BJA

British Journal of Anaesthesia, xxx (xxx): xxx (xxxx)

CORRESPONDENCE

Development and validation of a predictive model for transfusion in major abdominal surgery: the case for nonparametric methods. Comment on *Br J Anaesth* 2025; 135: 623–31

Yoshiyasu Takefuji

Faculty of Data Science, Musashino University, Tokyo, Japan

E-mail: takefuji@keio.jp

Keywords: biological datasets; machine learning; nonparametric analysis; parametric methods; predictive modelling

Editor—Sim and colleagues[1] conducted a multicentre retrospective study to develop and validate a predictive model for intraoperative red blood cell (RBC) transfusion in major abdominal surgery. Their approach involved using multivariate logistic regression, a statistical method commonly applied in scenarios where the outcome variable is binary (e.g. whether or not a transfusion occurs). They incorporated the least absolute shrinkage and selection operator (LASSO) technique to optimise variable selection in constructing the prediction model. LASSO is particularly useful for addressing potential multicollinearity among predictors and helps identify significant variables by applying a penalty that reduces the influence of less relevant predictors, effectively shrinking their coefficients to zero. This penalty is controlled by a regularisation parameter that determines the strength of the shrinkage effect, allowing researchers to balance model complexity against predictive performance. The final predictive model was constructed by retaining only those variables with non-zero coefficients, which enhances model applicability and accuracy within clinical settings.[1] This rigorous variable selection process is critical to ensuring that the model is both interpretable and clinically useful, as it eliminates unnecessary variables that might introduce noise rather than meaningful predictive power.

However, it is important to note that Sim and colleagues[1] might not have fully understood the fundamental principles underlying machine learning methods. Individual data analysis tools, such as logistic regression and LASSO, are built on specific underlying assumptions against the data. Logistic regression assumes parametric distribution. Similarly, LASSO is also designed for linear relationships and inherits many of the assumptions of the underlying linear model it regularises. If these assumptions are violated, such as when applying linear models to nonlinear data or when using parametric methods (such as logistic regression or LASSO) on nonparametric data, the results can be inherently distorted. In nonlinear relationships, for instance, a linear model would fail to capture the true underlying pattern, potentially missing important interactions or threshold effects in the data. This can lead to inaccurate *P*-values and misinterpretations regarding predictor importance, ultimately resulting in flawed conclusions about which variables truly predict transfusion requirements. It is essential for researchers to consider the characteristics of their data and choose appropriate methodologies to ensure the validity of their findings, perhaps by incorporating preliminary tests for linearity or considering more flexible modelling approaches when necessary.

Logistic regression is a parametric method; when it is applied to nonparametric data, as is often observed in biological analysis, the outcomes are potentially skewed, leading to erroneous conclusions.[2–8] Parametric methods such as logistic regression make strong assumptions about the underlying distribution of the data, typically assuming that relationships follow specific mathematical forms and that residuals are normally distributed. However, biological data frequently exhibit complex, non-normal distributions, threshold effects, and intricate interdependencies among variables that violate these assumptions. When these violations occur, the estimated coefficients, standard errors, and resulting *P*-values and feature importances can be severely biased, leading researchers to either miss truly important predictors or incorrectly identify spurious associations. LASSO assumes linear and parametric relationships between

DOI of original article: 10.1016/j.bja.2025.05.048.

predictors and outcomes, and when it is applied to nonlinear nonparametric data, violations are even more severe, leading to erroneous interpretations.[9–15] The penalty term in LASSO operates under the assumption that the true model is sparse and that relationships are linear, which might not hold in complex biological systems. When applied to nonlinear data, LASSO might erroneously eliminate important variables whose effects are nonlinear or manifest only in interactions with other variables, while retaining less important but linearly related predictors, thus compromising both model interpretability and predictive performance.

Currently, there are no algorithms to accurately calculate true associations between variables, so Sim and colleagues[1] advocate for use of multifaceted approaches using unsupervised machine learning models such as feature agglomeration and highly variable gene selection, followed by nonlinear nonparametric methods such as Spearman's correlation with *P*-values for monotonic relationships. Feature agglomeration is a dimensionality reduction technique that clusters similar features together, effectively reducing redundancy in the data while preserving its intrinsic structure. Unlike parametric methods, feature agglomeration makes no assumptions about linear relationships, instead grouping variables based on their natural similarities across the dataset. This approach is particularly valuable when dealing with high-dimensional biological data where complex interdependencies exist. Similarly, highly variable gene selection, although traditionally used in genomics, represents a concept applicable to identifying features with the highest information content across samples, focusing analytical attention on variables with the greatest potential predictive power. After this initial feature reduction, Spearman's rank correlation with *P*-values offers distinct advantages over parametric alternatives such as Pearson's correlation, as it assesses monotonic relationships (where variables tend to change together but not necessarily at a constant rate) rather than strictly linear ones. By converting values to ranks before calculating correlations, Spearman's approach becomes resistant to outliers and makes no assumptions about the underlying data distribution, making it particularly suitable for biological datasets where relationships often follow complex patterns but still exhibit important monotonic trends. This comprehensive approach acknowledges the inherent complexity of biological systems while providing robust statistical foundations for identifying meaningful associations.

Researchers should begin with a quality assessment of datasets, including evaluation of distributional properties, missingness, outliers, and redundancy. When considering logistic regression, unsupervised clustering can be used as a diagnostic to assess latent heterogeneity and overall data structure; evidence of multiple, well-separated subgroups can suggest the need for stratified analyses, interaction terms, nonparametric transformations, or mixture models, whereas a cohesive structure may support a single global model with appropriate diagnostics. When the optimum number of clusters is three or higher, logistic regression should not be applied, as forcing multiple latent states into binary outcomes can induce erroneous biases and model misspecification. Unsupervised methods can help mitigate label-driven bias and often yield more stable feature rankings when labels are limited or noisy. For example, feature agglomeration, highly variable gene selection, and rank-based measures such as Spearman correlation can reduce dimensionality, address multicollinearity, and provide stable feature prioritisation. In contrast, supervised methods such as LASSO and logistic regression can display variability in selected features across resamples because of model specification and sampling variation; reporting stability is therefore recommended. Finally, feature importance derived from supervised models should be interpreted as contributing to predictive performance within the specified model rather than as evidence of causal or mechanistic association. Supervised models involve two distinct notions of accuracy: target prediction accuracy (validation against given labels) and feature-importance reliability (in the absence of ground truth); high predictive accuracy does not guarantee reliable feature importances.

Declaration of interest

The author declares that they have no conflict of interest.

References

1. Sim JH, Oh AR, Kim S, et al. Development and validation of a predictive model for transfusion in major abdominal surgery: a multicentre retrospective study. *Br J Anaesth* 2025; **135**: 623–31
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* 2025; **25**: 15
3. 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**: 697–705
4. Wang T, Tang W, Lin Y, Su W. Semi-supervised inference for nonparametric logistic regression. *Statistics in Medicine* 2023; **42**: 2573–89
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. p. 84–130. Vol 0
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* 2019; **51**: 2378–89
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* 1989; **64**: 702–7
8. Zulfadhli M, Budiantara IN, Ratnasari V. Nonparametric regression estimator of multivariable Fourier series for categorical data. *MethodsX* 2024; **13**, 102983
9. Wüthrich K, Zhu Y. Omitted variable bias of lasso-based inference methods: a finite sample analysis. *Rev Econ Stat* 2023; **105**: 982–97
10. Freo M, Luati A. Lasso-based variable selection methods in text regression: the case of short texts. *AStA Adv Stat Anal* 2024; **108**: 69–99
11. Basu T, Einbeck J, Troffaes MCM. Uncertainty quantification in lasso-type regularization problems. In: Vasile M, editor. *Optimization Under Uncertainty With Applications to Aerospace Engineering*. Cham: Springer; 2021. p. 11–20

12. Fridgeirsson EA, Williams R, Rijnbeek P, Suchard MA, Reps JM. Comparing penalization methods for linear models on large observational health data. *J Am Med Inform Assoc* 2024; **31**: 1514–21
13. Hernández-Lemus E, Ochoa S. Methods for multi-omic data integration in cancer research. *Front Genet* 2024; **15**, 1425456
14. Li X, Jacobucci R. Regularized structural equation modeling with stability selection. *Psychol Methods* 2022; **27**: 497–518
15. Jain R, Xu W. HDSI: High dimensional selection with interactions algorithm on feature selection and testing. *PLoS One* 2021; **16**, e0246159

doi: 10.1016/j.bja.2025.08.059