In our simulation,

- The sample size is 173290
- We generate 500 data sets and 500 external study data, with the same sample size and data generation mechanism
- We assume all covariates are independent of each other
- There are no interaction effects between covariates
- To conduct cross-validations, we stratify the individuals by both race/ethnicity (white and non-white) and disease outcome.

For glm:

- We only include the covariates that have non-zero coefficients in the data generation

For lasso:

- We first conduct 5-fold cross validations to tune the penalty parameter λ based on the full data set. We select the value of λ that maximizes AUC and denote λ^* as this best value.
- Then, we fit a lasso model with all covariates in the data set and the penalty parameter of λ^*.
- We assess the model performance on the external study.
- Next, we conduct 5-fold cross-validations as an alternative to evaluate model performance. Again, we use the penalty parameter of λ^* to fit the lasso model.
- Question here: Suppose we are conducting 5-fold cross-validations. In the first round, we need to fit the model based on the first 4 folds and evaluate the model on the last fold. When fitting the model, should we also tune λ or should we just use λ^*?
- This is the same question for other machine learning models that need parameter tuning.

Results:

See the Excel file.

To-do:

- Look at the sandwich SE estimate using the sandwich function in the sandwich R package in both low-dimensional and high-dimensional settings.
- Do simulations with tree-based models.