Lab 8 - Predicting Breast Cancer Relapse (STAT 212)

NAME 1 – Varenya Jain; varenya3

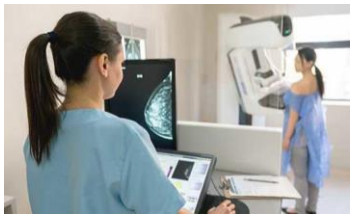
NAME 2 – NETID [if applicable]

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 (5pts): 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)*

The aim of this study is to identify a genetic predictor of the tamoxifen-treated primary breast cancers relapsing in order to help in the therapeutic management of estrogen receptor-positive cancers.

The "population" of interest is primary tumors from patients who received adjuvant tamoxifen, the analysis of which led to accurate classification of patients based on their relapse history by a gene-signature technique.

Essentially the purpose is to see whether a certain biotechnology can identify if a patient's tumor relapse correlates with tamoxifen treatment – an estrogen regulator to help reduce breast cancer in

women – and if analysis reveals that this method is a good model for classifying tumors and validating existing datasets.

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 and how representative they are of a larger population. A complete answer should include:

- How many cancer patients did they collect samples from for this study?
- What background or characteristics do these cancer patients have in common?
- What external validity threats might be relevant here?
- What questions could we ask the authors to better judge how well this sample generalizes to a larger population?

Thorough answers should use external validity threat terms from the notes. (suggested 120-200 words)

They collected 132 primary tumors, i.e. the first tumors identified to start the symptoms of cancer disease. These cancer patients all received adjuvant tamoxifen, meaning a hormone regulator to aid the main treatment and help suppress secondary tumors in the body. These Carcinomas were cancers that formed in epithelial tissue - the lining of most internal organs. The cancer patients had all undergone initial surgery between 1989 and 2001 at the Cancer Research Center of Val d'Aurelle in Montpellier (the Bergonie´ Institute in Bordeaux, or the Department of Obstetrics and Gynecology of Turin).

These biopsies were stained with Hematoxylin and Eosin, a gold standard for medical diagnosis, to select samples with at least 50% of tumor cells. One threat to external validity could be the unknown response effectiveness in cancerous tissue with less than 50% of tumor cells. Another external validity threat would be the nature of these cancers' status: none of these patients received neoadjuvant systemic therapy (chemotherapy, radiation therapy, and hormone therapy) to reduce the initial cancer. Since cancer is often caused by random mutations, we do not know how tumors which have already received treatment will respond to another suppressive measure.

A method to test these threats would be to select a larger population of primary tumors and select broader screening variables under microarray gene expression analysis, allowing for generalization to a larger population. Another study might consider the weaknesses of gene set enrichment under SAM analysis to achieve a difference between R and RF tumors with a lower false discovery rate (lower than 'below 5%'), increasing the accuracy of identification of the differentially expressed genes.

Question 3 (5pts): Did the researchers conduct an experiment or an observational study? Briefly explain your answer. *(suggested 25-40 words).*

This was a purely observational study on frozen samples taken from already treated patients. The methodology described in the paper is purely statistical in analysis and summarizes genetic correlation using common pathobiological tools alongside computational algorithms (e.g. SAM, K-nearest, etc.)

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: **When the analysis reveals that the predicted relapse-free cancer patients are actually free of tumor relapse after Tamoxifen treatment.**
- False Positive: **When the analysis reveals that the predicted relapse-free cancer patients actually have relapsed tumors after Tamoxifen treatment.**

Values are calculated from the short DNA sequence protospacer adjacent motif (PAM is about 2-6 base pairs long) that follows the DNA region targeted for cleavage by the CRISPR system (ex. CRISPR-Cas9 complex), along with 70-mer oligonucleotide microarrays and the K-nearest Neighbor classification function

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>>>** **Positive predictive value = A / (A + B)**
E. Negative Predictive Value

Skim the "Results" section

Question 6 (7pts): 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!)*

For the Validation set (83 patients), the Univariate analysis of the 36-gene signature: RF vs R (Relapse-Free vs Relapse) Odds Ratio is 3.96 with a 95% Confidence Interval of (1.56-10.05), while the Multivariate analysis is 3.01 with a 95% Confidence Interval of (1.01-9.14).

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. (**Bold** or clearly identify your answer(s))

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

NPI: V3.4 vs >3.4 5.96 (1.25-28.33), Adjuvant!: <20% vs z20% 4.90 (1.75-13.69)

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

The 'NOTE' at the bottom of Table 2 on page 1748 includes: "Significant values are in bold characters". In addition, the "NPI" and "Adjuvant!" Predictive Factors have P-values of 0.013 and 0.002, respectively. Each is less than the "standard" 0.05 alpha-significance level. Also, the authors alluded to their importance by discussing the use of these variables as predictors for treatment effectiveness earlier in the paper in the "*Patients and Methods*" section.

Skim the "Discussion" section

Question 7 (5pts): 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)*

In short: the classifiers can be used in prognoses, the 36 GS can help in tailoring the therapeutic decisions of particular patient subsets, and large-scale gene expression profiling has a use-case in investigating cancer progression and resistance to endocrine therapies.

Long Response: The authors note that the main problem they faced in gene expression profiling is the "relatively small overlap between independently reported molecular signatures", meaning that the gene expression classification may be of poor use in the prognosis of some breast cancer. Another observation is the mentioning of certain important Protein Kinases - enzymes that regulate the biological activity of proteins - which are critical to Signal Transduction Pathways (chain reactions which transmit signals through/between cells to elicit certain responses). This relates to the primary treatment observed – Tamoxifen – which itself regulates hormones which in turn cause certain cellular responses. The authors go on to discuss that they along with other researchers (Paik et al. (15)) can tell from their findings that the classifiers used were able to predict the clinical outcome of Tamoxifen treated breast cancers; The classifiers used are good indicators of the efficacy of Tamoxifen treatments on breast cancer and may be used in general prognoses. The 36-gene signature needs to be investigated more but it likely of significance due to

the number of estrogen-related genes, and can currently be used to distinguish those who cannot benefit from Tamoxifen treatment, meaning these people may be candidates for alternative endocrine or chemical therapies.