Asian Journal of Psychiatry 119 (2026) 104972

Contents lists available at ScienceDirect

Asian Journal of Psychiatry

journal homepage: www.elsevier.com/locate/ajp

Beyond feature importance: Validating true clinical associations in AI-driven psychiatric research through consistency and dose-response relationships

Yoshiyasu Takefuji [1,*]

SciencePark Corporation, 3-24-9 Iriya-Nishi, Zama-shi, Kanagawa 252-0029, Japan

ARTICLE INFO

Keywords:
Feature importance stability
Leave-top1-out validation
Supervised learning bias
Clinical association assessment
Unsupervised feature ranking

ABSTRACT

Current AI algorithms lack robust methods to validate genuine clinical associations, which require consistency and dose-response relationships while existing studies failed due to the absence of validation. This paper introduces a leave-top1-out validation approach that systematically removes the highest-ranked feature to assess ranking stability across supervised and unsupervised models. We demonstrate theoretically why supervised feature importances are unreliable due to target-driven biases and absence of ground truth validation. Using a publicly available sleep disorder dataset (374 samples, 13 features), we empirically show XGBoost achieves high accuracy (0.9198) but exhibits poor ranking stability, while Spearman correlation maintains perfect stability with competitive accuracy (0.9090). Unsupervised models demonstrate perfect consistency despite lower accuracy (0.8797–0.8850). Our findings suggest ranking stability, rather than predictive performance alone, better identifies true clinical associations, providing a rigorous multifaced framework for causal inference in medical AI applications.

1. Introduction

The intersection of artificial intelligence and psychiatric research is rapidly evolving, as evidenced by *Asian Journal of Psychiatry*'s publication trends: 125 articles on interpretations of sleep disorders (13 in 2025 and 2 in 2026), 10 articles on feature importance (3 in 2025 and 4 in 2026) and 23 articles using feature selection (9 in 2025 and 1 in 2026), highlighting a growing interest in identifying true causal associations. However, a concerning pattern of AI misapplications persist in scientific literature due to researchers' insufficient understanding of supervised machine learning models' fundamental principles. Many investigators erroneously interpret feature importance or variable selection metrics from these models as indicators of clinical associations—a critical methodological error arising from overlooking the absence of ground truth in these calculations. This misapplication leads researchers to present model-derived feature rankings as established biological or clinical relationships without proper validation. For such associations to be considered valid, two essential elements must be rigorously examined: consistency across different methodological approaches and the presence of dose-response relationships between variables. Unfortunately, current studies frequently lack these crucial validation steps, undermining the reliability of their conclusions about clinical associations derived from machine learning models.

Computational psychiatry has emerged as a pivotal framework for advancing data-driven analysis, offering novel methodological tools to deconstruct psychiatric heterogeneity and inform more precise diagnostic and therapeutic approaches (Tandon and Tandon, 2019; Tandon, 2024).

Wang et al. (2026) developed an interpretable machine learning approach for depression screening in geriatric primary care settings. After comparing multiple algorithms, they selected the XGBoost model for its superior predictive performance in identifying depression among elderly patients. The researchers enhanced clinical interpretability by implementing SHAP (SHapley Additive exPlanations) analysis, providing healthcare practitioners with transparent insights into the model's decision-making process and key predictive factors.

Researchers should recognize that supervised models operate with two distinct accuracy dimensions: predictive accuracy and feature attribution accuracy. While predictive performance can be validated against established ground truth labels, feature importance lacks

* Correspondence to: Emeritus Professor (Keio Univ.), AI laboratory, SciencePark Corporation, 3-24-9 Iriya-Nishi, Zama-shi, Kanagawa 252-0029, Japan.
E-mail address: takefuji@keio.jp.
[1] Yoshiyasu Takefuji completed this research and wrote this article

https://doi.org/10.1016/j.ajp.2026.104972
Received 22 February 2026; Received in revised form 4 April 2026; Accepted 7 April 2026
Available online 8 April 2026

comparable validation standards—creating a fundamental gap in verification methodology. This absence of ground truth for feature assessment can lead to misleading interpretations. Notably, when explainers like SHAP are applied to models (explainer=SHAP(model)), they necessarily inherit the underlying model's unvalidated feature importance hierarchy. SHAP's explanations, despite their mathematical elegance, cannot transcend the limitations of the base model and may instead amplify existing biases in feature attribution. This cascading dependency creates potentially unreliable explanatory outcomes that persist without rigorous cross-validation approaches specifically designed for feature importance verification.

Current AI algorithms lack robust methods to assess genuine associations, which require two mandatory critical criteria: consistency across multiple contexts and demonstrable dose-response relationships (Ioannidis, 2003; Roberts et al., 2019; Lai et al., 2025; Prada et al., 2025; Stamatakis et al., 2025; Ye et al., 2024). These criteria are frequently overlooked in existing studies, leading to potentially spurious findings. This paper introduces a leave-top1-out validation approach to identify true associations based on these established epidemiological criteria. We demonstrate theoretically why feature importances derived from supervised models are inherently unreliable due to the absence of ground truth labels and target-driven biases that can amplify spurious correlations. As a methodological alternative, we propose validation strategies using unsupervised models and non-target-prediction methods, validated through the leave-top1-out approach on publicly available sleep disorder datasets, to more accurately identify genuine clinical associations. While existing studies have discussed consistency using non-ordered feature sets, this paper advances the field by presenting ordered sets analysis and quantifying the strength of impact on feature ranking stability when systematically removing the highest-ranked feature, thereby providing a more rigorous framework for causal inference in clinical AI applications.

Landau & Nissim investigated the extraction of time-interval temporal patterns from multi-electrode, multi-wave electroencephalogram (EEG) data to improve both sleep stage classification performance and model interpretability (Landau and Nissim, 2025). Their study evaluated several machine learning algorithms, including Logistic Regression (LR), Artificial Neural Networks (ANN), K-Nearest Neighbors (KNN), Naïve Bayes (NB), along with two variants each of Random Forest (RF) and Support Vector Machine (SVM). They performed extensive hyperparameter tuning across three dimensions: (1) vectorization approaches for feature encoding (e.g., raw occurrence counts versus normalized proportions), (2) the dimensionality of the feature space, and (3) feature selection methodologies for identifying the most discriminative patterns. All reported results correspond to the optimal hyperparameter settings identified for each dataset (Landau and Nissim, 2025).

However, Landau & Nissim acknowledged a critical limitation: high predictive accuracy does not ensure reliable feature identification. This limitation arises from three fundamental issues. First, supervised learning models exhibit two distinct forms of accuracy: predictive accuracy for target outcomes and accuracy in identifying feature importance. Only the former can be validated against ground truth labels, while the latter lacks a comparable validation framework. Second, feature importance scores quantify contributions to predictive performance rather than representing true causal or associative relationships with the outcome. Third, validating genuine associations requires examining two essential criteria reproducibility across different samples or conditions (consistency) and the presence of systematic relationships between feature magnitude and outcome strength (dose-response relationships) (Ioannidis, 2003; Roberts et al., 2019; Lai et al., 2025; Prada et al., 2025; Stamatakis et al., 2025; Ye et al., 2024).

While existing studies on sleep disorders failed to validate clinical associations without assessing consistency and dose-response relationships, this paper challenges assessing consistency and dose-response relationships using a publicly available dataset on sleep disorders and empirically demonstrates inconsistency of feature importance ranking orders derived from supervised models and advocates for complementary methods including unsupervised models (feature agglomeration: FA, highly variable gene selection: HVGS) and nonparametric statistical methods like Spearman's correlation.

2. Methods

This paper advocates for a leave-top1-out approach to address these validation challenges: select the top n features from the full dataset (set1) where n is 4 for this experimnt, remove the highest-ranked feature to create a reduced dataset, re-select the top n-1 features from this reduced dataset (set2), and compare feature importance ranking orders between the two sets. This method enables consistency assessment by evaluating whether the remaining features maintain stable rankings across perturbations—if features in set2 remain in similar relative positions compared to set1, this demonstrates robust and reproducible feature identification. Simultaneously, the approach facilitates dose-response evaluation by examining whether the removal of the top feature produces a magnitude of ranking change proportional to that feature's importance score—substantial shifts in subsequent rankings would indicate a genuine dose-dependent relationship, where the top feature's contribution systematically influences the model's reliance on other features. The strategic removal of the highest-ranked feature creates a controlled perturbation that is sufficiently impactful to test ranking stability (consistency) while revealing whether feature importance translates into systematic predictive influence (dose-response)

We empirically demonstrate how the leave-top1-out method reveals that supervised models suffer from substantial instability in feature ranking orders, while unsupervised models and non-target prediction methods exhibit markedly stronger stability in feature rankings. Our analysis employs a publicly available sleep disorder dataset comprising 374 samples across three categories (None, Sleep Apnea, Insomnia) with 13 clinical features (Cutting, 2025). Detailed clinical features are summarized in Table 1.

Feature selection algorithms include diverse methods including supervised models (Random Forest: RF, XGBoost, Lasso, Ridge), unsupervised models (feature agglomeration: FA, highly variable gene

Table 1
Clinical features of the dataset.

Variable	Type	Role in Analysis	Description
Person ID	Identifier	Used in analysis	Unique number assigned to each individual
Gender	Categorical	Demographic predictor	Biological sex of the participant (Male/Female)
Age	Numerical	Demographic predictor	Age of the person in years
Occupation	Categorical	Demographic predictor	The persons job/profession
Sleep Duration	Numerical	Direct sleep metric	Average number of hours slept per day
Quality of Sleep	Ordinal	Direct sleep metric	Subjective rating of sleep quality (scale 1–10)
Physical Activity Level	Numerical	Lifestyle predictor	Minutes of physical activity per day
Stress Level	Ordinal	Lifestyle predictor	Self-reported stress rating (scale 1–10)
BMI Category	Categorical	Health condition predictor	Body weight classification (Underweight/Normal/ Overweight/Obese)
Blood Pressure	Numerical	Health condition predictor	Recorded as Systolic/ Diastolic (e.g. 120/80)
Heart Rate	Numerical	Health condition predictor	Resting heart rate in beats per minute
Daily Steps	Numerical	Lifestyle predictor	Number of steps walked per day
Sleep Disorder	Categorical	Target/Output variable	Diagnosis status (None/ Insomnia/Sleep Apnea)

selection: HVGS) and nonparametric statistical methods like Spearman's correlation. Cross-validation is used for assessing the quality of selected features and comparing two sets of features for consistency and dose-response relationships.

3. Results

Table 2 summarizes the cross-validation accuracy and feature importance ranking orders across all evaluated methods under two cross-validation configurations (CV4 and CV3).

Supervised nonlinear nonparametric models consistently achieved the highest prediction accuracies. XGBoost attained the highest CV4 accuracy (0.9198), followed closely by Random Forest (0.8956), both substantially outperforming the supervised linear parametric counterparts. Ridge regression yielded moderate CV4 accuracy (0.8370), while Lasso regression produced the lowest accuracy among all methods (0.7942), with a notable performance drop from CV4 to CV3 (0.7299), indicating poor generalizability. These results confirm that nonlinear nonparametric models hold a clear advantage over linear parametric models in predicting the target variable.

However, all supervised models exhibited marked instability in feature ranking orders across cross-validation folds. Random Forest identified Person ID, BMI Category, Diastolic_BP, and Occupation as the top four features under CV4, yet ranked BMI Category, Systolic_BP, and Occupation as the top three under CV3, with Person ID and Diastolic_BP dropping out of the top three entirely. Similarly, XGBoost ranked BMI Category, Systolic_BP, Occupation, and Heart Rate as the top four features under CV4, but shifted to Systolic_BP, Occupation, and Diastolic_BP under CV3, with BMI Category — previously the top-ranked feature — no longer appearing among the top three. Lasso and Ridge regression further demonstrated inconsistent feature prioritization across folds, with substantial reshuffling of rankings between CV4 and CV3. This instability in feature ranking orders across supervised models raises concerns about the reliability and interpretability of feature importance derived from target-predictive supervised learning frameworks.

In contrast, unsupervised models and non-target prediction nonparametric statistical methods demonstrated considerably stronger stability in feature ranking orders. Factor Analysis (FA) and High Variance Gene Selection (HVGS), both unsupervised approaches, produced fully consistent top feature rankings across CV4 and CV3, identifying Daily Steps, Person ID, Physical Activity Level, and Age as the top four features in both configurations, with the top three features — Person ID, Physical Activity Level, and Age — remaining completely stable. This consistency reflects the structural, data-driven nature of unsupervised methods, which are not influenced by target variable perturbations across folds.

Notably, Spearman rank correlation, a non-target prediction nonparametric statistical method, achieved the second highest CV4 accuracy (0.909) among all methods while simultaneously exhibiting perfect stability in feature ranking orders. Spearman consistently identified Physical Activity Level, Daily Steps, Gender, and Diastolic_BP as the top four features, and Daily Steps, Gender, and Diastolic_BP as the top three features across both CV4 and CV3 configurations, with no reordering observed. This combination of high predictive accuracy and complete ranking stability positions Spearman correlation as a particularly robust and interpretable method for feature selection in this context, suggesting that its non-parametric, rank-based association framework may offer a reliable alternative to supervised feature importance methods when stable feature identification is a priority.

For purposes of reproducibility and transparency, Python code, sleep2.py is publicly available at GitHub (GitHub, 2026).

Table 2
cross-validation accuracy and feature importance ranking orders per algorithm.

Method	CV4	CV3	Top4 Features	Top3 Features
RF	0.8956	0.9225	Person ID, BMI Category, Diastolic_BP, Occupation	BMI Category, Systolic_BP, Occupation
XGBoost	0.9198	0.9171	BMI Category, Systolic_BP, Occupation, Heart Rate	Systolic_BP, Occupation, Diastolic_BP
FA	0.885	0.885	Daily Steps, Person ID, Physical Activity Level, Age	Person ID, Physical Activity Level, Age
HVGS	0.8877	0.8797	Daily Steps, Person ID, Physical Activity Level, Age	Person ID, Physical Activity Level, Age
Spearman	0.909	0.9037	Physical Activity Level, Daily Steps, Gender, Diastolic_BP	Daily Steps, Gender, Diastolic_BP
Lasso	0.7942	0.7299	Quality of Sleep, Gender, BMI Category, Sleep Duration	Sleep Duration, Gender, Diastolic_BP
Ridge	0.837	0.7915	Diastolic_BP, Quality of Sleep, Systolic_BP, BMI Category	Quality of Sleep, Gender, BMI Category

4. Discussion

To establish true associations in any dataset, two mandatory components must be satisfied regardless of the analytical method employed: consistency of the association across independent investigations and the presence of a dose-response relationship between the feature and the outcome. These two criteria constitute the cornerstone of causal inference in epidemiological and clinical research. Importantly, consistency can be formally expressed through the stability of ordered sets of features, whereby the same ranked hierarchy of predictors reproduces reliably across independent samples or analytical contexts, providing quantifiable evidence of a genuine underlying association rather than a spurious artifact of a particular dataset or methodology. Critically, existing studies in sleep medicine have largely failed to rigorously demonstrate both criteria simultaneously, leaving the true causal structure of sleep disorder predictors insufficiently validated in the broader literature.

Furthermore, it is essential to distinguish between two conceptually distinct quantities that are frequently conflated in machine learning-based studies: feature importance and true association. Feature importance derived from supervised models such as Random Forest and XGBoost quantifies a feature's contribution to predictive accuracy within a specific model and dataset, rather than reflecting its true biological or clinical association with the outcome. In the absence of ground truth for feature importance, supervised models are inherently susceptible to instability in feature ranking orders, as demonstrated in our results. A feature that ranks highly in one model may rank poorly in another, not because of its clinical irrelevance, but because of differing model assumptions, optimization objectives, and sensitivity to data perturbations.

Therefore, cross-referencing algorithmically identified features with established sleep medicine pathophysiology, while intuitively appealing, risks conflating predictive utility with causal validity. We argue that true clinical validation must go beyond domain knowledge cross-referencing and instead formally examine consistency across independent cohorts and dose-response relationships between candidate features and sleep disorder outcomes. We have revised the manuscript to explicitly articulate this distinction and to propose these two criteria as the appropriate framework for future clinical validation of the identified stable features.

Our empirical findings reveal a striking dichotomy in feature ranking stability between supervised and unsupervised approaches, illuminating a fundamental limitation in the widespread practice of using supervised model outputs for clinical association inference—a practice that pervades the medical AI literature despite lacking proper methodological foundations.

The substantial instability observed in supervised models (XGBoost and Random Forest) is an inevitable consequence of their fundamental operating principles. Supervised models optimize feature importance through iterative target-driven learning processes, where feature contributions are continuously recalibrated based on their collective ability to minimize prediction error on the labeled outcome. This optimization introduces three interconnected sources of instability that fundamentally compromise the reliability of feature importance rankings for association inference.

First, label-driven error propagation creates a feedback loop where small perturbations—such as removing the top-ranked feature in our leave-top1-out validation—trigger cascading recalibrations throughout the entire feature hierarchy. Our results demonstrate this starkly: XGBoost maintained only two features (BMI Category and Occupation) consistently in the top-4 rankings and a single feature (Occupation) in the top-3 across validation folds. This instability reveals that feature importance in supervised models is fundamentally a relational property contingent upon the presence of other features and the specific distribution of training labels, rather than an intrinsic property reflecting genuine clinical associations.

Second, target-driven bias amplification causes supervised models to preferentially exploit spurious correlations when they improve predictive accuracy on the specific sample at hand. Because supervised models lack mechanisms to distinguish between causal relationships and coincidental associations, they treat all predictive signals equivalently during optimization. The dramatic ranking shifts we observed when removing top features suggest that many high-importance features in supervised models may represent such exploited spuriosities rather than robust clinical associations. Crucially, the very mechanism that enables high predictive accuracy—aggressive optimization toward training labels—simultaneously undermines the validity of derived feature rankings for inference purposes.

Third, the absence of ground truth for feature importance means that supervised models' feature rankings cannot be validated against any objective standard. While we can validate predictive accuracy against held-out outcome labels, no comparable validation exists for the correctness of feature importance scores. The high predictive accuracy achieved by XGBoost (0.9225) may create false confidence in its feature rankings, when in reality these rankings exhibit severe instability that would be unacceptable for clinical inference. This disconnect between verifiable predictive performance and unverifiable feature importance represents a dangerous conflation that has led to widespread misinterpretation in the medical AI literature.

In stark contrast, unsupervised models (feature agglomeration, HVGS) and non-target prediction methods (Spearman correlation) demonstrated perfect stability in feature ranking orders across all validation folds and leave-top1-out perturbations. This remarkable consistency derives directly from their fundamental independence from outcome labels and target-driven optimization processes.

Unsupervised models achieve stability through intrinsic feature quantification. Feature agglomeration and HVGS evaluate features based solely on their inherent statistical properties—variance patterns, distributional characteristics, and internal clustering structure—without reference to any outcome variable. These intrinsic properties remain invariant across data perturbations because they are computed independently for each feature. When we remove the top-ranked feature, the statistical properties of remaining features remain unchanged, and consequently, their relative rankings remain perfectly stable. The consistent ordering (Daily Steps, Person ID, Physical Activity Level, Age) across all validation folds reflects genuine intrinsic feature characteristics rather than artifacts of sample-specific label distributions.

Non-target prediction methods achieve stability by quantifying direct bivariate relationships. Spearman correlation measures the monotonic relationship between each individual feature and the outcome independently, without the complex multivariate optimization that characterizes supervised learning. This independence from multivariate model fitting eliminates the cascading recalibration effects that plague supervised models. When the top-ranked feature is removed, the Spearman correlations of remaining features with the outcome remain mathematically identical, producing perfect ranking stability. The consistent identification of Physical Activity Level, Daily Steps, and Gender as top features across all folds suggests these represent genuine, reproducible associations with sleep disorders rather than model-specific artifacts.

The perfect stability observed in both unsupervised and non-target prediction approaches provides strong evidence that these features represent true intrinsic properties or genuine bivariate associations rather than artifacts of optimization processes. The absence of label-driven errors means these methods cannot artificially amplify spurious correlations through iterative refinement, and the absence of multivariate recalibration means perturbations cannot trigger cascading instability.

Our leave-top1-out validation approach operationalizes two fundamental epidemiological criteria—consistency and dose-response relationships. Traditional studies have evaluated consistency using unordered feature sets, asking simply whether the same features appear repeatedly. Our ordered set analysis provides a more stringent test: features must maintain stable relative rankings. Supervised models fail this test catastrophically, while unsupervised and non-target prediction methods pass perfectly, maintaining identical rankings across all perturbations. This perfect consistency strongly suggests these methods capture genuine, reproducible associations.

The AI-driven medical literature has conflated two fundamentally distinct objectives: predictive modeling (forecasting outcomes) and association inference (identifying causal or clinical relationships). For predictive applications, supervised models with high predictive accuracy remain appropriate and valuable. However, researchers must resist interpreting feature importance scores from these models as representing true clinical associations. For association inference, our results demonstrate that methods with perfect ranking stability—unsupervised models and non-target prediction approaches—provide more reliable foundations despite their moderately lower predictive accuracy (0.8796–0.9090). The consistency of Spearman correlation in identifying Physical Activity Level, Daily Steps, and Gender, combined with its competitive accuracy (0.9090), suggests this approach captures genuine associations that both predict outcomes and maintain reproducibility.

The proliferation of AI-driven feature importance studies in medical journals—with this journal alone publishing 13 articles on feature importance and 22 on feature selection in recent years—demands immediate methodological scrutiny. Our findings suggest that studies relying exclusively on supervised model feature importance without stability validation may have promulgated numerous spurious associations into the medical literature. We call for methodological reforms: studies reporting feature importance should be required to demonstrate ranking stability across validation folds, with perfect or near-perfect stability as a prerequisite for association claims. Clear distinctions must be drawn between predictive models and inference models, with feature importance from the former explicitly discouraged for clinical inference.

The perfect stability demonstrated by unsupervised and non-target prediction methods reveals a fundamental truth: genuine associations should manifest as intrinsic feature properties or robust bivariate relationships that remain invariant across perturbations, not as conditional artifacts of complex multivariate optimization susceptible to label-driven error propagation. Supervised machine learning models suffer from fundamental instability in feature ranking orders due to label-driven error propagation, target-driven bias amplification, and cascading recalibration effects. These instabilities render supervised model feature importance metrics unreliable for clinical association inference despite excellent predictive accuracy. The medical AI literature must distinguish between prediction-optimized models and inference-appropriate methods, prioritizing ranking stability alongside

predictive performance when the research objective is identifying true clinical relationships rather than maximizing forecast accuracy.

According to ScholarGPS,

Yoshiyasu Takefuji holds notable global rankings in several fields. He ranks 25th out of 1287,415 scholars in life sciences, 22nd out of 805,705 in COVID-19, and 1st out of 109,919 in environmental sciences.

AI use

Not applicable

Authors' contributions

Yoshiyasu Takefuji completed this research and wrote this article.

CRediT authorship contribution statement

Yoshiyasu Takefuji: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization.

Consent to participate

Not applicable

Consent for publication

Not applicable

Ethics approval

Not applicable

Code availability

Not applicable.

Funding

This research has no fund.

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.

Acknowledgement

None.

Data availability

Not applicable

References

Cutting, B., 2025. Health and sleep relation 2024 [Data set]. Mendeley Data V1. https://doi.org/10.17632/46j8wrc7p7.1.

GitHub. (2026). sleep2.py. ⟨https://github.com/y-takefuji/sleep⟩.

Ioannidis, J.P.A., 2003. Genetic associations: false or true? Trends Mol. Med. 9 (4), 135–138. https://doi.org/10.1016/S1471-4914(03)00030-3.

Lai, Q., Dannenfelser, R., Roussarie, J.P., Yao, V., 2025. Disentangling associations between complex traits and cell types with seismic. Nat. Commun. 16 (1), 8744. https://doi.org/10.1038/s41467-025-63753-z.

Landau, O., Nissim, N., 2025. Mining multi-electrode and multi-wave electroencephalogram based time-interval temporal patterns for improved classification capabilities and explainability. Artif. Intell. Med. 170, 103269. https://doi.org/10.1016/j.artmed.2025.103269.

Prada, D., Ritz, B., Bauer, A.Z., Baccarelli, A.A., 2025. Evaluation of the evidence on acetaminophen use and neurodevelopmental disorders using the Navigation Guide methodology. Environ. Health 24 (1), 56. https://doi.org/10.1186/s12940-025-01208-0.

Roberts, M.R., Ashrafzadeh, S., Asgari, M.M., 2019. Research techniques made simple: Interpreting measures of association in clinical research. J. Invest. Dermatol. 139 (3), 502–511.e1. https://doi.org/10.1016/j.jid.2018.12.023.

Stamatakis, E., Ahmadi, M., Biswas, R.K., Del Pozo Cruz, B., Thøgersen-Ntoumani, C., Murphy, M.H., Sabag, A., Lear, S., Chow, C., Gill, J.M.R., Hamer, M., 2025. Device-measured vigorous intermittent lifestyle physical activity (VILPA) and major adverse cardiovascular events: evidence of sex differences. Br. J. Sports Med. 59 (5), 316–324. https://doi.org/10.1136/bjsports-2024-108484.

Tandon, N., Tandon, R., 2019. Using machine learning to explain the heterogeneity of schizophrenia: Realizing the promise and avoiding the hype. Schizophr. Res. 214, 70–75. https://doi.org/10.1016/j.schres.2019.08.032.

Tandon, R., 2024. Computational psychiatry and the Asian Journal of Psychiatry. Asian J. Psychiatry 95, 104055. https://doi.org/10.1016/j.ajp.2024.104055.

Wang, M., Luo, M., Li, L., et al., 2026. Primary-care-focused interpretable machine learning model for depression screening in geriatrics: a comparative study of multiple algorithms. J. Affect. Disord. 394 (Part A), 120551. https://doi.org/10.1016/j.jad.2025.120551.

Ye, M., He, Y., Xia, Y., et al., 2024. Association between bowel movement frequency, stool consistency and MAFLD and advanced fibrosis in US adults: a cross-sectional study of NHANES 2005-2010. BMC Gastroenterol. 24 (1), 460. https://doi.org/10.1186/s12876-024-03547-7.