Tessa Ferrari

Dr. Peter Calabrese

QBIO401

9 December 2022

**Final Project Written Report**

Uterine corpus endometrial carcinoma, or endometrial cancer, occurs when the tissue in the lining of the uterus, or the endometrium, experiences malignant and uncontrolled cell growth. It is the 9th most common form of cancer in the US, with 65,950 new cases in 2022 accounting for 3.4% of all new cancer cases in the country. For endometrial cancer, the death rate is 5.1 deaths per 100,000 people per year (National Cancer Institute, 2022). Because of the prevalence of this cancer, it’s important to understand the different risk factors that can affect endometrial cancer patients' survivorship of the disease. The Cancer Genome Atlas (TCGA) program provides publicly available cancer patient data that can help analyze these factors, including clinical, genomic, transcriptomic, and proteomic data. The age, race, and cancer stage of endometrial cancer patients as well as their gene expression (RNAseq) and mutation data (MAF) were analyzed in order to determine how they influence patient survivorship.

First 3 Kaplan–Meier plots (KM plots) were created, evaluating the survivorship of patients based on their age at diagnosis, race categories, and clinical cancer stages. To do this, clinical data of uterine cancer patients was downloaded from TCGA using the R package “TCGAbiolinks”. Death data was adjusted in order to account for non-reported death data, patients were categorized into 4 age groups (0-34, 35-54, 55-74, and 75+ years), and patient cancer stage was adjusted to the 4 main stages (rather than including substages). Then the R packages “survival” and “survminer” were used to create the final plots showing survival probability over time (in days). For the age group and cancer stage, the KM plot results were to be expected. Across the board, older age groups experienced lower survival probabilities over time than younger ones (with a p-value of p=0.0099). This is likely due to the fact that older patients are simply more prone to complications associated with age (whether or not those complications are cancer-related). Additionally, patients that were diagnosed with more advanced stages of cancer had a lower survival probability over time (with a p-value of p<0.0001). This is likely because more advanced cancer interferes more with the body, preventing life sustaining processes. Based on the race KM plot, most races seemed to have similar survivorship over time, with asian patients having slightly greater survivorship. However, this slight difference was determined to be statistically insignificant (with a p-value of p=0.59). The lack of conclusive results could be attributed to lack of comprehensive data points across all race categories, as the majority of patients self-identified as white.

In addition to looking into how clinical factors affected survivorship, genetic factors were also looked into. When picking which genes to look at for potential importance, two specific genetic syndromes were deemed as clinically significant to the incidence of uterine cancer in the population. The first is known as Cowden syndrome, which is a genetic disorder that causes high risk of various types of cancer and is associated with the genes *PTEN*, *KLLN*, and *WWP1* (ASCO, 2022). The second is known as Lynch syndrome, a genetic disorder that increases risk of developing colorectal, uterine, and other cancers and is associated with the genes *MLH1*, *MSH2*, *MSH6*, and *PMS2* (ASCO, 2021). These 7 genes were looked into to see how these 2 syndromes were related to clinical factors and survivorship.

In order to determine if there was any relationship between demographics/clinical data and the gene expression of uterine cancer patients, 3 boxplots were created of each of the 7 genes, comparing age, race and stage (21 boxplots total). To do this, transcriptome profiling data of uterine cancer patients was downloaded from TCGA using the R package “TCGAbiolinks”. Then, the overall RNAseq counts of each gene was compared between different groups using the R boxplot function. For most of the boxplots, not much difference was seen in gene expression between groups. However, there were some slight differences, though likely not significant enough to draw conclusions from. Overall the younger age groups had greater average gene counts of Cowden genes and more deviation in counts of Lynch genes. In addition, for all 7 genes black patients experienced much greater diversity in gene expression compared to other racial categories. This increase in diversity could be attributed to the out-of-Africa hypothesis, but it doesn’t tell us much about demographic association with the 2 aforementioned syndromes. There was no correlation between gene expression and cancer stage for any of the 7 genes, with all plots being virtually the same. This information tells us that both Cowden and Lynch syndrome aren’t more prevailant in certain categories of these 3 clinical factors.

Though Cowden and Lynch syndrome didn’t have demographic correlations, survivorship correlations were the next associations analyzed for. For each gene (except *KLLN* due to lack of data) KM plots were created, evaluating the survivorship of patients based on whether or not they had a mutation in that gene. To do this, simple nucleotide variation data of uterine cancer patients was downloaded from TCGA using the R package “TCGAbiolinks”. MAF data listing mutations was cross referenced with clinical data in order to determine which patients had mutations in the aforementioned genes. Again, death data was adjusted in order to account for non-reported death data and the R packages “survival” and “survminer” were used to create the final plots showing survival probability over time (in days). Contrary to expectations, for almost all genes (both Cowden and Lynch genes) patients with a mutation in the gene had a significantly *greater* survival probability than patients that didn’t (with plots having a p-value of p<0.05). The exceptions for this observation are the *MLH1* and *PMS2* plots, which though visually appeared to follow this trend, had insignificant p-values of p=0.17 and p=0.13 respectively (perhaps due to a lack of patients with the mutation in the dataset). One explanation for why patients with these mutations experienced increased survival probability is perhaps due to cancer screening. Patients with inherited diseases such as Cowden and Lynch likely have a long history of cancer in their family, and are therefore much more likely to undergo cancer screening regularly, starting at a younger age. This means that these patients get diagnosed with cancer younger, and based on the KM plot showing survivorship based on age at diagnosis, it’s known that patients who are diagnosed younger have greater survivorship.

To support this speculation, a final 6 boxplots were created, comparing the age at diagnosis for patients with each of these mutations (*KLLN* again disincluded due to lack of data) against that of patients without. This was done using the R boxplot function with previously calculated data. For all genes except *PMS2*, it was found that the average age at diagnosis for patients *with* the mutation was earlier than patients without (by about 5-10 years). This supports the idea that uterine cancer patients with heritable diseases such as Cowden and Lynch syndrome are diagnosed with cancer earlier than patients without, potentially resulting in greater survival probability. As for the other clinical factors, there was no found association between race, Cowden or Lynch syndrome, and survivorship of uterine cancer patients using this data. This doesn’t mean these factors are certainly not interrelated, just that the data may not be robust enough to create a valid conclusion, especially when it comes to the lack of significant data for patients of color.

**References**

“Cancer Stat Facts: Uterine Cancer.” Surveillance, Epidemiology, and End Results (SEER) Program, National Cancer Institute, 2022, https://seer.cancer.gov/statfacts/html/corp.html.

“Cowden Syndrome.” Cancer.Net, American Society of Clinical Oncology (ASCO), Jan. 2022, https://www.cancer.net/cancer-types/cowden-syndrome.

“Lynch Syndrome.” Cancer.Net, American Society of Clinical Oncology (ASCO), Sep. 2021, https://www.cancer.net/cancer-types/cowden-syndrome.

“The Cancer Genome Atlas - Uterine Corpus Endometrial Carcinoma.” GDC Data Portal, National Cancer Institute, 2022, https://portal.gdc.cancer.gov/projects/TCGA-UCEC.