**ALY6980: Experiential Capstone**

**Instructor’s name: Caroline Moore-Kochlacs**

**Assignment title: Project documentation**

**Student’s name: Le Ding**

**Academic term: 5th quarter**

**Completion date: June 29, 2019**

**The data we used**

The data we used is about the impact of drugs on the liver based on in vivo studies on animals and humans, and the data is collected from the existing researches which were sorted out through systematic review.

The original dataset was divided into two sheets; first is the sheet for the studies on human, second is the sheet for the studies on animals (e.g. “Rat”, “Mouse”, and “Monkey”). There are 27 columns in each sheet, and we are going to select some of them (e.g. “Bibliography”, “species”, “drug”, “endpoint name”, “endpoint units”, “daily dose”, “daily dose unit”, “time point”, “number of experimental units”, “effect value”, “variability”, and “unit of variability”), which are the critical conditions for liver toxicity measurement, to visualize the data and look for some meaningful insights.

**How we modified our data**

Because the original dataset was divided into two sheets, we had to bind them into one, which should be much more convenient for us to conduct the next visualization work. Firstly, we unified the column names of them, then bind them into a unified structure. Then, we had an intact dataset on the same page/sheet, and we could continue to separate them to create the subsets for different visualization work.

After we had the intact dataset, we inspected and found missing/unusual values in the data, so we reported them to our sponsor and made a backup for them. Hence, we had the clean data in that stage, and we could conduct the next steps.

1. **Data preparation for Part A (scatter plot & bubble plot):**

In this part, one of our instructors only asked for the visualization for the data under the endpoint name “ALT” (the data about other endpoint names could be easily reproduced after we completed the structure and code of our visualization), so we just sorted out the all the data whose endpoint names are related to ALT, and unified those diverse units which all actually stand for ALT. Of course, there are some ALT data which indicate that the experimental objects already have type 2 diabetes, so we have to clean them out.

Because there were various units of daily dose and effect values, we must unify them into one. Therefore, the units of daily dose were unified into “mg/kg”; the units of effect value were unified into “U/L”, “IUA”, and “counts” (those three could not be unified anymore), and audiences can change the unit of effect value on our dashboard.

1. **Data preparation for Part B (box plot & forest plot):**

This is a much more complicated process of data preparation, we spent a lot of time to construct the mindset and structure of the dataset for Part B. Since the data about treatment groups and control groups are mixed in the same dataset, so we had to separate them logically and reconnect the treatments and their corresponding controls. To achieve this goal, we added three levels of ID to do the recombination.

1. First, we added a “StudyID” column to the dataset, which is going to specify different species in the same study. (e.g. 75)
2. However, based on the specification of “StudyID”, there might be some different combinations of control and treatment in the same study and on the same species, so we had to add the second level of ID “CombinedID” to specify different combinations of control and treatment in the same study and on the same species. (e.g. 75.1)
3. Nevertheless, a single control might have several different treatments, so we added one more suffix to the “CombinedID” to create “ComparedID” which specifies each treatment to the same control. (e.g. 75.1.1)

In this way, the recombination was achieved, and we could calculate the effect size and CI for each pair of treatment and control. Hence, the data for the box plot and forest plot were prepared. In the end, we found that there are two kinds of control among the studies. One is setting control on the daily dose, the other one is setting control on time point; we separated the prepared dataset into two subsets which have different control IVs.

In addition, we added a column to convert the effect size into effect percentage for each experiment. The purpose of this action is to eliminate the disturbance of various endpoint units. The dependent variable of our box plots, which is the values on y-axis, uses the value of effect percentage.

During this whole preparation process, there were many unusual data which had diverse reasons. We had backed up them, and they were packed in our document of deliverables.

**Guide to using the visualization**

1. **Part A:**

Our visualization work was divided into two main parts, Part A and Part B. Part A contains the scatter plot and bubble plot, and both of them reflect the effect values with the change of daily dose under the endpoint name “ALT”. The main difference between them is, instead of observing the trend and pattern of effect value by daily dose, the bubble plot uses its bubble size to reflect the reliability of the corresponding experiment. The criteria of the bubble size are “SE” and “sample size”, and the rule for the bubble size is “the bigger the bubble is, the more reliable the corresponding experiment should be”. Thus, audiences can utilize the bubble plot to estimate the reliability of collected experimental data and sort out the high-quality data to conduct future studies.

1. **Part B:**

The visualization work for Part B consists of the box plot and forest plot, and the box plot has two parts: First one is the box plot reflecting effect percentage by species, the second one is the box plot reflecting effect percentage by the levels of the cumulative daily dose. Thus, audiences can observe the distribution of effect percentage on different species and levels.

The function of forest plot should be clear to us, it is going to show whether there is a significant difference between the treatment groups and control groups under the specific conditions, so that we can determine if the drug really has the significant side effect on the experimental objects under the specific units and conditions.

**Guide for EBTC to modify the visualization**

We create all the visualization with R, and the relevant R codes were packed along with the corresponding visualization. Thus, the sponsor and future researchers just need to review our codes and change the imported dataset, and the visualization will be automatically shown. Besides, all the visualization works, instead of forest plot, are designed into R shiny dashboard, so the function and feature of the visualization can be directly used or changed.

However, there it is essential to prepare dataset and make it fit the conditions written in the code, so the review of our codes for data preparation is necessary for future modification.