Interpreting Variable Importance in Metabolically Co-Regulated Systems: Considerations for Metabolic Dysfunction–associated Steatotic Liver Disease Risk Prediction

Ho *et al.* provide an important contribution by integrating multimodal UK Biobank data to develop supervised learning models for metabolic dysfunction–associated steatotic liver disease (MASLD) risk prediction.[1] Their study incorporated magnetic resonance imaging–proton density fat fraction (MRI-PDFF), electronic health records, and extensive metabolomic profiling from 82,638 participants. Comparison across supervised learning algorithms, including partial least squares discriminant analysis (PLS-DA), Random Forest (RF), support vector machine (linear kernel; SVM-linear), and logistic regression, yielded area under the curve (AUROC) values 0.80–0.85. Nevertheless, several methodological considerations related to metabolic co-regulation, nonlinear effects, and feature prioritization may influence interpretability. In particular, whether the analytical framework fully captures the hierarchical, co-regulated, and nonlinear architecture of metabolic dysfunction warrants further discussion.

Model-derived feature importance may obscure interdependent relationships among predictors. In linear classifiers such as SVM-linear and logistic regression, metabolic and clinical features relevant to MASLD are unlikely to act as independent risk factors and instead reflect shared latent processes linked to insulin resistance. This is evident in Table 1, where body mass index (BMI) (25.08–30.09), waist–hip ratio (0.50–0.58), triglycerides (1.30–1.94 mmol/L), high-density lipoprotein cholesterol (HDL-C) (1.44–1.20 mmol/L), and hepatic injury markers—including alanine aminotransferase (ALT), aspartate aminotransferase (AST), and gamma-glutamyl transferase (GGT)—shift in a coordinated manner between groups. They also note that 'the metabolites measured consisted of amino acids, lipid, lipoproteins and their compositions,' supporting strong biochemical co-regulation. Consequently, many predictors likely share variance from latent structure, and interpreting variable importance within linear modeling frameworks misattributes shared variance to individual predictors rather than underlying metabolic axes.[2,3]

A second consideration relates to nonlinear and threshold relationships. MASLD itself is defined using a threshold (MRI-PDFF ≥5%), and many metabolic traits—including BMI, glycated hemoglobin (HbA1c), triglycerides, and blood pressure—exhibit clinically recognized nonlinear or saturating effects. Nonlinear behavior is a core feature of metabolic regulation, and overlooking it can materially alter the apparent contribution of individual predictors. Yet, the present study does not report the use of splines, generalized additive models, or other approaches to explore nonlinear associations.[4] While the primary objective is prediction, they subsequently interpret selected features in biological terms. In this context, explicit evaluation of nonlinear and threshold effects would enhance robustness and interpretability, as AUROC optimization alone does not necessarily ensure that model behavior aligns with underlying biological relationships.

Third, the feature selection strategy may be influenced by collinearity and model-specific biases. They state that 'variable importance analyses of PLS-DA and Random Forest were used to select important features,' while also acknowledging that 'collinearity of variables may create clustering bias.' In high-dimensional metabolomic data, strongly correlated variables—such as triglycerides, HDL-C, ALT, GGT, BMI, and waist–hip ratio—compete for prioritization within RF-based feature selection procedures, leading to the preferential selection of certain representatives.[5] Such competition among correlated predictors results in unstable feature rankings across resampling or model specifications. These instabilities are especially relevant given that the selected features were subsequently used for clustering, where model-driven variability influences subgroup definitions. Moreover, RF Gini importance is known to favor continuous variables with broader ranges.[6] As a result, the selected predictors reflect model-dependent prioritization rather than stable biological signals, which has implications for downstream interpretation and clustering analyses.[7]

To complement the current framework without substantially altering the modeling pipeline, we suggest incorporating lightweight unsupervised and nonparametric analytical strategies that better reflect the latent structure and nonlinear behavior of metabolic systems. Feature Agglomeration can cluster highly correlated variables and derive representative features, reducing redundancy and stabilizing feature selection.[8] An adapted Highly Variable Gene Selection can function as a general variance--

https://doi.org/10.1016/j.jceh.2026.103648

© 2026 Indian National Association for Study of the Liver. Published by Elsevier B.V. All rights reserved, including those for text and data mining, AI training, and similar technologies.

Please cite this article as: Oka et al., Interpreting Variable Importance in Metabolically Co-Regulated Systems: Considerations for Metabolic Dysfunction–associated Steatotic Liver..., Journal of Clinical and Experimental Hepatology, https://doi.org/10.1016/j.jceh.2026.103648

based unsupervised filter for metabolomic and biochemical data, enabling extraction of high-information features without imposing model-specific biases.[9] For nonlinear and rank-based relationships, Spearman's rho and Kendall's tau serve as complementary exploratory tools to identify monotonic relationships that may be overlooked by model-dependent variable attribution.[10] Incorporating these approaches would not replace the supervised models but would provide a more faithful representation of metabolic structure and enhance interpretability.

In conclusion, Ho *et al.* provide an important framework for multimodal MASLD risk prediction using large-scale population data. However, the co-regulated, hierarchical, and nonlinear nature of metabolic dysfunction suggests that additional safeguards may be needed when interpreting variable importance biologically. Ensuring that feature attribution remains robust is particularly important when predictive models are also used to generate pathophysiological insights. Incorporating complementary unsupervised and flexible analytical strategies improves the stability and interpretability of identified predictors, thereby strengthening both the predictive utility and the biological credibility of the proposed framework.

CREDIT AUTHORSHIP CONTRIBUTION STATEMENT

Souichi Oka: Writing – original draft, Conceptualization. **Maito Suzuki**: Investigation. **Yoshiyasu Takefuji**: Writing – review and editing, Supervision, Project administration.

FUNDING

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

DECLARATION OF COMPETING INTEREST

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

DATA AVAILABILITY

No new data were created or analyzed in this study.

REFERENCES

1. Ho B, Xu Y, Goldring CE, Goodacre R, Pirmohamed M. Delineating metabolic dysfunction-associated steatotic liver disease and comorbid risk with clinical, blood and metabolomic parameters: a UK biobank analysis. *J Clin Exp Hepatol*. 2026;16:103537. https://doi.org/10.1016/j.jceh.2026.103537.
2. Anandhi P, Nathiya E. Application of linear regression with their advantages, disadvantages, assumption and limitations. *Int J Stat Appl Math*. 2023;8:133–137. https://doi.org/10.22271/maths.2023.v8.i6b.1463.
3. Pinheiro-Guedes L, Martinho C, Martins MR. Logistic regression: limitations in the estimation of measures of association with binary health outcomes. *Acta Med Port*. 2024;37:697–705. https://doi.org/10.20344/amp.21435.
4. 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*. 2022;2668070013 https://doi.org/10.1063/5.0111771.
5. Faragalli A, Ferrante L, Angelakopoulos N, Cameriere R, Skrami E. Do machine learning methods solve the main pitfall of linear regression in dental age estimation? *Forensic Sci Int*. 2025;367112353 https://doi.org/10.1016/j.forsciint.2024.112353.
6. Strobl C, Boulesteix AL, Zeileis A, Hothorn T. Bias in random forest variable importance measures: illustrations, sources and a solution. *BMC Bioinf*. 2007;8:25. https://doi.org/10.1186/1471-2105-8-25.
7. 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. https://doi.org/10.48550/arXiv.1801.01489.
8. Zhang J, Wu X, Hoi SCH, Zhu J. Feature agglomeration networks for single-stage face detection. *Neurocomputing*. 2020;380:180–189. https://doi.org/10.1016/j.neucom.2019.10.087.
9. Xie Y, Jing Z, Pan H, Xu X, Fang Q. Redefining the high variable genes by optimized LOESS regression with positive ratio. *BMC Bioinf*. 2025;26:104. https://doi.org/10.1186/s12859-025-06112-5.
10. Okoye K, Hosseini S. Correlation tests in R: Pearson Cor, Kendall's Tau, and Spearman's Rho. In: Okoye K, Hosseini S, eds. *R Programming: Statistical Data Analysis in Research*. Singapore: Springer Nature; 2024:247–277. https://doi.org/10.1007/978-981-97-3385-9_12.

Souichi Oka
Faculty of Data Science, Musashino University, 3-3-3 Ariake, Koto-ku, Tokyo, 135-8181, Japan

Maito Suzuki, Yoshiyasu Takefuji
Science Park Corporation, 3-24-9 Iriya-Nishi, Zama-shi, Kanagawa, 252-0029, Japan

Address for correspondence: Souichi Oka, Faculty of Data Science, Musashino University, 3-3-3 Ariake, Koto-ku, Tokyo 135-8181, Japan.
E-mail: s-oka@musashino-u.ac.jp (S. Oka)

Received: 3 August 2026.
Accepted: 20 August 2026.

© 2026 Indian National Association for Study of the Liver. Published by Elsevier B.V. All rights are reserved, including those for text and data mining, AI training, and similar technologies.

Please cite this article as: Oka et al., Interpreting Variable Importance in Metabolically Co-Regulated Systems: Considerations for Metabolic Dysfunction–associated Steatotic Liver..., Journal of Clinical and Experimental Hepatology, https://doi.org/10.1016/j.jceh.2026.103648