Breast cancer is one of the most common cancers in the United States with invasive breast cancer affecting an estimated 12% of women and has the second-highest mortality of any cancer that affects women (Breastcancer.org, 2019). However, death rates vary widely based on time of detection with 90% of patients surviving if it is detected early and 15% for those diagnosed in late stages (cancerresearch.uk, 2018). Unfortunately, early detection can be difficult as breast cancer is often symptomless in its earliest stages (Webmd, 2019). Regular physician check-ups and radiology scans for all women is not a viable solution as average costs for diagnostics are $400 for mammograms, $130 for ultrasounds, and over $2000 for biopsies. Thus, 53% of surveyed women do not have regular screenings due to financial barriers (McAlearney, 2007).

Research to determine the most common risk factors for breast cancer can alleviate some of these concerns as healthcare and charitable organizations would be able to determine more high-risk patients. Charities such as the National Breast Cancer Foundation can use the information to understand which patients need more immediate or regular access to screenings. Research organizations, hospitals, and pharmaceutical companies may find the information useful for a better understanding of the cause of breast cancer and possible treatments. For example, if a higher level of leptin appears to correlate with a higher risk of cancer, researchers can focus on whether or not there is a potential relationship between the protein level and cancer. Also, if a certain trend such as higher BMIs is shown to correlate with a higher incidence of breast cancer, physicians may choose to have those patients do more regular screenings or can advise patients that the screening may be worthwhile because they are of higher risk.

In order to find these risk factors, I will be analyzing the Breast Cancer Coimbra Data Set. This set has data for both breast cancer and non-cancer patients and has a variety of characteristics that can be identified easily with a blood test. Factors include glucose, insulin, leptin, BMI, age, etc. The data currently do not have any missing values, so it is already relatively clean and easy to use. I will start by making plots between the measurements and the number of patients that have breast cancer. This will give me a rough outline of potential correlations and if a relationship is present, can help show whether it is linear, polynomial, etc. Once I can find the most appropriate model for the characteristics with the most promising correlation, I can determine which algorithm to use to better quantify this trend such as linear regression and may use factor analysis to see which attributes are more useful. Once, I understand the common risk factors that can lead to breast cancer, I will summarize my findings in a PowerPoint Presentation. The presentation will include more information on the dangers of breast cancer and the usefulness of the experiment. Plots such as 2d color plots or box plots will also be present in the presentation based on which will convey the statistical analysis more appropriately. This will be the first of many data science challenges I will undertake and will be an exciting opportunity to bridge my previous experiences in research, healthcare, and cancer with my new career in data science.

Sources

“Breast Cancer Symptoms and Early Warning Signs.” *WebMD*, WebMD, www.webmd.com/breast-cancer/understanding-breast-cancer-symptoms.

McAlearney, Ann Scheck, et al. “Cost as a Barrier to Screening Mammography among Underserved Women.” *Ethnicity & Health*, U.S. National Library of Medicine, Apr. 2007, www.ncbi.nlm.nih.gov/pmc/articles/PMC4465254/.

Nbcf. “Information, Awareness & Donations :: The National Breast Cancer Foundation.” *Www.nationalbreastcancer.org*, www.nationalbreastcancer.org/.

“U.S. Breast Cancer Statistics.” *Breastcancer.org*, 13 Feb. 2019, www.breastcancer.org/symptoms/understand\_bc/statistics.

“Why Is Early Diagnosis Important?” *Cancer Research UK*, 4 Mar. 2019, www.cancerresearchuk.org/about-cancer/cancer-symptoms/why-is-early-diagnosis-important.