Beyond Parametric Assumptions: Unsupervised Methods Enhance DNA Repair Gene Discovery in *SMARCAL1* Identification

TO THE EDITOR:

Oak et al[1] investigated DNA damage response genes in pediatric cancers and validated the central role of DNA repair in cancer risk, identifying *SMARCAL1* as a novel osteosarcoma predisposition gene. Using logistic and Firth regression, they found genes with predisposing variants enriched in tumors and replicated these signals in 1,497 additional childhood cancer cases across three independent cohorts.[1] These results underscore the value of rigorous statistical replication in genetic predisposition studies.

However, it is important to recognize the limitations of logistic and Firth regression for interpreting feature importance because of parametric nature of both tools. These parametric models can yield misleading inferences when labels are noisy or when assumptions are not met. Common issues include dependence among observations, misspecification of the link function (eg, nonlinearity in the logit), multicollinearity, and an inadequate event per variable ratio. Violations can distort coefficient estimates, standard errors, z statistics, *P* values, CIs, and odds ratios. Given their widespread use in clinical research, careful diagnostics, sensitivity analyses, and complementary nonparametric or machine learning approaches should be part of a robust analytic workflow.[2-14]

A further conceptual point is the distinction between target prediction accuracy and feature importance accuracy in supervised models like logistic regression. Target prediction accuracy can be judged against ground truth labels, but feature importance does not have a direct ground truth, and coefficients primarily reflect contributions to the model's predictions under its assumptions rather than true biological associations. As a result, high predictive performance does not guarantee reliable or stable feature importance rankings.

To strengthen inference about biologic relevance, researchers can adopt multifaceted strategies that reduce reliance on parametric assumptions and label-driven artifacts. Unsupervised methods such as feature agglomeration and highly variable gene selection can reveal structure in the data independent of outcomes. These can be followed by nonparametric, nonlinear association measures like Spearman correlation with *P* values to assess relationships without imposing strict model forms. While supervised models like logistic regression may exhibit instability in feature ranking because of label-driven errors, approaches such as feature agglomeration, Highly Variable Gene Selection, and Spearman correlations can provide more stable rankings by decoupling feature evaluation from outcome labels. Together, these complementary methods can improve robustness and interpretability in genomic analyses.

Yoshiyasu Takefuji, PhD
Faculty of Data Science, Musashino University, Tokyo, Japan

AUTHOR'S DISCLOSURES OF POTENTIAL CONFLICTS OF INTEREST

Disclosures provided by the author are available with this article at DOI https://doi.org/10.1200/JCO-25-02471.

ACKNOWLEDGMENT

According to ScholarGPS, Yoshiyasu Takefuji holds notable global rankings in several fields. He ranks 54th out of 395,884 scholars in neural networks (AI), 23rd out of 47,799 in parallel computing, and 14th out of 7,222 in parallel algorithms. Furthermore, he ranks the highest in AI tools and human-induced error analysis, underscoring his significant contributions to these domains.

REFERENCES

1. Oak N, Chen W, Blake A, et al: Investigation of DNA damage response genes validates the role of DNA repair in pediatric cancer risk and identifies SMARCAL1 as novel osteosarcoma predisposition gene. J Clin Oncol 43:3833-3843, 2025
2. Dey D, Haque MS, Islam MM, et al: The proper application of logistic regression model in complex survey data: A systematic review. BMC Med Res Methodol 25:15, 2025
3. Pinheiro-Guedes L, Martinho C, Martins MRO: Logistic regression: Limitations in the estimation of measures of association with binary health outcomes. Acta Med Port 37:697-705, 2024
4. Wang T, Tang W, Lin Y, et al: Semi-supervised inference for nonparametric logistic regression. Stat Med 42:2573-2589, 2023
5. Osborne J: A practical guide to testing assumptions and cleaning data for logistic regression, in A Practical Guide to Testing Assumptions and Cleaning Data for Logistic Regression. SAGE Publications, Ltd, 2015, pp 84-130
6. van Maanen L, Katsimpokis D, van Campen AD: Fast and slow errors: Logistic regression to identify patterns in accuracy–response time relationships. Behav Res Methods 51:2378-2389, 2019
7. Work JW, Ferguson JG, Diamond GA: Limitations of a conventional logistic regression model based on left ventricular ejection fraction in predicting coronary events after myocardial infarction. Am J Cardiol 64:702-707, 1989
8. Zulfadhli M, Budiantara IN, Ratnasari V: Nonparametric regression estimator of multivariable Fourier series for categorical data. MethodsX 13:102983, 2024
9. Akturk B, Beyaztas U, Shang HL, et al: Robust functional logistic regression. Adv Data Anal Classif 19:121-145, 2025
10. Rifada M, Chamidah N, Ningrum RA: Estimation of nonparametric ordinal logistic regression model using generalized additive models (GAM) method based on local scoring algorithm. AIP Conf Proc 2668:070013, 2022

Corresponding author: Yoshiyasu Takefuji, PhD; e-mail: takefuji@keio.jp.

11. Suliyanto, Rifada M, Tjahjono E: Estimation of nonparametric binary logistic regression model with local likelihood logit estimation method (case study of diabetes mellitus patients at Surabaya Hajj General Hospital). AIP Conf Proc 2264:030007, 2020
12. Wibowo W, Amelia R, Octavia FA, et al: Classification using nonparametric logistic regression for predicting working status. AIP Conf Proc 2329:060032, 2021
13. Steyerberg EW, Schemper M, Harrell FE: Logistic regression modeling and the number of events per variable: Selection bias dominates. J Clin Epidemiol 64:1464, 2011
14. Özkale MR: Iterative algorithms of biased estimation methods in binary logistic regression. Stat Pap 57:991-1016, 2016

DOI: https://doi.org/10.1200/JCO-25-02471; Published at ascopubs.org/journal/jco on March 12, 2026.

AUTHOR'S DISCLOSURES OF POTENTIAL CONFLICTS OF INTEREST

Beyond Parametric Assumptions: Unsupervised Methods Enhance DNA Repair Gene Discovery in *SMARCAL1* Identification

The following represents disclosure information provided by the author of this manuscript. All relationships are considered compensated unless otherwise noted. Relationships are self-held unless noted. I = Immediate Family Member, Inst = My Institution. Relationships may not relate to the subject matter of this manuscript. For more information about ASCO's conflict of interest policy, please refer to www.asco.org/rwc or ascopubs.org/jco/authors/author-center.

Open Payments is a public database containing information reported by companies about payments made to US-licensed physicians (Open Payments).
No potential conflicts of interest were reported.