https://doi.org/10.29289/259453942025V35S1007

Machine learning-based transcriptomic model enhances prognostic stratification in breast cancer

Valbert Oliveira Costa Filho[1], Eduarda Severo Alvarenga[1], Carlos Alberto Barbosa Neto[1], Fabrícia Cardoso Marques[1], Gabriel Maciel Almeida[1], João Luiz Lima Pinheiro[1], Mariana Macambira Noronha[1], Anelise Poluboiarinov Cappellaro[2]

[1]Universidade Federal do Ceará – Fortaleza (CE), Brazil.
[2]Centro Universitário Maurício de Nassau – Barreiras (BA), Brazil.

Objective: To identify prognostic genes and test machine learning models based on transcriptomics to predict overall survival rate of breast cancer patients. **Methods:** A systematic review of transcriptomic datasets was conducted and registered on DOI:10.17605/OSF.IO/65F87. A gene was classified as a core prognostic gene if it consistently indicated either good or poor prognosis in at least 50% of the datasets without conflicting outcomes. These core prognostic genes were analyzed to predict patient prognosis using ten different machine learning models: CoxBoost, Elastic-Net, GBM, Lasso, plsRcox, Ridge, Random Survival Forest, StepCox, SuperPC, and SVM. Models were trained on the largest dataset, and the others were used for validation. C1q/TNF-related protein (CTRP) data were gathered to predict drug sensitivity across all patients. **Results:** Individual patient data from a total of 2,380 breast cancer cases from ten worldwide datasets was included. A set of 44 core prognostic genes was identified and used for subsequent machine learning analyses. CoxBoost demonstrated the highest C-index (0.7) and was selected as the final model. Pooling results with a random-effects model, patients classified as high-risk by our model had a hazard ratio (HR) of 3.1 (95% confidence interval [CI] 2.55–3.77) for overall survival and 3.7 (95%CI 2.67–4.25) for disease-free survival. The model achieved great area under the receiver operating characteristic (ROC) curve (AUC) values of 0.831, 0.721, and 0.724 for overall survival prediction at one, three, and five years, respectively. High-risk patients had higher TP53 mutations, while low-risk patients showed more PIK3CA and CDH1 mutations. High-risk tumors were enriched in Wnt/β-catenin and TGF-β pathways, while low-risk tumors had more TP53 pathway activity and immune complement function. High-risk tumors also showed reduced sensitivity to docetaxel, gemcitabine, and 5-fluorouracil. **Conclusion:** This study demonstrated the effectiveness of our CoxBoost-based model in predicting overall survival and disease-free survival. Patients classified as high-risk by our model exhibited markedly lower overall survival, distinct genomic alterations, and resistance to key chemotherapies. These findings highlight the potential of transcriptomic data for patient stratification.

Keywords: gene expression profiling; breast neoplasms; prognosis.