PharmaSUG 2016 - Paper DG10

Empowering Users By Creating Data Visualization Applications In R/Shiny

Sudhir Singh, Pharmacyclics LLC, CA

Brian Munneke, Pharmacyclics LLC, CA

Amulya R Bista, Pharmacyclics LLC, CA

Jeff Cai, Pharmacyclics LLC, CA

ABSTRACT

A statistical programmer often receives several requests from different functions for exploratory data analysis. This results in several programming hours spent on unplanned exploratory analyses. We believe a better and more efficient approach would be to build data visualization applications that allow our users to directly interact with the data. R + Shiny helps us develop scalable simple web based data applications that can access the data in real time [1]. This paper demonstrates several existing applications that we have developed to assist various functional groups. Our applications help us communicate information or provide analyses clearly and efficiently in real time. The system is scalable and more customized applications can be built within short development cycles.

INTRODUCTION

There is a great need by statisticians, clinical data managers and clinical scientists for real time data analysis and visualizations when conducting clinical trials data review. There are several tools in the market to accomplish some of these tasks, for example JReview. In this paper we will discuss some of the applications we have developed using R+Shiny to analyze and visualize data. The Shiny package, designed by RStudio, allows access to the diverse and powerful statistical methods implemented in R for users without programming experience or even statistical knowledge.

The Shiny package can reduce the user's interaction with complex statistical tools to simple interactive webpage GUIs. The ease of developing Shiny applications allows knowledge experts to easily design very specific GUI layouts that can ensure correct application of the statistical methods.

At Pharmacyclics we have implemented several Shiny applications, and here we will present three examples. The first is an application that provides data cleaning and review tools for clinical laboratory data management emphasizing data visualizations and outlier detection methods. The second is an application that allows for exact inference after modification of the second stage of a Simon two stage design [2]. The last example is an application that visually demonstrates differences between ranges of binomial confidence interval methodologies. The in-house design and programming of Shiny applications allows for quick deployment and updating cycles to address any user feedback when implementing the applications.

UNDERSTANDING THE USER NEEDS

The application we designed will primarily be used by clinical data managers, statisticians and medical monitors.

CLINICAL DATA MANAGERS

The primary Shiny application to be discussed, concerns a package of tools for data cleaning and review, specifically, when most of the variables under review are continuous rather than discrete. The laboratory clinical data management group has the task of reviewing and cleaning the laboratory-based data points collected at study sites and central labs during the course of the clinical trial. A single visit by a single patient can yield as many as 40 different hematologic and chemistry variables. Additionally, there are several scheduled collection visits during the trial, all of which result in multivariate repeated measure data. These hematologic and chemistry parameters are of interest because they are associated with physiologic changes (disease status, adverse events, etc) in the patients due to the use of study treatments or concomitant medications. These expected relationships can provide valuable insight into how to perform data cleaning tasks by suggesting important multivariate visualizations or outlier detection methodologies in addition to the well-known common univariate approaches.

STATISTICIANS

Another example is a Shiny application that was designed for biostatisticians to provide exact inference, estimates and confidence intervals, for a Simon two-stage phase 2 clinical trial design that has had unplanned modifications in the stage 2 size resulting from either over or under enrollment. The common Simon two stage design software

provides a stage 1 sample size, and a stage 2 sample size, with action limits for both stages. The first action limit acts to stop the study early if insufficient effectiveness is seen (a futility stopping criterion), and the second action limit when crossed proposes further development of a drug showing clinically sufficient effectiveness (a superiority criterion). Though the each individual patient can be modeled as a Bernoulli random variable, the total number of responders across both stages is a mixture distribution of binomial random variables. While the accompanying point estimate of the rate is a simple ratio of total responders to total sample size, the exact confidence intervals require determination of the mixture distribution of stages 1 and 2, not a straight forward calculation. The Shiny application is able to provide exact confidence intervals along with other relevant information about any Simon two stage design.

MEDICAL MONITORS/CLINICAL OPERATIONS

Our final example application was designed to assist the medical monitors in understanding the success criteria for hypotheses based on binomial endpoints. These hypotheses are generally one-sided (here assuming superiority) so that the parameter space for π = probability that a patient responds to therapy, which ranges in (0, 1), is partitioned into the null space of (0, θ_0], and the alternative space of (θ_0, 1). The test is conducted by estimating a confidence interval, which corresponds to an inversion of the hypothesis test, and in this case determining if the lower confidence bound exceeds θ_0 is equivalent to rejecting the null hypothesis. The medical monitors often want to know what count of responding patients is necessary for a successful trial, and why one confidence interval method is selected over other available methods. We created a Shiny application that provides answers to these two questions in a single visualization, in which the user selects confidence interval methods for a fixed sample size and the application provides confidence intervals for each count of responding patients. This tool aids the medical monitor in understanding the impact of the selection of the confidence interval method and the required number of responders for a successful trial outcome.

SETUP

In order to build the application you will need a Linux server, Rstudio, Samba and R + Shiny installation. For details about R+Shiny please refer to Amulya et al paper [1]. Samba software was used to read windows file from Linux systems. Development and testing of R program was done in R Studio development environment. Linux, R, Shiny, Samba and R Studio are all open source software so our cost of setup was minimal.

Our IT supported us in setting up a Red Hat Linux server. We installed the R and Shiny packages. We installed ggplot2, Shiny, Outliers, mvoutlier, reshape2, DT, rCharts, sas7bdat, dplyr and Hmisc libraries to aid our application development. The team decided to use SDTM SAS dataset as our source as they provides consistency with variable names. After the development was complete the program was deployed on Linux server. The final application resides on Red Hat Linux server.

The application is available on the intranet at http://10.10.10.218:3838/ShinyTest/ website. User can use IE or Chrome browser to access our application.

DESIGN

After several discussions with our lab data manager we came up with the following flow chart for our lab data application. The application has three tabs. The first window provides the summary of data, the second window visualizes the data and the last window provides change from baseline values.

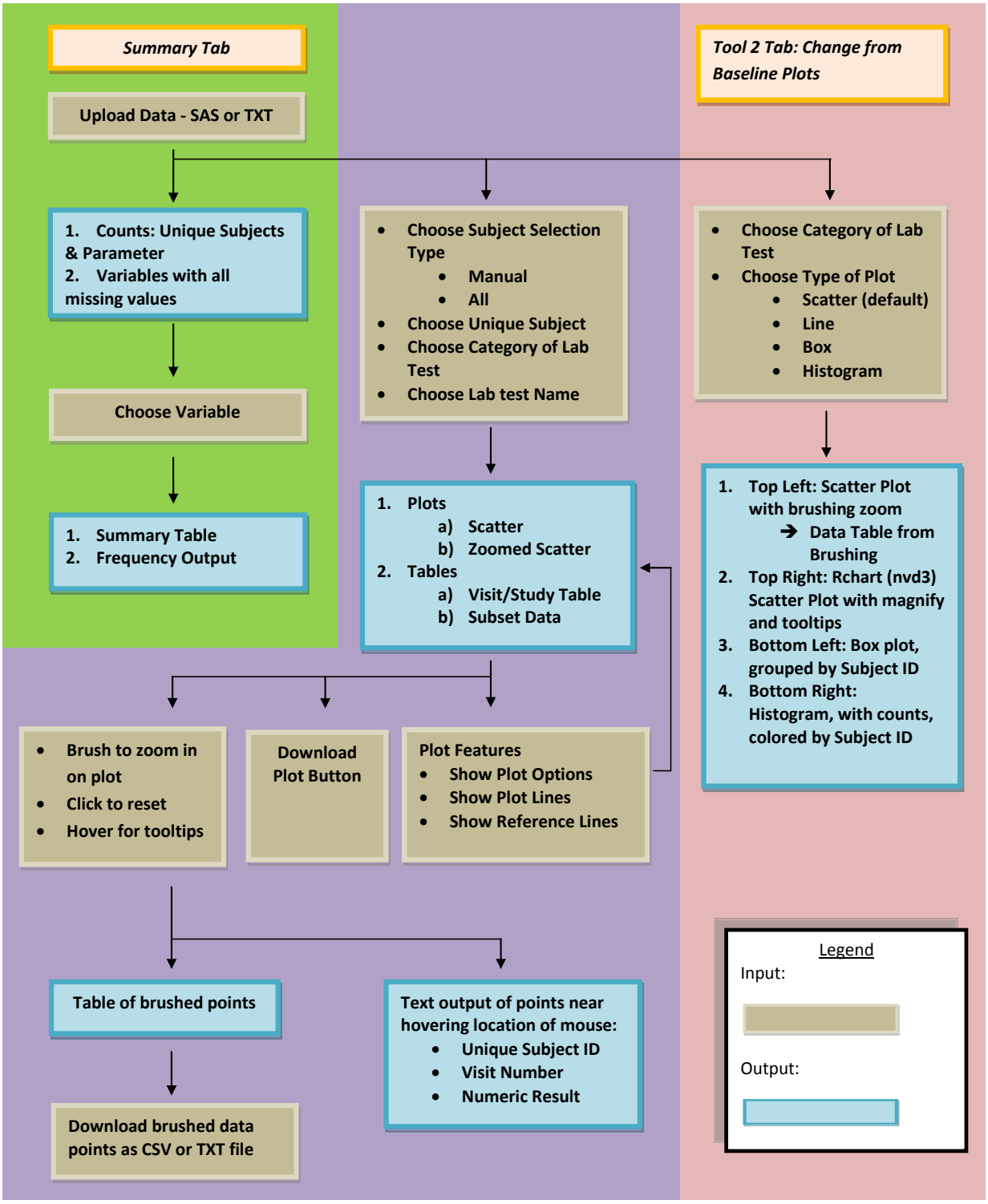


Figure 1. Flow chart for Lab data cleaning application.

APPLICATIONS

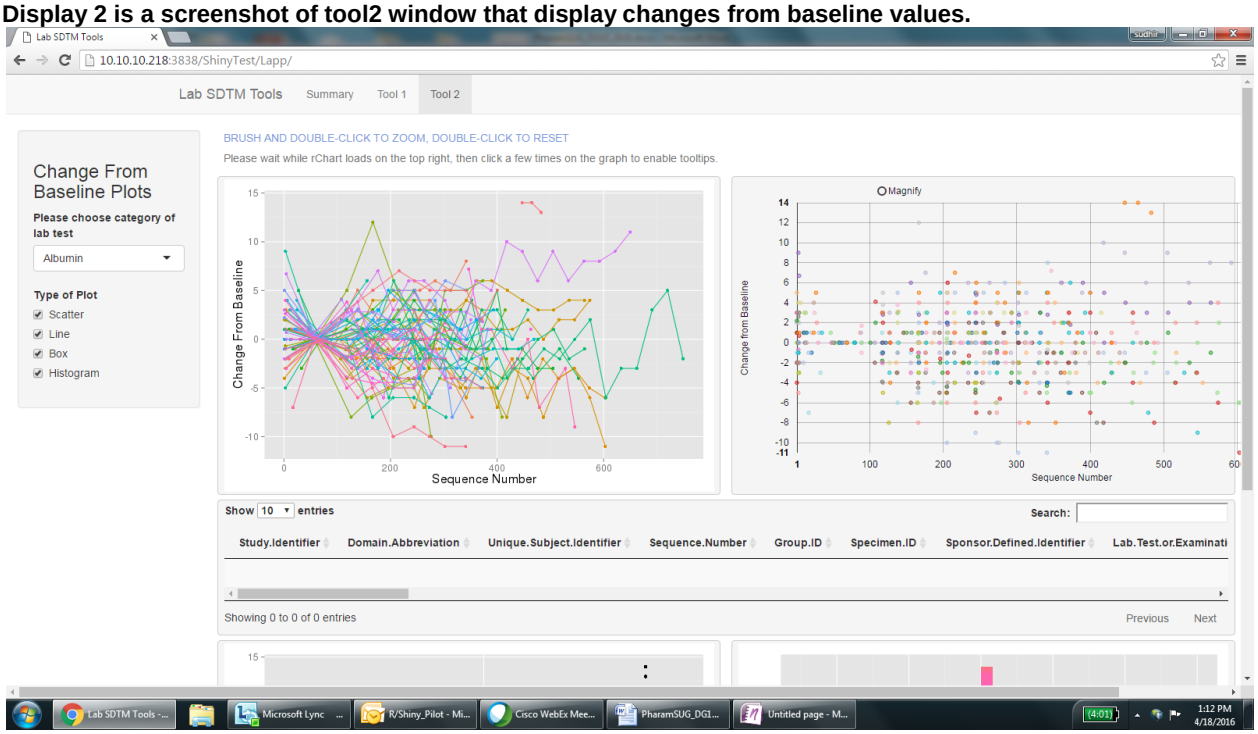
In this section we will provide the screenshot of our applications.

APPLICATION 1: CLINICAL LABORATORY DATA VISUALIZATION

The application has three windows. A summary window, a tool1 window and a tool2 window. Tool1 is a data review window. Tool2 provides some statistical analysis on the data. The user has an option to select individual patient or a group of patients or all patients. The user can also select a lab test category or an individual lab test. The graph is dynamically updated. Display 1 is a screenshot of laboratory data cleaning application.

Display 1. Main Interface for Lab data cleaning application.

Display 2 is a screenshot of tool2 window that display changes from baseline values.



Display 2. Screenshot of change from baseline values.

Display 3 is a screenshot of tool2 window that display bar chart and histograms.

Display 3. Screenshot of tool2 window.

Display 4 is a screenshot that displays how to select an outlier.

Display 4. Screenshot of tool2 window.

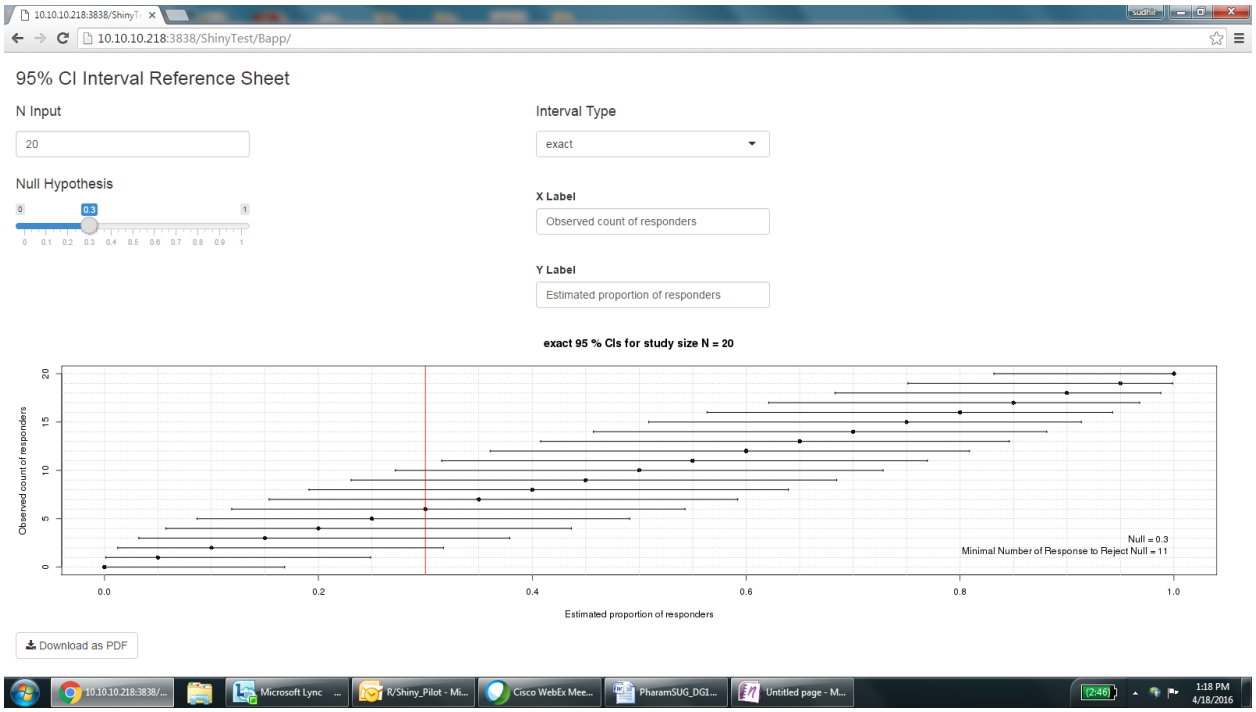
Graphical review helps the data manager to easily identify outliers and data that needs additional query. The application has several features: user can display reference ranges, show plot lines, zoom in, zoom out and download selected data points into spreadsheets.

APPLICATION 2 : HYPOTHESIS TESTING FOR DICHOTOMOUS ENDPOINTS BY INVERSION OF A BINOMIAL CONFIDENCE INTERVAL

The application for the binomial confidence interval comparison tool makes use of the binom package (author/ maintainer: Sundar Dorai-Raj) available from CRAN. This package is set-up as a simple sidebar UI layout with graphical output provided in the main plotting region. Separate tabs are provided for the two scenarios of:

1. Fix sample size N and response count k, and view all confidence interval methods, and

2. Fix confidence interval method and sample size N, and vary responder counts displaying the confidence intervals for each level of response count.



Display 5. Confidence interval application.

With this application the user provides the sample size, then selects a confidence interval method. The slider for the null hypothesis boundary allows the determination of the count of responders (x-axis) that yields a lower confidence bound exceeding the red line corresponding to the null hypothesis. Confidence intervals that lie entirely above the red line correspond to rejecting the null hypothesis.

APPLICATION 3: EXPLORATION AND INFERENCE FOR SIMON TWO-STAGE PHASE II CLINICAL TRIAL DESIGNS

The application for the Simon two stage design is a multiple tab layout that provides a table of all two stage futility design meeting the four specified constraints:

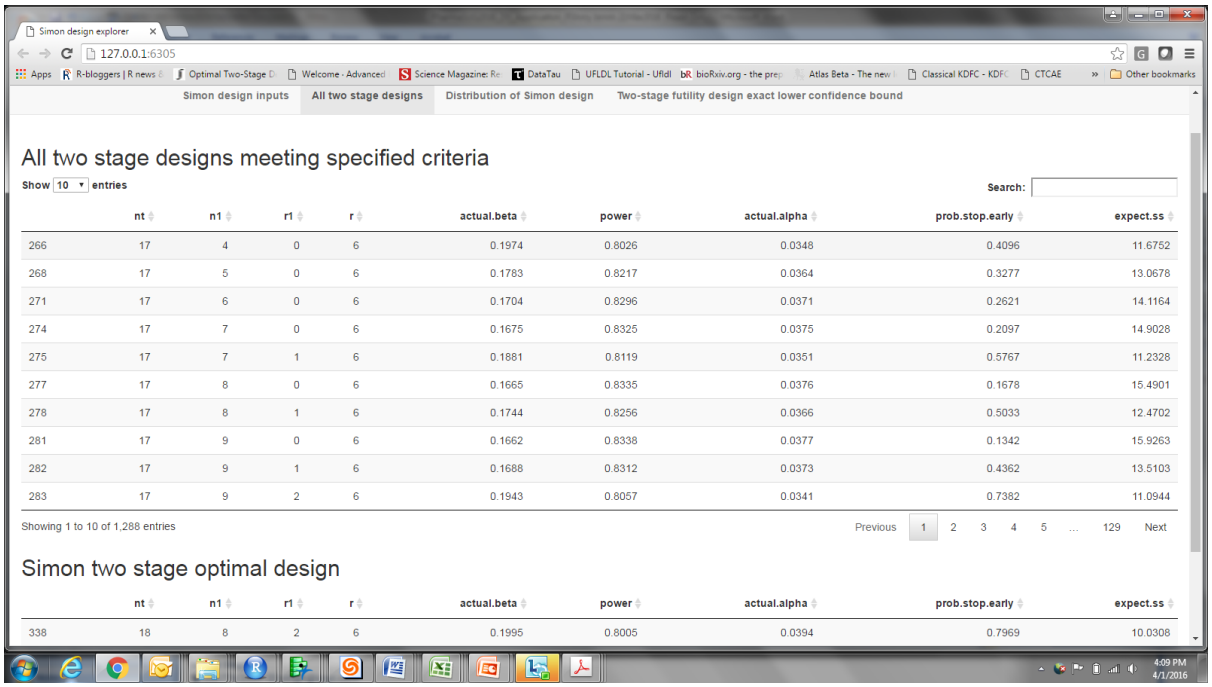
1. Type 1 error α,

2. Type 2 error β,

3. π_0, the response rate under the null hypothesis (an ineffective response rate), and

4. π_1, the response rate of a successful drug candidate in the region of the alternate hypothesis.

A table of the optimal and minimax Simon two stage designs are provided from all candidate designs, and another tab provides an exact confidence interval for any specified design days, allowing for modifications in stage 2 enrollment.



Display 6. Simon two-stage design application..

The Simon two stage application allows the user to explore all possible two stage designs meeting the 4 design constraints. The Simon optimal and minimax design are displayed below the table of all designs. The distribution tab allows a visualization of the binomial mixture distribution for the total number of responders, and the last tab provides an exact lower confidence bound allowing for a Stage 2 sample sizes different than originally planned.

CONCLUSION

When we started the project it seemed difficult to build the whole application using open source technology in two months. Application building was new for our team and there was a learning curve for all of us. After we got over the system setup we found the development environment to be quite user friendly. The three applications discussed here allowed our users to quickly visualize the data and identify outliers and review other data of interest without requesting statistical programmers. This has helped in improving the productivity of our group and better utilization of resources. All the code resides within the company server and any changes or updates to the application can be done quickly. Since all the software used to build these applications are open source the cost of maintaining these application is just the time spent by the programmers. We would encourage SAS programmers to take a look at R+Shiny development tools and hopefully build better applications to bring data to life.

REFERENCES

1. J Cai, A Bista PharmaSUG 2015 R you ready to show me Shiny
2. Simon, R. (1989) Optimal two-stage designs for Phase II clinical trials. Controlled Clinical Trials 10:1-10
3. Gonzalez, S. (2013, 04 12). Data Visualization: Reactive Functions in Shiny. Retrieved 03 12, 2013, from Data Community DC: http://www.datacommunitydc.org/blog/2013/04/data-visualization-reactive-functions-in-shiny
4. RStudio Inc. (n.d.). Shiny by RStudio. Retrieved 02 20, 2015, from RStudio: http://shiny.rstudio.com/gallery/
5. RStudio, Inc. (n.d.). Shiny by RStudio. Retrieved 02 20, 2015, from RStudio: http://shiny.rstudio.com/tutorial/lesson1/

ACKNOWLEDGMENTS

We would like to thank Linda Gau for providing us the opportunity to develop this application, Benjamin Hsieh and Grace Chang for their help in building this application, Shiquan Wu and Leo Cheung for providing us the test server and sharing their experiences with R+Shiny, Abu Mushayeed from IT for his help in setting up the server and Sauji Yachamaneni for reviewing the paper.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Sudhir Singh
Pharamcyclics LLC
999 East Arques Avenue
Sunnyvale, CA USA
408-774-3309
Ssingh@pcyc.com

Brian Munneke
Pharamcyclics LLC
999 East Arques Avenue
Sunnyvale, CA USA
408-215-3543
Bmunneke@pcyc.com