# BIOS 6611 Final Project

**Due by Thursday, December 10, 2020 by 10:00 pm to Canvas Assignment Basket**

**The Overview**

This project is intended to give you hands-on experience with a topic that you will choose from the list of six below that involve analysis of data from actual studies. Alternatively, if you have a topic or dataset of interest, you can email/talk with Alex about using that for your project instead.

For your chosen topic, you'll prepare a 3-page summary (single spaced; strict maximum; references can extend beyond the 3-page limit) in the form of a brief scientific report. For the data analysis projects, you'll find a general outline for this format on p. 3 of this document. An example is posted on the Final Project page of the Canvas course.

The rubric that will be used for grading is posted on our Canvas Final Project assignment page. In general, you will be expected to write in the style of a journal article/brief scientific report (i.e., avoiding overly casual language and overuse of "I/we feel"). The following sections/subsections will be expected (although you don't have to necessarily use these names):

- Introduction or Background
- Materials and Methods, Study Design, etc.
- Statistical Methods, Statistical Analyses, etc.
- Results (including a "Table 1" summarizing descriptive features of the data, at least one figure, a table of the analytic results or results in-text if brief, etc.)
- Discussion or Conclusion

The structure and a brief outline is discussed a little more on Page 3, with some resources highlighted.

The six available datasets have brief descriptions at the end of the document and in separate files, but come from either UCD studies or real studies with data contributed by the American Statistical Association's TSHS Section:

**UCD Datasets:**
- Parkinson's Exercise Trial
- Mole Development in Colorado Kids

**TSHS (Teaching Statistics in the Health Sciences) Datasets:**
- Colorectal Cancer: Seasonal Effect and Vitamin D Study
- Surgery Timing and Outcomes
- Prostatectomy and Blood Transfusion
- Licorice Gargle for Thoracic Surgery

**Analysis of a UCD/TSHS Dataset (or propose a different dataset)**
You'll formulate one research question, develop an analytic plan to address the question, and present the results and interpret them. You can choose one of the datasets provided for your project, or propose a different dataset you are working with or find interesting.

The UCD datasets have the file names, variable names, and descriptions at the end of this document. The TSHS datasets have their own PDFs with background and study variables. Each of the data analysis projects has its own subsection on the Canvas page for the Final Project. On Canvas, you will find the data and the resources available for each project. You can use either R or SAS, or some of each, to conduct your analysis.

**Develop one question and analyze data to answer it**
The resource material for each project will give you an idea of the potential interrelationships among the variables included. Use those materials to develop one question that could be answered with the data you have.

Many of these datasets include longitudinal data. If you choose to use longitudinal data you should choose two time points of your choice to compare (e.g., baseline and 1 year). You will learn to work with the longitudinal data in detail next semester. For now, you will use methods appropriate for uncorrelated data, e.g. by taking differences and treating the difference as the outcome of interest. For any observational study it will be important to identify *potential* confounders of the relationships between the variables you choose to focus on.

**Your project must include the following aspects:**

**Apply resampling to understand sampling variability of parameters of interest *or* to calculate a permutation test p-value**
You should investigate and summarize the bootstrap sampling distribution of at least one parameter estimate of interest related to your study question or calculate the permutation test p-value for your comparison.

**Apply Reproducible Research (RR) principles for reporting**
In addition to the 3-page summary, you should include a brief Appendix with the R or SAS code that you used to perform the analysis. This will ensure that anyone looking at your analysis and summary will be able to know exactly what you did.

***Be sure to*:** Comment your program, attach useful names to any variables you create and analyze, and attach labels to variable values using formats.

## Example Outline of Brief Scientific Report

Introduction with brief background – 0.5 page

Question or hypothesis – 0.125 page

Materials and Methods – 0.25 page

Analysis Plan – 0.25 page

Results – 1.5 page
- “Table 1”
- Descriptive or other plot(s)
- Analytic results table

Discussion and Conclusion – 0.375 page
- Include limitations

References (not included in 3-page limit)

Appendix for Reproducible Research Goal (not included in 3-page limit):
- SAS or R code

**Refer to example:** AJPH Example Brief Report.pdf, located on the Final Project Canvas page

**General Writing Resource:** Tips on Writing Results for a Scientific Paper Amstat News - Cummins - Sept 2009.pdf, located on the Final Project Canvas page

## Reproducible Research

You will need to follow Reproducible Research principles in carrying out your project. Adherence to the principles of reproducible research is essential to the reporting of scientific results.

**Be sure to familiarize yourself with some of the resources on this topic listed below.**

An outstanding example of reproducible research by Keith Baggerly and Kevin Coombes of the MD Anderson Cancer Center Bioinformatics and Computational Biology department is posted in the Paper Repository ("Deriving Chemosensitivity from Cell Lines: Forensic Bioinformatics and Reproducible Research in High-Throughput Biology" in the *Annals of Applied Statistics* by Baggerly and Coombes, and *Cancer Letter* 1, 2, and 3 by Paul Goldberg).

*RR Resources:* (Canvas Paper Repository)
Gentleman and Temple Lang paper – Statistical Analysis and Reproducible Research; from the Bioconductor Project

Roger Peng 2009 Editorial in *Biostatistics*

Roger Peng 2011 article in *Science*

Frank Harrell presentation on the use of R in clinical trials research

IOM Report on Evolution of Translational Omics

Lehrer 2010 New Yorker The Truth Wears Off

Dynamic/computable documents using R
http://yihui.name/knitr/
http://rmarkdown.rstudio.com/

Dynamic/computable documents using SAS
http://sites.northwestern.edu/stattag/

Recommended, but not required:

Use one of these (or another) computable document applications.

Git and GitHub – version control, collaboration tools
https://git-scm.com/
https://github.com/

## UCD Data Analysis Project – A Randomized Clinical Trial of Exercise in Early to Mid-Stage Parkinson's Disease

**Lead Investigator:** Margaret Schenkman PhD

**Data are in a .csv file:** PD Exercise RCT Selected Secondary Outcomes - Wide.csv

**Summary:**

121 participants randomized into one of 3 exercise groups
33 variables
Repeated outcomes: Baseline (0 months), 4 Months, 10 Months, 16 Months; some missing data; suffix for outcome measures below denotes when the measure was taken

**Data Dictionary:**

| Variable/Field Name | Label/Attributes |
|---|---|
| Participant | id number |
| Group | 4 = Home Exercise<br>5 = Flexibility, Balance, and Functional Training<br>6 = Aerobic Conditioning |
| Gender | 1 = Male<br>2 = Female |
| Age | years |
| YearsDx | Years with PD |
| HYStage0, HYStage4, HYStage10, HYStage16 | Hoehn and Yahr stage of PD – scale from 1 (lowest) to 4 (worst) in increments of 0.5 at baseline, 4, 10, and 16 months |
| FiveM_Wk0, FiveM_Wk4, FiveM_Wk10, FiveM_Wk16 | Five meter walk in number of steps at baseline, 4, 10, and 16 months |
| FiveM_Tm0, FiveM_Tm4, FiveM_Tm10, FiveM_Tm16 | Five meter walk in seconds at baseline, 4, 10, and 16 months |
| TUG0, TUG4, TUG10, TUG16 | Timed Up and Go in seconds at baseline, 4, 10, and 16 months |
| UPDRS0, UPDRS4, UPDRS10, UPDRS16 | Total score on UPDRS (see scale information in resource document) at baseline, 4, 10, and 16 months |
| SixMn_Wk0, SixMn_Wk4, SixMn_Wk10, SixMn_Wk16 | Six Minute Walk in meters at baseline, 4, 10, and 16 months |
| LEDD0, LEDD4, LEDD10, LEDD16 | Levodopa equivalents (mg/day) (commonly prescribed medication for PD) at baseline, 4, 10, and 16 months |

**Additional Resources:**

UPDRS,H&Y, S & E_MedEl_tool.doc
UPDRS Background Paper.pdf
Schenkman at al 2012 PTJ Exercise Early MidStage PD 16 Month RCT.pdf

## UCD Data Analysis Project - Mole Count Study in Colorado Children

**Lead Investigator:** Lori Crane PhD

**Data are in an Excel spreadsheet:** Mole Count Data 2004-2008.xls

**Summary:**

472 children age 6 followed from baseline to age 10<br>
15 variables<br>
Longitudinal study of mole development over five years

**Data Dictionary:**

| **Variable/Field Name** | **Label/Attributes** |
|---|---|
| Respondent Code Number | id number |
| oca2 status | 0 = gg<br>1 = ga<br>2 = aa<br>9 = missing |
| gender | 1 = Female<br>2 = Male |
| Hispanic | 0 = No<br>1 = Yes |
| molecount2004 | Number of moles in 2004 |
| molecount2005 | Number of moles in 2005 |
| molecount2006 | Number of moles in 2006 |
| molecount2007 | Number of moles in 2007 |
| molecount2008 | Number of moles in 2008 |
| eyecolor | 1 = blue, green or combo<br>2 = light/dark brown<br>3 = hazel |
| baseskincolor | Skin color based on a continuous score, higher is darker |
| haircolor | 1 = blonde<br>2 = red<br>3 = brown<br>4 = black |
| number vacs birth thru 2005 | Total number of waterside vacations from birth through 2005 |
| number vacs birth thru 2006 | Total number of waterside vacations from birth through 2006 |
| number vacs birth thru 2007 | Total number of waterside vacations from birth through 2007 |

**Additional Resources:**

Mole Study R01 July 2009 Draft.doc<br>
Crane et al. 2009 – Nevus development in children (pdf)<br>
Pettijohn et al. 2009 – Waterside vacations and nevus development (pdf)<br>
Crane et al 2012 AJPM Mailed Intervention Sun Prot Children RCT (pdf)

## TSHS Data Analysis Project – Colorectal Cancer and Seasonal Effect

**Data are in a .csv file:** Seasonal_Effect.csv

**Summary:**

Prospective cohort study
2919 participants divided into seasons (spring, summer, fall, winter)
14 variables
Repeated outcomes: none

**Additional Resources:**

Data dictionary: SeasonalEffect_dictionary.pdf
Introduction/Background: SeasonalEffect_Introduction.pdf

## TSHS Data Analysis Project – Surgery Timing and Outcomes

**Data are in a .csv file:** Surgery_Timing.csv

**Summary:**

Retrospective cohort study
32001 participants
25 variables
Repeated outcomes: none

**Additional Resources:**

Data dictionary: Surgery Timing Data Dictionary.pdf
Introduction/Background: Surgery Timing Dataset Introduction.pdf
Published Paper: Sessler et al., "Operation Timing and 30-Day Mortality After Elective General Surgery" (pdf)

## TSHS Data Analysis Project – Prostatectomy and Blood Transfusion

**Data are in a .csv file:** Blood_Storage.csv

**Summary:**

Retrospective cohort study
316 participants
20 variables
Repeated outcomes: none

**Additional Resources:**

Data dictionary: Blood Storage Data Dictionary.pdf
Introduction/Background: Blood Storage Dataset Introduction.pdf
Published Paper: Cata et al., "Blood Storage Duration and Biochemical Recurrence of Cancer After Radical Prostatectomy" (pdf)

## TSHS Data Analysis Project – Licorice Gargle for Thoracic Surgery

**Data are in a .csv file:** Licorice_Gargle.csv

**Summary:**

Randomized control trial
236 participants divided into seasons (spring, summer, fall, winter)
19 variables
Repeated outcomes: none

**Additional Resources:**

Data dictionary: Licorice Gargle Data Dictionary.pdf
Introduction/Background: Licorice Gargle Dataset Introduction.pdf
Published Paper: Ruetzler et al., "A Randomized, Double-Blind Comparison of Licorice Versus Sugar-Water Gargle for Prevention of Postoperative Sore Throat and Postextubation Coughing" (pdf)