

Consent Form

Thank you for visiting!

This is a study to collect initial data for a project that aims to create a model of chronic pain in adolescents. We hope to use the data you and others provide to help develop a model that can aid in identifying more effective methods of treating adolescents in chronic pain,

What will you be asked to do?

On the intake questionnaire, you will be asked to enter your email (and phone number if you would like text reminders) as well as a few demographic questions. After that, you will be asked to provide some background on your pain history.

After the initial questionnaire, you will be asked to complete a brief survey every day for a period of 28 days. The questionnaire should only take approximately 5 minutes per day. The daily questionnaire asks questions about mood, pain, and functioning.

The whole study will occur over a 28 day period.

What are the risks and benefits of participating?

This study is classified as having 'minimal risk' – in other words, it is no more risky than things you might experience in everyday life.

You will receive \$7.50 for each week you remain in the study. Every week in which you complete a survey every day will earn you a \$5 bonus. You stand to make \$50 for completing 28 surveys.

Your end reward will be sent via Tango, where you have the option of having a check mailed to your address, certain gift cards, or a donation to charity.

Beyond the financial compensation, your data can help in the creation of models that will help identify the best courses of treatment for adolescents in chronic pain.

Is my information confidential?

Yes. Your ID codes are generated by Qualtrics and your emails/phone numbers will be stored on a secure server.

What will happen to my data?

At the conclusion of the study, all identifying information about you (email/cellphone number) will be destroyed. While your data will become part of a validation set for a model, you will have the option to submit or withhold data during the final survey.

Still interested? Thank you!

By clicking here you are affirming that (a) you are 18 years old or older and (b) you have read the study description above and (c) you are voluntarily participating in this study.

More Questions?

If you have questions about the study, please contact Max Kramer (mkramer@oberlin.edu or 773-318-5225)

If you have questions about your rights as a research participant, please contact Associate Dean Daphne John, Office of the Dean of Arts and Sciences, Cox 101 (djohn@oberlin.edu or 440-775-8410).

-
-
-

- ☐ YES! I affirm I am 18 years old or older, have read the study description, and am participating voluntarily.
- ☐ No, I don't want to participate.

Default Question Block

Welcome to the intake questionnaire! Over the next four weeks, you will be asked to take a short survey each day.

At the end of each week, if you complete a survey every day of that week you will get a \$5 bonus!

If you enter your cellphone number, you can get text reminders and take the survey straight from your phone! Otherwise, you will receive notifications via email.

Block 1

Enter your cellphone number (or leave blank to only get email alerts).

Block 2

Please enter your email address below.

Block 3

Demographic Information

How old are you?

What is your gender identity? You may select more than one option.

☐ Female

☐ Male

☐ Gender non-conforming

☐ Transgender

☐ Other: I identify as

☐ Prefer not to respond

How would you describe your ethnicity and/or racial background? You may select more than one option.

☐ American Indian or Alaska Native (e.g., Navajo Nation, Blackfeet Tribe, Mayan, Aztec, Nome Eskimo Community, etc.)

☐ Asian (e.g., Chinese, Filipino, Asian Indian, Vietnamese, Korean, etc.)

☐ Black (e.g. African American, Jamaican, Haitian, Nigerian, Ethiopian, Somalian, etc.)

☐ Latinx, Hispanic, or Spanish origin (e.g., Mexican, Mexican American, Puerto Rican, Cuban, Salvadorian, Dominican, Colombian, etc.)

☐ Middle Eastern or North African (e.g., Lebanese, Iranian, Egyptian, Syrian, Moroccan, Algerian, etc.)

☐ Native Hawaiian or Pacific Islander (e.g., Samoan, Chamorro, Tongan, Fijian, Marshallese, etc.)

☐ White (e.g., German, Irish, English, Italian, Polish, French, etc.)

☐ Other: I identify as

☐ Prefer not to respond

Pain History

In the past month, how would you rate your pain on average?

No pain

012345678910

Worst imaginable pain

Last semester, about how many times did you experience:

	Never	Rarely	Once or Twice a Month	Almost Every Week	Almost Every Day	I choose not to answer this question
Mild Pain	<div></div>	<div></div>	<div></div>	<div></div>	<div></div>	<div></div>
Moderate Pain	<div></div>	<div></div>	<div></div>	<div></div>	<div></div>	<div></div>
Severe Pain	<div></div>	<div></div>	<div></div>	<div></div>	<div></div>	<div></div>

How often did you experience pain due to behaviors like engaging in sports, playing an instrument, hangovers, or another specific activity you enjoyed?

	Never	Rarely	Once or Twice a Month	Almost Every Week	Almost Every Day	I choose not to answer this question
Mild Pain	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Moderate Pain	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Severe Pain	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

How often did you experience pain due to a chronic condition such as migraines, fibromyalgia, PID, arthritis, a back injury, or joint inflammation?

	Never	Rarely	Once or Twice a Month	Almost Every Week	Almost Every Day	I choose not to answer this question
Migranes	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Fibromyalgia	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
PID	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Arthiris	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Back Injury	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Joint Inflammation	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

Please respond to each question or statement by marking one box per row.

	Without any difficulty	With a little difficulty	With some difficulty	With much difficulty	Unable to do	I choose not to answer this question
Are you able to do chores such as vacuuming or yard work?	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Are you able to go up and down stairs at a normal pace?	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Are you able to go for a walk of at least 15 minutes?	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Are you able to run errands and shop?	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

Please respond to each question or statement by marking one box per row.

	Never	Rarely	Sometimes	Usually	Always	I choose not to answer this question
I have trouble doing all of my regular leisure activities with others	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
I have trouble doing all of the family activities that I want to do	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
I have trouble doing all of my usual work (include work at home)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

	Never	Rarely	Sometimes	Usually	Always	I choose not to answer this question
I have trouble doing all of the activities with friends that I want to do	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

What activity caused you pain most frequently?

In the past month...

	Very poor	Poor	Fair	Good	Very good	I choose not to answer this question
My sleep quality was	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

Please respond to each question or statement by marking one box per row.

In the past month...

	Not at all	A little Bit	Somewhat	Quite a bit	Very much	I choose not to answer this question
I feel fatigued	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
I have trouble starting things because I am tired	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
How run-down did you feel on average?	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
How fatigued were you on average?	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

Please respond to each question or statement by marking one box per row.

In the past month...

	Never	Rarely	Sometimes	Often	Always	I choose not to answer this question
I felt fearful	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
I found it hard to focus on anything other than my anxiety	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
My worries overwhelmed me	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
I felt uneasy	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

Please respond to each question or statement by marking one box per row.

In the past month...

	Never	Rarely	Sometimes	Often	Always	I choose not to answer this question
I felt worthless	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
I felt helpless	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
I felt depressed	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
I felt hopeless	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

Please respond to each question or statement by marking one box per row.

In the past month...

	Not at all	A little bit	Somewhat	Quite a bit	Very much	I choose not to answer this question
My sleep was refreshing	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
I had a problem with my sleep	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
I had difficulty falling asleep	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

Please respond to each question or statement by marking one box per row.

In the past month...

	Not at all	A little bit	Somewhat	Quite a bit	Very much	I choose not to answer this question
How much did pain interfere with your day to day activities?	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
How much did pain interfere with work around home?	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
How much did pain interfere with your ability to participate in social activities?	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
How much did pain interfere with your household chores?	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

Please respond to each question or statement by marking one box per row.

	Strongly disagree	Somewhat disagree	Neither agree nor disagree	Somewhat agree	Strongly agree	I choose not to answer this question
No one's been able to tell me exactly why I'm in pain.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
My pain is confusing me.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
I don't know enough about my pain.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

	Strongly disagree	Somewhat disagree	Neither agree nor disagree	Somewhat agree	Strongly agree	I choose not to answer this question
I can't figure out why I am in pain.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

In general, how would you rate your ability to tolerate moderate pain?

Worse than most people

012345678910

☐☐☐☐☐☐☐☐☐☐☐

Better than most people

Default Question Block**Welcome Back!**

This is the daily questionnaire for 4/27/2020 !

You will first be asked to enter the ID you were provided by email, then you will be asked to complete a short survey about your mood, pain intensity, and other factors.

Voluntary Participation.

Your participation is completely voluntary and you are free to withdraw at any time. If you choose to participate, you may skip any other questions without penalty. In addition, on the last day of the survey, we will ask you to confirm that we may use your data.

Questions?

If you have questions about the study, please contact Nancy Darling (ndarling@oberlin.edu or 440-775-8363) Max Kramer (mkramer@oberlin.edu or 773-318-5225).

If you have questions about your rights as a research participant, please contact Associate Dean Daphne John, Office of the Dean of Arts and Sciences, Cox 101 (ocirb@oberlin.edu. or 440-775-8410).

Block 9

Please Enter your ID code from the Intake Survey

Mood Scale

Indicate if you felt each of the following since your last survey

	No	Yes
Interested	<input type="radio"/>	<input type="radio"/>
Distressed	<input type="radio"/>	<input type="radio"/>
Excited	<input type="radio"/>	<input type="radio"/>
Upset	<input type="radio"/>	<input type="radio"/>
Strong	<input type="radio"/>	<input type="radio"/>
Guilty	<input type="radio"/>	<input type="radio"/>
Scared	<input type="radio"/>	<input type="radio"/>
Hostile	<input type="radio"/>	<input type="radio"/>
Enthusiastic	<input type="radio"/>	<input type="radio"/>
Proud	<input type="radio"/>	<input type="radio"/>

	No	Yes
Irritable	<input type="radio"/>	<input type="radio"/>
Alert	<input type="radio"/>	<input type="radio"/>
Ashamed	<input type="radio"/>	<input type="radio"/>
Inspired	<input type="radio"/>	<input type="radio"/>
Nervous	<input type="radio"/>	<input type="radio"/>
Determined	<input type="radio"/>	<input type="radio"/>
Attentive	<input type="radio"/>	<input type="radio"/>
Jittery	<input type="radio"/>	<input type="radio"/>
Active	<input type="radio"/>	<input type="radio"/>
Afraid	<input type="radio"/>	<input type="radio"/>

Morning routine

Indicate if you did any of the following since your last survey

	No	Yes
Got out of bed	<input type="radio"/>	<input type="radio"/>
Got dressed	<input type="radio"/>	<input type="radio"/>
Left for school/class	<input type="radio"/>	<input type="radio"/>
Ran errands	<input type="radio"/>	<input type="radio"/>
Prepared food	<input type="radio"/>	<input type="radio"/>
Brushed teeth	<input type="radio"/>	<input type="radio"/>
Showered	<input type="radio"/>	<input type="radio"/>
Brushed hair	<input type="radio"/>	<input type="radio"/>
Took medications	<input type="radio"/>	<input type="radio"/>
Washed up	<input type="radio"/>	<input type="radio"/>

Activity checklist

Indicate if you did any of the following since your last survey

	No	Yes
Read a book	<input type="radio"/>	<input type="radio"/>
Created art	<input type="radio"/>	<input type="radio"/>
Played a game	<input type="radio"/>	<input type="radio"/>
Watched TV	<input type="radio"/>	<input type="radio"/>
Wrote in a journal	<input type="radio"/>	<input type="radio"/>
Played music	<input type="radio"/>	<input type="radio"/>
Played with pets	<input type="radio"/>	<input type="radio"/>

	No	Yes
Enjoyed my hobbies	<input type="radio"/>	<input type="radio"/>
Played Sports	<input type="radio"/>	<input type="radio"/>

Helping others

Indicate if you did any of the following since your last survey

	No	Yes
Was supportive to a friend	<input type="radio"/>	<input type="radio"/>
Did homework	<input type="radio"/>	<input type="radio"/>
Helped out family/friends	<input type="radio"/>	<input type="radio"/>
Volunteered	<input type="radio"/>	<input type="radio"/>
Fed pets	<input type="radio"/>	<input type="radio"/>

Diet/Hydration

Indicate if you did any of the following since your last survey

	No	Yes
Ate healthy	<input type="radio"/>	<input type="radio"/>
Followed diet	<input type="radio"/>	<input type="radio"/>
Stayed hydrated	<input type="radio"/>	<input type="radio"/>

Exercise

Indicate if you did any of the following since your last survey

	No	Yes
Exercised	<input type="radio"/>	<input type="radio"/>
Took a stroll	<input type="radio"/>	<input type="radio"/>
Got my heart rate up	<input type="radio"/>	<input type="radio"/>

Mindfulness/Stress Reductions

Indicate if you did any of the following since your last survey

	No	Yes
Was thankful	<input type="radio"/>	<input type="radio"/>

	No	Yes
Mindfulness	<input type="radio"/>	<input type="radio"/>
Meditated	<input type="radio"/>	<input type="radio"/>
Yoga/Tai Chi	<input type="radio"/>	<input type="radio"/>
Prayed	<input type="radio"/>	<input type="radio"/>
Biofeedback	<input type="radio"/>	<input type="radio"/>

Sleep

Indicate if you did any of the following since your last survey

	No	Yes
Went to bed on time	<input type="radio"/>	<input type="radio"/>
Rested during the day	<input type="radio"/>	<input type="radio"/>
Got up on time	<input type="radio"/>	<input type="radio"/>
Got to sleep	<input type="radio"/>	<input type="radio"/>

Pain

How is your pain today?

Contact

More Questions?

If you have questions about the study, please contact Nancy Darling (ndarling@oberlin.edu or 440-775-8363) Max Kramer (mkramer@oberlin.edu or 773-318-5225).

If you have questions about your rights as a research participant, please contact Associate Dean Daphne John, Office of the Dean of Arts and Sciences, Cox 101 (djohn@oberlin.edu or 440-775-8410).

Closing

MK Honors Thesis Dataset Generation

1.0 Setup

In order to synthesize an artificial dataset, we must first load our seed data into R.

```
require(readxl)
```

```
## Loading required package: readxl
```

```
library(readxl)  
library(descr)
```

```
## Warning: package 'descr' was built under R version 3.6.3
```

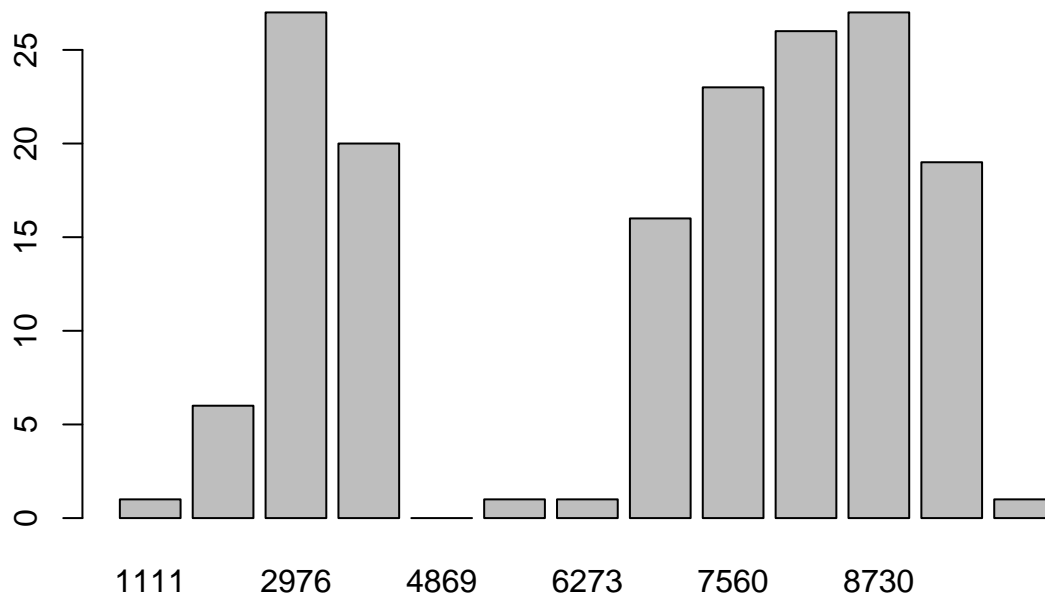
```
Intake <- read_excel("Intake+Questionnaire+for+MK+Honors+Thesis_December+3,+2019_12.02.xlsx")  
Daily <- read_excel("Daily+Questionnaire+for+MK+Honors+Thesis_December+3,+2019_11.54.xlsx")
```

The data are read in as character vectors by default rather than categorical factors. We must convert the data typing of each variable to factor before proceeding.

```
Intake2 <- as.data.frame(unclass(Intake))  
Daily2 <- as.data.frame(unclass(Daily))
```

Finally, labels and metadata should be stripped from the dataframe. For convenience, we also relabel the Intake2 and Daily2 datasets to i and d respectively.

```
i <- Intake2[-1,] # renaming and parsing factor labels  
d <- Daily2[-1,] # renaming and parsing factor labels  
  
Daily2$Q15[Daily2$Q15 == 4869] <- 4689  
  
freqlist <- freq(Daily2$Q15)
```

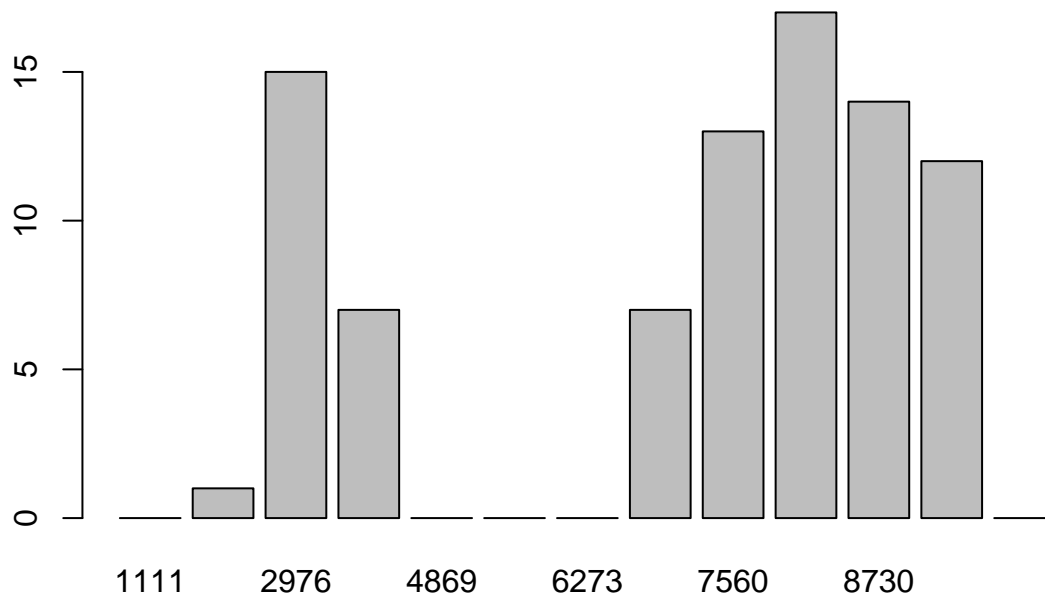


```
f <- which(freqlist[,1] > 1)
f <- names(f)[1:(length(f)-2)]

d <- Daily2[as.character(Daily2$Q15) %in% f,]
d <- d[d$Finished == "True",]

d_drop <- d[!is.na(d$Q17),] # removing NA values

freqlist <- freq(d_drop$Q15)
```



```
f <- which(freqlist[,1] > 1)
f <- names(f)[1:(length(f)-1)]
d_drop <- d_drop[as.character(d_drop$Q15) %in% f,]
d_main <- d_drop[,19:79] # parse metadata
d_main <- na.omit(d_main) ## THIS LINE REMOVES ALL INCOMPLETE CASES ##
d_label <- d_main[,61] # label vector

l_counts <- matrix(0,ncol=5,nrow=5)
for (i in f){
  d_sub <- d_drop[d_drop$Q15 == i,]
  for (row in 2:nrow(d_sub)){
    pastL <- d_sub$Q17[row-1]
    L <- d_sub$Q17[row]
    l_counts[pastL,L] <- l_counts[pastL,L] + 1
    #print(paste(i,row,pastL,L,nrow(d_sub)))
  }
}
roulette_l <- prop.table(l_counts,1)
```

1.1 Artificial Dataset Synthesis

The artificial dataset will be generated using joint distribution roulette wheel sampling.

Variables on daily: 19:78

MOOD: 1 - 20 DAILY FXN: 21 - 30 ACTIVITY FXN: 31 - 39 RELIABILITY FXN: 40 - 44 DIET/HYD
FXN: 45 - 47 EXERCISE FXN: 48 - 50 MINDFULNESS/BIOFEED FXN: 51 - 56 SLEEP FXN: 57 - 60
PAIN LVL: 61

```
sampleLabel <- function(day,pastL,d_label,roulette_l){
  if (day == 1){
    rand <- sample(1:length(d_label),1) # randomly sample a label from labels
    L <- d_label[rand] # sample that entry of d_label
    return(L)
  }
  else{
    probs <- roulette_l[pastL,]
    rand <- runif(1)
    for (L in 1:length(probs)){
      if (probs[L] >= rand){
        return(L)
      }
      else{
        rand <- rand - probs[L]
      }
    }
    return(length(probs))
  }
}

createDay <- function(L, d_main, scales, nscales) {
  inst <- c(L) # create new artificial instance with label L
  d_subset <- d_main[d_main$Q17 == L,] # only select cases with label
  for (s_i in 1:nscales){ # instance generation
    start = scales[2*s_i-1]
    end = scales[2*s_i]
    rand <- sample(1:nrow(d_subset),1)
    for(att in start:end){
      inst <- c(inst,as.integer(d_subset[rand,att]) - 2)
    }
  }
  return(inst)
}

rouletteWheelSampling <- function(n)
{
  #Variables
  art <- data.frame()
  scales <- c(c(1,20),c(21,30),c(31,39),c(40,44),c(45,47),c(48,50),c(51,56),c(57,60))
  nscales <- length(scales)/2

  paste('Now Generating',n,'samples',sep=" ")

  for (person in 1:n) {
    L <- 1 # arbitrary
    for (day in 1:28) {
```

```

    L <- sampleLabel(day,L,d_label,roulette_1)
    meas <- createDay(L, d_main, scales, nscales)
    inst <- c(person, day, meas)
    art <- rbind(art, inst)
  }
}
my_names <- names(Daily)

for (i in 1:ncol(Daily)) {
  my_names[i] <- paste0(my_names[i], ": ", Daily[1, i])
}
my_names <- my_names[19:79]

names(art) <- c("participant","day","label",my_names[1:60])
return(art)
}

art <- rouletteWheelSampling(100)
write.csv(art,'HonorsData.csv',row.names = FALSE)

```


"""
*Developing a Machine Learning Algorithm to Predict Daily Functioning in a Population
of Adolescents Living With Chronic Pain*

Psychology Honors Thesis
Max Kramer, Oberlin College Class of 2020

This code is designed to be run after the R markdown file generates the dataset
"""

```
#  
# EXTERNAL LIBRARY IMPORTATION  
#
```

```
import sys  
import csv  
from math import sqrt  
import pandas as pd  
import numpy as np  
import matplotlib.pyplot  
from sklearn import preprocessing, metrics, model_selection
```

```
#  
# MODELS  
#
```

```
from sklearn.naive_bayes import GaussianNB  
from sklearn.neural_network import MLPClassifier  
from sklearn.neural_network import MLPRegressor  
from sklearn.tree import DecisionTreeClassifier  
from sklearn.tree import DecisionTreeRegressor  
from sklearn.ensemble import RandomForestClassifier  
from sklearn.ensemble import RandomForestRegressor  
from sklearn.linear_model import LogisticRegression  
from sklearn.linear_model import LinearRegression
```

```
#  
# Helper Functions  
#
```

```
def readin(path, verbose=False): # takes verbose flag  
    dataset = pd.read_csv(path) # read csv from supplied filepath  
    if verbose: # for diagnostics  
        print('dataset contains {} instances and {} attributes'.format(dataset.shape[0], dataset.shape[1] - 1))  
    return dataset  
  
def instanceFormat(dataset,k,n):  
    X = []  
    y = []  
    participants = np.arange(1,101)  
    for participant in participants:  
        for label_day in range(k+n,29):  
            label = (dataset.label)[(dataset.participant == participant) & (dataset.day == label_day)].to_numpy()[0]  
            y.append(label)  
            inst = []  
            for day_offset in range(k-1,-1,-1):  
                attribute_day = label_day - n - day_offset  
                original_row = dataset[(dataset.participant == participant) & (dataset.day == attribute_day)].to_numpy()[0].tolist()  
                if day_offset == 0 and n == 0:  
                    inst.extend(original_row[3:])  
                else:  
                    inst.extend(original_row[2:])  
            X.append(inst)  
    return np.asarray(X),np.asarray(y)
```

```

def split(X, y, train_percent, seed, verbose=False): # Split dataset
    X_train, X_test, y_train, y_test = model_selection.train_test_split(X, y, train_size=train_percent,
                                                                    random_state=seed) # creates test and train set

    if verbose: # for diagnostics
        print('training set contains {} instances'.format(X_train.shape[0]))
        print('test set contains {} instances'.format(X_test.shape[0]))
        print('split complete')
    return X_test.shape[0], X_train, X_test, y_train, y_test

def ConfidenceInterval(acc, testset_size): # generate 95% CI with Bonferroni Correction for 4 comparisons per dataset
    CI = 2.39 * sqrt((acc * (1 - acc)) / testset_size)
    return CI

```

CLASSIFICATION MODELS

```

#
# ASSUMPTIONS TEST: Naive Bayes
#

```

```

def NaiveBayes(k,N,testset_size, dataset, X_train, X_test, y_train, y_test):
    clf = GaussianNB()
    clf.fit(X_train, y_train)
    acc = clf.score(X_test, y_test)
    predicted = clf.predict(X_test)
    summary = metrics.classification_report(y_test, predicted)
    conmat = metrics.confusion_matrix(y_test, predicted)
    with open('results_NB_{_}.csv'.format(k,N), mode='w') as csvout:
        writer = csv.writer(csvout, delimiter=',')
        writer.writerows(conmat)
    CI = ConfidenceInterval(acc, testset_size)
    print('Naive Bayes: {} accuracy 95% CI : [{}, {}]'.format('%0.3f' % acc, '%0.3f' % (acc - CI), '%0.3f' % (acc + CI)))
    print()
    print(summary)
    return acc

```

```

#
# BASELINE: Logistic Regression
#

```

```

def LR(k,N,testset_size, dataset, X_train, X_test, y_train, y_test,seed):
    clf = LogisticRegression(random_state=seed)
    clf.fit(X_train, y_train)
    acc = clf.score(X_test, y_test)
    predicted = clf.predict(X_test)
    summary = metrics.classification_report(y_test, predicted)
    conmat = metrics.confusion_matrix(y_test, predicted)
    with open('results_LR_{_}.csv'.format(k,N), mode='w') as csvout:
        writer = csv.writer(csvout, delimiter=',')
        writer.writerows(conmat)
    CI = ConfidenceInterval(acc, testset_size)
    print('LR: {} accuracy 95% CI : [{}, {}]'.format('%0.3f' % acc, '%0.3f' % (acc - CI), '%0.3f' % (acc + CI)))
    print()
    #debug = 1
    #if debug:
        #print(clf.intercept_)
        #print(clf.coef_[0].size)
        #print(clf.coef_)
        #print()
    print(summary)
    return acc

```

```

#
# Decision Tree
#

```

```

def DecisionTree(k,N,testset_size, dataset, X_train, X_test, y_train, y_test, seed):

```

```

clf = DecisionTreeClassifier(random_state=seed)
clf.fit(X_train, y_train) # fit model to data
acc = clf.score(X_test, y_test)
predicted = clf.predict(X_test)
summary = metrics.classification_report(y_test, predicted)
conmat = metrics.confusion_matrix(y_test, predicted)
with open('results_DT_{k}_{N}.csv'.format(k,N), mode='w') as csvout:
    writer = csv.writer(csvout, delimiter=',')
    writer.writerow(conmat)
CI = ConfidenceInterval(acc, testset_size)
print(
    'Decision Tree: {} accuracy 95% CI : [{}, {}]'.format('%0.3f' % acc, '%0.3f' % (acc - CI), '%0.3f' % (acc + CI)))
print()
print(summary)
return acc

```

```

#
# Random Forest
#

```

```

def RandomForest(k,N,testset_size, D, dataset, X_train, X_test, y_train, y_test, seed):
    clf = RandomForestClassifier(n_estimators=100, random_state=seed)
    clf.fit(X_train, y_train) # fit model to data
    acc = clf.score(X_test, y_test)
    predicted = clf.predict(X_test)
    summary = metrics.classification_report(y_test, predicted)
    conmat = metrics.confusion_matrix(y_test, predicted)
    with open('results_Forest_{k}_{N}.csv'.format(k,N), mode='w') as csvout:
        writer = csv.writer(csvout, delimiter=',')
        writer.writerow(conmat)
    CI = ConfidenceInterval(acc, testset_size)
    print(
        'Random Forest: {} accuracy 95% CI : [{}, {}]'.format('%0.3f' % acc, '%0.3f' % (acc - CI), '%0.3f' % (acc + CI)))
    if k == 1 and N == 0:
        importance = pd.DataFrame(clf.feature_importances_, index = D.columns[3:],
            columns=['importance']).sort_values('importance', ascending=False)
        #debug = 1
        #if debug:
            #print(importance)
        importance.to_csv('importance_RF.csv')
        print()
    print(summary)
    return acc

```

```

#
# Neural Network
#

```

```

def shallowNN(k,N,testset_size, dataset, X_train, X_test, y_train, y_test, seed):
    clf = MLPClassifier(hidden_layer_sizes=(150,150),solver='adam', max_iter=5000, learning_rate_init=0.001,
        random_state=seed)
    clf.fit(X_train, y_train)
    acc = clf.score(X_test, y_test)
    predicted = clf.predict(X_test)
    summary = metrics.classification_report(y_test, predicted)
    conmat = metrics.confusion_matrix(y_test, predicted)
    with open('results_NN_{k}_{N}.csv'.format(k,N), mode='w') as csvout:
        writer = csv.writer(csvout, delimiter=',')
        writer.writerow(conmat)
    CI = ConfidenceInterval(acc, testset_size)
    print('Shallow NN: {} accuracy 95% CI : [{}, {}]'.format('%0.3f' % acc, '%0.3f' % (acc - CI), '%0.3f' % (acc + CI)))
    print()
    print(summary)
    return acc

```

```

##### REGRESSION MODELS #####

```

```

#
# BASELINE: Linear Regression
#

def LinReg(testset_size, dataset, X_train, X_test, y_train, y_test):
    reg = LinearRegression()
    reg.fit(X_train, y_train)
    R2 = reg.score(X_test, y_test)
    predicted = reg.predict(X_test)
    MSE = metrics.mean_squared_error(y_test, predicted)
    #CI = ConfidenceInterval(acc, testset_size)
    print('Linear Regression: R^2 = {} MSE = {}'.format('%0.3f' % R2, '%0.3f' % MSE))
    print()
    return R2, MSE

#
# Decision Tree Regressor
#

def DecisionTreereg(testset_size, dataset, X_train, X_test, y_train, y_test, seed):
    reg = DecisionTreeRegressor(random_state=seed)
    reg.fit(X_train, y_train) # fit model to data
    R2 = reg.score(X_test, y_test)
    predicted = reg.predict(X_test)
    MSE = metrics.mean_squared_error(y_test, predicted)
    #CI = ConfidenceInterval(MSE, testset_size)
    print(
        'Decision Tree: R^2 = {} MSE = {}'.format('%0.3f' % R2, '%0.3f' % MSE))
    print()
    return R2, MSE

#
# Random Forest Regressor
#

def RandomForestreg(testset_size, dataset, X_train, X_test, y_train, y_test, seed):
    reg = RandomForestRegressor(n_estimators=100, random_state=seed)
    reg.fit(X_train, y_train) # fit model to data
    R2 = reg.score(X_test, y_test)
    predicted = reg.predict(X_test)
    MSE = metrics.mean_squared_error(y_test, predicted)
    #CI = ConfidenceInterval(acc, testset_size)
    print(
        'Random Forest: R^2 = {} MSE = {}'.format('%0.3f' % R2, '%0.3f' % MSE))
    print()
    return R2, MSE

#
# Neural Network Regressor
#

def shallowNNreg(testset_size, dataset, X_train, X_test, y_train, y_test, seed):
    reg = MLPRegressor(solver='adam', max_iter=1000, alpha=1e-3, random_state=seed)
    reg.fit(X_train, y_train)
    R2 = reg.score(X_test, y_test)
    predicted = reg.predict(X_test)
    MSE = metrics.mean_squared_error(y_test, predicted)
    #CI = ConfidenceInterval(acc, testset_size)
    print('Shallow NN: R^2 = {} MSE = {}'.format('%0.3f' % R2, '%0.3f' % MSE))
    print()
    return R2, MSE

```

```

##### MAIN #####

```

```

def main():

```

```

seed = int(sys.argv[1])
LR_table = np.zeros((7,8))
NB_table = np.zeros((7,8))
DT_table = np.zeros((7,8))
RF_table = np.zeros((7,8))
NN_table = np.zeros((7,8))
# Linreg_R2_table = np.zeros((7,8))
# DTreg_R2_table = np.zeros((7,8))
# RFreg_R2_table = np.zeros((7,8))
# NNreg_R2_table = np.zeros((7,8))
# Linreg_MSE_table = np.zeros((7,8))
# DTreg_MSE_table = np.zeros((7,8))
# RFreg_MSE_table = np.zeros((7,8))
# NNreg_MSE_table = np.zeros((7,8))
Dataset = 'Honors'
dataset = readin('./HonorsData.csv')
for k in range(1,8):
    for N in range(8):
        X , y = instanceFormat(dataset,k,N)
        testset_size, X_train, X_test, y_train, y_test = split(X,y,0.85,seed)
        # print('K is {}, N is {}'.format(k,N))
        # R2_LR, MSE_LR = LinReg(testset_size, dataset, X_train, X_test, y_train, y_test)
        # Linreg_R2_table[k-1,N] = R2_LR
        # Linreg_MSE_table[k-1,N] = MSE_LR
        # print('K is {}, N is {}'.format(k,N))
        # R2_DT, MSE_DT = DecisionTreereg(testset_size, Dataset, X_train, X_test, y_train, y_test, seed)
        # DTreg_R2_table[k-1,N] = R2_DT
        # DTreg_MSE_table[k-1,N] = MSE_DT
        # print('K is {}, N is {}'.format(k,N))
        # R2_RF, MSE_RF = RandomForestreg(testset_size, dataset, X_train, X_test, y_train, y_test, seed)
        # RFreg_R2_table[k-1,N] = R2_RF
        # RFreg_MSE_table[k-1,N] = MSE_RF
        # print('K is {}, N is {}'.format(k,N))
        # R2_NN, MSE_NN = shallowNNreg(testset_size, dataset, X_train, X_test, y_train, y_test, seed)
        # NNreg_R2_table[k-1,N] = R2_NN
        # NNreg_MSE_table[k-1,N] = MSE_NN

        print('K is {}, N is {}'.format(k,N))
        LR_acc = LR(k,N,testset_size, Dataset, X_train, X_test, y_train, y_test,seed)
        LR_table[k-1,N] = LR_acc
        print('K is {}, N is {}'.format(k,N))
        NB_acc = NaiveBayes(k,N,testset_size, Dataset, X_train, X_test, y_train, y_test)
        NB_table[k-1,N] = NB_acc
        print('K is {}, N is {}'.format(k,N))
        DT_acc = DecisionTree(k,N,testset_size, Dataset, X_train, X_test, y_train, y_test, seed)
        DT_table[k-1,N] = DT_acc
        print('K is {}, N is {}'.format(k,N))
        RF_acc = RandomForest(k,N,testset_size, dataset, Dataset, X_train, X_test, y_train, y_test, seed)
        RF_table[k-1,N] = RF_acc
        print('K is {}, N is {}'.format(k,N))
        NN_acc = shallowNN(k,N,testset_size, Dataset, X_train, X_test, y_train, y_test, seed)
        NN_table[k-1,N] = NN_acc
        print("-----")

# np.savetxt('Linreg_R2.csv',Linreg_R2_table,delimiter=',',fmt='%f')
# np.savetxt('Linreg_MSE.csv',Linreg_MSE_table,delimiter=',',fmt='%f')
# np.savetxt('DTreg_R2.csv',DTreg_R2_table,delimiter=',',fmt='%f')
# np.savetxt('DTreg_MSE.csv',DTreg_MSE_table,delimiter=',',fmt='%f')
# np.savetxt('RFreg_R2.csv',RFreg_R2_table,delimiter=',',fmt='%f')
# np.savetxt('RFreg_MSE.csv',RFreg_MSE_table,delimiter=',',fmt='%f')
# np.savetxt('NNreg_R2.csv',NNreg_R2_table,delimiter=',',fmt='%f')
# np.savetxt('NNreg_MSE.csv',NNreg_MSE_table,delimiter=',',fmt='%f')

np.savetxt('LR.csv',LR_table,delimiter=',',fmt='%f')
np.savetxt('NB.csv',NB_table,delimiter=',',fmt='%f')
np.savetxt('DT.csv',DT_table,delimiter=',',fmt='%f')

```

```
np.savetxt('RF.csv',RF_table,delimiter=',',fmt='%f')
np.savetxt('NN.csv',NN_table,delimiter=',',fmt='%f')
```

```
if __name__ == '__main__':
    main()
```