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

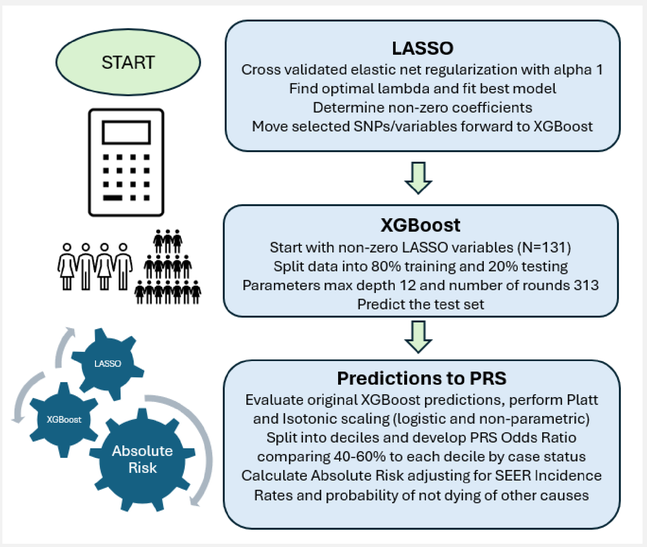
BACKGROUND

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. Among familial rare pathogenic gene carriers, breast cancer PRS predictions may change screening recommendations in the US up to 27% (1). Traditional breast cancer risk models are enhanced by Polygenic Risk Scores (PRS), yet often fail to integrate multiethnic data and complex SNP interactions. Machine learning (ML) approaches combining multiethnic samples, clinical risk factors, rare pathogenic variants, and common variants perform better than PRS alone (2, 3).

METHODS

- **Patients:** Ethnically diverse patients who underwent a multi-gene panel test for cancer predisposition in a single clinical diagnostic laboratory (N=10,880) from March 2018 to March 2021.
  - **Case/Control:** 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.
  - Caucasian (67.2%), Ashkenazi Jewish (5.8%), African American (10.5%), Hispanic (11.1%), Asian (5.5%) (**Figure 2**).
- **Genetics:** Rare Pathogenic Variants (PVs) among one or more of five breast cancer predisposition genes (*ATM*, *BRCA1*, *BRCA2*, *PALB2*, and *CHEK2*) which were aggregated as 5 indicator variables (**Figure 3**).
  - Common SNPs (860) across the genome known to be associated with breast and other cancers.

Figure 1. Workflow of SNP Selection and Model Construction



RESULTS

Figure 2. Breast cancer case status by ethnicity

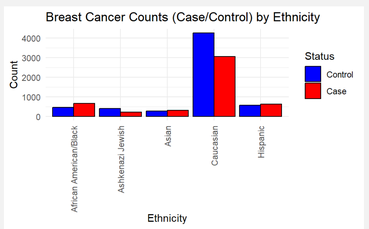


Figure 3. Pathogenic mutation carrier proportions by case status and ethnicity

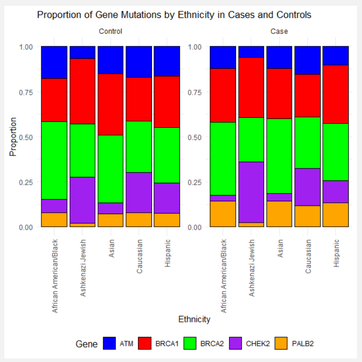


Figure 4. LASSO Top 15 Feature Importance Plot

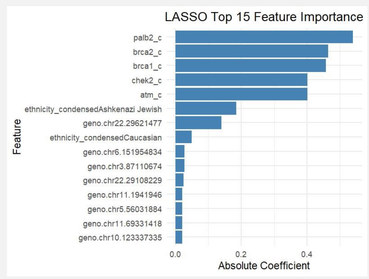
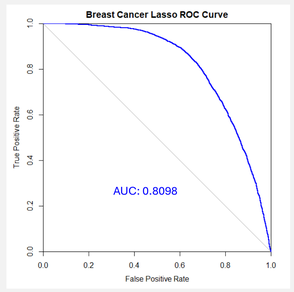


Figure 5. Breast Cancer LASSO ROC Curve



LASSO selected 131 SNPs/variables to proceed to the XGBoost method which yielded an AUC of 0.81 (Mean Square Error: 0.18; r^2: 0.26). Top 15 features from LASSO are shown in **Figure 4** and ROC in **Figure 5**. The Top 15 features from XGBoost training set are shown in **Figure 6**. The AUC for XGBoost training data (**Figure 7**) was 0.827 and on the XGBoost test set (N=2176) Percent Variance Explained: 28.6%; r^2: 0.40.

Figure 6. LASSO Top 15 Feature Importance Plot

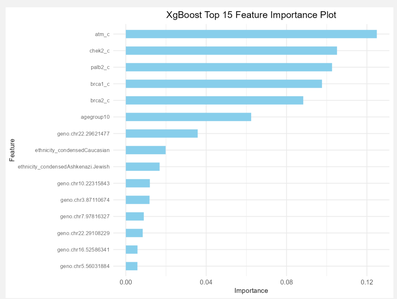


Figure 7. Breast Cancer LASSO ROC Curve

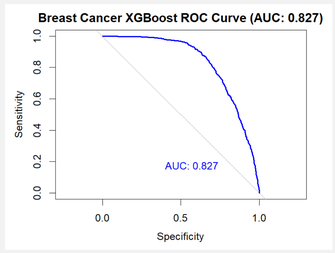
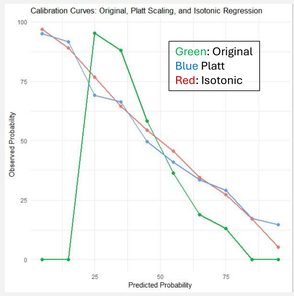


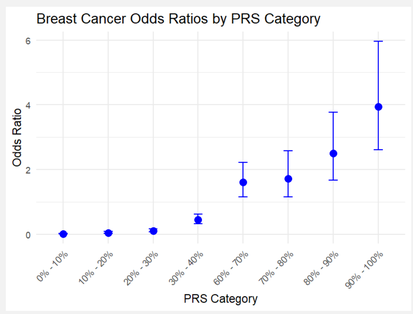
Figure 8. XGBoost observed versus expected probabilities compared to Platt and Isotonic scaled



The raw predictions from XGBoost represent the probability of an individual developing breast cancer based on the features used in the model, non-linear interactions of rare PVs in 5 genes with selected LASSO SNPs. Both Platt scaling and Isotonic regression were compared (**Figure 8**). Based on the conservative smoothing of the Isotonic regression (shown in red), these predictions were utilized for PRS modelling.

- **Platt Scaling:** Logistic regression approach when the model tends to be overconfident or underconfident in its predictions.
- **Isotonic Regression:** Non-parametric method (maps predictions to a monotonic function) when there are non-linearities that need to be corrected in the probability space.

Figure 9. Isotonic PRS Odds Ratio of Breast Cancer status versus each decile (referent 40-60%)



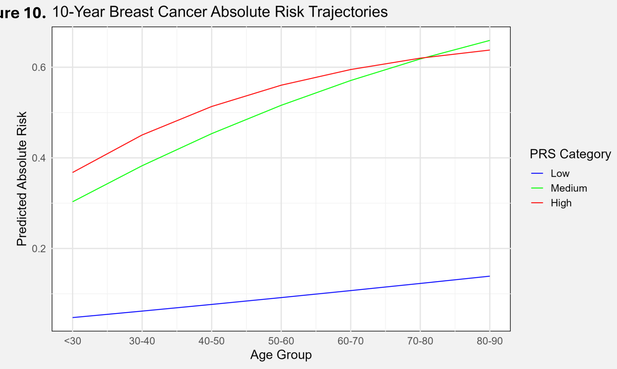
The predicted probabilities from the isotonic regression, split into deciles, compared the middle (40-60%) of the distribution to construct odds ratio by breast cancer case/control status (**Figure 9**).

Comparing the middle of the distribution to 90-100% decile, the odds of having breast cancer with selected ML features were nearly 4 times compared to those without breast cancer (OR=3.94; 95% confidence interval 2.60-5.97).

ABSOLUTE RISK FIGURE

- For Low, Medium and High categories, cut points were defined at 0.33 and 0.66 of the Isotonic scaled PRS distribution (**Figure 10**).
- Ten-year Absolute Risk trajectories for individuals surviving up to an age group and adjusted for the probability of not dying of another cause (female total US mortality minus female breast cancer deaths 2020) and female SEER breast incidence rates.
- Low group represents mainly non-carriers and Medium and High groups risks are similar in older age groups.

Figure 10. 10-Year Breast Cancer Absolute Risk Trajectories



TAKE HOME POINTS

- Common variants interacting with rare PVs in high-risk genes benefit from ML approaches in a multiethnic setting to create a 'PRS'
- Attenuation of 10-year Absolute Risk trajectories by SEER incidence rates and mortality describes additional risk due to genetics, extending most ML approaches
- Older individuals with higher PRS scores are at greater risk, however other risk factors such as PR or ER +/- status, family history, breast density, obesity, or lifestyle were not accounted for
- Absolute Risk Plot communicates how genetic risk (PRS) and age contribute to cancer risk, useful for risk stratification and personalized medical recommendations
- Future directions will incorporate a Tyrer-Cuzick score and determine changes in screening recommendations in PV carriers and non-carriers

REFERENCES

1. Lakeman, et al. **Clinical applicability of the Polygenic Risk Score for breast cancer risk prediction in familial cases. J Med Genet.** 2023 Apr;60(4):327-336.
2. Elgart, et al. **Non-linear machine learning models incorporating SNPs and PRS improve polygenic prediction in diverse human populations** Commun Biol . 2022 Aug 22;5(1):856.
3. Ke, et al. **An Integrative Pancreatic Cancer Risk Prediction Model in the UK Biobank.** Biomedicines. 2023 Dec 1;11(12):3206.