**Intermediate- Public Health/Cancer**

Urinary Biomarkers for Pancreatic Cancer

Project Description: This is a dataset from an open-access paper published December 10, 2020. Pancreatic cancer is an extremely deadly type of cancer. Once diagnosed, the five-year survival rate is less than 10%. However, if pancreatic cancer is caught early, the odds of surviving are much better. Unfortunately, many cases of pancreatic cancer show no symptoms until the cancer has spread throughout the body. A diagnostic test to identify people with pancreatic cancer could be enormously helpful.

In a paper by Silvana Debernardi and colleagues, published this year in the journal PLOS Medicine, a multi-national team of researchers sought to develop an accurate diagnostic test for the most common type of pancreatic cancer, called pancreatic ductal adenocarcinoma or PDAC. They gathered a series of biomarkers from the urine of three groups of patients:

* Healthy controls
* Patients with non-cancerous pancreatic conditions, like chronic pancreatitis
* Patients with pancreatic ductal adenocarcinoma

When possible, these patients were age- and sex-matched. The goal was to develop an accurate way to identify patients with pancreatic cancer. The key features are four urinary biomarkers: creatinine, LYVE1, REG1B, and TFF1.

* Creatinine is a protein that is often used as an indicator of kidney function.
* YVLE1 is lymphatic vessel endothelial hyaluronan receptor 1, a protein that may play a role in tumor metastasis
* REG1B is a protein that may be associated with pancreas regeneration
* TFF1 is trefoil factor 1, which may be related to regeneration and repair of the urinary tract

Age and sex, both included in the dataset, may also play a role in who gets pancreatic cancer. The dataset includes a few other biomarkers as well, but these were not measured in all patients (they were collected partly to measure how various blood biomarkers compared to urine biomarkers).

The goal in this dataset is predicting diagnosis, and more specifically, differentiating between 3 (pancreatic cancer) versus 2 (non-cancerous pancreas condition) and 1 (healthy). The dataset includes information on stage of pancreatic cancer, and diagnosis for non-cancerous patients, but remember—these won't be available to a predictive model.

Project source: <https://www.kaggle.com/datasets/johnjdavisiv/urinary-biomarkers-for-pancreatic-cancer>

Data file: debernardi-data.csv

Data dictionary: debernardi-dictionary.csv