**Marco Landicino**

**Yorktown High School**

**Research Plan**

1. **Rationale**
   1. Pancreatic cancer has the lowest 5-year survival rate among all cancers at 9.3% and is the third leading cause of all cancer deaths (SEER, 2019). Most patients diagnosed with pancreatic cancer die within the first 2 years as it is often diagnosed in stage 4 (SEER, 2019). By the time the cancer has reached stage 3 or 4 it has already metastasized and become lethal. The average time between diagnosis and death is about 4 to 6 months, which hasn’t improved over the past 20 years. Most pancreatic cancer, when diagnosed, is already metastatic and radiation therapy and chemotherapy is often ineffective (SEER, 2019). However, patients who are diagnosed with early stage pancreatic cancer, have a 5-year survival rate of 40% (SEER, 2019). The only plausible way to increase the dismal survival rate of pancreatic cancer is to surveil at-risk individuals and diagnose the disease before it reaches the late stages. The way we can determine at-risk individuals is through cancer epidemiology. Cancer epidemiology is the study of distribution and determinants of health-related states or events in specific populations in the application of the study to control health problems (Silva, 1999). Using cancer epidemiology, there have been many environmental risk factors determined that increase the risk for patients including age, smoking, obesity, and type 2 diabetes (American Cancer Society, 2017). A risk factor that is being studied in greater depth through cancer epidemiology is the effect of sex hormones, specifically estrogen, on pancreatic cancer. The pancreas also contains receptors for both estrogen and androgen. Estrogen is the hormone that is responsible for female features and reproduction. A possible connection lies in women with polycystic ovary syndrome (PCOS). PCOS is often diagnosed early in a woman’s life, around the time of her first menstrual cycle but can also arise later in life. PCOS creates a great imbalance of hormones in females, specifically an increase in androgen, although it is currently unknown why this occurs. An increase in androgen, the male reproductive hormone, often leads to facial acne or unwanted facial hair. Not only does PCOS lead to a great imbalance of hormones, but it also leads to a decrease in insulin released by the pancreas (Mayo Clinic, 2017). Many PCOS patients also have increased insulin resistance by their cells. This creates an increase in the woman’s blood sugar and triggers the release of more insulin by the pancreas. The increase of insulin in the body often leads to type 2 diabetes. Type 2 diabetes is a proven risk factor of pancreatic cancer and PCOS could be a confounder (Endocrine Society, 2017). There has been very little research done showing the correlation between PCOS and pancreatic cancer, but research regarding acne as well as estrogen mention PCOS as a possible confounding factor. There has been one study done to show the correlation between PCOS and all cancers and the results show promise for specifically pancreatic cancer.

Raising the extremely low survival rates is the main goal of all cancer epidemiology research. For pancreatic cancer, this goal is even more important because it is estimated that over 45,000 people will die from pancreatic cancer in 2019 (SEER, 2019). Determining a risk factor that has a great effect on people's risk of being diagnosed with pancreatic cancer is difficult. Being able to determine a new risk factor, such as polycystic ovary syndrome, would help to determine high-risk individuals. Identifying high-risk individuals will allow for the early surveillance and screening with the hope to be able to diagnose pancreatic cancer in the early stage. Polycystic ovary syndrome is easily diagnosable and if it is determined as a risk factor of pancreatic cancer, it would help to increase the stagnant 5-year survival rate.

1. **Research Question(s), Hypothesis(es), Engineering Goal(s), Expected Outcomes**
   1. **Research Question**
      * Polycystic ovary syndrome is a plausible risk factor of pancreatic cancer and determining the possible correlation may help identify an individual who is at high risk of being diagnosed with the disease. However, it is currently unknown exactly what effect polycystic ovary syndrome has on the risk of being diagnosed with pancreatic cancer.
   2. **Hypothesis(es)**
      * If a woman is diagnosed with polycystic ovary syndrome then their risk of pancreatic cancer increases because polycystic ovary syndrome creates insulin resistance which often leads to type 2 diabetes, an established risk factor of pancreatic cancer.
   3. **Engineering Goal(s)**
      * This is not  an engineering project so there are no engineering goals.
   4. **Expected Outcomes**
      * It is expected that PCOS patients will have an increased risk of pancreatic cancer in the future. This could be explained by the high levels of hormones and low levels of insulin in PCOS patients.
   5. **Procedure**
      * Role of Mentor vs. Role of Student
        1. The mentor will provide the student with de-identified data from the Pancreatic Tumor Registry at MSKCC which includes the coded answers to all questions from the main questionnaire given to 1,989 patients.
        2. The student will then take the data from the 1,989 patients and manipulate the data to fit the experiment
        3. **The data will have already been de-identified by the mentor and given to the student who will have no way of re-identifying the data.**
        4. All patients have previously been questioned using the main questionnaire, created in 2003, for previous lifestyle history and disease history by the mentor. The data is part of a registry and has info on patients over a period of 10 years (2003-2013).
        5. The mentor will also helped guide the student in choosing variables from the data as well as helping the student with learning programming and interpreting the data/analysis.
      * Role of Student
        1. Create a retrospective case-control study using the Pancreatic Tumor Registry data from MSKCC.
        2. The data will be provided to the student in a .sas7bdat file which the student could not edit but only input as a dataset into Statistical Analysis Software (SAS).
        3. Statistical Analysis System (SAS) will be chosen because it provides statistical analysis as well as the ability to create a logistic regression model with covariates which will be the main analysis of the project.
        4. The dataset includes the coded responses to each question answered by the patients who answered the main questionnaire between 2003 and 2013.
        5. All men will be excluded from the experiment as polycystic ovary syndrome is only present in females as well as if they did not answer the polycystic ovary syndrome (how many missing will be specified).
        6. A question on the questionnaire will establish cases vs. controls however, spousal and in-law controls will be counted as controls to add to the numbers in the experiment. A severity analysis will be done to ensure these control do not affect the results.
        7. The criteria for a control will include patients without any previous history of cancer except non-melanoma skin cancer. The patient will be excluded if they are considered an at-risk relative or IPMN.
        8. The patient will also be removed if the patient left an answer blank to any of the covariate questions.
        9. The total number of patients included in the analysis will be 555 with 370 cases and 185 controls.
        10. Polycystic ovary syndrome is a yes/no question within the questionnaire so there will be no need for manipulation of this variable.
        11. The original coding for the case/control status of the patient will be changed as it originally included 6 option: cases, spouses, in-laws, at-risk relatives, controls, and IPMNs. This will be grouped into cases (cases) and controls (spouses, in-laws, and controls) with relatives and IPMNs being removed.
        12. Covariates will be chosen and included in the fully-adjusted analysis if they are either a known pancreatic cancer risk factor, a symptom of polycystic ovary syndrome, or a common covariate in all experiments (Ex. race).
        13. Variables will include: Age, Race/Ethnicity, Education, Smoking status, Body Mass Index, Diabetes, Alcoholism, Pancreatitis, Hysterectomy, Oophorectomy, Menopause, Estrogen use, Acne, Facial hair, Thin hair, and Irregular periods.
        14. Table 1 will include the baseline characteristics frequencies/percentages for the cases and controls. Some variables will be represented differently in the table then in the analysis.
        15. For the variables in Table 1: Age (mean+-SD), Race (Categorical), Education (Categorical), Smoking (Binary), BMI (Categorical), Diabetes (Binary), Alcoholism (Binary), Pancreatitis (Binary), Hysterectomy (Binary), Oophorectomy (Binary), Menopause (Binary), Estrogen Use (Binary)
        16. Some variables will need to be manipulated because the questionnaire does not have a yes/no answer for the question.
        17. Race will be split into 4 categories as patients could answer any race they wanted: White, Black/African American, Asian, Other)
        18. Education will be split into 4 categories depending on how many years of schooling the patient completed (<12 or less than High school, 12-16 or High school, 17-20 or College, and >20 Graduate)
        19. Smoking will be considered a yes/no based on if the patient smoked more or less than 100 cigarettes in their lifetime.
        20. BMI will be calculated based on kg/m2. BMI will be categorized into 6 groups (Underweight or <18.5, Normal or 18.5-24.9, Overweight or 25-29.9, Obese Class 1 or 30-34.9, Obese Class 2 or 35-39.9, and Obese Class 3 or >39.9).
        21. A patient will be considered in menopause if they gave an age for their last menstrual cycle and will be considered not in menopause if they did not answer the age for last menstrual cycle question.
        22. A Minimally-adjusted analysis will be performed using age and race to obtain a correlation between polycystic ovary syndrome and pancreatic cancer
        23. A Fully-adjusted analysis will be performed adding in all covariates listed above, however some variables will be represented differently. Age (Continuous), Race (Categorical), Education (Categorical), Smoking (Binary), BMI (Continuous), Diabetes (Binary), Alcoholism (Binary), Pancreatitis (Binary), Hysterectomy (Binary), Oophorectomy (Binary), Menopause (Binary), Estrogen Use (Binary), Acne (Binary), Facial Hair (Binary), Thin hair (Binary), and Irregular periods (Binary).
        24. BMI will be used as continuous for the analysis as there was no reference category that could give accurate results.
        25. Within the acne variable only 2 things will be looked at: Did the patient ever have severe acne and did the patient ever have cystic acne? The times in which the patient had the acne will not be accounted for.
        26. Statistical Analysis Software (SAS) will be used to create new or clean variables as well as run PROC FREQ and PROC LOGISTIC commands for analysis
   6. **Risk and Safety**
      * There were no risk or safety concerns during the experiment because all data was de-identified by the mentor and will be given to the student without the rights of leaving the building.
   7. **Data Analysis**
      * Using Statistical Analysis Software (SAS), compare cases and controls by using the already created dataset
      * Through SAS programming and the already created dataset, using the PROC LOGISTIC command to create a logistic regression model of the data and compare cases vs. controls to polycystic ovary syndrome
      * A comparison between PCOS and pancreatic cancer using minimally-adjusted odds ratios will be performed.
      * A fully-adjusted odds ratios, using the variables specified in the procedure, will be calculated comparing the data of patients with PCOS and those without to whether they were diagnosed with pancreatic cancer
      * A high odds ratio will be considered statistically significant as long as the 95% CI does not contain the number 1
      * A p-value of <.05 will be considered statistically significant
   8. **Bibliography**
      * Olson, S. H., Xu, Y., Herzog, K., Saldia, A., DeFilippis, E. M., Li, P., … Kurtz, R. C. (2016). Weight Loss, Diabetes, Fatigue, and Depression Preceding Pancreatic Cancer. *Pancreas*, *45*(7), 986–991
      * Harris, H. R., Titus, L. J., Cramer, D. W., & Terry, K. L. (2017). Long and irregular menstrual cycles, polycystic ovary syndrome, and ovarian cancer risk in a population-based case-control study. *International journal of cancer*, *140*(2), 285–291. doi:10.1002/ijc.30441
      * Ding, D. C., Chen, W., Wang, J. H., & Lin, S. Z. (2018). Association between polycystic ovarian syndrome and endometrial, ovarian, and breast cancer: A population-based cohort study in Taiwan. *Medicine*, *97*(39), e12608. doi:10.1097/MD.0000000000012608
      * Barry, J. A., Azizia, M. M., & Hardiman, P. J. (2014). Risk of endometrial, ovarian and breast cancer in women with polycystic ovary syndrome: a systematic review and meta-analysis. *Human reproduction update*, *20*(5), 748–758. doi:10.1093/humupd/dmu012
      * Yin W, Falconer H, Yin L, Xu L, Ye W. (2019). Association Between Polycystic Ovary Syndrome and Cancer Risk. *JAMA Oncol.*;5(1):106–107. doi:10.1001/jamaoncol.2018.5188