

PRESENTATION: ORAL PRESENTATION (FAST PRESENTATION)

<https://doi.org/10.29289/259453942025V36S10885>

885 – Machine learning prediction of disease-free survival in breast cancer using blood cell counts and clinicopathological features: a multicentre, retrospective cohort study

Sara Socorro Faria¹, Maria Sílvia Petty Moutinho¹, Ugo de Giorgi¹, Michal Mego¹, Pablo Mandò¹, Daniel Guimaraes Tiezzi¹, Sumadi Luckman Anwar¹, Afonso Celso Nazario¹

¹Universidade Federal de São Paulo – São Paulo (SP), Brazil.

Introduction: Blood-based laboratory parameters, such as the neutrophil–lymphocyte ratio, have emerged as valuable markers in several cancer types. Machine learning models have shown potential for developing predictive and prognostic biomarkers.

Objective: To assess the prognostic value of blood cell count and clinicopathological data when applied to machine learning models to predict disease-free survival in patients with breast cancer.

Methods: This multicenter, retrospective cohort study collected clinicopathological and complete blood count data from 2,864 patients diagnosed with breast cancer in six countries (Argentina, Brazil, Italy, Indonesia, Belgium, and Slovakia) between January 1, 1999, and December 31, 2013. The database was split into training and validation sets at an 8:2 ratio. A Cox proportional hazards model with LASSO regularization was used to select the prognostic features. Four survival models were trained (RSF, XGB, GBM, and COX-LASSO) to estimate disease-free survival. This study was approved by the Institutional Ethics Committee of the Federal University of São Paulo (decision number 7.205.965).

Results: Seven key characteristics were selected for model development: age, neutrophil–lymphocyte ratio, T3 stage, positive lymph nodes, histological grade 3 (G3), estrogen receptor positivity, and human epidermal growth factor receptor 2 (HER2). Among the four machine learning models, the Gradient Boosting Machine (GBM) showed the best performance, achieving an area under the receiver operating characteristic curve (AUROC) of 0.610 (95% confidence interval 0.303–0.630). In the training cohort, the Random Survival Forest (RSF) model demonstrated the lowest overall Integrated Brier Score (IBS=0.075).

Conclusion: A large and diverse cohort of women with breast cancer was analyzed, including clinical, pathologic, and laboratory data. The final GBM model demonstrated consistent predictive performance across various cohorts. Further prospective randomized controlled studies are needed to enhance the accuracy and clinical applicability of the model.

Ethics: Research Ethics Committee Approval (CAAE 80509824.9.1001.5505).

Registry: ClinicalTrials.gov (NCT06447532).

Keywords: breast cancer; blood cell count; machine learning; prognosis.