A Three-Layered Framework Integrating Classical, Multivariate, and Machine-Learning Methods for Systemic Treatment Effect Detection in High-Dimensional Biomarker Data

Monika Ošmianskienė

Vilnius University, Faculty of Medicine,
M. K. Čiurlionio g. 21, LT-03101 Vilnius, Lithuania
monika.osmianskiene@mf.stud.vu.lt

Abstract. Longevity supplement trials often rely on single-biomarker tests, which can miss distributed systemic effects. This paper proposes a three-layer analytical framework: (i) classical biostatistics, (ii) multivariate systems biology, and (iii) machine learning with responder analysis. Applied to a 99-participant trial with 20 biomarkers, Layer 1 found few isolated effects, Layer 2 detected significant multivariate separation, and Layer 3 supported reliable directional effects for nicotinamide adenine dinucleotide (NAD+) and low-density lipoprotein cholesterol (LDL-C). Together, the layers provide complementary evidence beyond any single method.

Keywords: biomarker analysis, framework, systems biology, machine learning, PCA.

1 Introduction

Evaluating whether an intervention produces a coherent systemic effect across many biomarkers is a recurring problem in biomedical informatics. In small, non-randomized studies univariate per-biomarker testing often fails to detect small, distributed signals, while purely multivariate or machine-learning analyses may overfit or lack interpretability [1]. Dietary supplement research, in particular, has been criticized for relying almost exclusively on classical biostatistics, which rarely captures system-level responses [2].

The aim of this work is to design and evaluate a three-layered analytical framework that integrates classical, multivariate and machine-learning methods into a single pipeline for detecting systemic treatment effects in high-dimensional biomarker data. The framework was applied to a 99-participant non-randomized study (79 treatment, 20 control) of a five-compo-

DOI: https://doi.org/10.15388/LMITT.2026.21

nent anti-ageing supplement (NAD+, resveratrol, berberine, quercetin, fisetin) collectively named as TLM01. 20 blood and epigenetic biomarkers were measured before (T0) and after a three-month intervention (T1). The central idea is methodological triangulation - three layers, each grounded in a different statistical paradigm, provide converging evidence that is more robust than any single method.

2 Three-Layered Framework

The framework processes a participant × biomarker change matrix $\Delta = T1 - T0$ and produces three complementary views of the treatment effect. All analyses were implemented in Python 3.14 using pandas, NumPy, SciPy, scikit-learn, statsmodels and scikit-bio.

Layer 1. Classical biostatistics

Layer 1 treats each biomarker independently. For each Δ the framework automatically selects a test based on Shapiro–Wilk normality and Levene's variance checks: Student's t-test, Welch's t-test or Mann–Whitney U. Strongly skewed biomarkers (C-reactive protein (CRP), insulin, homeostatic model assessment of insulin resistance (HOMA-IR)) are log-transformed. To account for baseline imbalance in non-randomized data, covariate-adjusted analysis (ANCOVA) T1 ~ Treatment + T0 is fitted [3], with optional sex and age covariates. Biomarkers are organized into primary, secondary and exploratory tiers.

Layer 2. Multivariate systems biology

Layer 2 treats the biomarker panel as a correlated system. All Δ values are standardized (z-scores) and reduced with principal component analysis (PCA). Group separation is tested (i) per principal component with Mann–Whitney U and (ii) globally in principal component (PC) space with a 10,000-permutation test on the Euclidean centroid distance. After collinearity filtering ($|r| > 0.9$), the squared Mahalanobis distance D^2_M between group centroids is computed from the pooled within-group covariance matrix [4]. Significance is assessed by 5,000 permutations. Nearest-centroid classification provides an interpretable accuracy measure.

Layer 3. Machine learning and responder analysis

Layer 3 shifts the question from "does the intervention work on average?" to "for whom does it work, and can treatment be distinguished at the individual level?". Random Forest, Gradient Boosting and Logistic Regression classifiers

are trained on 17 domain-relevant Δ features using stratified 5-fold cross-validation, with area under the receiver operating characteristic curve (AUC-ROC) compared against a 200-permutation null [5]. Multi-domain responder status is defined across five clinically coherent domains (inflammation, glycemic, lipid, epigenetic ageing, cellular energy) with strict (5/5) and lenient (≥3/5) thresholds. Propensity-score inverse probability weights adjust for non-randomized assignment, and 2,000 bootstrap resamples yield 95% CIs for each effect.

3 Case Study and Results

The framework was applied to the TLM01 trial (n = 99) covering NAD^+, inflammation, glycemic, lipid, tissue-damage and four epigenetic ageing biomarkers. A post-hoc power analysis for the primary endpoint (biological age Δ) indicated ~36% power, confirming the study is underpowered for univariate detection of typical supplement effects (d ≈ 0.3) and motivating the multi-layer approach.

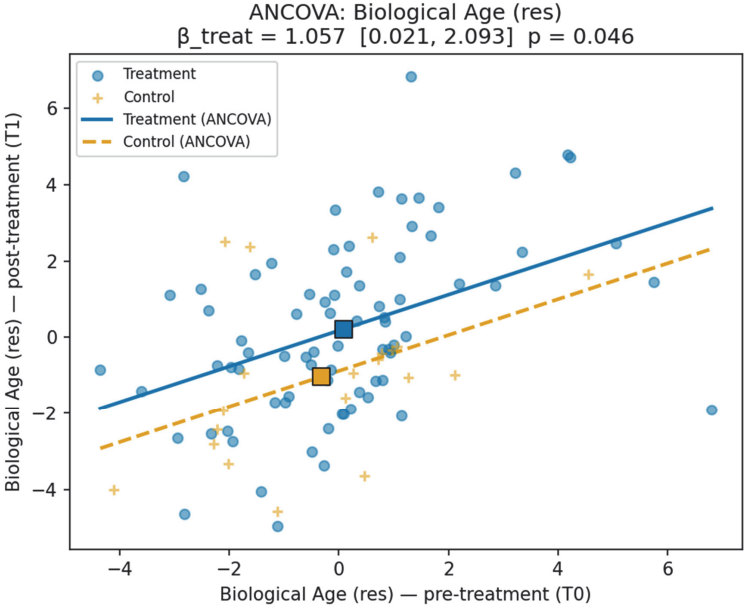

Figure 1. ANCOVA plot for Biological Age. The treatment group showed a significantly higher follow-up Biological Age residual than the control group, indicating an unfavorable adjusted treatment-associated difference.

Layer 1 yielded only one significant, but unfavorable direction (Figure 1). ANCOVA coefficient, biological age residual, β = +1.06 years, p = 0.046. No secondary endpoint was significant after baseline adjustment. LDL-C and non-high-density lipoprotein cholesterol (non-HDL-C) showed favorable but non-robust differences (Cohen's d ≈ −0.5).

Layer 2 produced a stronger signal. PCA required 11 components to explain 80% of variance (Figure 2), indicating highly distributed variation. Only PC1 (cholesterol-driven) separated groups significantly (Mann–Whitney p = 0.028). The PC-space permutation test was borderline (p = 0.070), but the Mahalanobis analysis on 17 decorrelated biomarkers was clearly significant (D^2_M = 2.09 versus a permutation null mean of 1.07 (p = 0.005)). Nearest-centroid classification correctly assigned 62/79 treatment and 14/20 control participants (76.8% accuracy).

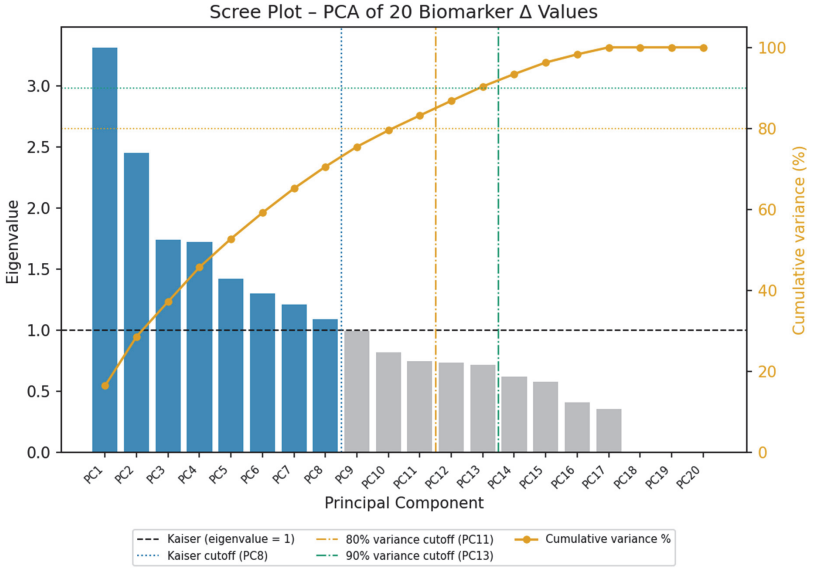

Figure 2. Scree plot of principal component eigenvalues for 20 standardized biomarker Δ values. Blue bars indicate components retained by the Kaiser criterion (eigenvalue > 1).

Layer 3 confirmed and extended these findings. No classifier exceeded chance significantly (best AUC 0.62, permutation p = 0.17). Bootstrap 95% confidence intervals (Cis) excluded zero for three biomarkers: NAD^+ (+1.98 [0.08, 3.92], favorable), LDL-C (−0.42 [−0.78, −0.04], favorable) and

interleukin-6 (IL-6) (+3.83 [0.47, 9.00], unfavorable). Multi-domain responder rates did not differ significantly (strict 45.6% vs 40.0%, p = 0.80). Table 1 summarizes each layer's contribution.

Table 1. Evidence produced by each layer of the framework.

Layer	Primary output	Key result on TLM01 data
1. Classical	ANCOVA β, per-biomarker p-values	ANCOVA showed only biological age significant (β = +1.06 y, p = 0.046, unfavorable). Univariate analysis of biological age showed a borderline and unfavorable pattern for TLM01.
2. Multivariate	PCA, Mahalanobis D^2_M, permutation p	D^2_M = 2.09, permutation p = 0.005, 76.8% nearest-centroid accuracy showed significant multivariate separation. PC1 (lipid-loaded, 16.4% of variance) differed between groups (p = 0.028) revealed a modest lipid-driven separation.
3. ML / responder	AUC-ROC, bootstrap CI, responder rate	Classifiers at chance. Reliable favorable effects for NAD+ and LDL-C; IL-6 unfavorable.

4 Discussion

The three layers produced evidence that no single layer could. Layer 1 alone suggested the intervention was ineffective for primary endpoint (biological age). Layer 2 revealed a moderate systemic separation in the full biomarker space mainly driven by lipids. Layer 3 localized it to two favorable (NAD^+, LDL-C) and one unfavorable (IL-6) directionally reliable effects. Significant multivariate separation is compatible with non-significant univariate and classifier results when effects are small, distributed and partly confounded by baseline imbalance [1]. To the author's knowledge, no published supplement study combines per-biomarker hypothesis testing, PCA/Mahalanobis-based multivariate analysis and ML-based responder classification in a single framework. The framework is domain-agnostic and reproducible. The main limitation is sample size with n = 99 and a 79:20 split, Layers 2 and 3 operate near the lower bound of or below published recommendations.

5 Conclusions

A three-layered analytical framework integrating classical biostatistics, multivariate systems biology and machine-learning responder analysis was designed and applied to a 20-biomarker supplement trial. The multivariate layer detected a systemic separation invisible to univariate tests, and the ML layer identified two directionally reliable favorable effects. The three layers ask different questions: "which biomarker moved?", "is the system as a whole different?", "who responded?". Their combined output provides a more honest picture of the intervention than any single analysis. The framework is reproducible and applicable to any high-dimensional biomarker dataset in which small-to-moderate systemic effects are plausible.

References

[1] M. R. Munafò and G. Davey Smith, "Robust research needs many lines of evidence," *Nature*, vol. 553, no. 7689, pp. 399–401, Jan. 2018, doi: 10.1038/d41586-018-01023-3.

[2] R. L. Bailey *et al.*, "Major Gaps in Understanding Dietary Supplement Use in Health and Disease," *Annu. Rev. Nutr.*, vol. 43, no. 1, pp. 179–197, Aug. 2023, doi: 10.1146/annurev-nutr-011923-020327.

[3] A. J. Vickers and D. G. Altman, "Analysing controlled trials with baseline and follow up measurements," *BMJ*, vol. 323, no. 7321, pp. 1123–1124, Nov. 2001, doi: 10.1136/bmj.323.7321.1123.

[4] "Reprint of: Mahalanobis, P.C. (1936) 'On the Generalised Distance in Statistics.,'" *Sankhya A*, vol. 80, no. S1, pp. 1–7, Dec. 2018, doi: 10.1007/s13171-019-00164-5.

[5] G. Varoquaux, "Cross-validation failure: Small sample sizes lead to large error bars," *NeuroImage*, vol. 180, pp. 68–77, Oct. 2018, doi: 10.1016/j.neuroimage.2017.06.061.

During the preparation of this manuscript, the author used ChatGPT (OpenAI) for English language polishing, including grammar, clarity, and style. All suggestions were critically reviewed and edited by the author, who takes full responsibility for the final content of the manuscript.