ISSN 2282-0930 • Epidemiology Biostatistics and Public Health - XIII[th] SISMEC Congress - Vol. 20 Suppl. 1

Statistical Strategies for Olink Proteomics Data: A Comparative Approach and Future Directions

Galotta Arianna[(1)], Mattio Maria Francesco[(1)], Morocutti Chiara[(1)], Brambilla Marta[(1)], Camera Marina[(1)], Bonomi Alice[(1)]

(1) IRCCS Centro Cardiologico Monzino, Milan, Italy

CORRESPONDING AUTHOR: *Galotta Arianna, arianna.galotta@cardiologicomonzino.it*

INTRODUCTION

Olink® proteomics platforms offer a powerful tool for high-throughput biomarker discovery through multiplexed protein quantification. Their application in cardiovascular research provides novel opportunities to identify predictive biomarkers, but the complexity and dimensionality of the resulting Omics data require tailored statistical methodologies for robust analysis and interpretation.

OBJECTIVES

This study aimed to compare multiple statistical techniques to analyze Olink data from coronary artery disease patients, with the goal of identifying plasma biomarkers associated with cardiovascular mortality.

METHODS

We analyzed 69 plasma samples from patients with coronary artery disease, of whom 17 (24.6%) experienced cardiovascular mortality. Protein expression was assessed using four Olink Target 96 panels (cardiometabolic, cardiovascular II and III, inflammation), yielding 333 Normalized Protein eXpression (NPX) values. A multi-method analytical pipeline was employed, including univariate t-tests, principal component analysis (PCA), Gene Set Enrichment Analysis (GSEA), heatmap visualization, Boruta feature selection, and multivariate logistic regression with stepwise variable selection. Analyses were conducted using SAS v9.4 and R v4.3.1, including the OlinkAnalyze R package [1].

RESULTS

Initial univariate analyses did not identify statistically significant differences between outcome groups after multiple testing correction. Volcano plots of adjusted p-values confirmed this lack of significance. PCA revealed low explanatory power of the first two components, suggesting limited separation between cases and controls based on the protein profiles. GSEA and heatmap analyses failed to detect any significant enrichment patterns. In contrast, the Boruta algorithm identified several relevant features, which were further evaluated in a multivariate logistic regression model. Stepwise selection based on unadjusted p-values led to the development of a predictive model with good performance (AUC = 0.89, 95% CI: 0.81–0.97). Clinical collaboration played a key role in contextualizing these findings.

CONCLUSIONS

This study highlights the importance of integrating diverse statistical methodologies for the analysis of high-dimensional Olink proteomics data. While no single approach yielded definitive results, the combination of techniques allowed for the identification of promising biomarkers and construction of a performant predictive model. However, the small sample size remains a major limitation, affecting the robustness and reproducibility of the findings. Future research should explore the integration of synthetic data generation techniques to simulate larger datasets. This could enhance the stability of statistical inferences and allow more confident identification of clinically relevant biomarkers in small-scale Omics studies.

DOI: 10.54103/2282-0930/29393

ISSN 2282-0930 • Epidemiology Biostatistics and Public Health - XIII^th SISMEC Congress - Vol. 20 Suppl. 1

REFERENCES

1. Nevola K., Sandin M., Guess J. et al. OlinkAnalyze: Facilitate Analysis of Proteomic Data from Olink. R package version 4.2.0, 2025

DOI: 10.54103/2282-0930/29393