Interactive Visualization on the CNS HIV Anti-Retroviral Therapy Effects Research

Kylie Schiermann* Xiaoyue Cheng* Howard Fox† Steven Totusek†
Abigail Heithoff† Allison Dye† Mahbubul Majumder*

Abstract
In collaboration with the University of Nebraska Medical Center department of pharmacology and experimental neuroscience and the National NeuroAIDS Tissue Consortium - Data Coordinating Center, we develop a data visualization application designed to enable users of the CNS HIV Anti-Retroviral Therapy Effects Research web portal the ability to graphically manipulate and explore clinical and experimental data collected and hosted therein. We utilize open-sourced statistical software, namely the Shiny package in R, to create an interactive dashboard that allows users to generate plots and tables based on the user's selection criteria. This application can be used to subset the data into a group of interest and to explore and uncover potentially meaningful relationships of variables within the data.

Key Words: Data visualization, interactive dashboard, HIV Anti-Retroviral Therapy

1. Introduction

The National NeuroAIDS Tissue Consortium - Data Coordinating Center (NNTC-DCC) web page[1] contains information related to the CNS HIV Anti-Retroviral Therapy Effects Research (CHARTER), which began in 2004. On this website, visitors can utilize the query tool[2] to download data from the CHARTER study. Users of the website can select certain participant demographics and download a CSV file to their machine.

The limitation of the query tool is that it leaves the user to perform all data exploration, visualization, and analysis through his or her own means. The University of Nebraska Medical Center (UNMC) and NNTC-DCC wanted to provide users of the query tool with an application that gives them more insight into the data. The solution for this problem is to create a visualization tool that gives users the ability to explore the data through graphs and tables that are dynamically produced based on user-selected criteria.

In this project, we conducted exploratory data analysis on the CHARTER data, and used modern statistical software to design an interactive dashboard. The dashboard can satisfy the primary needs of data selection, exploration and visualization and also provides several basic statistical methods for variable selection and clustering.

Section 2 introduces the CHARTER data and discusses some key variables. Section 3 gives an introduction on the statistical methods and software used. Section 4 describes the design and functionality of the dashboard. Section 5 provides some

*Department of Mathematics, University of Nebraska Omaha, 6001 Dodge Street, Omaha, NE 68182
†University of Nebraska Medical Center, Durham Research Center 3008, 985800 Nebraska Medical Center, Omaha, NE 68198

[1]https://neuroaids-dcc.unmc.edu/Home
[2]https://neuroaids-dcc.unmc.edu/CharterQTools

examples of how the dashboard can assist data analysis. Section 6 summarizes the contribution and discusses the future work.

2. Data

The data set from the CHARTER study contains 6,383 observations and 153 variables. Each subject in the study was given a UNMC identification number (UNMCID), which uniquely identifies him or her as a participant in both CHARTER and NNTC. There are several variables describing participant visits throughout the study, including date, visit number (visitno), enrollment status (CEN), and visit sub-study classification (visubstudy). The visit sub-study classification details which longitudinal studies and cross-sectional studies the participant was enrolled in. The sub-studies were: Acute and Early HIV Infection, Anti-Retroviral (ARV), Imaging, Metabolic Complications, Peripheral Neuropathy, and Viral Genetics. Participant demographics such as age (curage), gender, race (raceth), and education (educat) are some of the categorical variables that are used to group participants into a subset of interest. As a measure of a participant's overall impairment, there are two primary variables that are used. One of these variables is a binary variable, average impairment (AVGIMP), and the participant is classified as either "impaired" or "unimpaired". The other is a numerical variable, average t-score (cavrgts), and is calculated by taking the average t-score across seven different domains, including abstract and executive functioning, information processing, attention and working memory, learning, memory, verbal fluency, and motor. The Frascati HAND diagnosis rating (HAND), which stands for HIV-associated neurocognitive disorder, gives information about the level of neurocognitive impairment a participant has. The levels of HAND ratings are: NP-Normal, ANI, MND, HAD, NPI-O. There are a few key variables that give information about the types of anti-retroviral drugs the participant is taking or has taken in the past. Current regimen (currentregimen) is an aggregated variable that lists all anti-retroviral drugs (ARVs) that a participant is currently taking. Some participants are on highly active anti-retroviral therapy drugs (HAART), and the variable HAART classifies a participant as being ARV naive, non-HAART, HAART, or having no current ARVs. Other important variables are totalCPE1, csfvl, and bloodvl. totalCPE1 gives the total central nervous system (CNS) penetration effectiveness score for all current ARVs. csfvl gives the total central spinal fluid (CSF) viral load in IU/mL, and bloodvl gives the plasma viral load in IU/mL.

3. Methods

3.1 R, shiny, and shinydashboard

R (R Core Team, 2018) is an open-sourced statistical software that has capabilities for data manipulation, calculation, and graphical display. R comes with several base packages; however, one may download and install additional packages as necessary. These packages have been written by R users and statisticians alike to perform various calculations and provide easy to implement functions. In this project, shiny (Chang et al., 2018) and shinydashboard (Chang and Borges Ribeiro, 2018) are two of the primary packages that are utilized to create a polished, dynamic interface for users to explore the CHARTER data.

The shiny package (Chang et al., 2018) in R is designed to make building interactive web applications simple and attractive. Shiny applications have two major

components. The first is the user interface (UI) and the second is the server script. Together, these two components contain everything that is needed to support the web application.

In addition to shiny, shinydashboard (Chang and Borges Ribeiro, 2018) is a package that contains several functionalities for creating a "dashboard" style web application. Some of the features it includes are menu bars, side bars, tabs, and boxes. These features work along with the basic shiny functions and are incredibly convenient for organizing the dashboard layout.

3.2 Plotly and ggplot2

Two key data visualization tools for the project are plotly (Sievert, 2018) and ggplot2 (Wickham, 2016). Plotly is a data analytics and visualization tool produced by a Canadian tech company. There are graphing libraries available for a variety of programs, including R. The plotly package in R contains functions that smoothly implement beautiful interactive graphics into the dashboard environment. In addition, the ggplot2 package, which features a wide variety of options for creating static plots, can be implemented along with plotly. The ggplotly function from the plotly package brings together the features of ggplot2 and plotly, which results in elegant, highly customizable, reactive graphics.

3.3 k-means Clustering

k-means (Lloyd, 1982 and James et al., 2013) is a clustering algorithm that partitions the data into k distinct clusters, or groups. The number of clusters is predetermined before applying the algorithm. The algorithm assigns each data point to exactly one cluster. For effective k-means clustering, the within-cluster variation should be minimized. The squared Euclidean distance is the most commonly used choice for minimizing the within-cluster variation. Mathematically, we may define this problem as follows, where $|C_k|$ is the number of observations in cluster k.

$$\min_{c_1,\ldots,c_K} \left\{ \sum_{k=1}^{K} \frac{1}{|C_k|} \sum_{i,i' \in c_k} \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2 \right\}$$

A random cluster is initially assigned to each observation. Then the cluster centroid is calculated. The cluster centroid is the vector containing the mean value of each of p variables. Observations are assigned to the cluster whose centroid is closest. The algorithm terminates when cluster assignments stop changing.

3.4 Linear Regression Model

A linear regression model is a statistical model in which a response variable is modeled as a linear function of some set of predictors, or explanatory variables (Fox, 2015). A linear regression equation has the form:

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k$$

Where Y is the response variable, the X_i's are the explanatory variables, and the β_i's are the regression coefficients.

There are several assumptions of multiple linear regression. For a linear regression model to be an appropriate model, it is best if these assumptions are

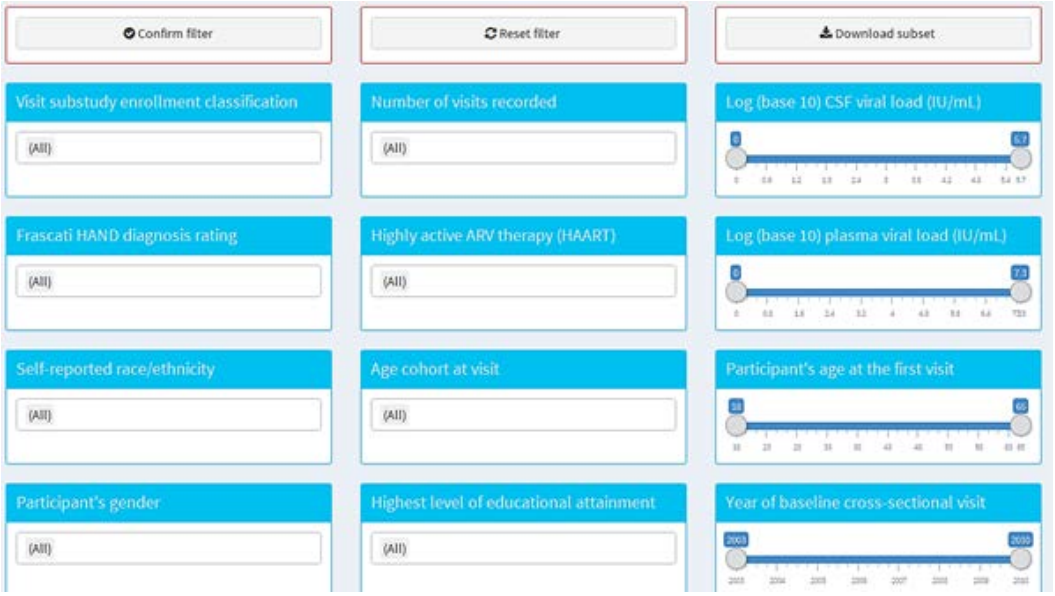


Figure 1: The initial page of the dashboard is the filter tab. There are twelve categories that can be used to subset the data into a group of interest. These filters apply to all other tabs in the dashboard.

not violated. The assumptions are linear relationship between the dependent and independent variables, no multicollinearity, normally distributed residuals, and homoscedasticity.

3.5 Random Forest

A random forest (Breiman, 2001 and James et al., 2013) is an ensemble method using decision tree as the basic model unit. The decision trees are created from a set of bootstrapped training samples. Each time a split in the decision tree is considered, a random sample of m (out of p) predictors is chosen as split candidates. A new random sample is selected for each split. Random forests can be used for prediction as well as assessing variable importance. When making predictions, the predicted value is the average over all trees in the random forest. Depending on whether the response variable is numerical or categorical, variable importance can be measured using the Gini Index or the mean square error.

4. Dashboard Functionality

The web application created for CHARTER is a dashboard with five main tabs. The five tabs were organized in an intuitive layout, each with a specific purpose.

4.1 Data Filter

The first tab is for filtering the data (Figure 1). Users can subset the data by selecting specific categories of interest. If the subset is too small, meaning it contains fewer than ten individuals, then the filter is automatically reset. Once an appropriate filter is selected, the user must confirm the filter. There is also an option to download the subset data. When the filter has been applied, all subsequent operations are applied to the filtered data.

4.1.1 Demographical and Longitudinal Summary

The second tab is the Overall Summary. This tab contains barcharts and histograms displaying participant demographics such as the year of the first visit, number of visits, age, race, gender, and education (Figure 2). The Overall Summary tab also contains a widget that allows users to select one variable of interest to compare the distributions at the first and last visits (Figure 3), and another tab to display the longitudinal time series plot for selected variable (Figure 4). To view an individual participant, the Individual Report tab will give summary information by a participant's UNMC ID. The user selects a participant ID and a variable of interest. The visualizations displayed are a time series, where the selected participant's information is highlighted (Figure 5), and a dot plot which shows the participant's ARV regimen and their corresponding Frascati HAND rating (Figure 6).

Figure 2: The Patient Demographics page of the Overall Summary tab shows histograms and bar charts of six different demographics.

4.2 Modeling

The Modeling tab allows users to create either a linear model or a random forest by selecting a response variable and a set of explanatory variables. The options for the response variable are the average t-score (cavrgts; numerical) and average impairment (AVGIMP; categorical). If the user selects a numerical response and a linear model, a basic linear model is applied. If the user selects a categorical response and a linear model, a logistic regression model is applied. A pairwise plot is displayed to show the relationship of each chosen explanatory variable to the chosen response variable (Figure 7). If the user selects a random forest model, the variable importance plot is displayed (Figure 8). Mean Decrease Gini Index is used to quantify importance for the categorical response model, whereas Percent Increase Mean Square Error is used to quantify importance for the numerical response model. In addition to the plots, the model outputs are displayed in a results tab (Figures 9 and 10). Here, users see the model summary from R.

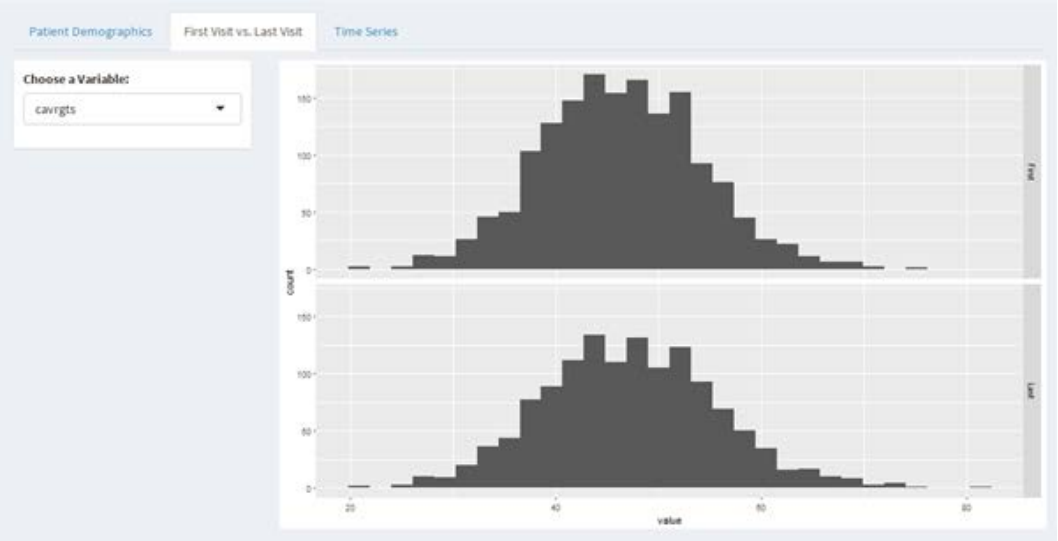


Figure 3: The First Visit vs. Last Visit page of the Overall Summary tab allows users to compare distributions of a specified variable at the first and last visits.

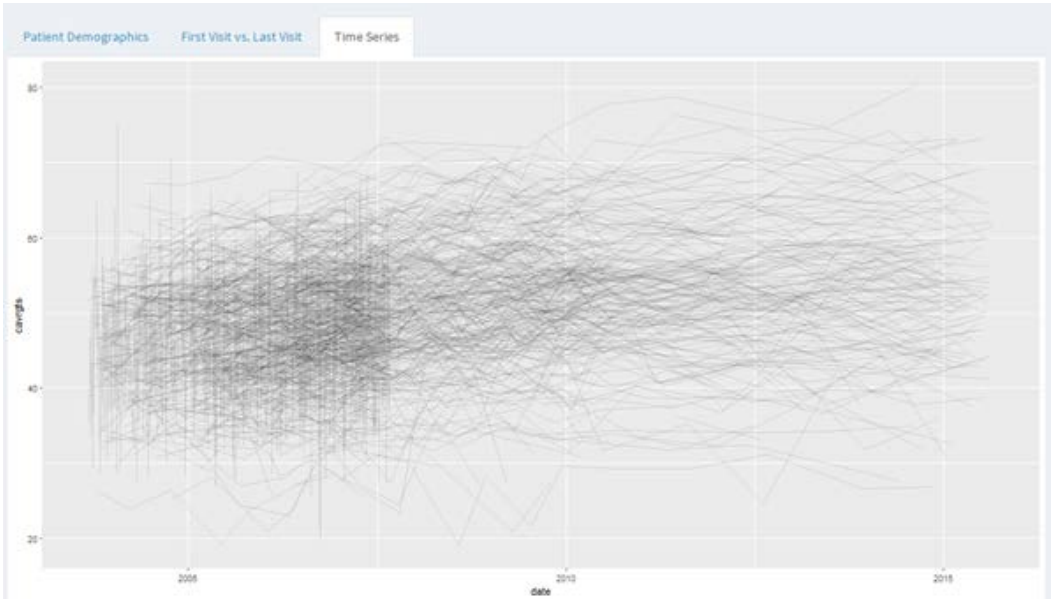


Figure 4: The Time Series page of the Overall Summary tab shows the longitudinal time series graph of the variable chosen in the First Visit vs. Last Visit page.

4.3 Clustering on Participants

Before clustering algorithms can be applied, the data needs to be reshaped. Some participants have multiple visits recorded in the CHARTER data. To ensure that each patient is assigned to only one cluster, the data must be reshaped to contain only one row for each patient in the study. To do this, the data was grouped by the UNMCID. For each patient, only the maximum number of visits was kept. The last recorded t-score (cavrgts) was kept for each patient. To summarize all average t-scores recorded for each patient, a linear regression of cavrgts on date was created, and the fitted slope was recorded as a new variable. For patients that had only one visit recorded, the slope was assumed to be zero. For the variables gender, raceth,

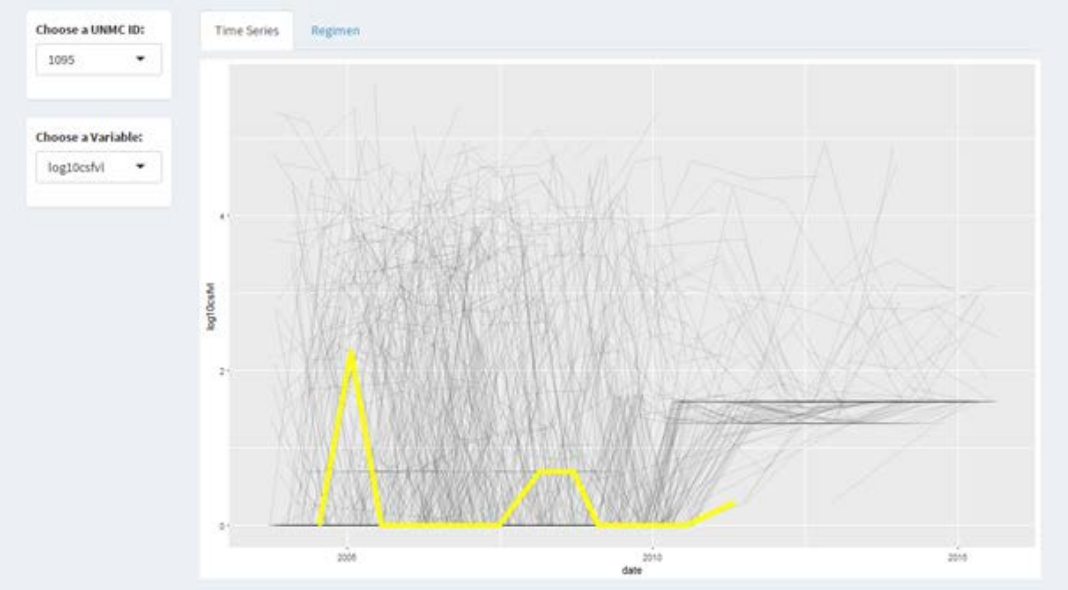


Figure 5: The Time Series page of the Individual Report tab produces a time series graph of the chosen variable and highlights the selected patient.

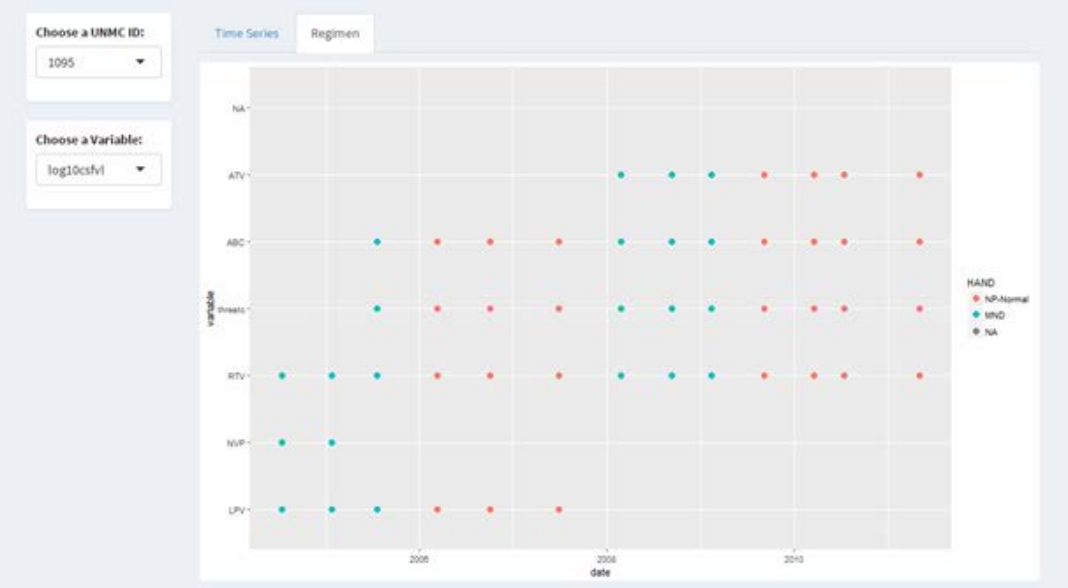


Figure 6: The Regimen page of the Individual Report tab shows the ARV's that the selected patient was taking and their corresponding Frascati HAND rating.

curage, and AVGIMP, the first recorded values for each patient were kept. All values were then standardized to avoid uneven weighting of variables during clustering.

The Clustering tab gives users the ability to apply k-means clustering to the data (Figure 11). The user can select a set of variables that will be used to cluster the data, as well as how many clusters, k, the k-means clustering algorithm should use. The resulting output is a time series display of the data, which has been color-coded by cluster assignment. Below the graph, a table summarizes the average value for each variable by cluster number. The options for clustering variables are: average t-score (cavrgts), slope, visit (visitno), gender, race (raceth), current age (curage),

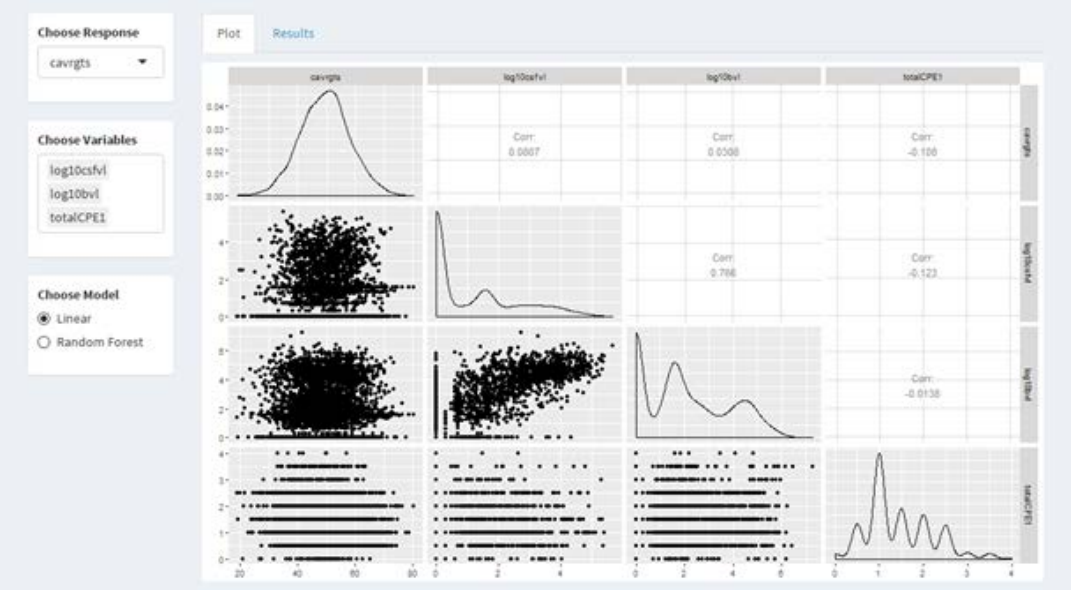


Figure 7: When a linear model is chosen, the Plot page of the Modeling tab displays a pairwise plot which shows the relationships of the selected variables.

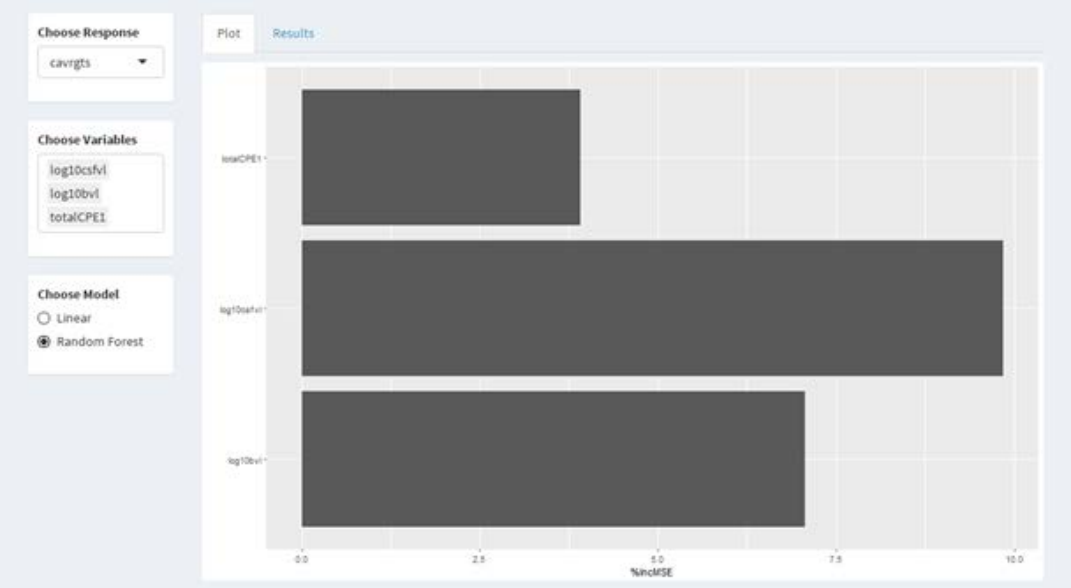


Figure 8: When a random forest is selected, the Plot page of the Modeling tab shows a bar chart of the variable importances.

and average impairment (AVGIMP).

5. Results

The dashboard's primary purpose is data exploration and visualization. As with any data set, analyzing the CHARTER data begins with exploratory data analysis. Using the app to explore the data visually can lead to the discovery of some interesting trends hidden in the data. Here, we discuss some example findings that were revealed by the data visualization tools within the dashboard.

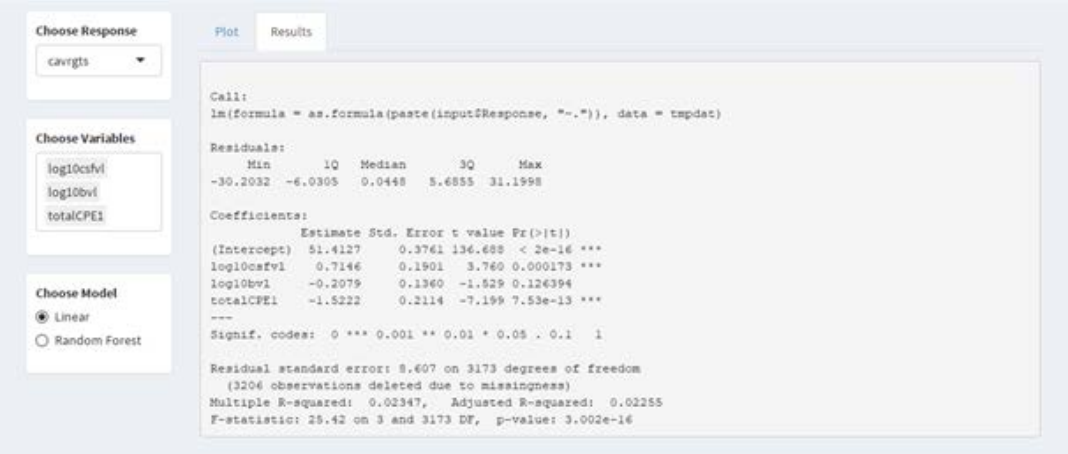


Figure 9: When a linear model is chosen, the Results page of the Modeling tab shows a summary of the model, including regression coefficients and p-values.

Figure 10: When a random forest is chosen, the Results page of the Modeling tab displays a summary of the model, including the number of trees and the mean of the squared residuals.

5.1 Individual Regimen Change over Time

In Figure 12, a participant with UNMCID 4032 has an ARV regimen which includes enfuvirtide, ritonavir, stavudine, and atazanavir and is diagnosed with a Frascati HAND rating of Mild Neurocognitive Disorder. In early 2007, the participant's regimen changes to include saquinavir mesylate, lamivudine, ritonavir, and atazanavir and is diagnosed as having Neuropsychological Impairment due to other causes. In late 2007, the participant's ARV regimen includes tenofovir disoproxil fumarate, raltegravir, emtricitabine, darunavir, and ritonavir; and the participant's diagnosis remains NPI-O. In 2008, the regimen does not change, but the Frascati HAND rating changes to NP-Normal. After 2009, the participant stops taking tenofovir disoproxil fumarate and emtricitabine and remains NP-Normal for the remainder of the study. The progression from impaired to normal could be due to the combination of ARV's in the participant's regimen and may warrant further investigation.

Figure 11: The Clustering tab displays a time series of the average t-score (cavrgts) where each line has been colored according to clustering assignment.

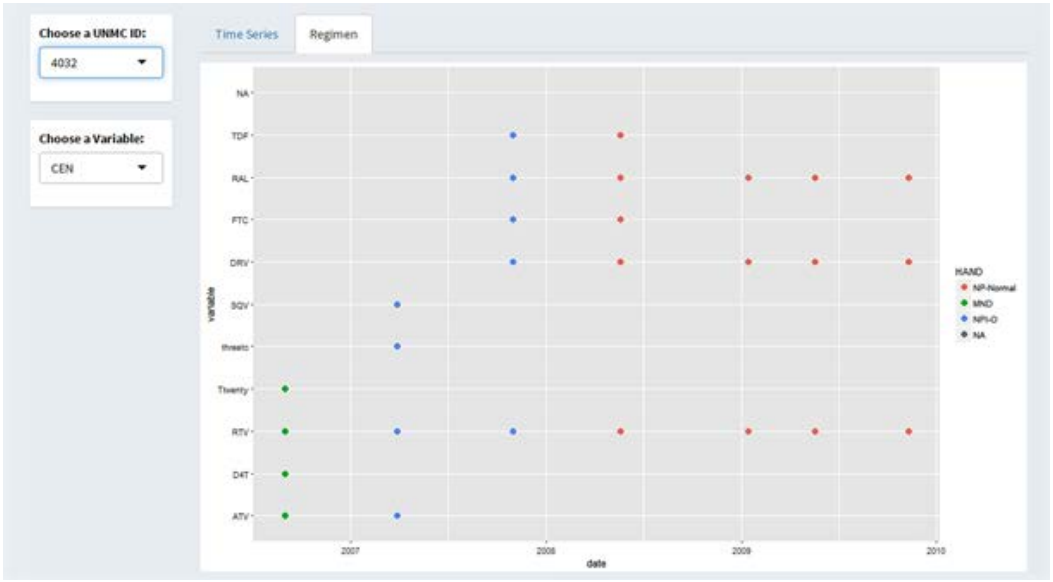
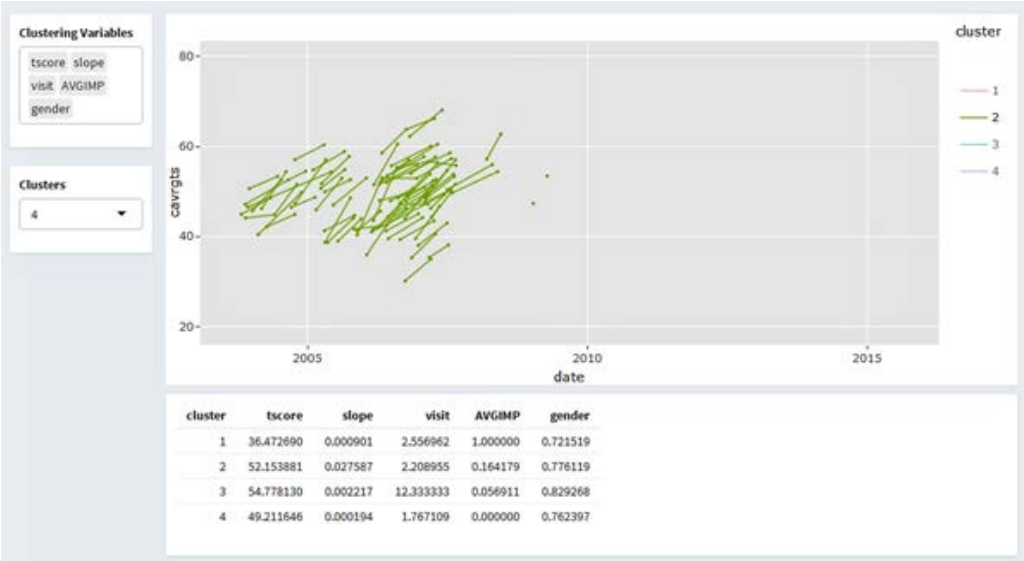


Figure 12: Regimen plot for UNMCID 4032.

5.2 Participant Groups

Using the Clustering tab in Figure 13, we choose to cluster the participants by t-score, slope, visit, average impairment, and gender. We also specify the number of clusters to be four. The resulting plot is a time series that has been color-coded by cluster number. Here, we have utilized one of the features of plotly to view only one of the clustering groups. Cluster number 2 is interesting because there appears to be an upward trend (positive slope) over time. The summary table below the graph also verifies that the average slope for cluster number 2 is positive, and larger than the averages of the other three clusters. It could be worthwhile to investigate the participants in cluster number 2 to see if there are any common attributes that

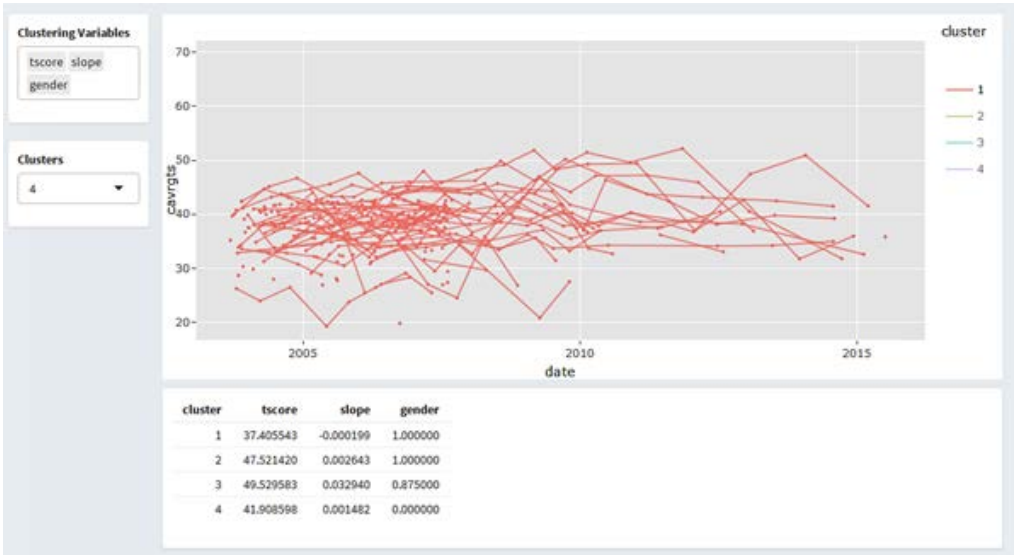
could be contributing to this trend.



cluster	tscore	slope	visit	AVGIMP	gender
1	36.472690	0.000901	2.556962	1.000000	0.721519
2	52.153881	0.027587	2.208955	0.164179	0.776119
3	54.778130	0.002217	12.333333	0.056911	0.829268
4	49.211646	0.000194	1.767109	0.000000	0.762397

Figure 13: A group of participants who had improvement on t-score.

Figure 14 shows another example of participant clustering. We again make use of the Clustering tab to uncover an interesting result. This time, we have applied a filter on the data with the Filter on Participants tab. Using the filter, we have chosen to select only participants with Frascati HAND diagnosis ratings of HAD, MND, and ANI (HIV Associated Dementia, Mild Neurocognitive Disorder, and Asymptomatic Neuropsychological Impairment, respectively). In the Clustering tab, we select t-score, slope, and gender as the clustering variables and 4 as the number of clusters to use for k-means. One of the resulting clusters, cluster number 1, contains only males and has a negative average slope. The average t-score for cluster number 1 is relatively low at 37.405543. It may appear that participants who are already impaired below a certain threshold do not show a significant amount of improvement over time.



cluster	tscore	slope	gender
1	37.405543	-0.000199	1.000000
2	47.521420	0.002643	1.000000
3	49.529583	0.032940	0.875000
4	41.908598	0.001482	0.000000

Figure 14: A group of participants who had decline on t-score.

As we illustrated with these examples, the dashboard is a useful tool for exploring the CHARTER data and can lead to some intriguing discoveries that may have otherwise gone unnoticed. Having visual evidence of trends or abnormal instances in the data allows us to ask questions that may not have been considered initially.

6. Summary

Data visualization is an important preliminary step in data analysis. The query tool available on the National NeuroAIDS Tissue Consortium - Data Coordinating Center website gives users access to the CHARTER data. However, prior to this project, there were no tools readily available for data visualization and exploratory data analysis. We utilized open-source statistical software including R, shiny, and plotly to create an application that enhances the experience for users of the query tool. The dashboard application has several features including tabs for filtering the data, clustering, modeling, and viewing overall or individual summaries. The visuals produced within the dashboard provide insight into the data, while allowing the user to select variables and demographics that are of interest.

6.1 Future Work

There are many possibilities for improving and enhancing the dashboard. We plan to implement all visualizations as plotly graphics. This would increase the interactivity of all graphics and provide a consistent aesthetic across all tabs in the dashboard. To improve the readability of the graphics, we would also aim to use variable names that are interpretable and concise, rather than the original names from the data set. Lastly, the Modeling tab has several opportunities for improvement. We would like to provide a cleaner output of the model results, increase user interaction, and select more appropriate models for the data.

Acknowledgments

We want to thank all the investigators and participants of the CHARTER study. NNTC is supported by the NIMH and NINDS grants: U24MH100931, U24MH100930, U24MH100928, U24MH100925.

REFERENCES

Breiman, L. (2001). "Random forests." Machine learning, 45(1), 5-32.

Chang, W. and Borges Ribeiro, B. (2018). "shinydashboard: Create Dashboards with 'Shiny'," R package version 0.7.0, URL: https://CRAN.R-project.org/package=shinydashboard

Chang, W., Cheng, J., Allaire, J., Xie, Y. and McPherson, J. (2018). "shiny: Web Application Framework for R," R package version 1.1.0, URL: https://CRAN.R-project.org/package=shiny

Fox, J. (2015). *Applied regression analysis and generalized linear models*, Sage Publications.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning*, New York: springer.

Lloyd, S. (1982). "Least squares quantization in PCM." IEEE transactions on information theory, 28(2), 129-137.

R Core Team (2018). "R: A language and environment for statistical computing." R Foundation for Statistical Computing, Vienna, Austria. URL: https://www.R-project.org/

Sievert, C. (2018). "plotly for R". URL: https://plotly-book.cpsievert.me

Wickham, H. (2016). *ggplot2: elegant graphics for data analysis*, Springer-Verlag New York.