Clustering and Predicting Vital Status from Breast Cancer Clinical data

Breast cancer is one of the most common cancer and cause of death among women [1]. The survival rates are dependent on stage of cancer, the more advanced the cancer, the lower the survival rate is. Luckily, with efforts of research and novel approved drugs, survival rates continue to improve, making the study of breast cancer an evolving area of interest and research. Survival rates are dependent on stage of tumor and are specific for each patient. According to a 2018 study [2], survival rates in terms of tumor stage. We note that as the tumor stage advances, the survival rate percent of 5-year estimate decreases.

Breast Cancer	**5-Year Breast Cancer-Specific Survival Estimate**
Stage I	98-100%
Stage II	90-99%
Stage III	66-98%

Despite the necessity and importance of genomic analyses, our approach will provide a prediction of survival rate according to clinical data, which are available in electronic health records (EHR), connecting genomic analysis with clinical cases and making the prediction readily quicker to obtain for both research and healthcare providers [3].

In our research, we propose that the use of clustering and prediction data mining methods will uncover correlations and patterns among the variables available in clinical data in relation to vital status. We anticipate that the outcome will answer questions such as: what clinical variables are mostly associated with survival or death of a group of patients? How accurate is the model built to predict vital status in the given dataset? How does the clustering and prediction of clinical data compare to genomic data?

The source of the dataset used in this research is obtained from The Cancer Genome Atlas (TCGA), a program that collects and analyzes genomic data samples from more than 30 cancer types [4]. This program started on 2007 and samples in this dataset were first published in 2012 by The Cancer Genome Atlas Network as part of TCGA, along with genomic analysis [5]. Additional samples were added along the years, adding up to 2,182 breast cancer samples.

The dataset consists of thirty-three variables, thirty-two predictors and one response variable (vital status: alive or dead). As part of the data processing, we are expecting to work on handling variables with missing values, removing uninformative variables, such as ID variables, converting age from days to years, and lastly, scale the dataset.

Reference List

[1] Czajka ML, Pfeifer C. Breast Cancer Surgery. 2020 Sep 14. In: StatPearls [Internet]. Treasure Island (FL): StatPearls Publishing; 2020 Jan–. PMID: 31971717.

[2] Weiss A, Chavez-MacGregor M, Lichtensztajn DY, Yi M, Tadros A, Hortobagyi GN, Giordano SH, Hunt KK, Mittendorf EA. Validation Study of the American Joint Committee on Cancer Eighth Edition Prognostic Stage Compared With the Anatomic Stage in Breast Cancer. JAMA Oncol. 2018 Feb 1;4(2):203-209. doi: 10.1001/jamaoncol.2017.4298. PMID: 29222540; PMCID: PMC5838705.

[3] Institute of Medicine (US). Integrating Large-Scale Genomic Information into Clinical Practice: Workshop Summary. Washington (DC): National Academies Press (US); 2012. 3, The Analysis of Genomic Data. Available from: https://www.ncbi.nlm.nih.gov/books/NBK92085/

[4] "The Cancer Genome Atlas Program." *National Cancer Institute*, https://www.cancer.gov/about-nci/organization/ccg/research/structural-genomics/tcga

[5] Cancer Genome Atlas Network, 2012. Comprehensive molecular portraits of human breast tumours. *Nature*, *490*(7418), p.61.