572–579, 2020.

37. Tansel Yilmaz, Gerwin Roks, Myrthe de Koning, Myrthe Scheenen, Harm van der Horn, Gerben Plas, Gerard Hageman, Guus Schoonman, Jacoba Spikman, and Joukje van der Naalt. Risk factors and outcomes associated with post-traumatic headache after mild traumatic brain injury. *Emergency medicine journal*, 34(12):800–805, 2017.

38. Jin-zhou Feng, Yu Wang, Jin Peng, Ming-wei Sun, Jun Zeng, and Hua Jiang. Comparison between logistic regression and machine learning algorithms on survival prediction of traumatic brain injuries. *Journal of critical care*, 54:110–116, 2019.

39. Alan O Sykes. An introduction to regression analysis. 1993.

40. Michael A Poole and Patrick N O'Farrell. The assumptions of the linear regression model. *Transactions of the Institute of British Geographers*, pages 145–158, 1971.

41. Xiaogang Su, Xin Yan, and Chih-Ling Tsai. Linear regression. *Wiley Interdisciplinary Reviews: Computational Statistics*, 4(3):275–294, Feb 2012.

42. Sandro Sperandei. Understanding logistic regression analysis, Feb 2014.

43. Gary B. Wilkerson, Dustin C. Nabhan, and Ryan T. Crane. Concussion history and neuromechanical responsiveness asymmetry. *Journal of Athletic Training*, 55(6):594–600, 2020.

44. Jeffrey J. Bazarian, Robert J. Elbin, Douglas J. Casa, Gillian A. Hotz, Christopher Neville, Rebecca M. Lopez, David M. Schnyer, Susan Yeargin, and Tracey Covassin. Validation of a machine learning brain electrical activity–based index to aid in diagnosing concussion among athletes. *JAMA Network Open*, 4(2), Feb 2021.

45. Karen Rose, Scott Eldridge, and Lyman Chapin. The internet of things: An overview. *The internet society (ISOC)*, 80:1–50, 2015.

46. Mohammed Aledhari, Rehma Razzak, Reza M. Parizi, and Fahad Saeed. Federated learning: A survey on enabling technologies, protocols, and applications. *IEEE Access*, 8:140699–140725, 2020.

47. Raed Kontar, Naichen Shi, Xubo Yue, Seokhyun Chung, Eunshin Byon, Mosharaf Chowdhury, Judy Jin, Wissam Kontar, Neda Masoud, Maher Nouiehed, Chinedum Emmanuel Okwudire, Garvesh Raskutti, Romesh Saigal, Karandeep Singh, and Zhisheng Ye. The internet of federated things (ioft): A vision for the future and in-depth survey of data-driven approaches for federated learning. *CoRR*, abs/2111.05326, 2021.

48. Nicholas C. Zakas. The evolution of web development for mobile devices. *Commun. ACM*, 56(4):42–48, apr 2013.

49. Rayhan A Tariq and Pamela B Hackert. Patient confidentiality, Sep 2022.

50. Jie Xu, Benjamin S Glicksberg, Chang Su, Peter Walker, Jiang Bian, and Fei Wang. Federated learning for healthcare informatics. *Journal of Healthcare Informatics Research*, 5:1–19, 2021.

51. Li Huang, Andrew L Shea, Huining Qian, Aditya Masurkar, Hao Deng, and Dianbo Liu. Patient clustering improves efficiency of federated machine learning to predict mortality and hospital stay time using distributed electronic medical records. *Journal of biomedical informatics*, 99:103291, 2019.

52. Tian Li, Anit Kumar Sahu, Ameet Talwalkar, and Virginia Smith. Federated learning: Challenges, methods, and future directions. *IEEE signal processing magazine*, 37(3):50–60, 2020.

53. Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-Efficient Learning of Deep Networks from Decentralized Data. In Aarti Singh and Jerry Zhu, editors, *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, volume 54 of *Proceedings of Machine Learning Research*, pages 1273–1282. PMLR, 20–22 Apr 2017.

54. Guillem Barroso. Admore itn, May 2018.

55. Moloud Abdar, Farhad Pourpanah, Sadiq Hussain, Dana Rezazadegan, Li Liu, Mohammad Ghavamzadeh, Paul Fieguth, Xiaochun Cao, Abbas Khosravi, U. Rajendra Acharya, and et al. A review of uncertainty quantification in deep learning: Techniques, applications and challenges. *Information Fusion*, 76:243–297, Nov 2020.

56. Matias Valdenegro-Toro and Daniel Saromo Mori. A deeper look into aleatoric and epistemic uncertainty disentanglement. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1508–1516. IEEE, 2022.

57. Eduard Hofer, Martina Kloos, Bernard Krzykacz-Hausmann, Jörg Peschke, and Martin Woltereck. An approximate epistemic uncertainty analysis approach in the presence of epistemic and aleatory uncertainties. *Reliability Engineering & System Safety*, 77(3):229–238, 2002.

58. Jishnu Mukhoti and Yarin Gal. Evaluating bayesian deep learning methods for semantic segmentation. *CoRR*, abs/1811.12709, 2018.

59. Bradley Efron and Robert Tibshirani. The bootstrap method for assessing statistical accuracy. *Behaviormetrika*, 12:1–35, 1985.

60. Steven P Broglio, Barry P Katz, Shi Zhao, Micahel McCrea, and Thomas McAllister. Test-retest reliability and interpretation of common concussion assessment tools: Findings from the ncaa-dod care consortium, Nov 2017.

61. Christopher J Recklitis, Jaime E Blackmon, and Grace Chang. Validity of the brief symptom inventory-18 (bsi-18) for identifying depression and anxiety in young adult cancer survivors: Comparison with a structured clinical diagnostic interview, Oct 2017.

62. Orthotoolkit. Free online standardized assessment of concussion score calculator, 2023.

63. Paul McMahon, Allison Hricik, John K Yue, Ava M Puccio, Tomoo Inoue, Hester F Lingsma, Sue R Beers, Wayne A Gordon, Alex B Valadka, Geoffrey T Manley, and et al. Symptomatology and functional outcome in mild traumatic brain injury: Results from the prospective track-tbi study, Jan 2014.

64. Anne Mucha, Michael W Collins, RJ Elbin, Joseph M Furman, Cara Troutman-Enseki, Ryan M DeWolf, Greg Marchetti, and Anthony P Kontos. A brief vestibular/ocular motor screening (voms) assessment to evaluate concussions: Preliminary findings, Aug 2014.

65. Jaclyn B Caccese, James T Eckner, Lea Franco-MacKendrick, Joseph B Hazzard, Meng Ni, Steven P Broglio, Thomas W McAllister, Michael McCrea, and Thomas A Buckley. Clinical reaction-time performance factors in healthy collegiate athletes, Jun 2020.

66. Shania Kennedy. How to improve data normalization in healthcare, Jan 2023.

67. Shichao Zhang. Nearest neighbor selection for iteratively knn imputation. *Journal of Systems and Software*, 85(11):2541–2552, 2012.

68. Ibtissam Abnane, Mohamed Hosni, Ali Idri, and Alain Abran. Analogy software effort estimation using ensemble knn imputation. In *2019 45th Euromicro Conference on Software Engineering and Advanced Applications (SEAA)*, pages 228–235, 2019.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704–0188

1. REPORT DATE *(DD–MM–YYYY)*	2. REPORT TYPE	3. DATES COVERED *(From — To)*
23–03–2023	Master's Thesis	September 2021 — March 2023

4. TITLE AND SUBTITLE

Uncertainty Quantification in Federated Learning for Persistent Post-Traumatic Headache

5a. CONTRACT NUMBER

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)

Kim, Byungmoo, B, 2d Lt, USAF

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

Air Force Institute of Technology
Graduate School of Engineering and Management (AFIT/EN)
2950 Hobson Way
WPAFB OH 45433-7765

8. PERFORMING ORGANIZATION REPORT NUMBER

AFIT-ENS-MS-23-M-132

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)

Intentionally Left Blank

10. SPONSOR/MONITOR'S ACRONYM(S)

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION / AVAILABILITY STATEMENT

DISTRIBUTION STATEMENT A.
APPROVED FOR PUBLIC RELEASE; DISTRIBUTION UNLIMITED.

13. SUPPLEMENTARY NOTES

14. ABSTRACT

A post-traumatic headache (PTH), resulting from a mild traumatic brain injury (mTBI), potentially develops into persistent post-traumatic headache (PPTH). Although no known cure for PPTH exists, research has shown that receiving treatment at earlier stages of PTH lowers the risk of patients developing PPTH. Previous studies have shown machine learning (ML) models capable of predicting a patient's PTH progression, but none have considered the issue of protecting patient privacy. Due to patient privacy, ML models only have access to data within the institution. Federated learning (FL) harnesses data from separate institutions without sacrificing patient privacy as institutions can run ML models on their own private dataset and share the trained model parameters without sharing data. Additionally, quantifying uncertainty of model parameters associated with key features of interest in predicting PTH progression has not been explored in the context of FL. Uncertainty Quantification in Federated Learning (UQFL) combines FL and uncertainty quantification to protect patient privacy and provide a measure of uncertainty for each model parameter.

15. SUBJECT TERMS

Federated Learning (FL), Uncertainty Quanitification (UQ), Uncertainty Quantification Federated Learning (UQFL), Post-Traumatic Headache (PTH), Traumatic Brain Injury (TBI)

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			Dr. Nathan B. Gaw, AFIT/ENS
U	U	U	UU	115	19b. TELEPHONE NUMBER *(include area code)* (937) 255-6565 x4791; Nathan.Gaw@afit.edu