**ADVANCED DATA ANALYSIS**

**PROJECT**

**CONCEPT PROPOSAL INSTRUCTIONS**

**TITLE OF PROJECT:**

**STUDENT(S):** \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

*The proposal should not exceed two pages (one page is fine), single spaced*

**BACKGROUND** (Why your project is important 1-2 paragraphs)

* Provide a brief but well-documented background support (literature review) for the importance of your project
* Be sure to answer “so what?” – who should/would care about this, to whom it will be important
* Be sure to use references properly (reference list does not count toward 2-page limit)

**OBJECTIVES** (1 paragraph)

* State objectives *(“The objectives of this project are to: (a), (b), (c).”).* These are the research question or questions that will guide the selection of your dataset and that will help you focus your analysis.

**Examples using cancer data**

**The objectives of this project are to:**

*1. Determine whether insurance status at diagnosis is associated with higher cancer stage at diagnosis*

*2. Determine whether insurance status at diagnosis is associated with worse survival*

*3. Determine whether race modifies the association between insurance status at diagnosis and stage at diagnosis*

*4. Determine whether race modifies the association between insurance status at diagnosis and survival*

**APPROACH** (1-3 paragraphs)

* Provide a *detailed* description of the methods for your project
  + Include the following:
    - **Data source** (include a description)

**Example using cancer data:**

*We will use the Surveillance, Epidemiology and End Results 9 results database (ref).*

* + - **Study population** (what people will you select from the data source?)

**Example using cancer data:**

*Our study will include females diagnosed with first primary malignancies of the breast who were 40-65 years old at diagnosis between 2009 and 2010.*

* + - **Variables.** What variables will you use in your analysis? Will you do any recoding of variables? How will your dependent and independent variables be defined?

**Example using cancer data:**

*We will use the following variables in our analysis: survival months, sex, race, age at diagnosis, and stage at diagnosis. Our dependent variable will be time to cancer death and our independent variable will be insurance status at diagnosis.*

* + - **Reproducibility plan.** How will you ensure reproducibility of your analysis when it is completed?

**Example:**

We will use R Markdown to document our coding and our dataset will be stored at a location that is accessible to all project members. To ensure reproducibility, we will have two people write (or run and check) the analysis code and make sure that they achieve similar results.

* + - **Statistical analysis.** What statistical software, models, and tests will you use to conduct your analysis?

**Example using cancer data:**

For all analyses, we will use R software (version number). We will perform descriptive statistics on our study population to describe their characteristics (i.e. total number in study population, sex, race, education level). For survival analyses, we will use Cox proportional hazards regression to model the time to cancer death in those who are insured at diagnosis compared to those who are not insured at diagnosis.

**APPENDICIES**

* References
* Project timeline (Break down tasks to get you to the point of having a finished project for presentation by 4/22). All groups should finish by 4/22 even if they do not present until 4/29
* Roles (Identify who will do what to meet the deadline for the project)

**NOTE:** All individuals/groups are encouraged to make an appointment with me to go over your concept proposal once you have a draft to ensure feasibility/clarity.