

## **Application using R: Assignment Document –Module 8**

Again, you are working for Doctors without Borders (Médecins Sans Frontières) in northern Uganda and have collected information from a cohort of 160 adults evaluated at primary regional health clinics to understand the epidemiology of malaria and HIV infection in this region. You and your team have created new research questions and you have once again consulted with your team's biostatistician. Following the analysis in R, your consultant presents you with the output in HTML format and supplies the R Markdown file as well.

Your task:

Using the HTML formatted output, apply concepts of basic epidemiology and descriptive and analytic biostatistics to address the questions that follow. Students are invited to run the code as well and generate the same output using the .rmd R Markdown file, but this is not required. Students who choose to run the R Markdown code in R Studio must save the .rmd file and the Uganda Malaria HIV Data.xlsx file in the same directory! For example, if you would like your R Studio working directory to be your Desktop, you must save all files to Desktop. You can choose your own location for the working directory, but the .rmd file and the .xlsx file must be in the same location. Additionally, for the student interested in running the program in R Studio, one of the final research questions requires saving and loading another data frame, ugandarct.RData. In following the assignment as written, students will note the appropriate point for loading this data frame.

### **A. Incidence Rates**

1. Malaria Incidence Rate:  
Your biostatistician has calculated a malaria incidence rate of 65.7/1000 person-years(rounded). A Northwestern University MSGH student is collaborating with your team and asks you for assistance in interpreting this value. Share your interpretation. (1 point)
2. HIV Incidence Rate:  
The MSGH student asks you why your team is interested in the HIV incidence rate. Share one reason why your team might want to know this rate. (1 point)
3. Death Incidence Rate:  
The MSGH student recalls learning about adjusted death rates. While adjusted rates are not reported for this data, the student asks you to share what factor is generally adjusted for in this process and under what circumstances adjusted death rates are likely to be reported. In other words, when do epidemiologists generally require adjusted rates? Share your response. (1 point)

## B. QQ Plots

1. Your MSGH student is very interested in recalling the fundamentals of creating and using QQ plots. In planning your response, you review a [video](#) explaining QQ plots. The student has the afternoon off and wants to recreate one of the plots by hand as demonstrated in the video. The student asks how many quantiles should be created for our data and if the comparison normal curve should be divided into the same number of quantiles. Next, the student asks how to interpret the plot when the created points do not follow the line. Provide your response to these questions. (1 point)
2. Exhausted by the tedious task of manually creating a QQ plot, the student inquires why the biostatistician would be interested in determining if variables are normally distributed. Specifically, how might this information impact the planned analysis? Provide your response. (1 point)
3. Before the student inquires further, you preemptively review parametric and nonparametric statistical tests. Share one point of comparison and one point of contrast. (1 point)

## C. Correlation for Continuous Variables

1. HIV RNA and CD4:  
The MSGH student has completed the internship with your team, thanked you for sharing your insight, and leaves you on your own for the next questions. For the HIV infected patients, evaluate the correlation between HIV RNA (viral load) and CD4 cell count ( a measure of immune function). What is the correlation coefficient? Is this association statistically significant? Support your conclusion. How do you interpret this? (1 point)
2. Hemoglobin and BMI:  
For all the subjects, evaluate the correlation between hemoglobin (hgb) (measure of anemia) and BMI. Note that your consultant first confirmed the distribution of the BMI variable before proceeding with the analysis. Does the BMI variable appear approximately normally distributed? Does the consultant use a parametric test? What is the correlation coefficient? Is this association statistically significant? Support your conclusion. How do you interpret this? (1 point)

## D. Associations for Categorical Variables

1. Condoms and HIV Infection:  
Is condom use (dichotomized as EVER/NEVER or YES/NO) associated with HIV infection? Support your conclusion. Report and interpret the relative risk, attributable risk, and number needed to treat. While some of

the programming steps are different from the code created by your biostatistician, you may find this [video](#) helpful for interpreting the R output. (1 point)

2. Bed Nets and Malaria Infection:

Is bed net use (dichotomized as EVER/NEVER or YES/NO) associated with malaria infection? Support your conclusion. Report and interpret the relative risk, attributable risk, and number needed to treat. (1 point)

## E. Logistic Regression

1. Factors Associated with HIV Infection:

Using multiple logistic regression, determine which of the following exposures (or risk factors) are independently associated HIV infection: age, sex, village, marital status, BMI category, number of sex partners, condom use. Support your conclusion. You may find the Module 7 Optional Videos on interpreting regression helpful. (1 point)

2. Factors Associated with Malaria Infection:

Using multiple regression, determine which of the following exposures (or risk factors) are independently associated with malaria infection: age, sex, village, BMI, hemoglobin, bed net use (days). Support your conclusion. (1 point)

You are planning to perform a large clinical trial of a novel insecticide treated bed net for malaria prevention in your region of northern Uganda. You will randomize healthy adults in the community 1:1 to the new bed net intervention or the standard of care, then follow them for one year. You hypothesize that the new bed net intervention will decrease one-year malaria infection cumulative incidence by 50%. Assume that the yearly cumulative incidence of malaria in this region is 15%, type 1 error ( $\alpha$ ) = 0.05 and power ( $1 - \beta$ ) = 0.80.

## F. Power Analysis

1. Sample Size:

You and your colleagues know that it is very important to get together with your biostatistician consultant early in the planning stages of a clinical trial. Together, you review your research question and determine the sample size needed for this clinical trial. How many individuals do you need to enroll in each group (one group randomized to receive the bed net intervention and one group randomized to receive the standard of care) if you hypothesize a decrease in the one-year malaria infection cumulative incidence of 50%? How would the sample size for each group change if you hypothesize a 75% decrease in one-year

cumulative incidence? How many individuals are needed in each group if you hypothesize a 25% decrease? Note that 'n' in the output represents the sample size needed for each group (bed net intervention and standard of care). (1 point)

For your next research question, you are interested in studying the effects of a new anti-malarial medication. You are pleased to learn that your team will be working with a Northwestern University MSGH student again. You wonder if this student will have as many questions as the last student. You will soon learn that, yes, this student has plenty of questions for you!

From your cohort in northern Uganda, you enrolled 50 malaria-infected subjects into a randomized clinical trial (RCT) of this new anti-malarial medication. The subjects were randomized to receive either this medication or a placebo pill along with the standard of care malaria treatment in this region. The study is testing whether the new medication in addition to the standard of care improves clearance time (in days) of malaria parasites from the blood. You have collected data on the 50 individuals enrolled in this RCT and your biostatistician colleague has created a data frame, `ugandarct`, with the data on these subjects. In `ugandarct.RData`, you have the following:

18 variables on 50 subjects:

`Subj_no` -- Subject number

`DOB` -- Date of birth

`Age` -- age at baseline visit in years

`entry_dt` -- Date of study baseline visit

`DOD` -- Date of death

`marital_status` -- Marital status (Married, Single, Divorced or Widowed)

`height` -- Height in inches at baseline

`height_m` -- Height in meters at baseline

`weight` -- Weight in kg at baseline

`BMI` -- body mass index at baseline

`BMI_cat` -- BMI category at baseline

sex -- Biological sex (Female or Male)

village -- Location of primary health clinic (Yumbe, Noko, Omugo, Kara, or Lori)

hgb -- Hemoglobin in mg/dL at baseline

parasitemia -- % level of malaria parasitemia in the blood before treatment (baseline)  
(measure of infection burden)

mal\_dx\_dt -- Malaria diagnosis date

RCT\_treat -- Study treatment group (1=active antimalarial study drug or 0=placebo)

time\_clear -- time (days) from starting study treatment until clearance of parasitemia

## G. Baseline Characteristics of Research Subjects

### 1. Categorical Variables:

Your new student asks you why you are concerned about checking for balance in baseline characteristics and if the baseline characteristics are indeed balanced between the active study treatment and placebo groups. Evaluate the variables: marital status, sex, village, BMI category. Explain to the student why an early step in RCT data investigation involves analyzing if baseline characteristics are balanced between the randomized groups. Determine if the listed variables are balanced and support your conclusions. (1 point)

### 2. Continuous Variables:

For the continuous variables, are the baseline characteristics balanced between the active study treatment and placebo groups? Evaluate the variables: age, height, weight, BMI, hemoglobin, parasitemia. Support your conclusions. (1 point)

## H. Addressing Your Research Question about New Anti-malarial Medication

### 1. Parasitemia:

Your biostatistician applied statistical tests to see if your 2 randomly created study cohorts had balanced baseline characteristics and now, you're ready to begin to address the research question about the new anti-malarial drug. You wonder if you will find evidence of efficacy. Your research team measured the number of malaria parasites in the blood (parasitemia) in the research subjects. Your

MSGH student excelled in biostatistics and shares that your consultant began this section of the analysis by checking to see if the variable of interest, time\_clear, or time to clearance, (how much time does it take for the malaria parasites to clear from the blood?) is normally distributed. The student helpfully points out that the consultant used a histogram and plotted the variable against the normal distribution on a QQ plot. This clever student realizes that the continuous variable, time\_clear, that we are interested in comparing between our study groups, is not normally distributed. Therefore, analysis proceeds using nonparametric methods. The student decides to test your recall of data analysis and biostatistics and asks you if there is a statistically significant difference in time to clearance between the active treatment (drug) group and the placebo group. The student points out that your consultant also created a boxplot that presents the data in a different way. Share your response to your student and support your conclusion. (1 point)

2. Clearance time:

As another way to study if treatment with the new anti-malarial medication influenced the amount of time it takes for subjects to clear malaria parasites from the bloodstream (time to clearance, the dependent, or outcome, variable of interest), your biostatistician set up a linear regression equation. The equation contains the primary independent variable of interest, RCT\_treat (study treatment group- active drug or placebo) along with a few other variables (potential confounders). Your student has forgotten why we would use linear regression here versus logistic regression. Share your response to the student's question and help the student interpret the linear regression output. Did the new anti-malarial medication have an effect on the outcome that was independent of other potential confounders? Support your conclusion. (1 point)

## I. Power

1. Power of the Study:

Your student asks you about the power of the study and additionally, asks you to interpret the power of the study. What does the power of the study measure? (1 point)

## J. You're All Done 😊

Your MSGH student thanks you for providing the opportunity to review epidemiology and biostatistics concepts by addressing these research questions!