Lab 8 - Predicting Breast Cancer Relapse (STAT 212)

NAME 1 – NETID

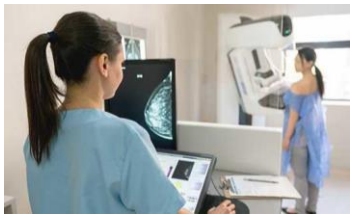
NAME 2 – NETID [if applicable]

NAME 3 – NETID [if applicable]

Formatting Requirements:

- Please submit your lab report as a **pdf** to Gradescope.
- When you upload to Gradescope, please **match pages** with the **question number**.
- Be sure that all **group members** are **added** in your submission to Gradescope (click view/edit group on the top right of the page once shown your final submission after matching pages).

Assignment Overview:



- In this assignment, you will be reading and summarizing key points from the article titled: "A Gene Expression Signature that Can Predict the Recurrence of Tamoxifen-Treated Primary Breast Cancer."
- The goal of this lab is to identify the aims of this study, the design, the statistical results, and the claims they are making from those results.

Tips for reading research articles:

- You won't understand a lot of what is being said (perhaps even half of what is being said), and that's ok! Research articles are often full of jargon, especially with regard to instrumentation and software use. Focus instead on making sense of the study's primary aims and contributions.
- Look for key words like "aim" or "goal" or "objective" or "contribution" and take notice—these are the milestones to guide you through.
- Whenever you see a term used multiple times, but aren't sure what it is, take a few seconds and search it online!
- Abstracts are great at helping you pull out key details. You should read this first, then at various stages of reading the rest of the paper, come back and read it again! Each time, it will make a little more sense.

Extra things you might want to know from *this* article

- This study talks about "Training" data and "Validation" data. That means that they used the training data to create a model to predict the likelihood of relapse given certain information. Then they used the validation data to test out this model on an independent sample to see how effective their predictions were.

Read the Abstract and the Introduction (from beginning until the section "Patients and Methods")

Question 1 (4pts): Briefly discuss the **aims** of this study. A complete answer will 1) identify the research problem being addressed, 2) describe the population of interest (who can we generalize our findings to), and 3) describe what the researchers hope to contribute with this study. *(suggested 70-110 words)*

By this point, you should know what the authors mean by "gene signature." If you are still fuzzy about what that means, do a quick web search of this term.

Skim the "Patients and Methods" section

Note that there is a lot of jargon and additional details that go beyond what we can really understand and discuss. Don't worry if you don't understand a lot of it! Just focus on these questions:

Question 2 (6pts): Focusing on the subsection titled: "*Patients and treatment*," describe the sample used in this study. How many cancer patients did we collect samples from? What do we know about these patients and what might we like to ask? What external validity threats might be relevant here? *Hint: Thorough answers will use terms from Chapter 3 related to external validity. (suggested 120-200 words)*

Question 3 (3pts): Did the researchers conduct an experiment or an observational study? Briefly explain your answer. *(suggested 25-40 words).*

Check the table on page 1746, including the description below it.

Question 4 (4pts): Briefly describe which situations would be labeled as what in sentence form. *NOTE: that **being "relapse free" is the "condition of interest"** here, which might feel strange because in class, we usually describe the illness or problematic result as the condition of interest, so just be aware they did it kind of "backwards" to most of our examples (suggested 12-25 words each, in sentence form.)*

- True Positive:
- False Positive:

Question 5 (3pts): Check the caption below the table on page 1746. Which measure represents the proportion of patients with the relapse-free gene signature who actually remained relapse free? (This is a multiple choice question. **Bold** or clearly identify your answer)

A. Sensitivity
B. Specificity
C. Accuracy
D. Positive Predictive Value
E. Negative Predictive Value

Skim the "Results" section

Question 6 (6pts): As you can see in Figure 1, the researchers identified 23 genes that were associated with patients who were relapse-free and 13 genes that were associated with patients who experienced relapse. Using these signatures, the researchers sorted all patients as either best fitting the "non-relapse" gene signature group or "relapse" signature group. Now, look at Table 2, Univariate Analysis. This table is displaying **the odds of being relapse-free** given different given conditions *(for each line, the condition listed before "vs" would be the group that has the higher odds in each comparison).*

Focus on the **Validation set (part B),** as that tells us the model's predictive power with the validation set of 83 tumor samples.

Part a) What is the estimated odds ratio of being relapse-free for someone identified as having the relapse-free gene signature? *(just listing the odds ratio value is enough!)*

Part b) According to the table, which of the following conditions are the researchers at least 95% confident are associated with higher odds of staying relapse free? This question may have one or more answers. (This is a multiple choice question. **Bold** or clearly identify your answer)

A. Being 55 or over
B. having a tumor size <20 mm
C. having an NPI less than or equal to 3.4

Part c) Briefly explain how you chose your answer(s) for part b *(suggested 20-40 words)*

Skim the "Discussion" section

Question 7 (4pts): Briefly describe the contributions this article made. What implications do the researchers suggest from these findings? *The 3 paragraphs on page 1751 are a nice place to focus. (suggested 60-100 words)*