QBIO490: Multi-omic Data Analysis
Fall 2022 Clinical Data Partner Activity

Due: Meeting 6.1 (Monday, 9/26)

Deliverables:

1. Submit your answers to the written activity by adding a pdf document to your GitHub (and sharing a google doc with Nicole + TAs).
2. Submit your coding responses by creating an R script (not a Notebook!) called `partner_activity_yourname.R` and adding it to your GitHub.
    a. Because this is a script, you will need to charge your working directory via `setwd("/PATH/TO/ANALYSIS/DATA")`
3. Though this is a partner activity, each student needs to push both their written answers and their code to GitHub. It is okay if your answers are the same as you are working together!

A Few Steps Before Starting

1. Ensure that your data is in your `analysis_data` folder.
2. Read in `brca_clinical_data.csv` file from your local machine. This is the file that you should have saved at the end of the clinical tutorial. Saving and reading in this file allows you to skip the `GDCquery` step every time you want to read in your data.

```
clinical <- read.csv("PATH/TO/brca_clinical_data.csv")
```

3. We will also be working with drug and radiation data. The data were already queried and downloaded, just prepare them with the following two commands:

```
clinical.drug <- GDCprepare_clinic(query = clinical_query,
                                   clinical.info = "drug")

clinical.rad <- GDCprepare_clinic(query = clinical_query,
                                  clinical.info = "radiation")
```

Written Activity

1. Define the following: *categorical variable, discrete variable, continuous variable.* Provide examples of each.

Categorical Variable = Falls into mutually exclusive categories - for example age group.

Discrete Variable = They are countable in a finite amount of time, they are whole numbers - for example number of patients

Continuous Variable = Data that can theoretically take on any value, for example time.

2. Look at the different column names of the `clinical` dataframe. Choose one that is interesting to you and your partner. Ensure that there are not too many NAs in this column by using `is.na(clinical$COLUMN_NAME)`. Remember that in coding, TRUE is equal to 1 and FALSE is equal to 0. You can then use the `sum()` function to find how many TRUEs exist. Which variable have you chosen?

Progesterone receptor level cell percent

3. Google your chosen variable. How is your variable measured or collected? Is your variable categorical, discrete, or continuous?

The test is called immunohistochemistry, it detects progesterone receptors in cancer cells from a sample of tissue.

The data is continuous, however progesterone receptor positive is categorical.

4. Find two research articles that mention your clinical variable. Provide the links and a brief description of the findings.

https://aacrjournals.org/cancerres/article/50/21/7057/496180/Immunocytochemical-Localization-of-Estrogen-and

This research article found that cancer tumour cells that are oestrogen and progesterone receptor positive had significantly higher survival rates in comparison to oestrogen and progesterone negative tumour cells. They also discovered that the greater the progesterone receptor level cell percentage, the greater the chances of survival.

https://www.proquest.com/docview/212478810/fulltext/16797DAC68254078PQ/1?accountid=14749

In the lobular structure of the breast, the content of both oestrogen and progesterone receptors is proportional to the rate of cell division and growth. They also found that the cell proliferation is controlled by oestrogen through an indirect mechanism.

5. Look at the different column names of the `clinical.drug`, `clinical.rad`,

and `clinical` dataframes. Choose a variable from one of these data frames. Ensure there are not too many NAs (there will likely be more NAs in the drug and radiation dfs than in the patient data, don't worry about it too much). Which variable have you chosen? Provide a brief description of the variable.

Breast carcinoma progesterone receptor status. This is a categorical variable with positive and negative.

6. Scientists generate hypotheses before experimenting or exploring data. Generate three hypotheses: (1) Relate your variables to each other, (2) Relate your first variable to survival in breast cancer, (3) Relate your second variable to survival in breast cancer.

1. The negative status of breast carcinoma progesterone receptor corresponds to low progesterone receptor level cell percent.

2. The higher the progesterone receptor cell percentage, the greater the survival rate in breast cancer.

3. The patient with negative status of breast carcinoma progesterone receptor would have higher survival rate in breast cancer.

7. Summarize what you learned from your graphs! What is the significance of these findings? (Answer this question after you finish your analyses)

The histogram shows that all patients with negative breast carcinoma progesterone receptor status have less than 10% of the progesterone receptor level cell percentage, while those with positive status have either extremely high percentage (90%) or low percentage (10%).
The graph KM1, shows that survival probability is affected by the percentage of progesterone receptors and the graph KM2 similarly shows that the presence of progesterone receptors positively affects the survival rate. Low progesterone receptor cell level percentage would lead to lower survival rate, which is also the reason why patients with negative progesterone receptor status have a lower survival probability.
Overall the three graphs all indicate that a higher progesterone receptor cell level percentage, or more generally the presence of progesterone receptors positively impacts the survival rate.

Coding Activity:

1. Perform an analysis looking at the two variables that you chose. First brainstorm and sketch out a plot that contains both variables. Feel free to get creative, if you are struggling, feel free to ask for ideas! (Helpful functions/packages: plot(), hist(), boxplot(), pairs(), ggplot2 package + associated functions)
    - TIP: Sometimes it can be hard to plot a continuous variable with another variable. You can convert a continuous variable to a categorical one. For example, we previously defined age < 50yrs old as "Young" and age ≥ 50 yrs old as "Old." Here we have converted age, a continuous variable, to young and old, a categorical variable.

2. Perform a survival analysis, following the steps of the clinical data tutorial with the first variable.
    - As with the previous tip, the survival analysis needs a categorical variable. If you have a continuous variable, use an `ifelse()` statement to create a new column with a categorical version of the variable.

3. Repeat with the second variable. Note that for drug and radiation data, there might be many categories in one column. Try to keep the KM plot simple by limiting the data to ~5 stratification categories.

4. For an extra challenge (optional) perform a survival analysis where survival is stratified by *both* variables.

5. Save your plots and write any data frames you used to your local machine.

<u>Check before submitting:</u>

You **must** include informative comments throughout your code.

```
str(clinical) # view structure of clinical data frame
head(clinical) # view first few rows of clinical data frame
```

You **must** install and load all necessary packages at the top of your coding fall.

```
if (!require(package)){
install.packages("package")
}

library(package)
```

You **must** change your working directory at the top of your coding file.

```
setwd("/Users/nicoleblack/Desktop/QBIO/qbio_nicole/analysis_data")
```

You **must** be able to run your script from top to bottom (with a clean environment) without any issues.

- Before turning it in, hit the broom in the top right corner of Environment to clear all values and data. Then run the entire script by hitting the run button in the top right of your source panel. Your code should run all the way through with no errors.