Clinical Data Partner Activity (Written)

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

   Categorical variable: not quantitative/numbers, typically are limited to a few groups/types, used in demographic-based analysis (e.g. race, sex, etc.)

   Discrete variable: variables that can be measured by counting, and that aren't necessarily changing constantly/continuously (e.g. number of people in a room, amount of money paid in a transaction)

   Continuous variable: variables that can be measured at various time points, but that typically always change; it is usually not possible to capture the full scope of a continuous variable unless you are constantly monitoring it within a certain period (e.g. height of a plant, speed of an accelerating car).

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?

   Our chosen variable is the age at initial pathologic diagnosis. There are no N/A values in the data frame for this specific variable.

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

   Even though age is a continuous variable in the real-world (e.g. we are constantly aging and our age is changing all the time), because the age is measured in years in this data frame, it is considered a discrete variable. This variable is collected at checkups in clinical settings.

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

   Link 1: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6108139/
   Description: This article examined patients at varying age-at-diagnoses, looking at reported quality of life and symptoms of prostate cancer. Symptoms included fatigue, pain, and nausea, and quality of life factors included physical functioning, cognitive abilities, and social functioning. The article found that patients who were older at diagnosis experienced harsher symptoms from prostate cancer, and have lower quality of life.

   Link 2: https://bmccancer.biomedcentral.com/articles/10.1186/s12885-022-09454-y
   Description: This article examined post-metastasis mortality in women with breast cancer, relating it to their age at initial diagnosis. This article looked at 1636 women at the West

Chian Hospital between 1989 and 2020, and found that elderly patients face a much higher risk of post-metastasis mortality compared to patients with younger age-at-diagnoses.

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.

   Our chosen variable is drug_name, from the clinical.drug data frame. This variable is categorical, and documents the type of drug that was used to treat each patient.

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.

   Hypothesis 1: The type of drug depends on the age at diagnosis of a patient – we hypothesize that older patients will likely be given a different drug during their course of treatment compared to younger patients, because different age groups may react differently to various drugs.

   Hypothesis 2: The age at diagnosis leads to a worse breast cancer survival and outcome.

   Hypothesis 3: The type of drug used to treat a patient is correlated with the outcome of a patient's 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 graphs that were generated confirmed the hypotheses put forth in the previous question. For example, if the age at diagnosis was considered "young" (e.g. < 50 years old, the overall survival of the patient was much greater than the age at diagnosis being "old." Additionally, it was found that some drugs are correlated with certain survival outcomes. Because the total number of drug names exceeded 400, only the top 9 drugs were analyzed in our graph – and we found that drugs like Doxorubicin and Cyclophosphamide typically correlated with poorer survival, compared to drugs like Taxol and Adriamycin that had a higher survival rate. This is further correlated with the age category of the patients, as the survival time links both of these factors. The significant of our findings lays in the fact that we showed some drugs being correlated with worse survival, which was also correlated with an older population of patients; ultimately, this means that there is a clear link between these three variables that may be explored at a greater scale in the future.