npj Digital Medicine

https://doi.org/10.1038/s41746-026-02491-y

Article in Press

Algorithmic opacity in opioid risk scoring and the need for transparent AI regulation

Received: 1 December 2025

Accepted: 15 February 2026

Cite this article as: Wang, S.Y., Stofer, R., Chu, Z. *et al.* Algorithmic opacity in opioid risk scoring and the need for transparent AI regulation. *npj Digit. Med.* (2026). https://doi.org/10.1038/s41746-026-02491-y

Sherry Yun Wang, Ryan Stofer, Zhouzhou Chu, Xiao Huang & Ang Li

We are providing an unedited version of this manuscript to give early access to its findings. Before final publication, the manuscript will undergo further editing. Please note there may be errors present which affect the content, and all legal disclaimers apply.

If this paper is publishing under a Transparent Peer Review model then Peer Review reports will publish with the final article.

Title:

Algorithmic Opacity in Opioid Risk Scoring and the Need for Transparent AI Regulation

Sherry Yun Wang*[1], Ryan Stofer[1], Zhouzhou Chu[1], Xiao Huang[2], Ang Li[3]
[1]School of Pharmacy, Chapman University, Irvine, California, United States
[2] Department of Environmental Sciences, Emory University, Atlanta, Georgia, United States
[3]Department of Electrical Engineering, University of Maryland, College Park, Maryland, United States
***Correspondence:**
Sherry Yun Wang, School of Pharmacy, Chapman University, Irvine, California, USA
RK 94-206, 9401 Jeronimo Road, Irvine, California 92618
Email: yunwang@chapman.edu

Abstract

NarxCare®, a proprietary opioid risk scoring system embedded in Prescription Drug Monitoring Programs (PDMPs), has generated significant patient complaints. We adhered to the technical specifications and applied them to PDMP and IQVIA PharMetrics® Plus Closed Health Plan claims database. Despite adding socioeconomic covariates, precision (0.01–0.32) was far below the reported benchmark of 0.75, and F1 scores (0.02–0.39) were also substantially lower than the benchmark value of 0.65, across all our reconstructed models.

Main Text

Machine learning (ML) tools are increasingly embedded in U.S. clinical workflows, yet their opacity raises persistent concerns about fairness, transparency, and regulatory oversight. One such class of tools, opioid risk scoring (ORS) systems, has become central to opioid stewardship strategies across the United States[1]. Among them, NarxCare®, a proprietary algorithm developed by Bamboo Health, is the most widely adopted and currently integrated into statewide Prescription Drug Monitoring Programs (PDMPs) in over 20 states[2]. Despite this broad integration and its growing influence over treatment decisions, NarxCare® remains a black box.

In April 2024, Bamboo Health released technical documentation[3] revealing that NarxCare®'s ORS calculates risk scores using a basic logistic regression model, with predictive features such as the number of prescribers, the number of dispensing pharmacies, and thresholds of daily Morphine Milligram Equivalents (MMEs) over varying time windows. This raises the broader question of whether commonly used prescription- and claims-based data sources[3], paired with publicly disclosed documentation for modeling pipelines, are sufficient to support reliable prediction of opioid-related harms across diverse real-world populations.

Reconstructing NarxCare's Published Model

To explore this, we attempted to approximate NarxCare®'s ORS using two independent datasets: the IQVIA PharMetrics® Plus (2006–2022) and California's PDMP (2010-2023). We engineered features consistent with NarxCare®'s documentation[3], including cumulative opioid dosage, prescriber count, and pharmacy switching behavior, and trained supervised ML models (logistic regression, random forest, XGBoost, and neural networks). Models were trained to predict either receipt of medication for opioid use disorder (MOUD), as a proxy for opioid use disorder (OUD) in PDMP data [4-6], or broader opioid-related adverse events defined by International Classification of Diseases (ICD) codes[7] in the IQVIA dataset (Supplementary Table 1-3). We used stratified 5-fold cross-validation, applied a range of class balancing techniques[8]

[Synthetic Minority Oversampling Technique (SMOTE), random undersampling (RUS), and edited nearest neighbors (ENN)]. We also incorporated expanded feature sets, including individual-level characteristics (e.g., age, gender, and payment type) and neighborhood-level social determinants of health (SDoH), e.g., median age, income, education, disability, and racial/ethnic composition. This study was deemed exempt from institutional review board approval because it involved unidentifiable data.

Benchmark Performance Discrepancies

Since NarxCare® trained its model on PDMP data from a Midwestern state[3], we anticipated comparable, or even improved performance when applying the same approach to our PDMP dataset. However, despite reproducing the published feature set, adding expanded SDoH covariates, and testing multiple modeling strategies (logistic regression, XGBoost, neural networks) with class-balancing techniques (SMOTE, RUS, ENN), all models showed only modest performance across both datasets. For each metric, we report the best-performing value within each model family across all hyperparameter configurations (Figure 1). The NarxCare® baseline model is shown for reference, with its missing F1 value manually derived from its reported precision and recall. Full results are available in Supplementary Table 3.

Our reconstructed models consistently yielded precision values substantially lower than those reported by NarxCare®, and we acknowledge several important limitations that constrain direct comparability. PDMP-based identification of opioid use disorder relies on initiation of medication for opioid use disorder (MOUD), which does not capture transitions into disease and primarily reflects observed treatment access and prescribing behavior, while claims-based datasets such as IQVIA do not include Medicaid beneficiaries. Algorithms trained on PDMP data may systematically underrepresent patients engaged in opioid treatment programs (OTP)-based care, which has implications for both performance estimates and equity considerations. Taken together, these structural limitations indicate that opioid risk algorithms intended for use across broad patient populations may underperform when trained on incomplete or nonrepresentative data. This discrepancy further suggests that NarxCare®'s reported performance may depend on additional engineered features, undocumented preprocessing steps, or proprietary training optimizations that are not publicly disclosed. Our analysis does not assert causal inference or claim incorrectness of NarxCare®'s outputs or training methods; rather, it highlights the broader risks associated with opaque clinical algorithms that influence high-stakes decisions without independent validation. NarxCare® thus exemplifies a structural transparency challenge: because key elements of the model are proprietary, independent researchers, clinicians, and state agencies cannot determine the sources of observed performance discrepancies or systematically assess generalizability, fairness, or safety.

Consequences for Clinical Algorithm Oversight

The U.S. Food and Drug Administration (FDA) has historically struggled to delineate its authority over Clinical Decision Support (CDS) tools that do not make explicit treatment recommendations. Tools like NarxCare® fall outside the current definition of Software as a Medical Device (SaMD), allowing them to bypass premarket review. Yet their influence is undeniable: reports of patients being denied pain treatment due to ORS flags are mounting, particularly in marginalized communities[2,9-11]. When the FDA cleared the Apple Watch in 2018 for detecting irregular heart rhythms, many worried the agency would get stuck trying to reverse-engineer the device[12]. Instead, Apple provided extensive firm-based validation data: large-scale clinical studies, clear descriptions of the underlying datasets, transparent reporting of model performance across demographic groups, and evidence that the device performed reliably in real-world conditions. This allowed the FDA to evaluate not only the product but also the company's development process, documentation practices, and data integrity, consistent with how the agency assesses Software as a Medical Device (SaMD[12]. A similar framework would benefit high-stakes Clinical Decision Support (CDS) systems like NarxCare®, which currently fall outside device regulation under section 520(o)(1)(E) of the FD&C Act[13] because they provide a risk score, i.e., a specific diagnostic or treatment-relevant output that clinicians may rely on, without enabling users to "independently review the basis for the recommendation," a key requirement for non-device CDS. In fact, the FDA explicitly states that software functions that provide a risk probability or risk score for a disease or condition do not meet Criterion 3 and are therefore device functions requiring oversight. This mismatch highlights the regulatory gap: NarxCare® has a device-like influence on clinical care, yet it is not necessary to demonstrate the type of transparent, evidence-based validation or a replicable evidence base that would enable independent assessment of safety, fairness, or generalizability.

While the 21st Century Cures Act of 2016 and the FDA's 2022 guidance attempted to clarify regulatory boundaries, ambiguity remains, and some developers have reportedly designed their tools to avoid triggering regulatory thresholds[14,15]. Meanwhile, federal policy is shifting. A January 2025 executive order directed agencies to remove barriers to AI innovation, and the White House's "Winning the AI Race: America's AI Action Plan[16]" now places healthcare at the center of national AI transformation efforts. Its provisions[16] include FDA regulatory sandboxes, healthcare AI testbeds, and Centers of Excellence designed to vet AI tools in real-world clinical environments. This moment presents a critical opportunity. A modernized, risk-based regulatory framework is needed to calibrate oversight to clinical impact and align with the SaMD model. Regulatory sandboxes should enable iterative development while ensuring transparency, reproducibility, and safety. Without such mechanisms, tools like NarxCare® may continue to influence clinical decision-making without sufficient transparency unless oversight mechanisms are implemented., raising concerns about potential algorithmic harm in the absence of transparent oversight, even when implemented as part of innovation efforts. Healthcare AI should not be exempt from scrutiny. As national infrastructure evolves to support innovation, we must also build the regulatory and ethical frameworks necessary to protect patients and uphold the integrity of clinical practice.

Methods

Dataset

This study leveraged two independent large-scale healthcare datasets to replicate and evaluate the NarxCare® opioid risk scoring algorithm: California's PDMP and the IQVIA PharMetrics® Plus Closed Health Plan claims database. The California PDMP, accessed via the Controlled Substance Utilization Review and Evaluation System (CURES), comprises de-identified patient-level dispensing records for all Schedule II–V controlled substances, including opioids, from 2015 to 2023. It captures prescription-level details such as drug name, strength, quantity dispensed, days' supply, prescriber and dispensing pharmacy identifiers, fill date, and patient demographics (age, gender, payment type, ZIP5). The IQVIA PharMetrics® Plus Closed Health Plan claims database contains longitudinal adjudicated medical and pharmacy claims from a national sample of commercially insured and Medicare Advantage enrollees, with information on National Drug Codes (NDCs), fill dates, quantities dispensed, prescriber identifiers, ICD-9/10 diagnostic codes, procedure codes, and insurance type. Unlike PDMP, IQVIA does not include pharmacy identifiers. For contextual enrichment, IQVIA ZIP3 codes were mapped to ZIP5 using the SimpleMaps crosswalk[17], and ZIP5-level socioeconomic indicators (i.e., median age, socioeconomic status, housing, education, employment, disability, and racial/ethnic composition) were appended from the SimpleMaps crosswalk[17]. This enabled integration of neighborhood-level SDoH as a proxy for individual SDoH[18] into predictive modeling (Supplementary Table 1 and Figure 1). Following data cleaning and preprocessing, the CURES dataset yielded 17.9 million training observations, 8.9 million validation observations, and 6.7 million testing observations. The IQVIA dataset yielded 1.03 million training, 256,122 validation, and 322,652 testing observations (Supplementary Table 2).

Feature Construction

Feature construction was aligned with Bamboo Health's published NarxCare Application Overview (2023)[3], implementing temporal-aggregate measures that mirror the original model's inputs. Core features[3] included cumulative MME over the past 365 days and 2 years; total MME dosage in the past 2 years and ≥1 year before the index date; the number of prescriptions with daily MME >120; and counts of unique prescribers over 2 years and the past 180 days. In the PDMP dataset, an additional behavioral feature, the number of distinct pharmacies dispensing opioids in the prior 2 years, was incorporated. All features were engineered relative to an index date defined as the most recent opioid prescription before the observation window, with sliding-window aggregation implemented using optimized, vectorized routines in Python. ZIP5-linked SDoH variables were appended to capture contextual socioeconomic influences. Continuous features underwent z-score normalization, while categorical variables were one-hot encoded before model training. NarxCare's baseline model, as disclosed in its technical documentation[3], was trained via logistic regression on a case-control dataset comprising over 5,000 autopsy-adjudicated unintentional overdose deaths matched by age and gender to 500,000 patients prescribed controlled substances who did not die from overdose in a Midwestern state's PDMP data. Models were trained to predict opioid-related outcomes available within each data source. In the PDMP dataset, receipt of medication for OUD (MOUD; e.g., buprenorphine, methadone) MOUD was used as a pragmatic proxy for OUD, consistent with prior literature[19-21], recognizing that PDMP data lack ICD-coded diagnostic information and therefore cannot directly capture clinically validated OUD or overdose outcomes. Because methadone administered through federally regulated OTP may not be comprehensively reported to PDMP systems due to 42 CFR Part 2 confidentiality protections, PDMP-observable MOUD primarily reflects pharmacy-dispensed treatments and likely under-ascertains OTP-based methadone treatment. In the IQVIA dataset, opioid-related adverse event labels were assigned based on the presence of ICD-9/10 diagnosis codes[7] for opioid-related adverse events recorded at any time following the index prescription. Our study cannot determine why NarxCare's reported performance exceeds what can be reproduced using the publicly stated feature set. Across both datasets, our results show that, even after incorporating additional covariates (e.g., SDoH), conducting extensive hyperparameter optimization, and applying multiple imbalance-mitigation strategies, no reasonable reconstruction of the published feature space achieves performance close to NarxCare's benchmarks. These discrepancies highlight the possibility that (a) additional, undisclosed features or preprocessing steps could have influenced NarxCare's reported performance, (b) preprocessing steps or transformations that are not documented, or (c) potential data leakage arising from the internal construction of case–control sets or temporal windows. To ensure a rigorous evaluation, we first trained models using PDMP data, where MOUD initiation served as a proxy for opioid use disorder due to the absence of ICD-based diagnostic information. We then trained models on the IQVIA dataset, which includes ICD codes for opioid-related adverse events, allowing us to evaluate the same feature family under a more clinically specific outcome definition. In both datasets, we implemented multiple strategies to address class imbalance and maximize model performance within the limits of the available information. Despite these efforts, the performance of all reconstructed models remained far below NarxCare's self-reported metrics. California's PDMP is not currently integrated with electronic health records (EHR), which limits its capacity to ascertain clinically validated opioid-related adverse events accurately. Given these limitations in outcome ascertainment, we retrained the model using the IQVIA dataset, which contains structured ICD-coded diagnostic data, enabling more robust and clinically grounded labeling of opioid-related adverse events. However, both datasets demonstrated significant class imbalance in the outcome. MOUD initiation can contribute to an underestimation of the actual probability of opioid-related adverse events. In the PDMP dataset, initiation occurred in approximately 1% of patients, indicating a highly imbalanced class distribution. The IQVIA dataset exhibited a more moderate imbalance, with a positive-to-negative case ratio of approximately 1:8. To address this, we employed a range of strategies, including class rebalancing techniques (e.g., oversampling of the minority class, under-sampling of the majority class), as well as deep learning models optimized for imbalanced data. Despite these efforts, model precision did not approach the reported NarxCare® benchmarks, though contextual differences in datasets and outcome definitions limit direct comparability.

Model Training

Following the NarxCare baseline design, a logistic regression model with L2 regularization was implemented as the primary replication model (Supplementary Table 3), providing methodological comparability to the original algorithm. To investigate whether alternative architectures could better capture non-linear interactions, we also trained Random Forests, Extreme Gradient Boosting (XGBoost), feedforward neural networks, wide and deep hybrid architectures, and self–attention–augmented networks. Training and testing sets were generated by randomly splitting the entire dataset to evaluate model performance. Hyperparameter optimization was conducted using grid search with Optuna as the primary tuning strategy, integrated with nested cross-validation within the training partition to ensure reliable and generalizable parameter estimates. We evaluated both hard voting (majority class selection) and soft voting (probability averaging) ensemble strategies, ultimately adopting soft voting due to its superior performance and finer discrimination from aggregating predicted probabilities rather than binary class outputs.

Evaluation Metrics

Model performance was assessed using precision, recall, specificity, negative predictive value (NPV), and F1 score (Supplementary Table 3). All metrics were selected to enable direct comparison with the baseline model reported by NarxCare®. Both datasets exhibited outcome class imbalance, particularly PDMP, where MOUD initiation occurred in approximately 1% of patients and IQVIA, with a more moderate imbalance for opioid-related adverse events (~1:8). To address this, we implemented multiple imbalance-mitigation strategies, including Synthetic Minority Oversampling Technique SMOTE, RUS, and ENN.

Data Availability: The CURES dataset is available upon request from the Department of Justice. The census data can be obtained from the US Zip Codes Database (Pareto SoftwareTM, version 2023). Concerning access to and use of the IQVIA PharMetrics® Plus for Academics dataset, which is licensed to Chapman University under the terms of its agreement with IQVIA Inc.

Code Availability: The code is publicly accessible at https://github.com/Sherry-Yun-Wang/Algorithmic-Opacity-in-Opioid-Risk-Scoring-Need-for-Transparent-AI-Regulation-in-Healthcare.

Acknowledgment: We extend our heartfelt gratitude to the California Department of Justice for their invaluable support in providing the data and their unwavering assistance throughout our research journey. We acknowledge that CURES is not associated with the NarxCare platform, and any opinions, findings, conclusions, or recommendations expressed in this publication are those of the authors and do not necessarily reflect the views of the California Department of Justice CURES Program or IQVIA Inc. This project was supported by funding from the National Institute of Health (NIH) AIM-AHEAD program.

Author Contributions Statement S.Y.W. conceptualized the research idea, designed the study, authored the primary manuscript, and secured funding as the Principal Investigator (PI). R. S. conducted the data analysis. A.L., C. Z., and X. H. contributed to the major revision. All authors edited and reviewed the manuscript.

Competing Interests The authors declare no competing interests.

References

1 Ardeljan, L. D. *et al.* Current state of opioid stewardship. *American journal of health-system pharmacy* **77**, 636-643 (2020).

2 Bhagwat, A. M., Ferryman, K. S. & Gibbons, J. B. Mitigating algorithmic bias in opioid risk-score modeling to ensure equitable access to pain relief. *Nature medicine* **29**, 769-770 (2023).

3 Bamboo Health. *NarxCare Application Overview*, <https://dopl.idaho.gov/wp-content/uploads/2024/03/BOP-PDMP-Overview-NarxCare.pdf> (2023).

4 Larochelle, M. R. *et al.* Medication for opioid use disorder after nonfatal opioid overdose and association with mortality: a cohort study. *Annals of internal medicine* **169**, 137-145 (2018).

5 Wakeman, S. E. *et al.* Comparative effectiveness of different treatment pathways for opioid use disorder. *JAMA network open* **3**, e1920622-e1920622 (2020).

6 Biondi, B. E., Zheng, X., Frank, C. A., Petrakis, I. & Springer, S. A. A literature review examining primary outcomes of medication treatment studies for opioid use disorder: what outcome should be used to measure opioid treatment success? *The American journal on addictions* **29**, 249-267 (2020).

7 Acharya, M. *et al.* Comparative study of opioid initiation with tramadol, short-acting hydrocodone, or short-acting oxycodone on opioid-related adverse outcomes among chronic noncancer pain patients. *The Clinical journal of pain* **39**, 107-118 (2023).

8 Xu, Z., Shen, D., Nie, T. & Kou, Y. A hybrid sampling algorithm combining M-SMOTE and ENN based on Random forest for medical imbalanced data. *Journal of Biomedical Informatics* **107**, 103465 (2020).

9 Buonora, M. J., Axson, S. A., Cohen, S. M. & Becker, W. C. Paths forward for clinicians amidst the rise of unregulated clinical decision support software: our perspective on NarxCare. *Journal of general internal medicine* **39**, 858-862 (2024).

10 Siegel, Z. In a World of Stigma and Bias, Can a Computer Algorithm Really Predict Overdose Risk?: A Machine-Learning Algorithm Is Being Deployed Across America to Prevent Overdose Deaths. But Could It Be Causing More Pain? *Annals of Emergency Medicine* **79**, A16-A19 (2022).

11 Szalavitz, M. The pain was unbearable. So why did doctors turn her away. *Wired. August* **11** (2021).

12 Gottlieb, S. in *JAMA Health Forum.* e242691-e242691 (American Medical Association).

13 Clinical Decision Support Software: Guidance for Industry and Food and Drug Administration Staff. (U.S. Food and Drug Administration, 2022).

14 Harvey, H. B. & Gowda, V. How the FDA regulates AI. *Academic radiology* **27**, 58-61 (2020).

15 Boubker, J. When medical devices have a mind of their own: the challenges of regulating artificial intelligence. *American Journal of Law & Medicine* **47**, 427-454 (2021).

16 House, T. W. *Winning the AI Race: America's AI Action Plan*, <https://www.whitehouse.gov/wp-content/uploads/2025/07/Americas-AI-Action-Plan.pdf> (2025).

17 SimpleMaps. *United States ZIP Code Database (Version 2025)*, <https://simplemaps.com/data/us-zip-codes> (2025).

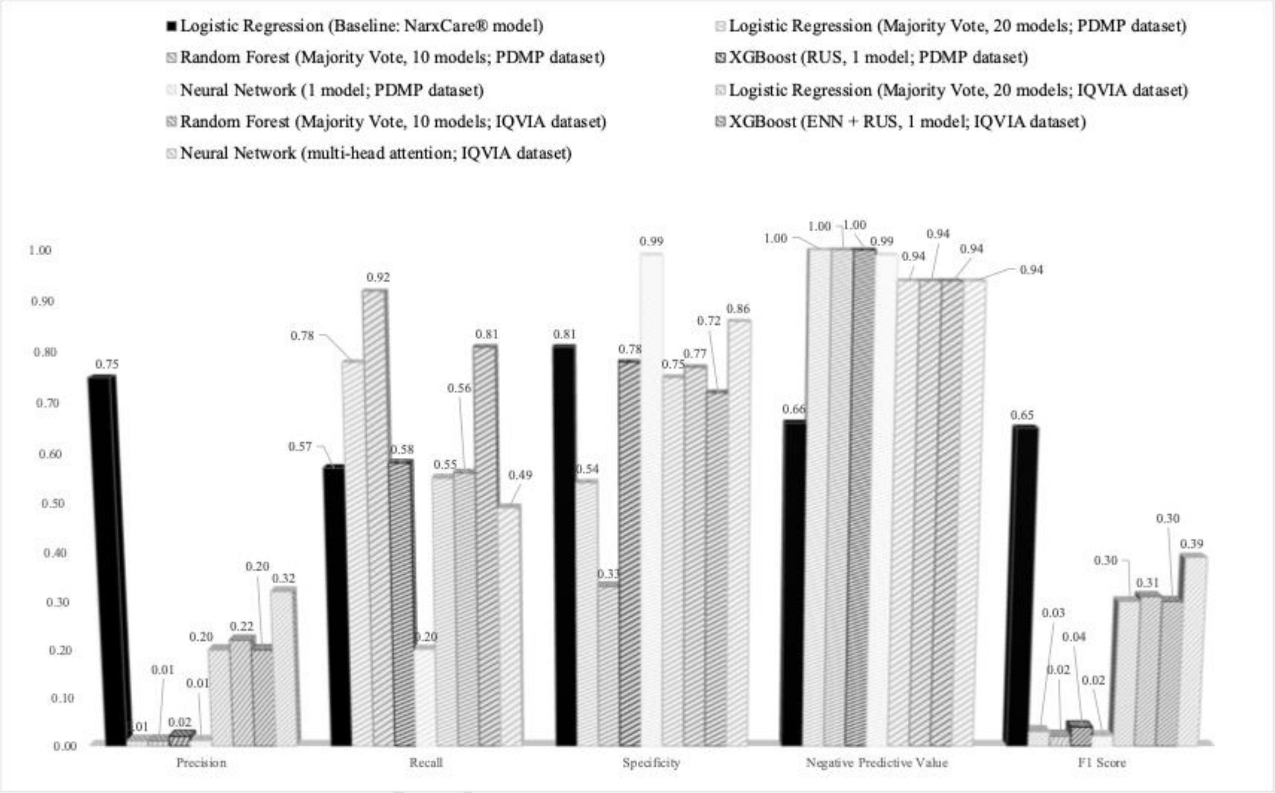
18 Li, C. *et al.* Realizing the potential of social determinants data in EHR systems: A scoping review of approaches for screening, linkage, extraction, analysis, and interventions. *Journal of Clinical and Translational Science* **8**, e147 (2024).

19 Hurley, R. W. *et al.* Evidence-based framework for identifying opioid use disorder in administrative data: A systematic review and methodological development study. *Pain Medicine*, pnaf116 (2025).

20 Elkington, K. S. *et al.* Examining the Impact of the Innovative Opioid Court Model on Treatment Access and Court Outcomes for Court Participants. *Journal of addiction medicine* **18**, 635-642 (2024).

21 Roy, P. J. *et al.* Impact of study design decisions on identification of treatment initiators of medications for opioid use disorder. *Addiction* (2025).

Figure 1. Comparative performance of reconstructed models across PDMP and claims datasets.



This figure compares precision, recall, specificity, NPV, and F1 scores across logistic regression, XGBoost, neural networks, and ensemble models trained on PDMP and IQVIA datasets using the NarxCare®-aligned feature set, with class-balancing methods including SMOTE, RUS, and ENN. Each panel presents the best-performing model from each model family, selected across all hyperparameter configurations; detailed results for all models are reported in Supplementary Table 3, enabling direct comparison with the NarxCare® benchmark metrics.