

The American Journal of Bioethics

ISSN: 1526-5161 (Print) 1536-0075 (Online) Journal homepage: www.tandfonline.com/journals/uajb20

Separating AI Prediction Accuracy from Feature-Importance Reliability in Clinical Decision-Making

Yoshiyasu Takefuji

To cite this article: Yoshiyasu Takefuji (2026) Separating AI Prediction Accuracy from Feature-Importance Reliability in Clinical Decision-Making, The American Journal of Bioethics, 26:11, 119-120, DOI: 10.1080/15265161.2026.2741286

To link to this article: https://doi.org/10.1080/15265161.2026.2741286

Published online: 01 Nov 2026.

Submit your article to this journal

View related articles

View Crossmark data

Full Terms & Conditions of access and use can be found at
https://www.tandfonline.com/action/journalInformation?journalCode=uajb20

THE AMERICAN JOURNAL OF BIOETHICS
2026, VOL. 26, NO. 11, 119–120
https://doi.org/10.1080/15265161.2026.2741286

Taylor & Francis
by informa

Check for updates

OPEN PEER COMMENTARIES

Separating AI Prediction Accuracy from Feature-Importance Reliability in Clinical Decision-Making

Yoshiyasu Takefuji

Science Park Corporation

Authors conflate two fundamentally different objectives of AI in medicine: predicting clinical targets and estimating feature importance, or the underlying associations between variables. Target prediction concerns how accurately a model forecasts an outcome, such as treatment response, whereas feature importance concerns whether the variables identified by the model represent genuine, clinically meaningful determinants of that outcome. These objectives require different forms of validation.

Predictive accuracy can be rigorously assessed against ground-truth labels using measures such as discrimination, calibration, and external validation. In contrast, feature importance generally has no universally accepted ground truth against which it can be directly evaluated. Consequently, a machine learning model may predict outcomes accurately while identifying features that are unstable, confounded, or merely predictive proxies rather than true causal or clinically relevant factors. This distinction highlights both the limitations of feature assessment and the need for independent external validation.

In clinical decision-making, target prediction can help estimate the likely effectiveness or safety of therapies, whereas reliable feature importance can help identify which therapy, biomarker, or patient characteristic is most relevant among competing alternatives, potentially leading to more individualized and effective care. At present, no single algorithm can reliably recover true associations or guarantee stable, clinically meaningful feature importance from observational or high-dimensional data alone.

The extended Bradford Hill considerations (Hill 1965)—including feature consistency and dose–response relationships (Hill 1965; Loos et al. 2026; Loughman et al. 2026; Thapa et al. 2026; Zhao et al. 2026)—offer useful evidence for evaluating the plausibility and robustness of identified associations, but these are necessary rather than sufficient conditions and cannot, by themselves, establish causality. Many existing AI studies therefore exhibit serious methodological weaknesses: they emphasize predictive performance while treating feature importance as though it were automatically trustworthy, rely on data-driven associations without adequate control for confounding and bias, and lack independent validation of the reported features.

As a result, clinicians may currently outperform AI in certain decision-making contexts—not because clinical judgment is inherently superior across all settings, but because AI-derived feature assessments remain vulnerable to instability, spurious correlations, and poor generalizability. Even state-of-the-art AI systems have yet to solve the fundamental problem of reliably distinguishing true clinical associations from features that are merely useful for prediction. Underlying much of this literature is an implicit and largely untested assumption: that high target-prediction accuracy guarantees reliable feature importance, despite the complete absence of ground truth against which feature assessments can be verified.

To address this gap, this research advocates a leave-top-feature-out (LTFO) approach serving as an external feature validation framework for operationalizing the Bradford Hill criteria. LTFO proceeds as follows: (1) select the top n features from the full feature set (Set 1); (2) remove the single highest-ranked feature to create a reduced dataset; (3) re-fit the model and re-select the top n-1 features from this reduced dataset (Set 2); and (4) compare the resulting ordered feature rankings across Set 1 and Set 2. Removing the highest-ranked feature exerts the strongest perturbation on subsequent rankings, as it typically contributes most to prediction in supervised models, whereas removing the lowest-ranked feature exerts the weakest perturbation—together allowing systematic examination of feature consistency and dose–response relationships. LTFO has been validated successfully across multiple domains and disciplines using diverse datasets, demonstrating its methodological robustness (Takefuji 2025, 2026a, 2026b, 2026c, 2026d).

CONTACT Yoshiyasu Takefuji takefuji@keio.jp Science Park Corporation, Zama, Kanagawa Prefecture, Japan.
© 2026 Taylor & Francis Group, LLC

There is a significant difference between existing ablation studies and LTFO. Conventional ablation studies, including Jaccard-based comparisons and related methods, examine feature sets as unordered collections, thereby discarding ranking information entirely. LTFO, by contrast, treats feature sets as ordered sequences, preserving rank information that is essential for evaluating feature consistency and dose–response relationships—both central to the Bradford Hill considerations.

Because no algorithm can currently guarantee true associations or reliable feature importance, robust estimation requires triangulation across methodologically diverse approaches rather than reliance on any single model. We therefore recommend combining LTFO with complementary techniques spanning distinct modeling paradigms—nonparametric supervised models, nonparametric unsupervised models, and non-target-prediction statistical measures such as Spearman's correlation. Convergent feature rankings across these fundamentally different paradigms provide stronger evidence of genuine, stable associations, whereas divergence signals potential instability, confounding, or model-specific artifacts. This cross-paradigm consistency serves as an additional external safeguard against spurious or unreliable feature importance estimates.

AUTHOR CONTRIBUTIONS

CRediT: **Yoshiyasu Takefuji**: Conceptualization, Investigation, Writing – original draft, Writing – review & editing.

DISCLOSURE STATEMENT

No potential conflict of interest was reported by the author(s).

FUNDING

The author(s) reported there is no funding associated with the work featured in this article.

ORCID

Yoshiyasu Takefuji http://orcid.org/0000-0002-1826-742X

REFERENCES

Hill AB. 1965. The environment and disease: association or causation? Proc R Soc Med. 58(5):295–300. https://doi.org/10.1177/003591576505800503

Loos A, Khazneh E, Kübler J. 2026. Enhancing safety evaluations: a comprehensive framework for evidence-based safety assessment using the Bradford Hill criteria. Ther Innov Regul Sci. 60(2):349–359. https://doi.org/10.1007/s43441-025-00910-y

Loughman J, Bullimore MA, Flitcroft DI. 2026. Myopia and eye disease: from correlation to causation. Prog Retin Eye Res. 114:101497. https://doi.org/10.1016/j.preteyeres.2026.101497

Takefuji Y. 2025. Beyond XGBoost and SHAP: unveiling true feature importance. J Hazard Mater. 488:137382. https://doi.org/10.1016/j.jhazmat.2025.137382

Takefuji Y. 2026a. Evaluating linear parametric poisson regression vs nonparametric unsupervised learning in ulcerative colitis data. Gastroenterology. 171(2):384–386. https://doi.org/10.1053/j.gastro.2026.04.011

Takefuji Y. 2026b. Beyond prediction: assessing stability in feature selection methods for materials science applications. Comput Mater Sci. 267:114609. https://doi.org/10.1016/j.commatsci.2026.114609

Takefuji Y. 2026c. The absence of true association validation in supervised feature importance: evidence from battery capacity prediction for second-life applications. J Storage Mater. 180Article:124146. https://doi.org/10.1016/j.est.2026.124146

Takefuji Y. 2026d. Ensuring reliable feature importance in food chemistry AI. Food Chem. 509:148515. https://doi.org/10.1016/j.foodchem.2026.148515

Thapa SK, Li J, Mach J, Hilmer SN, Kirkpatrick CMJ. 2026. Applying the Bradford Hill criteria to assess the independent causal roles of aging and medication in frailty progression: a systematic review. Drugs Aging. 43(1):9–28. https://doi.org/10.1007/s40266-025-01273-7

Zhao J et al. 2026. Understanding senna risks: evidence from the FDA adverse event reporting system. J Ethnopharmacol. 361:121170. https://doi.org/10.1016/j.jep.2026.121170