Received: 10 October 2025 | Accepted: 15 October 2025

DOI: 10.1097/HEP.0000000000001604

CORRESPONDENCE

Letter to the Editor: Beyond AUROC—Critical evaluation of FSTL-1 as a liver fibrosis biomarker

To the editor,

Li et al.[1] investigated plasma "FSTL-1" as a non-invasive biomarker for advanced liver fibrosis. Using ordered logistic regression controlling for albumin, AST, BMI, and glucose, they found significantly elevated "FSTL-1" in liver tissue of patients with advanced fibrosis (AUROC = 0.83; $p < 0.0001$), strongly correlated with LSM from transient elastography. Peripheral "FSTL-1" showed similar diagnostic performance (AUROC = 0.85; $p < 0.0001$), indicating plasma levels mirror hepatic expression.[1] This suggests a meaningful advance over invasive biopsies and limited-accuracy methods.

However, key methodological issues in biomarker development were overlooked. Supervised models (eg, logistic regression) embed 2 different accuracy notions: target prediction accuracy and feature importance accuracy. Only prediction accuracy can be validated against labels; feature importance lacks direct validation. Thus, high AUROC does not guarantee stable or reproducible importance rankings.[2–4] Because AUROC is label-dependent, using it for feature selection without evaluating susceptibility to label noise, class imbalance, and confounding can compromise biomarker validity.

This issue is critical in clinical biomarker research, where relative feature importance drives marker inclusion in diagnostic panels. If importance rankings are unstable, panels may underperform across cohorts or settings, limiting translational utility despite strong initial AUROC.

To illustrate, as shown in Table 1 we analyzed a public liver biomarker dataset (OpenML ID 43403), comparing supervised methods (Random Forest, XGBoost, logistic regression), unsupervised methods (feature agglomeration "FA" and highly variable gene selection "HVGS"), and non-target approaches (Spearman correlation). We assessed cross-validated predictive performance and the stability of feature ranking orders under perturbation (removing top features) (Table 1).

Spearman correlation achieved both high accuracy (0.7136) and perfect order stability upon top-feature removal. "FA" and "HVGS" also maintained perfect ranking stability with slightly lower accuracy (0.6947). In

TABLE 1 Cross-validation accuracy and feature ranking order stability

Model	CV accuracy	Top 5 feature ranking orders	Top 4 feature ranking orders without the highest feature
Random forest	0.6964	Alkaline_Phosphotase, Aspartate_Aminotransferase, Age, Alamine_Aminotransferase, Total_Bilirubin	Alamine_Aminotransferase, Aspartate_Aminotransferase, Age, Total_Bilirubin
XGBoost	0.6878	Direct_Bilirubin, Total_Bilirubin, Alkaline_Phosphotase, Alamine_Aminotransferase, Age	Total_Bilirubin, Alamine_Aminotransferase, Albumin, Age
Logistic regression	0.7118	Albumin, Albumin_and_Globulin_Ratio, Total_Protiens, Direct_Bilirubin, Gender	Albumin_and_Globulin_Ratio, Direct_Bilirubin, Total_Protiens, Gender
FA	0.6947	Aspartate_Aminotransferase, Alkaline_Phosphotase, Alamine_Aminotransferase, Age, Total_Bilirubin	Alkaline_Phosphotase, Alamine_Aminotransferase, Age, Total_Bilirubin
HVGS	0.6947	Aspartate_Aminotransferase, Alkaline_Phosphotase, Alamine_Aminotransferase, Age, Total_Bilirubin	Alkaline_Phosphotase, Alamine_Aminotransferase, Age, Total_Bilirubin
Spearman	0.7136	Aspartate_Aminotransferase, Total_Bilirubin, Direct_Bilirubin, Alamine_Aminotransferase, Alkaline_Phosphotase	Total_Bilirubin, Direct_Bilirubin, Alamine_Aminotransferase, Alkaline_Phosphotase

contrast, supervised methods varied: Random Forest showed moderate instability with rank shifts among features, indicating narrow margins between adjacent ranks. Logistic regression was less stable in order and composition, dropping some features and reordering others after top-feature removal. Most concerning, XGBoost was highly sensitive to perturbation, introducing new features and substantially reshuffling importance rankings after removing its highest-ranked feature. Code for full reproducibility is available on GitHub.[5]

These results indicate supervised approaches can yield unstable feature importance rankings, consistent with sensitivity to label-driven errors and model specification. Unsupervised and non-targeted methods, lacking label dependence, produced more consistent ranking orders and may offer sturdier foundations for biomarker panel construction.

While Li et al.'s "FSTL-1" findings are promising for clinical application, biomarker development should integrate supervised and unsupervised strategies, emphasizing stability analyses and robustness to label noise, rather than relying solely on AUROC-based feature selection.[1–5]

CONFLICTS OF INTEREST

The author has no conflicts to report.

Yoshiyasu Takefuji

Faculty of Data Science, Musashino University, Tokyo, Japan

Correspondence
Yoshiyasu Takefuji, Faculty of Data Science, Musashino University, 3-3-3 Ariake Koto-ku, Tokyo 135-8181, Japan.
Email: takefuji@keio.jp

ORCID

Yoshiyasu Takefuji https://orcid.org/0000–0002–1826–742X

REFERENCES

1. Li W, Chi Y, Xiao X, Li J, Sun M, Sun S, et al. Plasma FSTL-1 as a noninvasive diagnostic biomarker for patients with advanced liver fibrosis. Hepatology. 2025;82:669–82.
2. Parr T, Hamrick J, Wilson JD. Nonparametric feature impact and importance. Inf Sci. 2024;653:119563.
3. Watson DS, Wright MN. Testing conditional independence in supervised learning algorithms. Mach Learn. 2021;110:2107–29.
4. Lipton ZC. The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. Queue. 2018;16:31–57.
5. GitHub. biomarker.py. https://github.com/y-takefuji/biomarker