Non-linear machine learning models to improve variant prediction among clinical breast cancer patients

Larah Siouffey[1], Adam Chamberlin[2], Amal Yusuf[3], Marcy Richardson[3], **Linda M. Polfus[2]**
[1]Keck Graduate Institute, Department of Human Genetics and Genomic Data Analytics, Claremont, California, U.S.A
[2]Ambry Genetics, Advanced Analytics, Aliso Viejo, California, U.S.A
[3]Ambry Genetics, Medical Sciences, Aliso Viejo, California, U.S.A

Approximately 13% of women will be diagnosed with breast cancer at some point during their lifetime. Women harboring pathogenic variants (PV) in key homologous repair deficiency genes have both increased breast cancer risk and respond more favorably to certain therapies, like platinum agents and PARP inhibitors. Traditional breast cancer risk models are enhanced by Polygenic Risk Scores (PRS), yet often fail to integrate multiethnic data and complex SNP interactions. We examined (N=10,880) ethnically diverse patients who underwent a multi-gene panel test for cancer predisposition in a single clinical diagnostic laboratory. Patients were selected for having a rare PV among one or more of five breast cancer predisposition genes (*ATM*, *BRCA1*, *BRCA2*, *PALB2*, and *CHEK2*) as well as 860 common SNPs known to be associated with breast and other cancers. Cases were breast-cancer-affected according to test requisition forms and clinical notes while controls unaffected by breast cancer and reported no family history of breast cancer. Diverse ethnic groups included Caucasian (67.2%), Ashkenazi Jewish (5.8%), African American (10.5%), Hispanic (11.1%), and Asian (5.5%). We applied LASSO regression to identify influential features based on age at diagnosis (binned by decade), self-reported ethnicity, PV status, and SNP data, achieving an AUC of 0.81 (Mean Square Error: 0.18; r2: 0.26). The LASSO model's significant predictors (N=139) were refined through XGBoost, which utilized an 80/20 training/testing split, resulting in a model AUC of 0.83. Predictions on the test set (N=2176) yielded an AUC: 0.83 (Percent Variance Explained: 28.6%; r^2: 0.40). XGBoost selected 134 features and ranked importance revealed 5 PV genes, age group, Caucasian, Ashkenazi Jewish ancestry, rs1329390, rs11814448, rs2660753, and rs6465657. The XGBoost model underscored several intergenic SNPs and others co-localizing with known breast cancer-associated regions. A compelling interaction was detected between *CHEK2* PVs and rs132390, an intronic variant within *EMD1*, a gene implicated in cancer metastasis. Future research will expand the dataset via ICD10 phecodes and incorporate additional machine learning features into an enhanced PRS framework, integrating the Tyrer-Cuzick model to refine risk predictions. Our approach aims to accurately identify individuals at increased risk, optimizing preventive and therapeutic interventions based on personalized genetic profiles.