Enhancing biological data analysis: A critical examination of nonlinear and nonparametric statistical methods versus linear approaches

To the Editor: Liu et al[1] investigated treatment outcomes of biologic therapy in super-responders and biologic-refractory psoriasis patients through a single-center retrospective study in China. Their retrospective analysis included psoriasis patients who initiated their first biologic therapy. Multivariable logistic regression analysis identified predictors for super-responders and biologic-refractory patients.[1]

This study highlights significant concerns regarding the use of logistic regression for determining feature importance, primarily due to its model-specific nature, which can lead to erroneous conclusions. Although supervised machine learning models like logistic regression have established ground truth values for validating target prediction accuracy, the feature importance derived from these models does not benefit from similar validation criteria. The absence of ground truth values means that various models may employ distinct methodologies for calculating feature importance, potentially resulting in biased interpretations. Notably, high target prediction accuracy does not inherently assure the reliability of feature importance. Over 100 peer-reviewed studies have documented substantial biases in feature importance derived from machine learning models, including logistic regression.[2-5]

The use of linear methods and parametric approaches on biological data, which are inherently nonlinear and nonparametric, can significantly distort analytical outcomes and result in misleading conclusions. Unlike supervised machine learning techniques, which can validate prediction accuracy against known target values, feature importance scores derived from these models often lack corresponding ground truth benchmarks for validation. This absence of reliable validation introduces inherent bias in the interpretation of feature importance, potentially compromising the integrity of research findings. This paper critically examines the application of linear and parametric methods, such as logistic regression, for calculating feature importance and analyzing complex biological data sets that may not conform to the assumptions of these methodologies.

To enhance clarity for diverse readers, this study defines key technical terms. Feature importance indicates how each patient characteristic, like body mass index or the presence of psoriatic arthritis, influences the prediction of biologic therapy outcomes in psoriasis treatment. Ground truth values refer to observed clinical outcomes, noting that 19.0% of the patients became super-responders and 4.6% biologic-refractory. Target prediction accuracy assesses how closely model predictions align with these outcomes. Additionally, a machine learning model analyzes patient data to predict treatment responses without explicit programming. Understanding these concepts is crucial for interpreting the multivariable logistic regression analysis presented in the study.

Logistic regression, a commonly used parametric method for binary outcomes, can yield distorted outcomes when applied to nonlinear and nonparametric biological data, leading to incorrect conclusions. Three critical components must be considered when calculating feature importance. First, analyze data distributions to evaluate normality, patterns, outliers, and both univariate and multivariate characteristics, including skewness and kurtosis. Second, assess statistical relationships between variables, covering linear and nonlinear associations, direct and indirect relationships, interaction effects, and multicollinearity. Third, validate statistical significance through comprehensive *P* value assessments, multiple testing corrections, effect size calculations, and CI estimations.

This paper advocates for the use of robust nonlinear and nonparametric methods, including Spearman's correlation, Kendall's tau, and Goodman-Kruskal gamma, accompanied by *P* values, whereas Mutual Information analysis provides complex interactions among multiple variables.

Yoshiyasu Takefuji, PhD

From the Faculty of Data Science, Musashino University, Tokyo, Japan.

Funding sources: None.

IRB approval status: Not applicable.

Key words: biological data modeling; feature importance analysis; nonlinear biological data; nonparametric statistics; robust statistical methods; statistical validation.

Correspondence to: Yoshiyasu Takefuji, PhD, Faculty of Data Science, Musashino University, 3-3-3 Ariake Koto-ku, Tokyo 135-8181, Japan

E-mail: takefuji@keio.jp

Conflicts of interest

None disclosed.

COR 5.7.0 DTD ■ YMJD19986_proof ■ 19 July 2025 ■ 11:05 am

REFERENCES

1. Liu Y, Hu K, Duan Y, Chen X, Zhang M, Kuang Y. Characterization and treatment outcomes of biologic therapy in super-responders and biologic-refractory psoriasis patients: a single-center retrospective study in China. *J Am Acad Dermatol*. 2025;93(1):46-54. https://doi.org/10.1016/j.jaad.2025.02.063
2. Fisher A, Rudin C, Dominici F. All models are wrong, but many are useful: learning a variable's importance by studying an entire class of prediction models simultaneously. *J Mach Learn Res*. 2019;20:177.
3. Steiner PM, Kim Y. The Mechanics of omitted variable bias: bias amplification and cancellation of offsetting biases. *J Causal Inference*. 2016;4(2):20160009. https://doi.org/10.1515/jci-2016-0009
4. Nazer LH, Zatarah R, Waldrip S, et al. Bias in artificial intelligence algorithms and recommendations for mitigation. *PLoS Digit Health*. 2023;2(6):e0000278. https://doi.org/10.1371/journal.pdig.0000278
5. Pinheiro-Guedes L, Martinho C, O Martins MR. Logistic regression: limitations in the estimation of measures of association with binary health outcomes. *Acta Med Port*. 2024;37(10):697-705. https://doi.org/10.20344/amp.21435

https://doi.org/10.1016/j.jaad.2025.04.091