Ensemble Learning with Recursive Partitioning Methods to Explore Relationships between Mental Health and Physical Activity

Jessica Gilmore

Follow this and additional works at: https://scholarworks.uark.edu/hhpruht

Part of the Community Health and Preventive Medicine Commons, Exercise Science Commons, and the Other Mental and Social Health Commons

Citation

This Thesis is brought to you for free and open access by the Health, Human Performance and Recreation at ScholarWorks@UARK. It has been accepted for inclusion in Health, Human Performance and Recreation Undergraduate Honors Theses by an authorized administrator of ScholarWorks@UARK. For more information, please contact scholar@uark.edu.
Ensemble Learning with Recursive Partitioning Methods to Explore Relationships between Mental Health and Physical Activity

An honors thesis submitted in partial fulfillment of the requirements for the honors program in Health, Human Performance, and Recreation

by

Jessica Gilmore

December 2020
University of Arkansas

This thesis is approved for submission.

Dr. Samantha Robinson, Ph.D.
Honors Thesis Director

Dr. Erin K. Howie Hickey, Ph.D.
Honors Committee Member
Abstract

Physical and mental health are imperative to maintaining a well functioning immune system, which is especially critical during a global pandemic. Moreover, physical and mental health contribute to the overall quality of life experienced by an individual. Consequently, it is important to explore factors that contribute to both physical and mental health. Physical activity has been previously shown to improve physical and mental health yet many individuals do not get enough physical activity daily. Using data collected during the larger Exercise is Medicine (EIM) study, the current study utilized ensemble learning with recursive partitioning methods to explore the relationships that exist between health as measured by the SF-12, various types of physical activity as measured by the International Physical Activity Questionnaire (IPAQ), and participant demographics. Results indicated that physical activity, especially in leisure, is an important variable contributing to both physical and mental health. Results also demonstrated the value in utilizing ensemble learning with recursive partitioning methods to study the effect of physical activity on overall health.

Acknowledgments

I would like to thank Dr. Samantha Robinson for her patience and dedication throughout this thesis project. I would like to thank Dr. Erin K. Howie Hickey for allowing us to use her Exercise is Medicine data to analyze such a pressing topic. Also, I appreciate all my friends and family who pushed me to continue and never give up.
# Table of Contents

Introduction .................................................................................................................. 1

Motivation of Study .................................................................................................. 1

Purpose of Study ....................................................................................................... 1

Literature Review ....................................................................................................... 2

Physical Health in the United States ........................................................................ 2

Mental Health in the United States ........................................................................... 2

Link between Physical Activity and Health .............................................................. 2

Method ....................................................................................................................... 3

Participants and Recruitment ................................................................................ 3

Instruments and Measures ....................................................................................... 3

Analytic Approach .................................................................................................. 5

Results ...................................................................................................................... 8

Participant Demographics ..................................................................................... 8

Physical Health ......................................................................................................... 10

Mental Health ........................................................................................................... 17

Discussion ................................................................................................................ 23

Summary of Results ................................................................................................ 23

Comparison of Findings with Literature ................................................................. 24

Limitations and Future Research .......................................................................... 25

References ............................................................................................................... 26
List of Tables and Figures

Table 1 Demographic features representative of the sample participants .............................................. 9
Table 2 Select physical activity measures and reported health of the sample participants .......... 10
Figure 1 Example of tree instability ........................................................................................................... 6
Figure 2 Un-simplified regression tree predicting the physical health component score .......... 11
Figure 3 Simplified regression tree predicting the physical health component score ................. 12
Figure 4 Variable importance in predicting physical health based upon bagging ....................... 13
Figure 5 Variable importance in predicting physical health based upon random forest .......... 14
Figure 6 Simplified regression tree predicting the age-specific mean difference score for physical health ........................................................................................................................................... 15
Figure 7 Variable importance in predicting age-specific mean differences in physical health based upon bagging ........................................................................................................................................... 16
Figure 8 Variable importance in predicting age-specific mean differences in physical health based upon random forest ........................................................................................................................................... 17
Figure 9 Simplified regression tree predicting the mental health component score ............... 18
Figure 10 Variable importance in predicting mental health based upon bagging ....................... 19
Figure 11 Variable importance in predicting mental health based upon random forest ............ 20
Figure 12 Simplified regression tree predicting the age-specific mean difference score for mental health ........................................................................................................................................... 21
Figure 13 Variable importance in predicting age-specific mean differences in mental health based upon bagging ........................................................................................................................................... 22
Figure 14 Variable importance in predicting age-specific mean differences in mental health based upon bagging ........................................................................................................................................... 23
Introduction

Physical and mental health are two of the most important aspects of health. Poor mental health can be damaging to physical health and vice versa. Exercise is vital for staying “physically healthy as well as improving your mental wellbeing” (Mental Health Foundation, 2016). Although always important, with the current COVID-19 pandemic happening, it is especially imperative that people stay healthy, physically, and mentally. During a pandemic, it is essential for individuals to make sure their immune systems are functioning well. A major component of a healthy immune system is being mentally and physically healthy and being physically active. Studies show that physical activity increases a person’s mental and physical health. In this era of the pandemic, physical activity could be a key factor in staying healthy.

Motivation of Study

This study is motivated by the current statistics and data regarding physical and mental health in the United States. With many people not getting enough physical activity and the prevalence of mental illness being substantial, it is important that we look at the relationship between the two. This study is also motivated by the COVID-19 pandemic. With people being at home more, it is possible that people are getting less exercise, which could cause a decline in health.

Purpose of Study

The purpose of the current study is twofold. The primary purpose of this study is to explore the relationships that exist between health, physical activity, and participant demographics using recursive partitioning statistical methods. The secondary purpose of this study is to illustrate the benefits of using ensemble learning with recursive partitioning methods.
Literature Review

This study seeks to analyze how an individual’s physical activity affects their mental and physical health. To provide the necessary, though not exhaustive background and motivation for the current study, we have reviewed and summarized the literature on this topic.

Physical Health in the United States

Physical health is important for people to live a long, healthy life. It is important that individuals participate in at least 150 minutes of moderate-intensity aerobic activity, 75 minutes of vigorous-intensity aerobic activity, or an equivalent combination of both per week (American College of Sports Medicine [ACSM], 2018). While these are the recommended amounts of physical activity, a large amount of the United States does not meet this measure. In 2018, 24.2% of adults did not engage in any leisure-time physical activity, and in 2017, only 50.3% of adults met the recommended guidelines (Centers for Disease Control, n.d.).

Mental Health in the United States

Mental health is an increasingly prevalent public health issue in the United States. In 2019, 20.6% of U.S. adults had a mental illness, an estimated 51.5 million adults (National Institute of Mental Health, 2020).

Link Between Physical Activity and Health

Research shows that physical activity can greatly impact one’s health. Higher levels of physical activity are associated with lower risks of chronic diseases such as diabetes, cardiovascular disease, or CVD, and various types of cancer (Miles, 2007). Some chronic diseases also follow a dose-response pattern, indicating that “more intense exercise, carried out more often and for longer episodes over a more prolonged period of months reduces CVD risk more than less intense exercise” (Miles, 2007). Additionally, it has been shown that regular
physical activity can decrease anxiety and depression and improve cognitive function (ACSM, 2018). In a 2008 study conducted in the United Kingdom, researchers concluded that participating in at least 20 minutes a week of any physical activity is associated with lower rates of psychological distress. Furthermore, the study found that at higher volumes and/or intensities of exercise, there a dose-response relationship between physical activity and psychological distress (Hamer, et al., 2008). Another European study resulted in similar findings. The experimental group with the most physical activity had significantly lower levels of anxiety and depression than the groups with less amounts (Tyson, et al., 2010). This is consistent with the dose-response pattern found in the first study. These findings lay the foundation for the purpose of our research.

**Method**

**Participants and Recruitment**

Participants were recruited from the University of Arkansas community through physical advertisements such as flyers and posters, through electronic advertisements such as through campus news that is emailed daily, and through resulting snowball sampling in the form of word-of-mouth recruitment. All participants in the current study were part of the larger, Institutional Review Board (IRB) approved Exercise is Medicine (EIM) study. It is assumed that most if not all participants were members of the UA community i.e., faculty, staff, or students.

**Instruments and Measures**

Participants completed an online questionnaire with demographic questions, questions about their level of physical activity within the last seven days, and questions that intended to measure both physical and mental health.
**SF-12.** The SF-12 is a 12-item survey that “is a generic measure and does not target a specific age or disease group” (Utah Department of Health [UDH], 2001). It consists of taking a composite score for mental and physical health aspects. Scores range from 0 to 100, “where a zero score indicates the lowest level of health measured by the scales and 100 indicates the highest level of health” (UDH, 2001). Because the survey is not age specific and scores can vary throughout an individual’s lifespan, the use of age-specific mean difference scores is helpful for interpretation. This “difference score” is determined by how much an individual’s score differs from the mean score of their age group. For example, an individual who has a difference score of -2.5 scored 2.5 points lower than the mean score for their age. This indicates slightly poorer health. Likewise, an individual receiving a difference score of 2.5 scored 2.5 points higher than the mean of the age group, indicating slightly better heath. The use of age-specific mean difference scores allows us to assess individuals across different age groups, because a difference score of -2.5 has the same meaning of health relative to an individual’s age group.

**IPAQ.** The International Physical Activity Questionnaire (IPAQ) is an instrument used to assess physical activity in adults aged 15-96 years old (The IPAQ Group, n.d.). The form assesses scores for walking, moderate-intensity activity, and vigorous-intensity activity in the domains of work, transportation, domestic activity, and leisure time. Individuals are asked to provide their duration and frequency of physical activity in the categories as previously mentioned. The answers are then converted into MET-minutes per week.

MET-minutes per week is calculated by multiplying the number of METs from the activity, the number of minutes, and the number of days. For example, the calculation of MET-minutes per week for an individual who walks 10 minutes on 5 days out of the week would be as follows: MET-minutes per week = 3.3 METs x 10 minutes x 5 days. This means the individual
would have 165 MET-minutes per week for the walking category. Once the individuals’ answers are converted to MET-minutes per week, the individuals are given scores for their total physical activity. The individuals are placed into a category of either Low, Moderate, or High, depending on the amount of physical activity completed weekly. Low individuals are in the the lowest level of physical activity. Moderate individuals participate in either 3 or more days of vigorous-intensity activity, 5 or more days of moderate-intensity activity, or 5 or more days of any combination of physical activity totaling to at least 600 MET-minutes per week. High individuals complete either 3 or more days of vigorous-intensity exercise totaling 1500 MET-minutes per week or 7 or more days of any combination of physical activity totaling at least 3000 MET-minutes per week. (The IPAQ Group, n.d.).

**Analytic Approach**

The purpose of the current study is to explore the relationships that exist between health, physical activity, and participant demographics using recursive partitioning statistical methods as well as illustrate the benefits of using ensemble learning with recursive partitioning methods. These methods are applied in four separate settings, as we consider separately the dependent variables of PCS, PCS age-specific mean difference, MCS, and MCS age-specific mean difference. For each analysis, only physical activity measures obtained from the IPAQ and demographics are considered as independent variables. A description of the analytic approach is provided below that explains the methods utilized in the current study in detail.

**Classification and Regression Tree Methods.** Classification and regression trees (CART) methods are a type of supervised learning technique used to predict a dependent variable using several independent variables. The resulting prediction can be visualized in the form of a decision tree. If predicting a categorical variable, a classification tree is created. If
predicting a continuous variable, as in the current study, a regression tree is created. Recursive partitioning is the statistical method for multivariable analysis that is utilized to create such a tree and just refers to the process of recursively splitting the feature space into several binary, if-then splits until a certain stopping criterion is met (Breiman et al., 1984). According to Stobl, Malley, and Tutz (2009), the main characteristic of this nonparametric regression approach is that the recursive partitions are made such that similar observations become grouped together.

Specifically, in un-pruned/un-simplified regression trees, we recursively partition the data until the terminal nodes (i.e., the last groupings created by the splits) have some minimum number of observations. The splits are made to minimize the Residual Sum of Squares (RSS), which is a type of predictive error i.e., $\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$. A great benefit of CART methods is the visualization of prediction in terms of if-then splits, which makes them very easy to explain.

In the current study, regression trees will be utilized to predict the four dependent variables using the physical activity and demographic variables. Regression trees will be pruned so that the tree stops if splitting groups in terminal nodes have 25 or fewer observations.

**Ensemble Methods.** These types of learning methods use multiple learning algorithms in aggregate e.g., a set of several regression trees to improve overall predictive performance.

![Figure 1. Example of tree instability. (Source: Dwyer & Holte, 2007)](image_url)
While there are advantages to CART methods, classification and regression trees suffer from instability. The high variability of single trees is easily illustrated by building trees with different bootstrap samples from the original data and observing the tree structure for each. As in this example from Dwyer and Holte (2007), only one observation differs between the two trees, yet the structure is very different. This is the reason ensemble methods are often employed. A set of trees can be built using random samples and, then, combined to reduce overall variance in prediction. Two commonly used ensemble methods are bootstrap aggregating (bagging) and random forests. Both will be utilized in the current study.

**Bootstrap Aggregating.** Bootstrap aggregating, also known as bagging, uses multiple bootstrap samples to create multiple regression trees, which will then be combined for prediction. Bootstrap samples are samples of the same size as the original data and are drawn from the original data with replacement. As cited in Strobl et al. (2009), Breiman, Dietterich, Bauer and Kohavi found that individual trees are not stable but, on average, are more stable and have better predictive performance. In the current study, bagging of 100 regression trees will be utilized to predict each dependent variable as well as to ascertain the importance of each independent variable.

**Random Forests.** Random forests produce a more diverse set of trees to aggregate because they involve both a random bootstrap sample and a random set of predictors for each individual tree in the forest. As noted by Strobl et al. (2009), this improves the predictive performance of the algorithm even further. Since the entire set of independent variables does not need utilized each time, the algorithm is quicker and more trees can be combined in the same amount of time. Thus, in the current study, 5000 trees are combined in each random forest used to predict each dependent variable. Also, by noting which variables occur with greatest
frequency in each tree and/or which variables improve predictive performance most, a measure of variable importance will be provided.

Ensemble learning methods improve predictive performance and allow for measures of variable importance. However, the visual benefit of CART methods is lost since it is difficult to visualize more than one tree and, in fact, hundreds of trees.

**Results**

**Participant Demographics**

Table 1 below shows the demographic features for our sample participants. Of the 398 participants, the vast majority were female (76.2%) and white (81.1%), with 94.4% of the individuals identifying as non-Hispanic. While many of the participants were undergraduates (42.4%), the mean age was 32.3 years with a standard deviation of 14.3 years. This is due to the ages of the staff members who made up 31.7% of the sample likely skewing the data. Many participants have a graduate school education (45.2%) or an undergraduate education (39.3%), meaning 84.5% of the participants had a higher level of education. The majority of the subjects’ mothers and fathers both had a high school or college education. 55.5% of the participants were never associated with Greek Life and 39.9% were associated with Greek Life at the time of the study.
Table 1. Demographic features representative of the sample participants.

<table>
<thead>
<tr>
<th>Demographic</th>
<th>N=398*</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Sex/Gender</strong></td>
<td><strong>Education</strong></td>
</tr>
<tr>
<td>Female</td>
<td>301 (76.2)</td>
</tr>
<tr>
<td>Male</td>
<td>94 (23.8)</td>
</tr>
<tr>
<td><strong>Ethnicity</strong></td>
<td><strong>Mother’s Education</strong></td>
</tr>
<tr>
<td>Hispanic</td>
<td>22 (5.6)</td>
</tr>
<tr>
<td>Non-Hispanic</td>
<td>370 (94.4)</td>
</tr>
<tr>
<td><strong>Race</strong></td>
<td><strong>Father’s Education</strong></td>
</tr>
<tr>
<td>Asian</td>
<td>25 (6.3)</td>
</tr>
<tr>
<td>Black</td>
<td>13 (3.3)</td>
</tr>
<tr>
<td>Multiple</td>
<td>18 (4.5)</td>
</tr>
<tr>
<td>White</td>
<td>321 (81.1)</td>
</tr>
<tr>
<td>Other</td>
<td>19 (4.8)</td>
</tr>
<tr>
<td><strong>Role</strong></td>
<td><strong>Greek Life</strong></td>
</tr>
<tr>
<td>Admin or Faculty</td>
<td>39 (9.9)</td>
</tr>
<tr>
<td>Graduate Student</td>
<td>63 (16.0)</td>
</tr>
<tr>
<td>Staff</td>
<td>125 (31.7)</td>
</tr>
<tr>
<td>Undergraduate</td>
<td>167 (42.4)</td>
</tr>
<tr>
<td><strong>Age (years)</strong></td>
<td>32.3 (14.3)</td>
</tr>
</tbody>
</table>

*For Categorical Variables, table values are Count (%); For Continuous Variables, table values are Mean (SD); Nonresponses/NAs are removed from individual descriptive statistic calculations
Table 2 below displays the means and standard deviations for a select number of the survey measure scores.

*For Categorical Variables, table values are Count (%); For Continuous Variables, table values are Mean (SD); Nonresponses/NAs are removed from individual descriptive statistic calculations.

The mean physical health score among the participants was 55.3 and the mean mental health score was 42.1. All activity domains of the survey had a mean duration of MET-minutes per week over 1,000. The most physical activity was found under the work domain with a value of 2,227 MET-minutes per week and the least physical activity was domestic activity (1,030 minutes). Active/Transport activity (1,216 minutes) and leisure activity (1,704 minutes) fell between work and domestic activity.

**Physical Health**

Physical health is the state of “optimal health and functioning” of an individual’s body (University of New Hampshire, 2020). This consists of fitness, cardiovascular endurance, muscular strength, etc. Many factors, such as genetics, age, diet, and physical activity, affect
one’s physical health. Physical activity, especially, is important to physical health, as many diseases and conditions can be controlled or eliminated with recommended levels of physical activity.

**SF-12 Physical Health Composite Score.** A regression tree was created to predict the PCS score using all physical activity and demographic variables.

The un-pruned regression tree can be seen in Figure 2 below.

![Figure 2](image-url)  
*Figure 2.* Un-simplified regression tree predicting the physical health component score.

The un-pruned tree splits the feature space into predictive groups based upon the description of the student (F, So, J, Sen, Non-Student, etc.), the amount of vigorous physical activity each day, race, age, the highest level of education for both an individual’s mother and father, etc. This un-pruned tree is very difficult to read since there are so many variables and so
many splits. Because of this, a simplified tree was created such that the recursive partitioning would stop if any terminal node had fewer than 25 observations.

The pruned and simplified regression tree predicting PCS score from physical activity and demographic variables is displayed in Figure 3 below. It can be seen that PCS (i.e., the physical health component score) for an individual can be adequately predicted by variables such as age, vigorous activity minutes, and individual description. If a participant is over 25 years of age, engages in vigorous physical activity for less than 5 minutes each day, and also has lower MET minutes weekly in active transport, their PCS score is the lowest predicted of all individuals in the sample.

![Simplified regression tree predicting the physical health component score.](image)

**Figure 3.** Simplified regression tree predicting the physical health component score.

Due to the expected unstable nature of any one tree, bagging was implemented with 100 regression trees. As displayed in Figure 4, physical activity attributed to leisure as well as age
were the most important variables when predicting the physical health of an individual, as measured by the PCS score.

To minimize the variance in prediction further, a random forest was utilized with 5000 trees. As displayed in Figure 5, physical activity attributed to all domains was the most important variable type when predicting the physical health of an individual, as measured by the PCS score.

Figure 4. Variable importance in predicting physical health based upon bagging.
Figure 5. Variable importance in predicting physical health based upon random forest.

SF-12 Physical Health Composite Age-Specific Mean Difference Score. The Physical and Mental Health Composite Scale scores have little intuitive meaning. This is due to the range of possible scores varying greatly. PCS and MCS scores tend to change over an individual’s life span and for different age groups as well, as “PCS tends to decrease with age, while MCS tends to increase” (UDH, 2001).

Since it is expected that PCS and MCS scores vary over an individual lifespan, with physical health tending to decrease with age, the analysis above was repeated using age-specific mean difference scores. These mean difference scores were found by subtracting the US age
group average PCS scores reported in 2001 from every observed PCS score in the current sample.

The pruned and simplified regression tree predicting PCS age-specific mean difference score from physical activity and demographic variables is displayed in Figure 6 below. It can be seen that PCS age-specific mean difference score for an individual can be adequately predicted by variables such as leisure activity, active transport activity, and age. If a participant has high activity levels, their PCS age-specific mean difference score is higher regardless of any other demographic features.

![Figure 6. Simplified regression tree predicting the age-specific mean difference score for physical health.](image-url)
Due to the expected unstable nature of any one tree, bagging was implemented with 100 regression trees. As displayed in Figure 7, physical activity attributed to leisure as well as age were the most important variables when predicting the physical health of an individual, as measured by the PCS age-specific mean difference score.

Figure 7. Variable importance in predicting age-specific mean differences in physical health based upon bagging

To minimize the variance in prediction further, a random forest was utilized with 5000 trees. As displayed in Figure 8, physical activity attributed to all leisure was the most important variable type when predicting the physical health of an individual, as measured by the PCS age-specific mean difference score.
Mental Health

Mental health is more than just being free of mental illnesses. It is a “state of well-being in which an individual realizes his or her own abilities, can cope with the normal stresses of life, can work productively and is able to make a contribution to his or her community” (World Health Organization, 2018). There are various factors that determine an individual’s level of mental health. Physical activity, in particular, promotes mental wellbeing and is seen in many studies, including our own.

**SF-12 Mental Health Composite Score.** The pruned and simplified regression tree predicting the MCS score from physical activity and demographic variables is displayed in Figure 9 below. It can be seen that the MCS score for an individual can be adequately predicted...
by variables such as age, leisure activity, domestic activity, and participant description as it relates to education. If a participant is younger, has lower leisure-related and is either in a doctoral program or is a first or second year undergraduate student, their MCS is lower. Individuals above 51 years of age have a higher MCS score regardless of any other physical activity measures or demographic features.

Figure 9. Simplified regression tree predicting the mental health component score.

Due to the expected unstable nature of any one tree, bagging was implemented with 100 regression trees. As displayed in Figure 10, physical activity attributed to leisure as well as age were the most important variables when predicting the mental health of an individual, as measured by the MCS score.
To minimize the variance in prediction further, a random forest was utilized with 5000 trees. As displayed in Figure 11, physical activity attributed to all leisure was the most important variable type when predicting the mental health of an individual, as measured by the MCS score.
SF-12 Mental Health Composite Age-Specific Mean Difference Score. Since it is expected that PCS and MCS scores vary over an individual lifespan, with mental health tending to increase with age, the analysis above was repeated using age-specific mean difference scores. These mean difference scores were found by subtracting the US age group average MCS scores reported in 2001 from every observed MCS score in the current sample.

The pruned and simplified regression tree predicting MCS age-specific mean difference score from physical activity and demographic variables is displayed in Figure 12 below. It can be seen that MCS age-specific mean difference score for an individual can be adequately predicted.
by variables such as leisure activity, education level, and walking. If a participant has high leisure activity levels, has a highest education level that is not college/university (indicating that these students must not be graduate students but, rather undergraduate students or faculty), and walks a great deal, their MCS age-specific mean difference score is higher regardless of any other demographic features and is the only age-specific mean difference score that is positive, reflecting values above the US average.

![Simplified regression tree](image)

And here.

*Figure 12.* Simplified regression tree predicting the age-specific mean difference score for mental health.

Due to the expected unstable nature of any one tree, bagging was implemented with 100 regression trees. As displayed in Figure 13, age and physical activity attributed to leisure were the most important variables when predicting the mental health of an individual, as measured by
the MCS age-specific mean difference score. It is worth noting how important age is when predicting overall mental health.

Figure 13. Variable importance in predicting age-specific mean differences in mental health based upon bagging.

To minimize the variance in prediction further, a random forest was utilized with 5000 trees. As displayed in Figure 14, physical activity attributed to all leisure was the most important variable type when predicting the mental health of an individual, as measured by the MCS age-specific mean difference score. In addition, variables such as age and participation in Greek life were also very important in predicting mental health.
Figure 14. Variable importance in predicting age-specific mean differences in mental health based upon bagging.

Discussion

The purpose of the current study was to determine the relationship between an individual’s health and physical activity. We also looked at how other factors can predict the one’s level of physical activity.

Summary of Results

When looking at physical health, we found that an individual’s amount of physical activity is a predictor of their physical health. It is shown through the regression trees that active transport and/or leisure physical activity is a predictor of both PCS and PCS age-specific mean difference scores. Individuals with a higher level of physical activity had a higher PCS age-
specific mean difference score, regardless of any other demographics. Age was also found to be an indicator of PCS and PCS age-specific means difference scores. This is expected since PCS scores tend to decrease with age.

For mental health, the same variables were found to be important predictors. Participants who had a high amount of leisure activity and walking and had a highest education level that was not college/university, meaning they are likely undergraduate students or faculty, had higher MCS age-specific mean differences scores regardless of any other factors. Leisure physical activity as well as age were found to be important variables for MCS and MCS age-specific mean differences scores, similarly to the physical component. This is expected as MCS scores tend to increase over one’s lifespan.

From this study, we can conclude that physical activity does in fact have an impact on both physical and mental health. Some type of physical activity had the highest importance for predicting all of the dependent variables related to physical and mental health, showing that physical activity may be the single most important factor to an individual’s overall health.

**Comparison of Findings with Literature**

These results are consistent with other studies conducted. Many other studies have shown that higher levels of physical activity are associated with better physical and mental health. Our current study suggests similar conclusions can be made. In both the regression trees for mental health and physical health, leisure activity was the most important variable in predicting mental and physical health. This means that continued research into physical activity and its affect on health is promising.
Limitations and Future Research

From the analysis of the study participants, many of the demographics are disproportional. 76.2% of the participants were females, 94.4% were non-Hispanic, and 81.1% were white. Many of the participants (42.2%) were also undergraduates. Because of this, the results may have limited external validity. Future studies should have a sample of participants with demographics more representative of the larger population of interest. This will help to ensure that the results are valid across different groups of people and are actually able to be used to make an assumption about a population.

The purpose of the current study was to explore the relationships that exist between health, physical activity, and participant demographics using recursive partitioning statistical methods while illustrating the benefits of using ensemble learning with recursive partitioning methods. Although ensemble learning with recursive partitioning methods appears to be beneficial for visualizing the relationships between physical activity and health, future work could explore additional analytic approaches. Additionally, while the SF-12 is a valid and reliable measure of health, additional instruments could be utilized as measures of health in future work. Instruments that specifically measure certain types of mental health would also be of interest in future studies.

Despite limitations, this study demonstrated that physical activity (especially in leisure) is a significant variable contributing to both physical and mental health. This study also demonstrated that recursive partitioning methods and ensemble learning with such methods can be an advantageous analytic approach when trying to capture the impact physical activity and demographics have on overall health.
References


