Independent Variable Analyses of Tuberculosis Drug Studies Using R (Group 2)

Alexia Alfano, Rebecca Foos, Colleen Brents, Mackenzie Fry

December 4, 2017

knitr::opts\_chunk$set(echo = TRUE, message = FALSE, warning = FALSE)

### Rationale and Idea Development for Independent Variable Analyses

This project investigates the independent variables from *Mycobacterium tuberculosis* (TB) drug studies, which use mouse species as animal models. The independent variables include variables that were measured both in-vivo and in-vitro. This group chose three different visual analysis methods to look at these independent varaibles including beeswarm plots, correlation plots, and hierarchial clustering plots. For each type of plot the goal was to help the researchers understand the connections between the different indepedent variables based on their data. As a group, we explored three different packages that would assist the researchers in understanding the independent data, these include ggbeeswarm for the beeswarm plots, ggcorplot and ggpolypath for the correlation plots, and ggdend for the dendrograms that represent hierarchial clustering. Beeswarm plots were made using the R package called ggbeeswarm (Clarke & Sherrill-Mix, 2017). A beeswarm plot, also commonly known as a “violin plot”, allows multiple groups of one dimensional data to be plotted in a way that is more representative of the underlying “distribution” of the data, it also helps prevent overlapping of data points that you see in more common types of plots. In this case, it helps to show a pattern of distribution of TB drugs, according to the dosage-interval of the drug given and according to the resulting quantities of drug found in different measurements of mouse/human factors.

**Add quick summaries about other plots here**

This group pushed forward with all the analyses initially proposed from the exploration stage; however, the biggest challenges were in attempting to understand the methodology of the different packages. Many of these packages use techniques that were, at times, more advanced than our R skills. These challenges were addressed using stack overflow and vignettes for the packages to see examples of how they are used.

*Describe the different ideas your group explored. What were the biggest challenges in this stage? For any ideas that didn’t pan out, what were the key constraints? Also describe how you would tackle this problem if you were starting over.*

*Key functions: Describe the final functions / app framework you decided on. Explain why you picked these. For functions, include documentation for the functions:*

### Beeswarm

#### Beeswarm Coding Methods and Plots

#### Beeswarm Functions

The beeswarm functions called invivo\_beeswarm\_function() and invitro\_beeswarm\_function() follow suit with the division of the independent variables between in-vivo and in-vitro. The functions are built to produce visual outputs for both. The two functions operate with very similar or almost identical logistics in terms of the code, so the following is an explanation of the main concepts used to design and build both functions. One goal of the functions is to allow the researchers to plot all of the independent variable data quickly and efficiently. Another goal of the functions is to provide the researchers with the power to control which data they want to visualize and/or compare across all of the independent variables measured. There are particular parts of the function code that accomplish these goals. By setting the statements for variables and drugs to NULL (line ), it tells the function to produce plots of all the variables and drugs UNLESS otherwise specified. If the researcher inputs invivo\_beeswarm\_function(efficacy\_summary) then all the small multiple plots and all the drugs will be returned. The following IF statements (lines ) in the function are what direct the function in that UNLESS otherwise specified situation. When the variables statement or drugs statement is NOT null, the function tells it to filter the variables or drugs as requested in the input from the “variables\_filtered” or “Drugs” columns of data. As an example, if the researcher types specific variables or drugs they want to plot together (line ), then the function will plot only what they ask it to (figure ). If you are wondering how the “variables\_filtered” differs exactly from the “variables”, this aspect is explained in detail below, under the topic “Problems We Faced”. The beeswarm functions were designed with these capabilities following discussion with the researchers who provided/and are familiar with the independent variable data. Because all of the data together can be quite overwhelming to look at, it became clear that the researchers were interested in the ability to separate out the plots as much as possible so comparisons of specific variables and drugs can be explored upon their discretion. The beeswarm functions employ plotly and ggplotly() in particular to add an interactive element to the plots, however the code lines for this needed to be left out of this writeup in order for Rmarkdown to knit this document. This is a key component of the code that was given to the group working on building the Shiny App, which allowed the final plots built in the app to have detailed descriptions appear when the mouse was placed over any point in any of the small multiple plots produced. The Shiny App group also contributed to the creation of radio (check) buttons to go along with my functions, which ultimately allow for the researchers to make selections of specific variables and drugs within the Shiny interface.

### Correlation Plots

key function stuff

### Hierarchial Clustering

key function stuff

### Problems We Faced

#### Mouse small plots probs/Correlation Plots:

#### Beeswarm

For the beeswarm plots and function, there were some problems that we faced that are important to note. It was not possible to use the data, as it was upon importing, to create small multiples. The gather() step in the function is an important step in particular, which modifies the dataframe to a format that allows for the creation of small multiples. Labeling in general became problematic because in many instances labels for the actual plots needed to be different or more specific than the variable names, which were minimized to an extent that would not be suitable for plot titles (i.e. PLA, SLU, huPPB, etc). This became a particular issue in the functions when markup language was used to format the spacing on the proper plot titles and was then required, as per the function code, to be typed as such in the variables statement input. The resolution to this problem was met by applying the minimized variable names to the variables statement and this is the reason why “variables\_filtered” is found as a reference in the function instead of “variable”, as would be suspected. It can be noted that in both functions, a new column “variables\_filtered” is created (lines ). By creating this new column, we can conserve the formal plot titles without having to reference them exactly in the function statement. It should be recognized that a set up like this might only be ideal for those who have had time to develop an understanding of both the shorthand (dataframe) and longhand (plot title) versions of these independent variables. The beeswarm plots, when plotted with all of the possible data, can appear pretty busy. This problem was touched upon in the description of the function as one of the reasons why the function was built as it was. These functions play a significant part in mitigating this “busy” problem by incorporating tools that help make the plots more specific and customizable by the researchers and other potential users. One aspect that might be noticed about the beeswarm plots is that the classic beeswarm or violin shape isn’t prominent. This might be related to the quantity of data. Analyses using beeswarm across multiple studies over more time, that collect the same independent variable data, might demonstrate a more clear beeswarm shape.

#### ggdend:

Aesthetics of ggdendrogram plots

### Input Parameters/Potential Errors

While writing functions for each of our plots, we considered potential user error for our functions. One consideration, is dealing with data if there is missing data within the columns for each of our plots there may be limitations if there is missing data in an entire column. For example, the beeswarm functions will still run with existing "NA" values (they are excluded from the plot automatically) but if an entire column of data is missing, that would cause an error because the function would not be able to find an object it needs. It is also important to consider that our group used the format of the efficacy\_summary dataframe as the basis for the functions created. If another dataframe is referenced instead, it could cause a number of errors. Another issue we thought about is the number of rows within the datasets for future use. We considered if there as just one row versus 500 rows, if our code would run the same as it is now. Another consideration is how the radio buttons might discord with the drug labels. For example, in the beeswarm functions, the functions should still work if the drugs are called different names (as long as the column name containing the drug information stays the same) but the radio buttons in the Shiny App would require an adjustment.

### Future Considerations

Some of the next steps in general, which touch upon all of our specific analyses, that were considered included being able to customize and munipulate the labels. We might also want to add descriptions of the different tests or measurements using plotly so the user can understand what the different measurements or tests are indicating about the TB drug. We also thought it might be useful to incorporate a one-click export to PDF feature for the plots, in order to save specific, useful outputs. Another thought was that it would be beneficial to include control data, to further clarify the relationship between the independent variables and the drugs. In addition to controls, it would be interesting to consider how to incorporate Cmax and Trough data into these functions and the use of combined-drug data that might be found in some TB drug trials.  
A future consideration for the beeswarm functions, in particular, would be including code that would plot blank small multiples for missing columns in a dataset. Also, it would be helpful to code a fix in the radio buttons for potential drug-name variation.

### References

Erik Clarke and Scott Sherrill-Mix (2017). ggbeeswarm: Categorical Scatter (Violin Point) Plots. R package version 0.6.0. <https://CRAN.R-project.org/package=ggbeeswarm>