Background Information

The dataset I chose comes from the *VectraPolaris BioConductor* package in Rstudio. The data represents tissue images of high-grade serous ovarian cancer (HGSOC) taken via multiplexed immunofluorescence (MXiF) imaging, where tissues are stained with cell biomarkers that show the phenotype of each cell within the tissue. Furthermore, MXiF imaging also provides the spatial location of each cell on the tissue, making this dataset bivariate in nature on the cellular level. As such, an important question that biologists ask is how do interactions between specific cell types influence probability of death in a patient? In order to approach this problem, a combination of clustering and classification is needed. More specifically, scientists can apply statistical indices like the nearest neighbor G functions to measure clustering between different cell phenotypes and then input the information from those indices into a classification model such as logistic regression to predict the probability of survival. In recent literature, cytotoxic T-cells (CD8+) and macrophages (CD68+) have been shown to cluster around one another to worsen survival outcomes for patients with HGSOC.

While the data is detailed, both stroma and tumor cells are included. Since our motive is to understand how cell clustering influences HGSOC progression, we must filter the cell types to be tumor cells. Furthermore, we must filter the type of tumor to be primary - or the original tumor - since the tumor microenvironment in primary tumor is an indicator of whether a tumor will become malignant and metastasize. Finally, fluorescence values are given for each phenotype and the classification of a cell's phenotype is stored as a string. To address this, a column for each phenotype of interest will be made, making the value of phenotype for any specific cell binary for one of cytotoxic T-cell (cd8), helper T-cell (cd3), macrophage (cd68), or B-cell (cd19).

Problem Statement

This study will serve as a proof of concept to assess how spatial clustering patterns of cell phenotypes can inform regression models to more accurately predict patient outcomes of death.

Hypothesis

Recent literature has found that cytotoxic T-cells and macrophages associate to worsen survival odds for patients with HGSOC; therefore, if a logistic regression model were to account for spatial clustering pattern between cytotoxic T cells/macrophages, it would have increased accuracy in predicting survival.

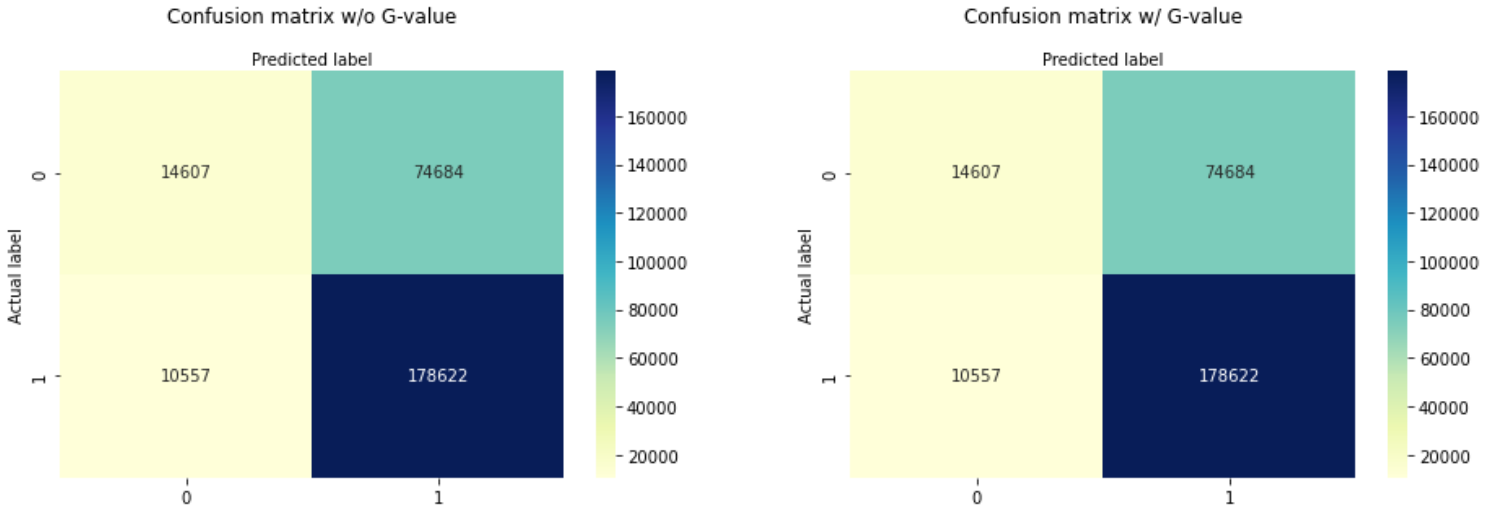
Methods

Here, a nearest neighbor G (NNG) function from the *SpatialTIME* package was used to assess cd8/cd68 clustering patterns. Here, the problem required an understanding of how frequently cd8/cd68 cells were clustering near each other in the HGSOC tumor microenvironment, and using the NNG function provided a proportion for the amount of cells that were cytotoxic T cells (cd8+) that had a nearest neighbor that was a macrophage (cd68+). From here, death was stored as a binary variable in the data, with '1' indicating death and '0' indicating survival. As such, a logistic regression was used to predict death in patients since the outcome of death was

binary and covariates of different structures ranging from the continuous NNG score to binary clinical variables such as BRCA mutations were used.

Results and Discussions

Ironically, the implementation of clustering a score had no impact on the performance of the logistic regression model. In order to assess the accuracy of the logistic regression, I implemented a confusion matrix to calculate how many times the death outcome was correctly predicted. The accuracy of prediction was the exact same for both models: survival was correctly predicted 16.4% (14607/89291) of the time and death was correctly predicted 94.4% (178622/198219) of the time. Here, a confusion matrix was the best tool used to assess the model as it succinctly displays the models predicting power. While the predictive power of the model doesn't appear to be strong, I find it interesting that the model biases towards predicting death, or '1', as shown by the stark contrast in prediction rates. Nonetheless, it appears that the spatial clustering of cd8/cd68 cells was non-informative to predicting survival outcomes of patients with HGSOC.



Sources

- Logistic Regression: https://www.datacamp.com/tutorial/understanding-logistic-regression-python
- spatialTIME: https://academic.oup.com/bioinformatics/article/37/23/4584/6420699
- Data: http://juliawrobel.com/MI_tutorial/MI_VectraPolarisData.html#Ovarian_cancer_data