# Homework 3

BIOST 544

Autumn 2024

Please submit your homework on Canvas, in a compiled R-markdown file (to pdf or html), as well as your code (as appendix).

All code in this assignment should be cleanly written and well commented, with appropriate use of functions/arguments. Imagine you need to give this code to someone else and they need to understand it (which you may need to do!)

## High-Dimensional Predictive Models

The existence of a large number of necrotic cells in a tumor can be indicative of a successfully mounted immune defense. We might be interested in understanding biomolecular pathways regulated/dysregulated in a tumor that make it more/less susceptible to the immune system. By identifying genes with expression (in the tumor microenvironment) related to quantity of necrotic tumor tissue, we might hope to a) build a better picture of the biology of immune regulation/dysregulation in cancer and/or b) find potential targets for therapy. In this homework we would like investigate the relationship between gene-expression values in the tumor and the existence and extent of necrotic tissue.

To evaluate this we will again work with the **NOAH** data. The data can be found on the course website in the following files:

- **clinical_data.csv** contains the clinical/phenotypic information.
- **expression_data_probeID.csv** contains the expression information (by probeset).
- **annotation.csv** contains the genename identifiers corresponding to each probeset.

To get more information on the probesets, feel free to look into the affymetrix human 133 plus 2.0 array annotation (on the affymetrix site).

In particular, the variable **necrotic_cells.pct** (the percentage of necrotic tissue in a tumor found by pathology) may be useful.

One approach for assessing the relationship between gene-expression values in the tumor and the existence and extent of necrotic tissue is to use a predictive model to identify genes that are associated with necrotic tissue, while avoiding over-fitting due to the large number of genes. To this end, first build a predictive model for `necrotic_cells.pct` as a function of gene expression values. Next, consider a categorical version of the same outcome, where `necrotic_cells.pct` either 0 or greater than 0. Compare the results of the two predictive models and discuss pros/cons of the two modeling approaches.