**Paper AS02**

# Big Lab Data at the Fingertips of the Statistical Reviewers Using R Shiny

Mukul K Mittal, Sarah Cannon Development Innovations, Nashville, TN, USA
Aditya Rachoori, Sarah Cannon Development Innovations, Nashville, TN, USA

## ABSTRACT

Laboratory data (LB) is one of the largest datasets in oncology clinical trials. Numerous parameters, tests along with multiple time points make analysis and visualization time consuming and cumbersome for both programmers and statisticians. To address these issues, we utilized the power of R statistical software to analyze the data and R Shiny package to make it a visual pleasure. Using interactive dashboard we can subset data as per parameter, test or the time range and visualize them for different cohorts/treatments and flags. This would give us flexibility to visualize data comprehensively or at patient level. These well-designed interactive graphs not only bring clarity to statistical results, they also save a lot of time for reviewers. In summary, R Shiny provides an elegant and powerful framework to turn complicated and large data analysis into interactive graphs at your fingertips.

## INTRODUCTION

Clinical trials are designed to evaluate the safety and effectiveness of a test compound. Among all the data collected, laboratory data is the most versatile and informative. Clinicians and statisticians closely monitor the laboratory data during the course of a trial to find the best treatment. Graphical representation of laboratory data is often used to provide visual information on trends or to compare laboratory tests among different treatment groups, as well as to present safety or abnormalities observed in a study.

Laboratory samples are routinely collected and assayed in clinical trials to ensure that patients are not experiencing any unexpected toxicities. These data are important because of their objectivity, their relationship to organ function and as a first clue to upcoming clinical signs or symptoms. Laboratory safety assessments are key inputs that help investigators or sponsors to make appropriate decisions on treatment strategies.

Presentation of routine laboratory data include tables, listings and figures. Tables may show the percentage of patients with each type of laboratory abnormality, and the frequency of patients change in status from normal to abnormal or from abnormal to normal. Listings may include all collected data organized by analysis days and visits. Figures mainly include line graphs showing analysis values over time and boxplots revealing overall distribution and skewness of the data.

Statistical programming traditionally relies on printable TLFs containing all the information of the study to be investigated. SAS ODS graphics and SGPLOT enable us to make beautiful static plots but these plots lack the power of interactivity. Good visualization can condense hundreds of pages of data/listings into one plot, and adding interactivity further help us to communicate the results in a form that is both engaging and usable for reviewers. Interactive visualization enable the reviewers to explore patterns and anomalies in the data, and thus better understand the messages that the data contain. There are many other software tools like Tableau/TIBCO Spotfire, which provide the options for interactive visualization but fall short by their inability to analyze the data. R Shiny enable us to build interactive web applications without knowing JavaScript, HTML or CSS. R Shiny is designed for fully interactive visualization, via instantaneous updates whenever the user makes a change. The analysis power of R combined with the interactive visualization capabilities of Shiny makes them an outstanding choice for the future.

This paper presents an R Shiny (1) application designed to quickly identify patterns, trends, and outliers by combining the ability to view either single or multiple laboratory tests with the advantage of selecting either a single subject or a user selected subset of subjects. This provides an opportunity for the reviewers and stakeholders to make early decisions on the fate of trial or on necessary course correction. In a nutshell, R statistical software provide an excellent opportunity for statistical programmers, statisticians and investigators to analyze, visualize and quickly deep dive into data in a user friendly format.

## DATA IMPORT

Three imaginary datasets were created in SAS, to mimic the standard ADSL, ADLB and DS datasets developed for phase I/II study. The datasets comprise approximately 100 subjects of different races and gender, divided into 13 cohorts and were designed to capture all the possible scenarios and test values that would occur in a real study. The main lab categories are “Chemistry”, “Hematology” and “Coagulation”, with each category containing several laboratory tests e.g. Albumin, Bilirubin,

Hemoglobin, Monocytes etc. Each test is performed at multiple time points during the trial and data is collected on many visits. Datasets were created in SAS and imported into R studio. The SAS datasets were imported into R using the package “sas7bdat”. Figure 1 displays all other packages that have been used for this R Shiny application.

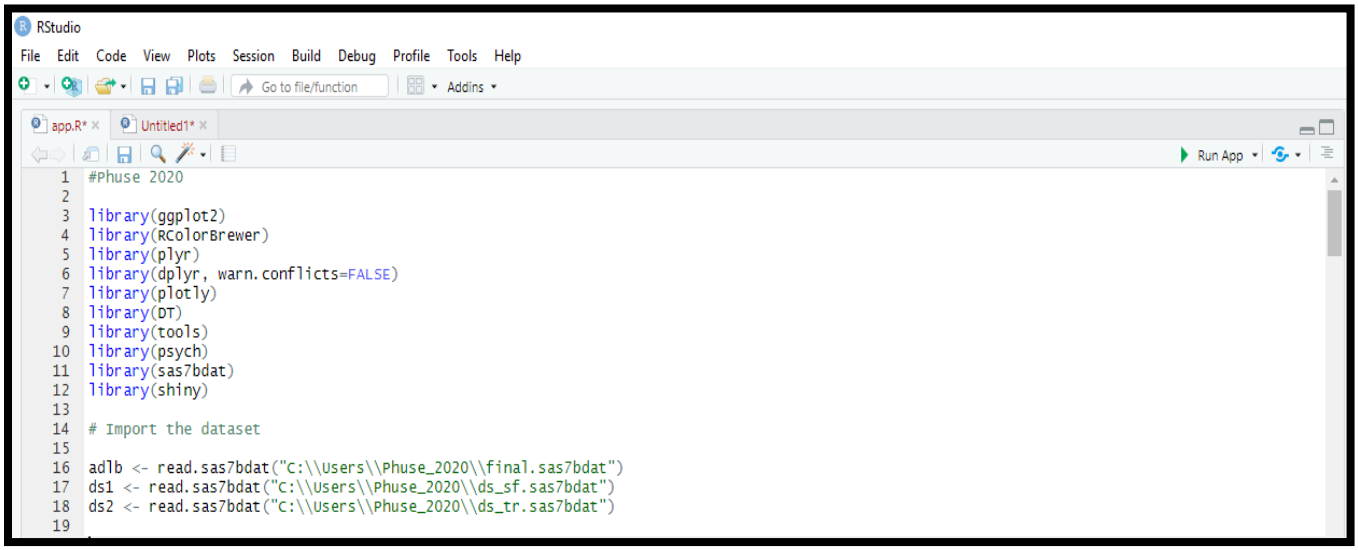



Figure 1

## USER SELECTIONS

The R Shiny application gives the flexibility to select different flags and cohorts. For this fluid interface was used. Here for this application, we used dynamic ui (user interface), which makes all the user provided selections interactive to each other. In this study, we made use of Enrollment Flag, Safety Flag, Cohort, Race, Sex, Lab category, Lab test, Analysis values range, Subject id and Analysis days (Figure 2).

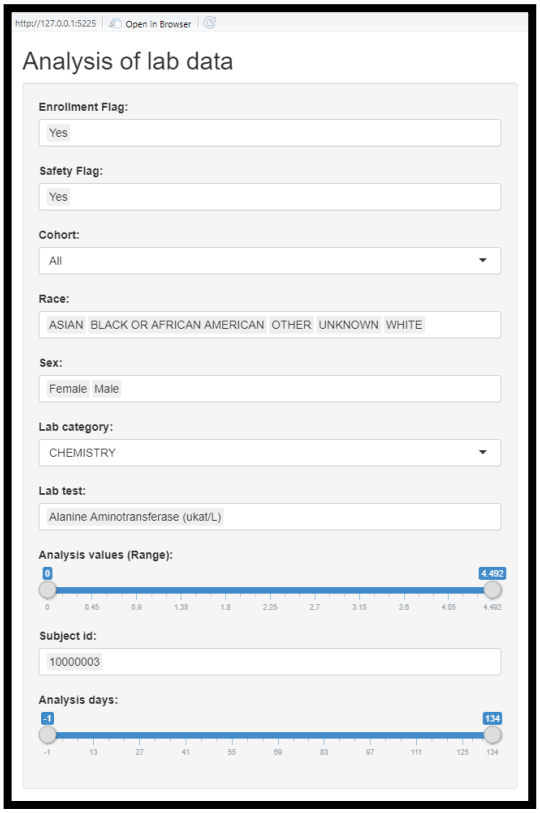



Figure 2

The R Shiny code segment used to generate the user selections is shown in Figure 3.

```
  # Sidebar with a slider inputs
  sidebarLayout(
     sidebarPanel(

       #Selecting Flag
       selectInput(inputId = "ENRFL", label = "Enrollment Flag:",choices = sort(unique(adlb$ENRFL)),selected = "Yes", multiple = T),

       uiOutput("saflUI"),
       uiOutput("cohUI"),
       uiOutput("raceUI"),
       uiOutput("sexUI"),
       uiOutput("catUI"),
       uiOutput("testUI"),
       uiOutput("avalUI"),
       uiOutput("SubjUI"),
       uiOutput("dayUI")
                    ),
```

Figure 3

## LINE PLOTS: CASE STUDY

Laboratory data analysis typically involve changes in test values over time and helps in the determination of the safety of the subject. Investigators analyze the data of one particular test or multiple tests of the same category to look for any abnormal shifts or organ dysfunction. Examples are Hy's law, where potential drug induced liver injury (DILI) is identified by the levels of ALT, AST and bilirubin and the complete blood count (CBC) is used to measure the levels of red blood cells, white blood cells, and platelets.

Laboratory analysis typically begins with a line plot, where analysis values change over time and reviewers access the effect of the test compound on the laboratory values. An example is shown in Figure 4, where the ALT values change over the course of a clinical trial. The ggplot2 package (2) was used to make these graphs.

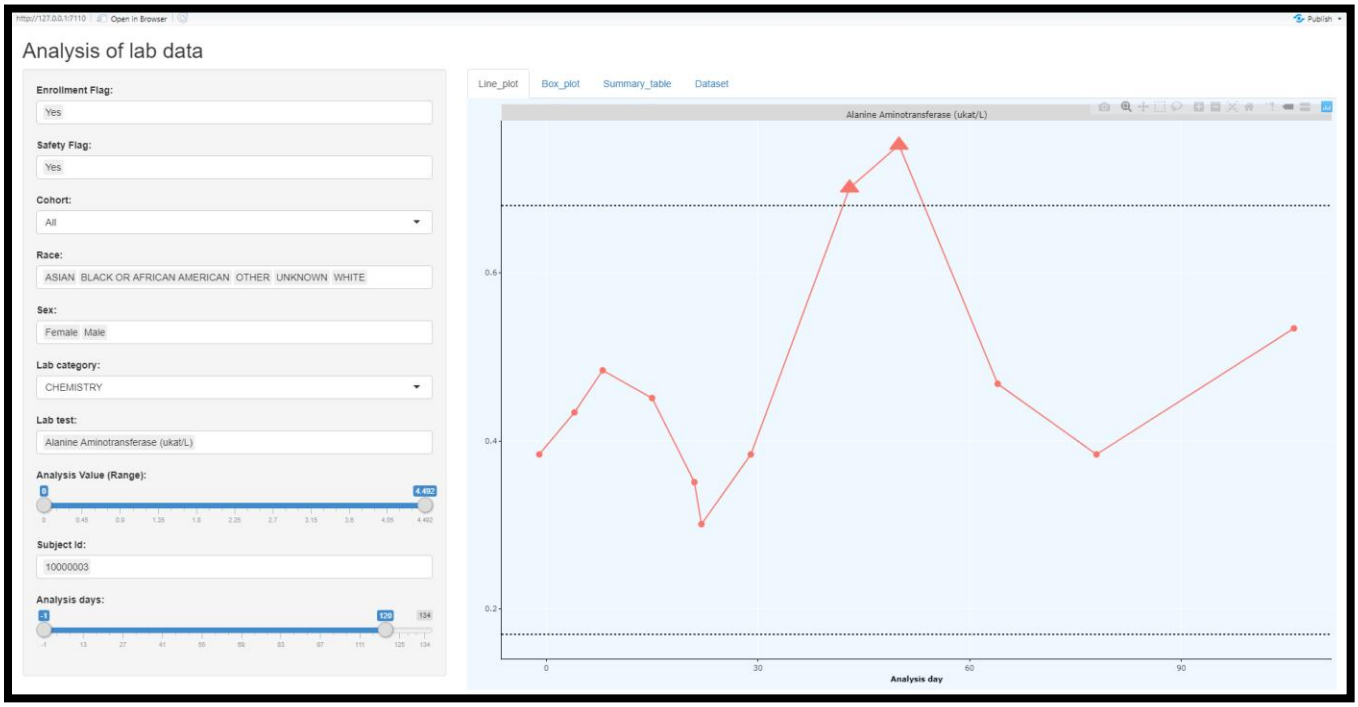



Figure 4

Note that the ALT levels in this example increase rapidly starting about day 40 and that the values displayed as triangles are above the upper range. This raises the question of whether this is an individual case or if it affects a wider group of subjects. To investigate this, the two other subjects in the cohort were added to the display as shown in Figure 5. Different subjects are represented by different colors.

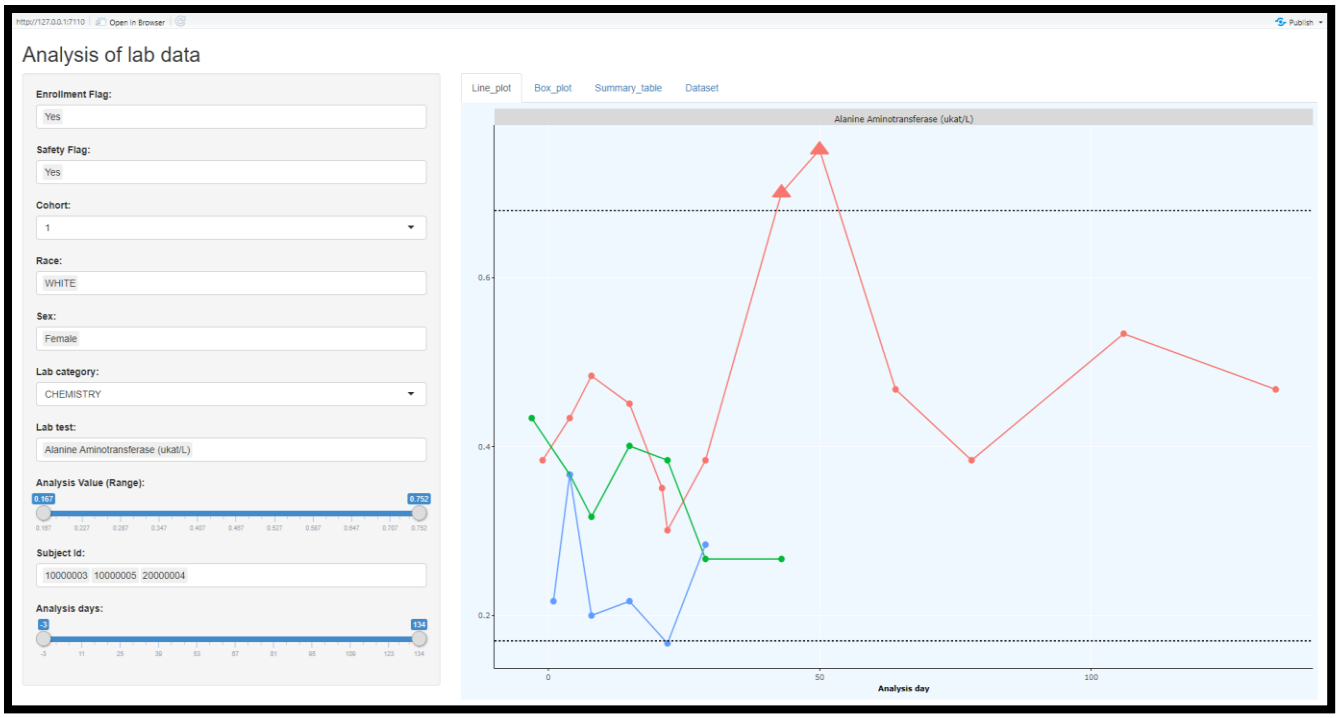



Figure 5

Note that the ALT values for these other patients remain in the normal range over the course of a treatment. However the presence of a high ALT value for one subject can still be indicative of a potential liver toxicity, so it is natural to inquire whether any other DILI lab values are also abnormally high. We looked for ALT, AST, Albumin and Bilirubin together. Figure 6 gives a simultaneous view at ALT, AST, Albumin and Bilirubin.

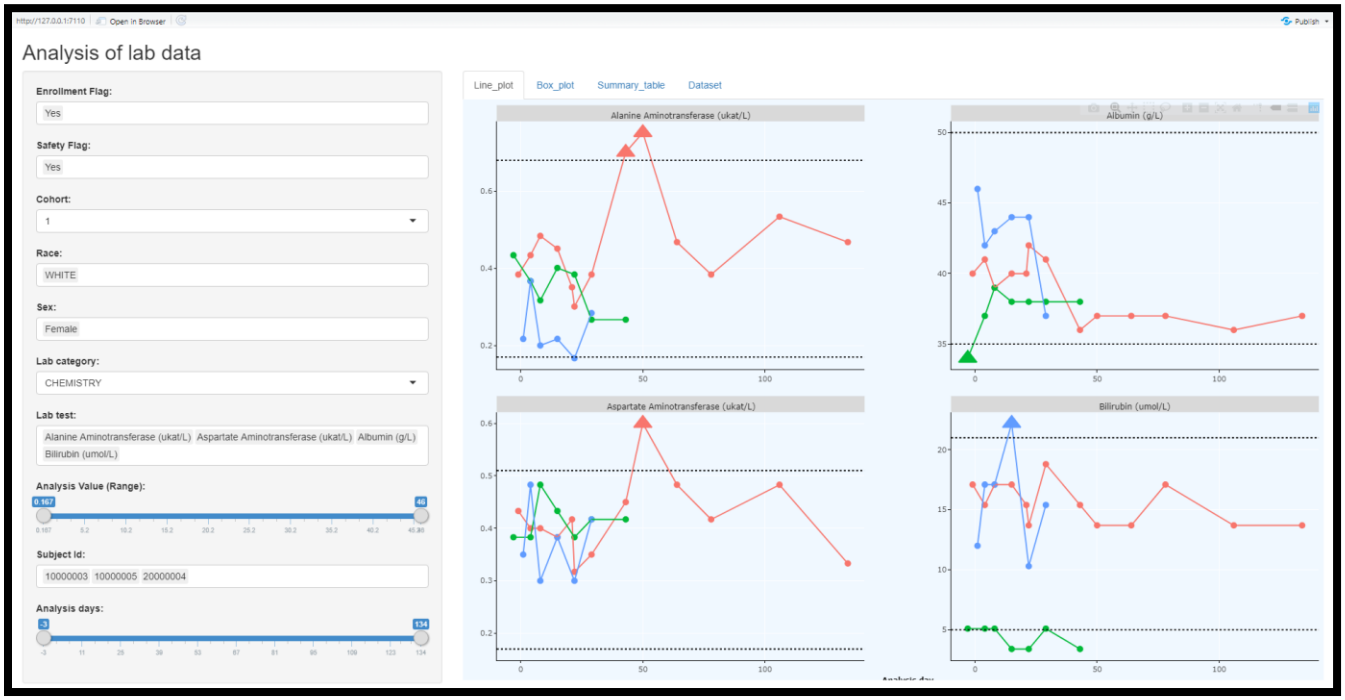



Figure 6

It can now be seen that the subject who had a high ALT also had a high AST at the same timepoint, but no other subjects in the cohort had high ALTs or ASTs. One subject had a high Bilirubin (subject with blue line), but all other DILI values for this subject in the cohort were below the upper limit of normal.

Note that the interactive nature of the R Shiny package has allowed us to quickly investigate in a series of steps a question initially raised by high ALT values observed for just one subject. The ability to expand the view first to multiple subjects and then to multiple lab tests means that an answer can be reached in a few minutes that could take weeks or months if the traditional back and forth between the clinical and programming groups was being employed.

## BOXPLOTS

Boxplots are another standard way of displaying the distribution of data based on summary statistics, i.e., minimum, q1, median, q3 and maximum. They can also tell you if your data is symmetrical, how tightly your data is grouped, and if and whether your data is skewed. Outlier values, very important in safety, can easily be detected by using boxplots. The boxplots in Figure 7 were plotted using ggplot2 and reveal some interesting patterns in the data. Red arrows indicate the outlier values in the given data.

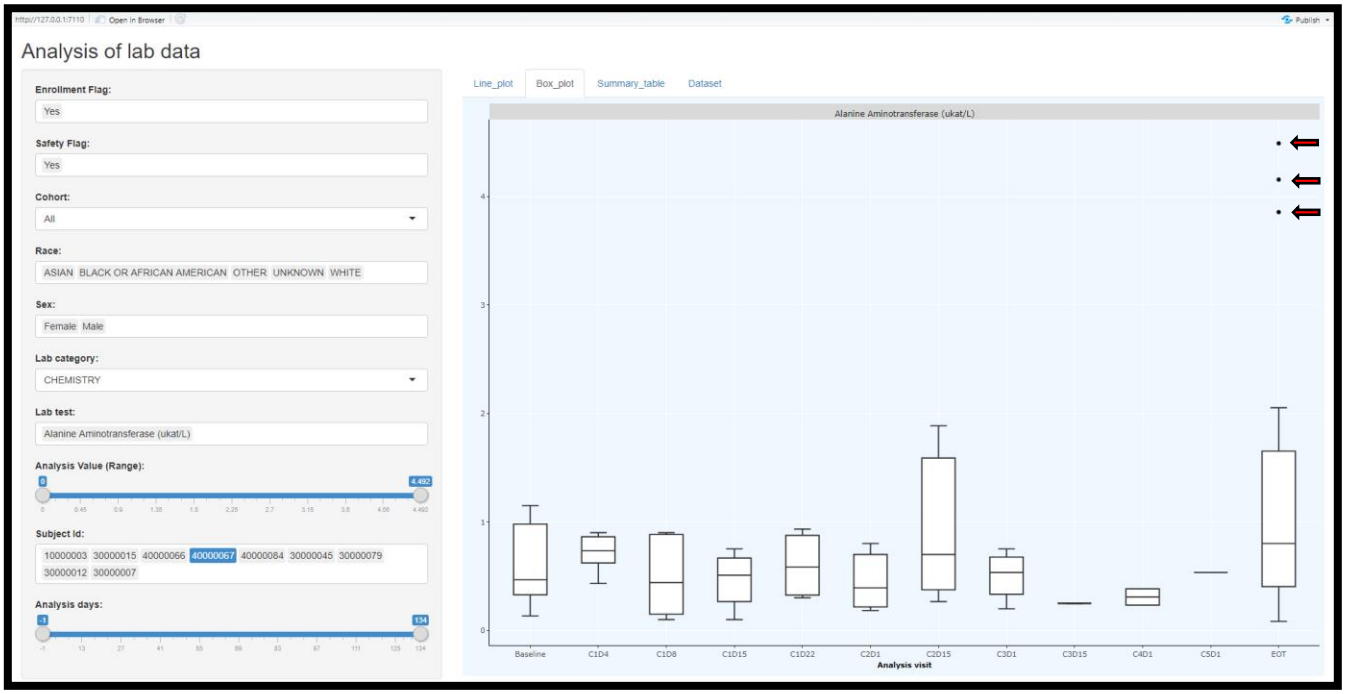



Figure 7

## SUMMARY STATISTICS

Tabulations of summary statistics are an essential part of any clinical data analysis. Choosing the Summary Table tab in the application yields the display in Figure 8.

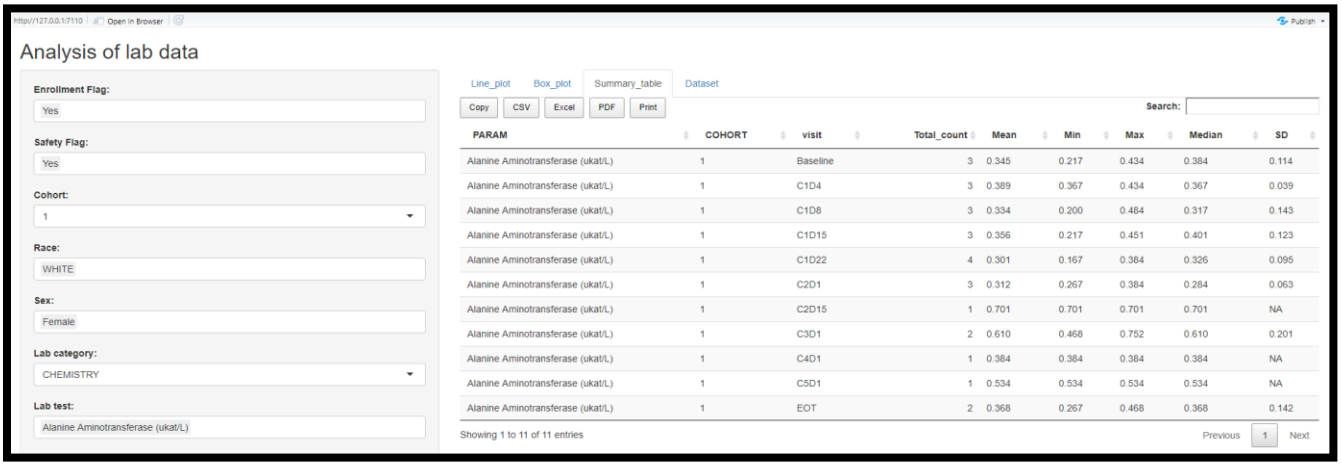



Figure 8

## ANOMALY DETECTION

Data is not always clean. It may have some inherent discrepancies or issues related to programming. It is difficult to figure out these anomalies in conventional method but in interactive application like this, we can easily find out the data anomaly in timely manner. Figure 9 shows two different high or low ranges of C reactive protein (CRP) in the same patient as can be seen by four dotted lines. This was further confirmed by the dataset attached to the last tab.

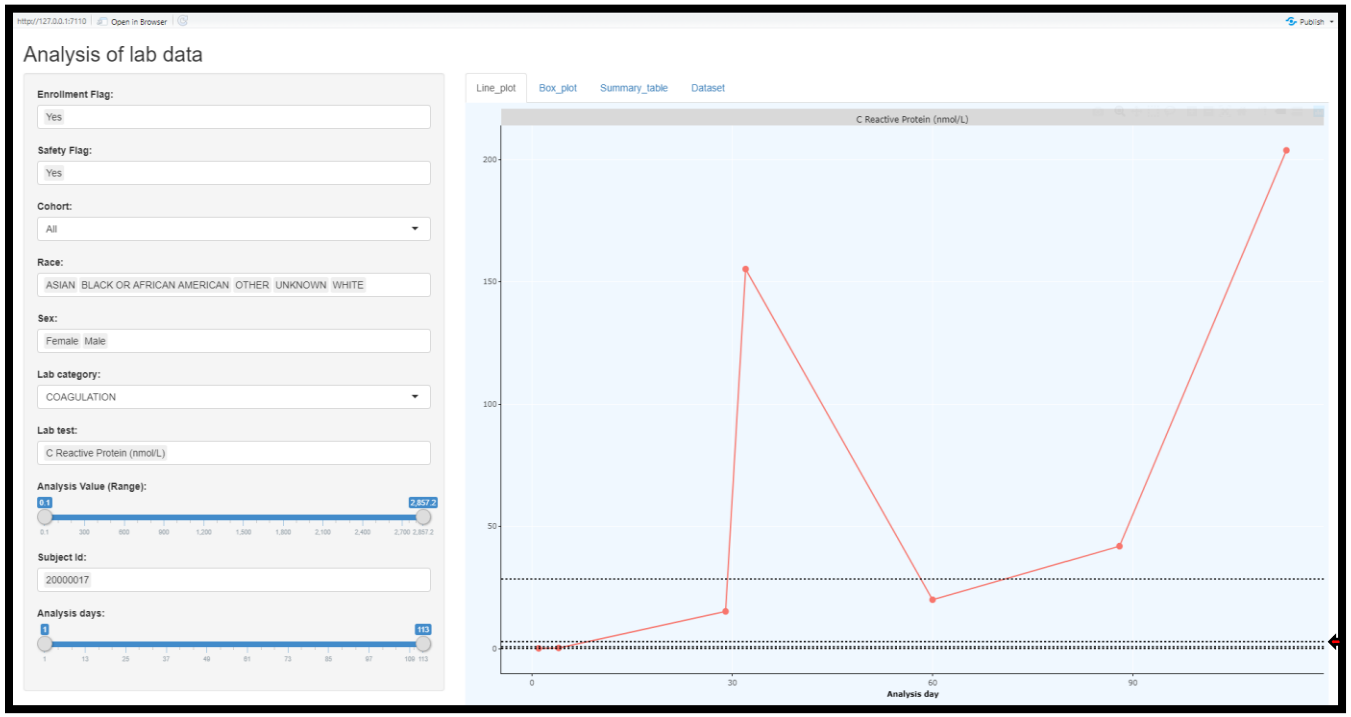
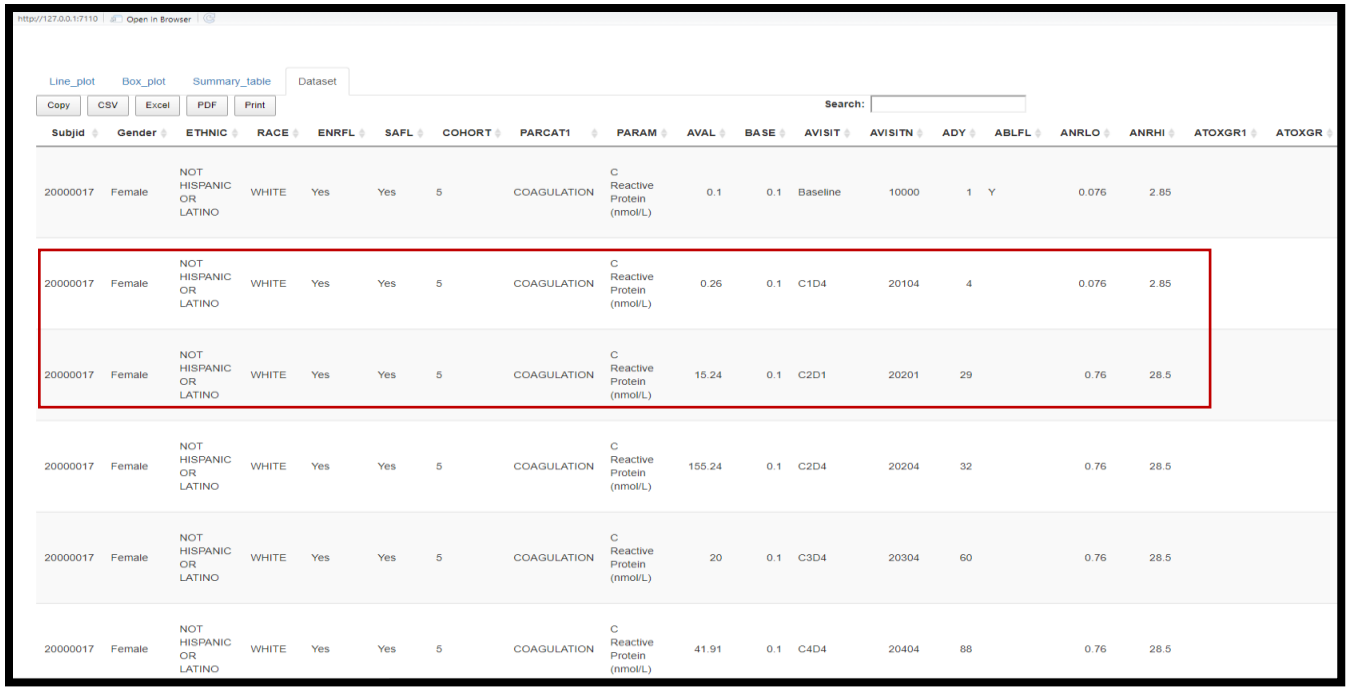



Figure 9

The dataset showing two different values of the CRP high and low ranges can be seen here in Figure 10.

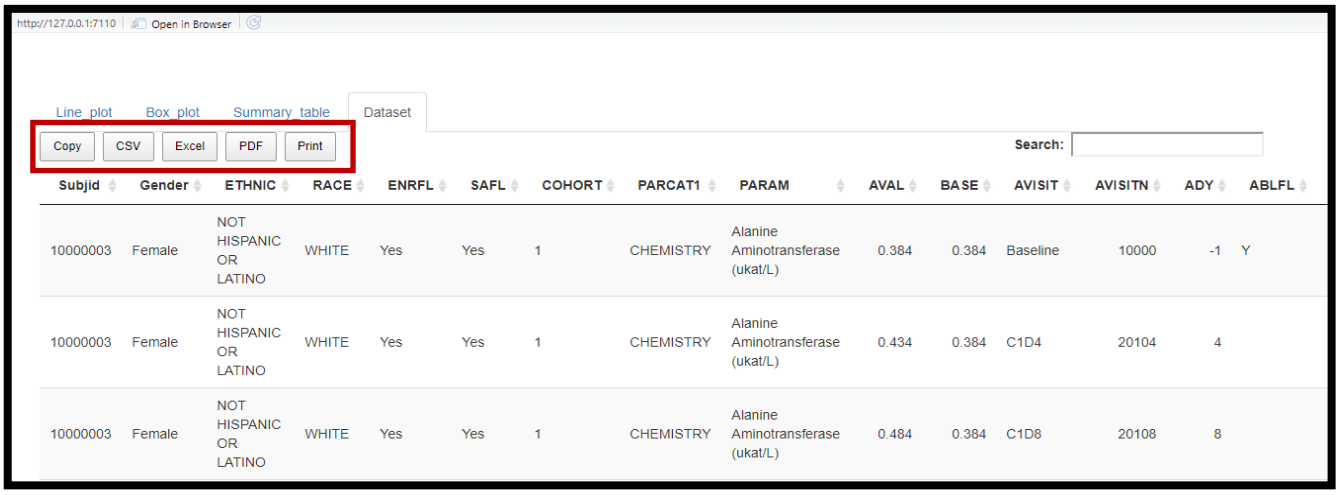



| Subjid | Gender | ETHNIC | RACE | ENRFL | SAFL | COHORT | PARCAT1 | PARAM | AVAL | BASE | AVISIT | AVISITN | ADY | ABLFL | ANRLO | ANRHI | ATOXGR1 | ATOXGR |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 20000017 | Female | NOT HISPANIC OR LATINO | WHITE | Yes | Yes | 5 | COAGULATION | C Reactive Protein (nmol/L) | 0.1 | 0.1 | Baseline | 10000 | 1 | Y | 0.076 | 2.85 | | |
| 20000017 | Female | NOT HISPANIC OR LATINO | WHITE | Yes | Yes | 5 | COAGULATION | C Reactive Protein (nmol/L) | 0.26 | 0.1 | C1D4 | 20104 | 4 | | 0.076 | 2.85 | | |
| 20000017 | Female | NOT HISPANIC OR LATINO | WHITE | Yes | Yes | 5 | COAGULATION | C Reactive Protein (nmol/L) | 15.24 | 0.1 | C2D1 | 20201 | 29 | | 0.76 | 28.5 | | |
| 20000017 | Female | NOT HISPANIC OR LATINO | WHITE | Yes | Yes | 5 | COAGULATION | C Reactive Protein (nmol/L) | 155.24 | 0.1 | C2D4 | 20204 | 32 | | 0.76 | 28.5 | | |
| 20000017 | Female | NOT HISPANIC OR LATINO | WHITE | Yes | Yes | 5 | COAGULATION | C Reactive Protein (nmol/L) | 20 | 0.1 | C3D4 | 20304 | 60 | | 0.76 | 28.5 | | |
| 20000017 | Female | NOT HISPANIC OR LATINO | WHITE | Yes | Yes | 5 | COAGULATION | C Reactive Protein (nmol/L) | 41.91 | 0.1 | C4D4 | 20404 | 88 | | 0.76 | 28.5 | | |

Figure 10

## DATA DOWNLOAD

R Shiny, plotly, and the DataTable package allow data and graphs seen on the screen to be downloaded for reference or for further use as shown in the Figure 11. Figures can be downloaded in .png format, while summary table/data can be copied, printed and downloaded in csv, excel or pdf format. This put a reviewer in the driving seat to see data interactively and to save it for future use.




| Subjid | Gender | ETHNIC | RACE | ENRFL | SAFL | COHORT | PARCAT1 | PARAM | AVAL | BASE | AVISIT | AVISITN | ADY | ABLFL |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 10000003 | Female | NOT HISPANIC OR LATINO | WHITE | Yes | Yes | 1 | CHEMISTRY | Alanine Aminotransferase (ukat/L) | 0.384 | 0.384 | Baseline | 10000 | -1 | Y |
| 10000003 | Female | NOT HISPANIC OR LATINO | WHITE | Yes | Yes | 1 | CHEMISTRY | Alanine Aminotransferase (ukat/L) | 0.434 | 0.384 | C1D4 | 20104 | 4 | |
| 10000003 | Female | NOT HISPANIC OR LATINO | WHITE | Yes | Yes | 1 | CHEMISTRY | Alanine Aminotransferase (ukat/L) | 0.484 | 0.384 | C1D8 | 20108 | 8 | |

Figure 11

The code for downloading data option is given below.

```
# Print data table
output$data <- DT::renderDataTable({

  datatable(
    filtered_data2(),
    extensions = 'Buttons',rownames = FALSE,
    options = list(
      pageLength = 5,
      dom = 'Blfrtip',
      buttons = list(
        list(extend = 'copy', title = "Dataset"),
        list(extend = 'csv', title = "Dataset"),
        list(extend = 'excel', title = "Dataset"),
        list(extend = 'pdf', title = "Dataset"),
        list(extend = 'print', title = "Dataset")
      )
    )
  )


})
```

Figure 12

## CONCLUSION

Nearly all analysis of laboratory data is exploratory in nature and unexpected behaviors in the data frequently lead to new questions, which can be challenging to answer through traditional outputs. The ability of data visualization tools to interactively select both measurements and subsets make it possible to quickly answer these critical questions and quickly make crucial clinical trial decisions. In summary, R Shiny application can be assumed a tool of future clinical trial data analytics.

## REFERENCES

1. R Shiny, https://cran.r-project.org/web/packages/shiny/shiny.pdf
2. H. Wickham. ggplot2, 2009. *Elegant Graphics for Data Analysis*. Springer, New York.

## ACKNOWLEDGMENTS

We gratefully acknowledge the inputs and critical reviewing of the manuscript by our Senior Biostatistician Bill Noble (william.noble@sarahcannon.com).

## CONTACT INFORMATION

Mukul K. Mittal
Sarah Cannon Development Innovations
1100 Martin Luther King Blvd
Nashville, TN, 37203
Work Phone: 615.524.4009
Email: mukul.mittal@sarahcannon.com

Aditya Rachoori
Sarah Cannon Development Innovations
1100 Martin Luther King Blvd
Nashville, TN, 37203
Work Phone: 615.341.7839
Email: aditya.rachori@sarahcannon.com