**Slide 1:** Good morning everyone. Thank you for being here and listening to an update of our project: Survival Analysis Study in Pancreatic Ductal Adenocarcinoma Patients. My name is Simran Samra and these are my fellow Group 7 members: Hassan Ali, Jingyiran Li, and Lily Xia.

**Slide 2:** Pancreatic cancer has a 5-year survival rate of less than 10 percent, and it is highly aggressive as 80-90% of patients present with surgically unresectable disease. Additionally, there is a lack of clinically actionable biomarkers, making it difficult to diagnose. Pancreatic cancer will soon be the third leading cause of cancer death in Canada.

The most prevalent form of pancreatic cancer is pancreatic ductal adenocarcinoma (PDAC).

**Slide 3:** Despite tremendous scientific research, survival rates have not changed much during the last 20 years, and our understanding of this disease remains incomplete.

Taking this information into account, we developed this current project to perform a survival analysis in patients with PDAC.

**Slide 4:** Our goal was to examine the relationship between gene expression and vital status (alive or deceased).

We hypothesized that certain genes are more heavily expressed in PDAC patients that survive while other genes seem to associate with PDAC patients that do not survive.

Identifying gene targets would enable us to identify ideal chemotherapeutic drug candidates and create a tailored therapy regimen.

**Slide 5:** Our hypothesis will be investigated using the TCGA-PAAD dataset from The Cancer Genome Atlas (TCGA).

The dataset contains the expression of 44,084 genes from 177 resected tumors of patients suffering from PDAC. The libraries were sequenced on the Illumina HiSeq 2000. Gene expression levels are measured using Reads Per Kilobase of transcript per Million (RPKM) mapped reads in the log10 scale.

This dataset additionally contains subject demographics, including, sex, age, race, vital status, and other survival-related information. No control subjects are used since this dataset only reflects the cancer patient’s genome.

**Slide 6:** Our workflow consisted of the following steps: data wrangling, exploratory data analysis, expression analysis, feature selection on genes, classification analysis (alive or deceased), and survival analysis.

**Slide 7:** First the gene expression dataset was crossed-referenced with a meta-data containing the covariates, and then data wrangling was performed. Data wrangling is the process of transforming the format of raw data into another format that is more appropriate for downstream analyses.

Sample clinical data had multiple unfilled columns and they were removed. We mainly focused on patients’ age, sex, race, tumor stages, treatments, survival time and vital status. To allow for comparability of the effects of these on gene expressions, expression data were standardized to generate heatmaps

**Slide 8:** This slide shows a list of covariates included in our meta-data.

The first four variables listed are part of the TNM Staging System developed by The American Joint Committee on Cancer (AJCC). The TNM staging system is currently the most widely used prognostic factor for predicting survival in patients with pancreatic cancer. This staging system describes the spread of cancer in a patients’ body. It entails three measurements:

1. **T** which stands for the extent of the tumor. How large is the tumor and has it grown outside the pancreas into adjacent blood vessels?
2. **N** which stands for lymph nodes. Has cancer spread to nearby lymph nodes? How many of the lymph nodes have cancer?
3. **M** which stands the presence of metastasis. Has cancer spread to distant lymph nodes or distant organs including the liver, the lining of the abdominal cavity, lungs or bones?

Numbers after T, N, and M provide more details about each of these variables. Higher numbers mean the cancer is more advanced. Once a patient’s T, N, and M categories have been determined, this information is combined and a stage is assigned to the patient. The six different staging options are listed on the right.

**Slide 9:** Next, a series of exploratory data analyses (EDAs) were performed to visualize features and relationships of the covariates, such as age, gender, race, and pathological stages using density plots.

The density plot on the right shows that the mean age of females and males at the time of the index was similar, approximately 66 years old.

**Slide 10:** This density shows the age distribution across race. This plot highlights that a majority of the patients were caucasian, then African American and then Asian. However, a large portion of the patients did not provide their race. Age distribution of not reported individuals and Asians are bimodal around 45 and 75 years old.

**Slide 11:** The figure on the right is a PCA plot of the top 5,000 most variable genes. Here we have grouped the data by the six pathologic stages, and the data is color-coded by vital status: orange indicating the patient is alive and green indicating the patient is deceased. Two of the 177 patients were not assigned a stage value, and they are shown in the first PCA plot.

The key takeaways from this plot are 1) there does not seem to be an effect of vital status on the gene expression data and 2) a majority of the tumors were at stage IIB. At this stage: the cancer is confined to the pancreas and is no bigger than 2 cm across, and it has spread to no more than 3 nearby lymph nodes.

**Slide 12:**To assess differential gene expression, limma and empirical Bayes were used to determine the top 10 most influential genes on the vital status of cancer subjects. This heatmap on the right shows the correlation among these 10 genes

**Slide 13:** A principal component analysis (PCA) was conducted among the top 10 genes. Approximately 63% of the variation was due to the first component and ~ 8.3% was due to the second component. The first two components are able to explain ~ 70% variation of the total. We also clustered pathologic stages based on principal components 1 and 2 for the top 10 genes.

**Slide 14:** We then clustered the top 10 genes using k-mean algorithm with k equal to 2, and these two clusters of genes were plotted in 2D PC coordinate system.

**Slide 15:** In the hierarchical dendrogram (clustering), the samples are not well-separated when we stop the clustering algorithm at k equal to 5. It might be because of the similarity of gene expressions among all pancreatic ductal adenocarcinoma patients.

**Slide 16:** Lasso is a shrinkage regression method that helps us to do feature selection. When lambda equals to zero, it is just a linear regression model. But as the tuning parameter, lambda, increases, more variable coefficients are set to zero and among non-zero coefficients, more shrinkage is employed. The most important variables selected based on Lasso were listed on the sides.

**Slide 17**: The accuracy for the random forest algorithm is 70% while that for the AdaBoost algorithm is 73%. The AUC for the random forest is 0.8825 while that for AdaBoost is 0.8194. Both algorithms yielded similar results.

**Slide 18:**  A comparison of the top 10 genes by vital status using a violin plot and two-sample t-test gave a p-value of 0.7, given selected alpha=0.05 we did not detect a significant difference. Furthermore, no clear distinction could be made from a scatter plot of vital status and samples.

**Slide 19:**  We then sought to investigate a difference in the probability of survival between treatments by means of Kaplan-Meier curves. The clinical meta-file had therapy (chemo or radiation) assigned to each patient, another column indicated if the treatment was actually given (provided no metastasis or death occurred). The probability of survival was plotted for whether treatment was given further stratified by the type of treatment originally assigned.

Maximum survival ranges from 5-13.75 years. Interestingly, in the treatment group, a higher survival probability was observed in the radiation group over the pharmaceutical group. Additionally, in the treatment not reported group, the opposite trend was observed although we can not be certain for this group.

**Slide 20:** Next, we pooled the data to detect survival differences between treatment types (pharma and rad). Similar to the treatment positive group we observed a higher survival probability in the radiation group than the pharmaceutical group. However, with a p-value of 0.086, this difference is considered insignificant.

**Slide 21:** Further analysis of the radiation group stratified by treatment and treatment-naive conditions showed higher survival in the treatment group and with a p-value of 0.018 this difference is significant.

**Slide 22/23:** To summarize, no clear gene expression could be inferred between current- and non- survivors. Our sub-analysis did show that cases, where radiotherapy is required, have a higher probability of survival.

**Slide 24:** A major limitation in our analysis was the lack of mutational signature data. If we had that data we could have selected for top mutated pdac genes and attempted to find a significant link between vital status and expression of those mutated genes. Furthermore, the classification accuracy was limited in the classifiers used; an optimization strategy is needed. Stratification was a challenge with some of the clinical covariates as they could not be used as cluster factors. Lastly, it is possible that individual genes may not impact survival times, rather, an effective strategy would be to look at the total number of genes mutated per case (tumor mutational burden).

**Slide 25:** Thank you for listening to an update of our work.