**Using R for Data Analysis in Pharmacokinetics**

**Data Set**: *medadmin.csv*, *pt\_observations.csv*

This dataset is a synthetic dataset similar to real-world electronic health record data. It describes patients treated with vancomycin at two different hospitals. In this assignment, you will demonstrate your ability to use R to perform typical data manipulations and to calculate clinical parameters with the end goal of assessing risk factors of vancomycin-induced acute kidney injury.

Please provide:

* Your written answers to the questions posed below, complete with any graphs or summary data tables you made to answer them. Written responses should not need to be more than one or two sentences each.
* The R code you used to analyze the data and create the graphs or summary tables.

You may find it convenient to combine both of the above deliverables into a single R markdown file. A well-commented R file(s) combined with a PDF or similar filetype is also acceptable.

Please use R for all analysis and graphing. If you use a package not presented in the R workshop, please justify why you chose to use that package.

Make sure your R scripts are *reproducible*. They should run without issue on other machines, assuming the user has set the working directory to point to a folder containing the provided datasets. Do not encode your personal directories (ex: setwd(~/Users/jasminehughes/Documents)) or install packages within your script (but do load any required packages).

Make sure your R code is *readable*. Use variable names that are easily understood. Use comments where appropriate to explain how your code works.

Make sure any graphs are communicative. Axes should have labels, and the data should be presented in a way that is intuitive and interpretable.

**Data Cleaning**

1. *Familiarize yourself with the dataset* 
   1. How many unique patients are there overall (you can assume each patient has one unique `blinded.id`)? How many patients were treated at each hospital?
   2. What was the mean, minimum and maximum number of TDM samples (vancomycin labs) collected for each patient?
2. *Check patient descriptors for missing values or incorrect values*
   1. Some patients have multiple measurements for weights/heights occurring at the same time. Impute these values with the mean value.
   2. Using either histograms or summary data tables, explore patient height, patient weight, patient age, serum creatinine and vancomycin levels.
   3. Do all the values seem reasonable?
   4. If any of these measurements seem like outliers, use your clinical judgement to either correct/impute these values or remove these patients from analysis. Justify your reasoning.
3. *Convert regimen data to an easier-to-use format*
   1. Regimens have been provided as character strings in a single column. Using either base R’s regex functions or the stringr package, extract the dose quantity, dose unit, dose interval, and dose interval unit.
   2. Use a histogram or summary table to check the values of the doses and intervals. Do all the values seem reasonable?
   3. If any doses seem like outliers, use your clinical judgement to either correct/impute these values or remove these patients from analysis. Justify your reasoning.
4. *Calculate the duration of stay* (i.e., the difference between the discharge time and the admission time).
   1. Hint: the lubridate package has many useful functions for performing calculations with dates. Base R also has useful functions like `difftime` and `as.POSIXct` (R’s datetime object).

**Calculating Pharmacokinetic Parameters**

1. *Write a function to calculate pharmacokinetic parameters (CL, V).*
   1. One function should take baseline serum creatinine, baseline weight, and age as arguments, and return clearance.
   2. One function should take baseline weight as an argument and return volume of distribution.
   3. Use the underlying pharmacokinetic equations found in this published article. (<https://www.ncbi.nlm.nih.gov/pubmed/23340565>)
2. *Write a function to estimate area under the curve (AUC) at steady-state, assuming a 1 compartment model.*
   1. The function should take clearance and the total quantity of drug administered in a 24 hour period as arguments, and return AUC.
   2. Hint: you can use
3. *Write a function to estimate glomerular filtration rate*.
   1. For pediatric patients, use the Revised Schwartz Equation. For adult patients, use the Cockcroft-Gault equation.
      1. Cockcroft-Gault equation:
         1. Age: years
         2. Weight: kg
         3. SCr: mg/dl
      2. Revised Schwartz equation:
         1. Height: cm
         2. SCr: mg/dl
4. *Use the functions you wrote above to estimate AUC for each regimen and eGFR for each patient’s maximum serum creatinine measurement.* 
   1. This step will require joining two datasets. Use the column ‘blinded.id’ to match observations in the two tables.

**Deriving The Outcome Variable (Acute Kidney Injury)**

1. *For each patient, calculate an Acute Kidney Injury (AKI) outcome variable* using the RIFLE Classification system (defined here: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5094385/#s2title>)
   1. Hint: Each patient will have 1 categorical outcome variable, one of c(“No\_Risk”, “Risk”, “Injury”, “Failure”).
   2. Hint: You may find it convenient to write a function to help with this calculation!
   3. Assume patient’s baseline GFR (Glomerular Filtration Rate) is 120ml/min
      1. GFRB = 120 ml/min
   4. Assess GFR decrease according to RIFLE criteria. Use maximum SCr lab value (or minimum GFR) obtained during a single course of treatment to classify each patient.

**Data Analysis**

*You may use graphs and/or regression models and/or other statistical tools to answer the following questions. The goal is to demonstrate your ability to use these tools rather than to test your understanding of statistical concepts.*

1. Do patients that develop an AKI typically have longer lengths of stay?
2. Is vancomycin exposure (AUC) associated with an increased risk of developing an AKI? What other patient characteristics or laboratory measurements are associated with developing an AKI (RIFLE score of Injury or Failure)?

**Data File Specifications**

*pt\_observations.csv*

* Facility two different institutions.
* blinded.id de-indetified patient ID, used to link the two data tables.
* Pt\_age age of patient in years
* Gender sex of patient, M = male, F = female.
* HOSP\_ADMSN\_TIME date and time of hospital admission
* HOSP\_DISCH\_TIME date and time of hospital discharge
* Pat\_Class Treatment setting.
  + Inpatient = Drug administered within the hospital
  + Hospital Ambulatory Surgery = After a surgical procedure for a patient (outpatient setting)
  + Observation = Unknown meaning
* Hosp\_Service Hospital Department
* Pt\_Weight weight of patient in lbs
* Pt\_Height height of patient in feet
* lab\_type type of lab; either vancomycin (drug) or serum creatinine
* Measurement\_TIME date and time of lab
* Result value of lab. For vancomycin, the result is in units of mg/L. For serum creatinine, the result is in units of mg/dL.

*medadmin.csv*

* blinded.id de-indetified patient ID, used to link the two data tables.
* Vanco\_sig Vancomycin dosing regimen, a string describing both dose amount and dosing interval
* Vanco\_Admin Vancomycin administration time.