

Same-wave cross-sectional risk stratification for parent-reported current anxiety among US children with ADHD who received behavioral treatment: development in the 2023 NSCH and evaluation in the 2024 NSCH

Chenglei Qin

The First Affiliated Hospital of Huzhou Normal University, Huzhou First People's Hospital

Yi Shi

The First Affiliated Hospital of Huzhou Normal University, Huzhou First People's Hospital

Qian Chen

The First Affiliated Hospital of Huzhou Normal University, Huzhou First People's Hospital

Junyu Xu

The First Affiliated Hospital of Huzhou Normal University, Huzhou First People's Hospital

Guo Chen

The First Affiliated Hospital of Huzhou Normal University, Huzhou First People's Hospital

Mengyu Wu

The First Affiliated Hospital of Huzhou Normal University, Huzhou First People's Hospital

Yiwei Fu

The First Affiliated Hospital of Huzhou Normal University, Huzhou First People's Hospital

Chen Yuan

1243660363@qq.com

The First Affiliated Hospital of Huzhou Normal University, Huzhou First People's Hospital

Research Article

Keywords: attention-deficit/hyperactivity disorder, anxiety, children, prediction model, complex survey, cross-sectional study

Posted Date: October 14th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10849068/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.
[Read Full License](#)

Additional Declarations: No competing interests reported.

Abstract

Background

Anxiety is common in pediatric attention-deficit/hyperactivity disorder (ADHD), including among children receiving behavioral treatment. We developed models for research-oriented, same-wave current-status classification of parent-reported current anxiety in US children aged 6–17 years with current ADHD and recent behavioral treatment, then evaluated locked models in an independent 2024 cross-sectional sample.

Methods

We conducted a cross-sectional secondary-data model-development and evaluation study using the 2023–2024 National Survey of Children’s Health (NSCH); annual samples comprised different respondents. Twenty prespecified predictors were used to develop logistic LASSO, bidirectional stepwise logistic regression, random-forest, and XGBoost models in 2023; locked predictions were evaluated in 2024. Models were fitted without sampling weights, whereas performance estimates incorporated the pooled survey weight and complex design. Outcomes included AUC, Brier score, calibration, and net benefit at exploratory thresholds. Survey-design 95% CIs were percentile intervals from 500 state-by-stratum HHID-cluster replicates with fixed predictions; model-development uncertainty was excluded.

Results

Among 106537 source records, 5088 children met the target-population criteria; 4973 had observed anxiety status, including 2922 with current anxiety (survey-weighted prevalence, 51.8%). The complete-case analysis included 4619 children (2075 in 2023; 2544 in 2024) and 2714 events. In 2024, weighted AUCs were 0.717 (95% CI, 0.678–0.758) for LASSO, 0.723 (0.678–0.766) for stepwise regression, 0.711 (0.665–0.751) for random forest, and 0.674 (0.627–0.724) for XGBoost. Corresponding Brier scores were 0.214, 0.212, 0.215, and 0.250; calibration slopes were 0.97, 0.85, 1.13, and 0.35. At a 0.50 threshold, net benefit was 0.217 for LASSO and 0.227 for stepwise regression versus 0.082 for treat-all. Outcome-informed multiple-imputation and alternative-coding sensitivities were broadly similar; the former was not deployment compatible.

Conclusions

Parsimonious regression models provided moderate discrimination and better calibration than XGBoost when evaluated in an independent 2024 cross-sectional sample. They classify parent-reported current anxiety within the same survey wave; they do not predict future onset and are not diagnostic

instruments. Independent clinical validation, fairness assessment, and prospective impact evaluation are required before practice use.

Background

Attention-deficit/hyperactivity disorder (ADHD) ranks among the most prevalent neurobehavioral conditions in childhood, often coexisting with emotional, behavioral, developmental, and learning disorders. According to the 2022 National Survey of Children's Health (NSCH), about 10.5% of U.S. children aged 3–17 years were reported by parents to have current ADHD, with most of these children also experiencing at least one additional disorder [1, 2]. Recent systematic reviews similarly highlight frequent psychiatric comorbidities in children and adolescents with ADHD, including frequent co-occurrence of anxiety disorders [3, 4].

Recognizing anxiety is crucial as comorbidity can impact clinical presentation, functioning, treatment planning, and the necessity for specialized care. The American Academy of Pediatrics advises screening children and adolescents being evaluated for ADHD for emotional and behavioral comorbidities, such as anxiety and depression [1]. Behavioral treatment for ADHD does not address the variability in mental health burdens. Even among children undergoing this treatment, clinicians and health systems must identify those who particularly require a focused anxiety assessment.

Population surveys are instrumental in characterizing heterogeneity, yet prediction research utilizing complex surveys demands specific methodological considerations. While model development often emphasizes individual-level predictive structures, performance estimates intended to describe the eligible survey population must incorporate sampling weights, strata, and clustered primary sampling units. Furthermore, relying solely on discrimination is inadequate; calibration and decision-curve measures are also informative, provided their assumptions and uncertainty are stated.

We developed four models to classify parent-reported current anxiety in the same survey wave among U.S. children aged 6–17 years with current ADHD who had received behavioral treatment in the past 12 months. Using 2023 data, we prespecified predictors, constructed the models, and evaluated locked predictions in the independent 2024 annual sample. Evaluation included survey-weighted discrimination, overall accuracy, calibration, classification performance, and exploratory decision-curve analysis. We hypothesized that regression-based models would perform at least as well as more flexible machine-learning approaches while offering more stable calibration.

Methods

Study design, intended use, and time origin

We conducted a cross-sectional secondary-data model-development and evaluation study using two consecutive annual NSCH waves comprising different respondents. Models were developed in 2023 and evaluated without updating in the independent 2024 annual sample from the same survey program. The

index time was completion of the parent/caregiver questionnaire; all predictors and the outcome were measured in the same questionnaire, so no future prediction horizon existed. The target estimand was the probability of parent-reported current anxiety in the same survey wave among eligible respondents. The models were not intended to predict future anxiety onset, provide a diagnosis, estimate treatment response, or support causal inference. The study does not establish that every predictor would be available before anxiety recognition in clinical practice. Reporting followed TRIPOD + AI, with STROBE as a secondary framework [5, 6].

Data source

The NSCH is an annual survey, based on addresses, sponsored by the Health Resources and Services Administration's Maternal and Child Health Bureau and conducted by the US Census Bureau. From each eligible household, one child is selected for a detailed topical questionnaire, which is completed by a parent or caregiver. The weighted overall response rates were 35.8% in 2023 and 36.3% in 2024, with 55,162 and 51,375 topical questionnaires completed, respectively [7, 8]. The authors did not develop or administer a study-specific questionnaire. This secondary analysis used variables from the pre-existing 2023 and 2024 age-specific NSCH topical questionnaires: NSCH-T2 for children aged 6–11 years and NSCH-T3 for those aged 12–17 years; official English-language versions are publicly available from the US Census Bureau [9, 10]. We utilized the combined 2023–2024 CAHMI Data Resource Center Topical file, which contained 106,537 records.

Target population

The study focused on children aged 6–17 years with current ADHD whose parents reported behavioral treatment for ADHD within the past 12 months ($ADHDBehTreat_{2324} = 1$). The eligible cohort consisted of 5088 children, with 2314 from 2023 and 2774 from 2024. Model development used 2075 complete cases with 1199 events in 2023. With 30 predictor terms plus an intercept (31 fitted parameters), the nominal development events-per-parameter ratio was $1199/31 = 38.7$. This descriptive ratio does not replace a formal prediction-model sample-size calculation. The 2024 sample was reserved for evaluation and was not included. All eligible records were used; no separate formal sample-size calculation was performed.

Outcome

The primary outcome focused on parent-reported current anxiety, identified as $anxiety_{2324} = 3$. This was determined when a parent or caregiver reported that a healthcare provider had diagnosed the child with anxiety, and the child was currently experiencing the condition. Anxiety status was assessed in 4,973 eligible children, with 2,255 in 2023 and 2,718 in 2024, including 1,298 and 1,624 events, respectively. Sensitivity outcomes included either current anxiety or current depression, as well as

current depression alone. It is important to note that the outcome was not based on an independently adjudicated clinical diagnosis or a symptom-scale assessment.

Candidate predictors and coding

Before developing the model, twenty candidate predictors were finalized: age3_2324, sex_2324, race4_2324, povlev4_2324, famstruct5_2324, AdultEduc_2324, ADHDSev_2324, ADHDMed_2324, autism_2324, behavior_2324, learning_2324, DevDelay_2324, ACEct_2324, FamResilience_2324, ParentMH_2324, ParAggrav_2324, bullied_2324, FoodSit_2324, HrsSleep_2324, and PhysAct_2324. Detailed variable definitions, raw-code audits, and missingness summaries are provided in Additional file 1.

The primary specification comprised 10 binary terms and five ordinal-linear terms, including bullied_2324, FoodSit_2324, PhysAct_2324, FamResilience_2324, and ParentMH_2324. Additionally, it incorporated dummy variables for race/ethnicity, family structure, family poverty level, and adult education. Adverse childhood experiences (ACEct_2324) were coded as 0, 1, or at least 2, with 0 serving as the reference. This configuration resulted in a design matrix with 30 predictor terms and an intercept. During an encoding sensitivity analysis, the five ordinal variables were converted into dummy variables, and ACEct_2324 was treated as a count ranging from 0 to 9.

To prevent target and service-utilization leakage, variables directly representing anxiety, anxiety severity, depression, mental health needs or care, psychotropic service use, or complex-needs status were excluded prior to model development. The excluded variables included anxiety_2324, AnxietSev_2324, nomAnxietyDep12to17_2324, depress_2324, MEDB10ScrQ5_2324, mhneeds_2324, MentHCare_2324, MedEmotion_2324, and CSHCN_2324.

Missing data

Before creating cohort and complete-case indicators, variable-specific out-of-scope and missing codes were converted to missing values. The primary analysis necessitated an observed outcome and complete data for all 20 locked predictors. This analysis encompassed 4,619 children and 2,714 events, with 2,075 children experiencing 1,199 events in 2023 and 2,544 children encountering 1,515 events in 2024.

For sensitivity analysis, we used multiple imputation by chained equations with 10 imputations and 10 iterations [11]. The imputation models included the outcome, fwc_2324, and FIPSST. A `mids` object was initialized on 2023 data and extended with `mice.mids(newdata=...)`; the 2024 rows were automatically marked `ignore=TRUE`, so they did not contribute to estimation of the imputation-model parameters but were imputed using the 2023-fitted equations together with their own covariates.

Within each imputation, LASSO penalty selection and stepwise term selection were repeated, whereas `ranger` and `XGBoost` reused the hyperparameters selected in the complete-case development analysis;

XGBoost recalculated `scale_pos_weight` in the imputed 2023 development sample. Each model was refitted on imputed 2023 development data and evaluated without updating in imputed 2024 data. All 2774 eligible 2024 records entered this joint-model sensitivity, including 56 records whose anxiety outcomes were imputed. Predictor imputation could use the row's observed or imputed outcome, and model performance was calculated against the observed or imputed outcome. Accordingly, this was an outcome-informed stability analysis under the joint imputation model, not deployment-compatible missing-predictor processing or a strictly outcome-independent locked evaluation.

For coefficient pooling, a common prespecified 30-term survey-weighted logistic model was fitted in every imputation, and coefficients and standard errors were combined by Rubin's rules. Results from model selection with varying term sets were not pooled. Performance across imputations is reported as means and ranges, not as Rubin-pooled confidence intervals.

Complex survey design and domain analysis

The survey design included all 106537 source records before subgroup restriction, following full-design-before-domain analysis [12]. We specified HHID as the primary sampling unit, `interaction(FIPSST, STRATUM)` as the stratum, `fwc_2324` as the weight, and `nest=TRUE`; HHID remained a character identifier. In the CAHMI 2023–2024 combined file, each official annual final child weight (FWC) was divided by two and supplied as `fwc_2324`; it was used directly. Thus pooled estimates target the average annual population across the two survey years rather than the sum of two annual populations. The full-design degrees of freedom were 106435. The analytic cohort covered 100 of 102 full-design strata and included four single-PSU strata; `options(survey.lonely.psu="adjust")` was applied. Cohort, year, complete-case, and sensitivity domains were defined with `subset(design_all, ...)`, rather than by deleting records before constructing the design. The weight transformation was independently audited against both official annual files; all annual records matched $FWC/2$ within a relative tolerance of 1×10^{-9} .

Model development

The primary strategy separated model fitting from estimation of performance in the survey-weighted eligible population. In 2023, four models were developed without survey sampling weights: logistic LASSO (`glmnet`, `alpha = 1`, `lambda.min`) [13], bidirectional AIC stepwise logistic regression, probability random forest (`ranger`, 500 trees) [14], and XGBoost [15]. Class weights for random forest and `scale_pos_weight` for XGBoost addressed outcome imbalance and were distinct from survey sampling weights. This strategy evaluates models optimized in the unweighted development sample within the survey-weighted target population; it does not estimate a survey-weighted risk equation.

Internal evaluation used outer 10-fold cross-validation. In each outer training set, LASSO selected `lambda.min` by inner 10-fold cross-validation; random-forest and XGBoost hyperparameters were chosen by inner 5-fold cross-validation; and stepwise selection used only the outer training data. For XGBoost,

the same held-out inner fold served as the early-stopping evaluation set (early_stopping_rounds = 20) and the source of tuning AUC. These inner scores were used only for tuning and may be optimistic; neither outer test folds nor the 2024 sample was used for tuning or early stopping. Fold assignments were stratified by state and outcome, and each HHID remained wholly within one outer fold. All preprocessing, coding, selection, tuning, and class-imbalance calculations used only development data. Final models were then fitted in all 2075 development records; LASSO retained 18 terms and stepwise regression retained 16 terms.

Validation measures

The independent 2024 annual sample from the same survey program was used only for locked evaluation; it comprised different respondents and was not longitudinal follow-up. Primary performance estimates were survey weighted, with unweighted estimates reported secondarily. Metrics included the area under the receiver-operating-characteristic curve (AUC), Brier score, calibration intercept and slope, sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), and decision-curve net benefit at prespecified exploratory threshold probabilities of 0.20, 0.30, 0.40, 0.50, 0.60, and 0.70 [16, 17]. These thresholds were not derived from clinician or family preferences; the treat action denotes hypothetical high-risk classification, not a validated clinical intervention. Calibration intercept and slope were estimated by survey-weighted logistic regression of outcome on predicted logit. Calibration plots used 10 model-specific survey-weighted prediction-decile bins; bin-level mean predictions and observed prevalence were estimated with svymean, and observed 95% Wald CIs were calculated as ± 1.96 SE and truncated to [0,1]. Weighted and unweighted Youden cutoffs were selected from 2023 out-of-fold predictions and fixed for 2024.

Survey-design 95% CIs were percentile intervals from 500 bootstrap replicates. Within each interaction(FIPSST, STRATUM) stratum of the full annual survey frame, HHID primary sampling units were sampled with replacement and replicate multiplicities were mapped to the analytic domain. Predictions and fitted-model parameters remained fixed. Identical replicate draws were used across models for paired differences. Seeds were 161803 for 2023 and 161804 for 2024; all 500 replicates produced finite estimates for every reported metric and strategy. Bootstrap calibration used replicate survey weights normalized by a constant for numerical stability, which does not alter binomial score equations. These intervals quantify complex survey-sampling uncertainty conditional on the fitted models or realized out-of-fold predictions; they do not include uncertainty from preprocessing, selection, tuning, coefficient estimation, or model refitting.

Sensitivity, subgroup, and transportability analyses

Prespecified analyses assessed multiple imputation with survey-weighted model fitting, alternative predictor coding, current anxiety or depression, current depression alone, age groups 6–11 and 12–17 years, and the 2024 subgroup not taking ADHD medication. Locked models were also applied separately to children with ADHD in 2023 and 2024 who did not receive behavioral treatment (ADHDBehTreat_2324 = 2). This analysis assessed transportability to a different target population and was not an evaluation

within the treated cohort. No multiplicity adjustment was made for exploratory sensitivity or subgroup analyses.

Model interpretation and software

The models were designed for risk stratification and research comparison, rather than for diagnostic purposes. Logistic coefficients were presented as exact values to ensure reproducibility. Since anxiety was prevalent, the exponentiated logistic coefficients were interpreted as odds ratios rather than prevalence ratios, and they were not considered as indicators of population prevalence effects. Post-selection coefficient confidence intervals (CIs) were not employed for confirmatory inference. SHAP values were utilized to explain prediction contributions in random forest and XGBoost models, rather than to infer causal effects [18]. The analyses were performed using R.

Patient and public involvement

Patients and the public were not engaged in the design, conduct, reporting, or dissemination planning of this secondary data analysis. This lack of involvement should be taken into account when determining clinical thresholds and implementation pathways.

Results

Participant flow and characteristics

Out of 106,537 source records, 5,088 children met the criteria for the target population. Anxiety status was unavailable for 115 of these children, resulting in 4,973 with observed outcomes. Additionally, 354 children were missing at least one locked predictor, leaving 4,619 complete cases (Additional file 2: Figure S1). The complete-case sample comprised 2,075 children in 2023 and 2,544 in 2024, with a total of 2,714 events recorded. The survey-weighted anxiety prevalence among children with observed outcomes was 51.8% overall, 49.3% in 2023, and 54.0% in 2024.

In the complete-case sample, 53.2% of participants were aged 6–11 years, and 64.4% were male. The weighted distribution of race/ethnicity was as follows: 53.4% White non-Hispanic, 19.1% Hispanic, 15.9% Black non-Hispanic, and 11.5% multiracial or other non-Hispanic. Overall, 19.6% of the sample fell below 100% of the federal poverty level, and 59.5% resided in two-parent married households. Moderate or severe ADHD was reported in 71.3% of the participants, with 64.1% taking ADHD medication. Current parent-reported conditions included autism or autism spectrum disorder in 23.5%, behavioral or conduct problems in 56.1%, learning disabilities in 43.8%, and developmental delays in 27.4% (Table 1).

Table 1
Selected complete-case sample characteristics (N = 4619)

Characteristic	Unweighted n	Survey-weighted %
Age 6–11 years	2193	53.2
Age 12–17 years	2426	46.8
Male	2941	64.4
Female	1678	35.6
Hispanic	534	19.1
White, non-Hispanic	3298	53.4
Black, non-Hispanic	291	15.9
Multiracial/other, non-Hispanic	496	11.5
< 100% federal poverty level	604	19.6
Two-parent married household	2983	59.5
Moderate/severe ADHD	3338	71.3
Taking ADHD medication	3161	64.1
Current autism/ASD	1001	23.5
Current behavioral/conduct problems	2518	56.1
Current learning disability	1873	43.8
Current developmental delay	1141	27.4
ACE count ≥ 2	1855	39.0
ACE: adverse childhood experience; ADHD: attention-deficit/hyperactivity disorder; ASD: autism spectrum disorder. Percentages use the pooled survey weight and complex design. Full distributions are provided in Additional file 1.		

Model development and internal evaluation

Each of the 2075 development records received one out-of-fold prediction from every model. In nested 10-fold cross-validation, survey-weighted AUCs were 0.769 (95% CI, 0.736–0.804) for LASSO, 0.764 (0.731–0.799) for stepwise regression, 0.743 (0.711–0.781) for random forest, and 0.729 (0.683–0.774) for XGBoost. Corresponding weighted Brier scores were 0.199 (0.186–0.211), 0.199 (0.185–0.212), 0.210 (0.198–0.221), and 0.214 (0.195–0.233). The 2023 intervals are conditional on the realized out-of-fold predictions and do not propagate uncertainty from repeating preprocessing, selection, tuning, or fitting.

Evaluation in the independent 2024 annual sample

In the independent 2024 annual sample, the survey-weighted AUCs were 0.717 (95% CI, 0.678–0.758) for LASSO, 0.723 (0.678–0.766) for stepwise regression, 0.711 (0.665–0.751) for random forest, and 0.674 (0.627–0.724) for XGBoost. Corresponding weighted Brier scores were 0.214 (0.200–0.227), 0.212 (0.197–0.227), 0.215 (0.204–0.227), and 0.250 (0.226–0.273). The unweighted AUCs were 0.728, 0.726, 0.717, and 0.677, respectively (Table 2). These intervals are conditional on fixed predictions from the locked fitted models and do not include model-development uncertainty.

Table 2
Performance in the independent 2024 annual sample (n = 2544; 1515 events)

Model	Weighted AUC (95% CI)	Weighted Brier (95% CI)	Unweighted AUC	Cal. intercept (95% CI)	Cal. slope (95% CI)	NB at 0.50 (95% CI)
LASSO	0.717 (0.678–0.758)	0.214 (0.200–0.227)	0.728	-0.02 (-0.22–0.17)	0.97 (0.76–1.20)	0.217 (0.148–0.273)
Stepwise logistic	0.723 (0.678–0.766)	0.212 (0.197–0.227)	0.726	0.05 (-0.14–0.22)	0.85 (0.64–1.06)	0.227 (0.175–0.276)
Random forest	0.711 (0.665–0.751)	0.215 (0.204–0.227)	0.717	-0.09 (-0.30–0.10)	1.13 (0.87–1.40)	0.195 (0.121–0.262)
XGBoost	0.674 (0.627–0.724)	0.250 (0.226–0.273)	0.677	0.16 (-0.03–0.33)	0.35 (0.25–0.47)	0.160 (0.099–0.217)
AUC: area under the receiver-operating-characteristic curve; Cal.: calibration; CI: confidence interval; NB: net benefit. Survey-design 95% CIs are percentile intervals from 500 of 500 finite state-by-stratum HHID-cluster bootstrap replicates (seed 161804) with fixed predictions. They quantify survey-sampling uncertainty conditional on the locked fitted models and do not include model-development uncertainty. Treat-all net benefit at 0.50 was 0.082 (95% CI, – 0.016 to 0.168); threshold analyses were exploratory.						

The paired weighted AUC difference between LASSO and stepwise regression was – 0.006 (95% CI, – 0.027 to 0.014). Compared with XGBoost, differences were 0.044 (0.001–0.090) for LASSO, 0.049 (0.010–0.092) for stepwise regression, and 0.037 (0.008–0.068) for random forest. These comparisons were interpreted together with calibration and decision-curve results. They were exploratory and not adjusted for multiplicity; their CIs should not be interpreted as confirmatory evidence of model superiority.

The weighted calibration slopes were as follows: 0.97 (95% CI, 0.76–1.20) for LASSO, 0.85 (0.64–1.06) for stepwise regression, 1.13 (0.87–1.40) for random forest, and 0.35 (0.25–0.47) for XGBoost. Corresponding weighted calibration intercepts were – 0.02 (– 0.22 to 0.17), 0.05 (– 0.14 to 0.22), – 0.09 (– 0.30 to 0.10), and 0.16 (– 0.03 to 0.33), respectively (Fig. 1). The slope for XGBoost suggested that its predictions were excessively extreme in 2024.

In 2024, using the weighted Youden thresholds derived from 2023 out-of-fold predictions, the LASSO cutoff was set at 0.603. At this cutoff, sensitivity was 0.551 (95% CI, 0.501–0.600), specificity reached 0.774 (0.721–0.821), the positive predictive value (PPV) was 0.742 (0.699–0.784), and the negative predictive value (NPV) was 0.594 (0.538–0.652). The fixed cutoffs for other methods were 0.604 for stepwise regression, 0.531 for random forest, and 0.575 for XGBoost.

At a threshold probability of 0.50, weighted net benefit was 0.217 (95% CI, 0.148–0.273) for LASSO, 0.227 (0.175–0.276) for stepwise regression, 0.195 (0.121–0.262) for random forest, and 0.160 (0.099–0.217) for XGBoost. Treat-all net benefit was 0.082 (– 0.016 to 0.168), and treat-none net benefit was 0 (Fig. 2). Under this hypothetical threshold preference, the models had greater estimated net benefit than treat-all; this exploratory comparison does not demonstrate clinical utility or patient benefit.

Sensitivity, subgroup, and transportability results

In the outcome-informed joint-model multiple-imputation sensitivity (2024 n = 2774, including 56 imputed anxiety outcomes), mean 2024 weighted AUCs were 0.712 for LASSO (range, 0.701–0.725), 0.717 for stepwise regression (0.711–0.723), 0.702 for random forest (0.693–0.711), and 0.681 for XGBoost (0.672–0.693). These estimates assess stability under the imputation model rather than deployment-compatible evaluation. With alternative coding, weighted AUCs were 0.712 (95% CI, 0.671–0.754) for LASSO and 0.720 (0.673–0.765) for stepwise regression. For current anxiety or depression, AUCs were 0.727 (0.684–0.767) and 0.719 (0.676–0.758); for current depression alone, they were 0.796 (0.759–0.830) and 0.791 (0.752–0.826) (Table 3).

Table 3
Selected sensitivity, subgroup, and transportability results

Analysis	LASSO AUC	Stepwise AUC	Interpretive note
MICE, 2024 (n = 2774; 56 imputed outcomes); mean (range)	0.712 (0.701–0.725)	0.717 (0.711–0.723)	Outcome-informed joint-model sensitivity; ranges are not CIs
Alternative coding, 2024	0.712 (0.671–0.754)	0.720 (0.673–0.765)	Weighted AUC (95% CI)
Current anxiety or depression, 2024	0.727 (0.684–0.767)	0.719 (0.676–0.758)	Weighted AUC (95% CI)
Current depression, 2024	0.796 (0.759–0.830)	0.791 (0.752–0.826)	Weighted AUC (95% CI)
Untreated ADHD transportability, 2023	0.692 (0.643–0.731)	0.691 (0.644–0.732)	Different target population
Untreated ADHD transportability, 2024	0.700 (0.662–0.730)	0.696 (0.657–0.729)	Different target population
Age 6–11 years, 2024	0.655 (0.585–0.716)	0.689 (0.620–0.751)	Exploratory subgroup
Age 12–17 years, 2024	0.722 (0.665–0.777)	0.704 (0.641–0.767)	Exploratory subgroup
No ADHD medication, 2024	0.720 (0.632–0.790)	0.727 (0.629–0.802)	n = 787; 431 events
CI: confidence interval; MICE: multiple imputation by chained equations. Except for MICE ranges, entries are survey-weighted AUCs with 95% CIs. The MICE row used all 2774 eligible 2024 records, including 56 imputed anxiety outcomes, and was an outcome-informed joint-model sensitivity rather than deployment-compatible evaluation. Sensitivity and subgroup analyses were exploratory and were not multiplicity adjusted. No formal between-age interaction or performance-difference test was prespecified.			

For children with ADHD who did not undergo behavioral treatment, the LASSO and stepwise AUCs were 0.692 (95% CI, 0.643–0.731) and 0.691 (0.644–0.732) in 2023, respectively, and 0.700 (0.662–0.730) and 0.696 (0.657–0.729) in 2024. These findings assessed transportability instead of validation within the treated target population.

In 2024, the LASSO AUC was 0.655 (95% CI, 0.585–0.716) for children aged 6–11 years and 0.722 (0.665–0.777) for those aged 12–17 years. Among 787 children not taking ADHD medication, with 431

events, the AUC was 0.720 (0.632–0.790) for LASSO and 0.727 (0.629–0.802) for stepwise regression. These subgroup estimates were exploratory and less precise; no formal between-age interaction or performance-difference test was prespecified, and the results were not used for model selection or updating.

Reproducibility and model explanations

The exact LASSO and stepwise coefficients used for generating evaluation predictions are provided in Additional file 1 and in Additional file 2: Figures S2 and S3; no separate model refitting was performed for the coefficient plots. SHAP additivity was confirmed for XGBoost on the log-odds margin scale, with a maximum absolute error of 6.07×10^{-6} and a tolerance of 1×10^{-5} , and for random forest on the probability scale, with a maximum absolute error of 6.05×10^{-15} and a tolerance of 1×10^{-6} . SHAP values, shown in Additional file 2: Figures S4 and S5, illustrate contributions to model predictions, not causal or population effect estimates. TRIPOD + AI and STROBE manuscript crosswalks are provided in Additional files 3 and 4.

Discussion

Principal findings

In a national probability survey of eligible children with ADHD who had received behavioral treatment, parents commonly reported current anxiety, with a survey-weighted prevalence of 51.8% among those with observed outcome data. When models developed in 2023 were evaluated without updating in the independent 2024 annual sample, LASSO, stepwise regression, and random forest showed moderate and broadly similar discrimination. LASSO and random forest had calibration slopes near 1, whereas XGBoost produced substantially overdispersed predictions and the highest Brier score. At prespecified exploratory thresholds, regression models had greater estimated net benefit than treating all children as high risk under the corresponding hypothetical preferences. These results do not establish clinical utility or real-world patient benefit.

The findings suggest three interpretations. First, substantial risk heterogeneity persisted within a subgroup already receiving behavioral treatment for ADHD. Second, flexible machine-learning algorithms did not outperform regression models in the independent later annual sample. Algorithmic complexity alone does not ensure improved transportability, particularly when predictors are structured survey variables and the effective signal-to-noise ratio is limited. Third, the decrease from internal cross-validation to the 2024 annual-sample evaluation shows why a random split of pooled data would have provided an incomplete assessment of generalizability.

Clinical and public-health interpretation

The AAP advises evaluating children and adolescents with ADHD for comorbid anxiety and other emotional or behavioral conditions [1]. The current models should not substitute for this assessment.

Their most justifiable role is research-focused risk stratification or prioritization; for example, identifying subgroups in which structured anxiety screening might merit study, rather than confirming or ruling out a diagnosis. The fixed LASSO classification cutoff illustrates this distinction: specificity and PPV were higher than sensitivity and NPV, so a negative classification would not reliably exclude anxiety.

The models should not be interpreted as showing that individual predictors cause anxiety. Predictors and outcome were reported by the same respondent within the same cross-sectional survey wave, and current anxiety required prior professional identification. Apparent performance may therefore reflect underlying psychopathology, common-method reporting, social conditions, differential access to diagnosis, or a combination of these factors. Logistic odds ratios are not prevalence ratios when the outcome is common, and post-selection coefficients are not suitable for confirmatory etiological inference.

Comparison across modeling strategies

The findings underscore the importance of assessing calibration alongside discrimination. A model may rank individuals effectively yet yield probabilities that are excessively extreme or systematically biased [16]. In 2024, XGBoost demonstrated both lower discrimination and a calibration slope of 0.35, suggesting possible overfitting or an overly broad prediction range. LASSO and stepwise regression showed comparable AUCs, Brier scores, and net benefits, with the paired AUC confidence interval including zero. Consequently, the decision between these two models should be based on factors such as reproducibility, stability, ease of implementation, and potential for future validation, rather than asserting the clear superiority of one over the other.

Exploratory decision-curve analysis showed potential net benefit under the prespecified hypothetical threshold preferences; however, no threshold was elicited from clinicians or families or tested in an implementation setting. The analysis therefore does not demonstrate clinical utility or patient benefit. Future work should define the intended decision, identify who would respond to a high-risk result, elicit stakeholder-relevant thresholds, and assess whether model use improves timely assessment or outcomes without unnecessary labeling or resource use.

Strengths

Strengths include use of a national probability survey, a prespecified target population and predictor set, and explicit exclusion of outcome-derived and service-utilization variables. Models were developed in 2023 and evaluated without updating in an independent 2024 annual sample, with nested internal cross-validation and survey-weighted evaluation using the full sampling design. The analysis preserved the full design before domain restriction, resampled household clusters within state-by-stratum cells, kept development-derived cutoffs fixed, and assessed discrimination, calibration, overall accuracy, classification performance, and exploratory decision-curve net benefit. Exact coefficients and supporting audit summaries are supplied in Additional file 1. Multiple-imputation, coding, outcome, medication, age, and untreated-population analyses provided complementary checks on model behavior.

Limitations

This study has several limitations. First, the NSCH is cross-sectional: predictors and outcome were reported within the same questionnaire, so the models classify current anxiety status rather than predict future onset, and temporal ordering and causal effects cannot be inferred. The annual samples comprised different respondents rather than longitudinal follow-up. Second, anxiety was parent reported as a current professionally identified condition without confirmation by diagnostic interview or validated symptom scale. Because predictors and outcome came from the same respondent and professional identification depends on access to care, common-method reporting and differential diagnostic access may influence apparent performance.

Third, the primary analysis was complete case. Multiple imputation yielded similar discrimination but cannot remove bias if missingness depends on unobserved factors or the imputation models are misspecified. Because the joint imputation model included the outcome and the MI performance analysis included imputed 2024 outcomes, that sensitivity analysis was outcome informed and did not emulate deployment-time missing-predictor processing or a strictly outcome-independent locked evaluation. Fourth, the 2023 and 2024 samples came from the same survey program; the 2024 analysis is evaluation in an independent later annual sample, not validation in an independent health system, country, data-collection platform, or prospective clinical cohort. Fifth, weighted overall response rates were 35.8% in 2023 and 36.3% in 2024; weighting can mitigate but not eliminate nonresponse bias [7, 8].

Sixth, the bootstrap intervals condition on fixed fitted models or realized out-of-fold predictions and do not propagate uncertainty from model development. With 500 replicates, percentile endpoints also retain finite Monte Carlo error. Seventh, paired model, subgroup, and alternative-outcome analyses were exploratory and not adjusted for multiplicity. No formal between-age interaction or performance-difference test was prespecified; subgroup estimates should not be treated as evidence of differential performance or fairness. Finally, implementation would require independent validation, recalibration if needed, assessment across demographic and clinical subgroups, stakeholder-defined thresholds, workflow and acceptability evaluation, and a prospective impact study.

Implications for future research

To ensure robust independent validation, clinically assessed anxiety outcomes and data collected outside the NSCH should be used, ideally with a prospective time origin to ensure predictors precede outcomes. Researchers should evaluate calibration-in-the-large and calibration slopes across various settings, investigate whether simpler models can maintain performance, and assess algorithmic fairness across age, sex, race and ethnicity, socioeconomic status, and geographic context. For models intended to trigger screening or referral, thresholds should be determined collaboratively with clinicians and families and tested in an impact study. Updating or recalibration must be distinct from evaluation and transparently reported following TRIPOD + AI guidelines [5].

Conclusions

In two independent cross-sectional samples from consecutive NSCH waves, regression and random-forest models developed in 2023 showed moderate discrimination for parent-reported current anxiety among eligible children with ADHD who had received behavioral treatment when evaluated, without updating, in the 2024 sample. These models classify same-wave current anxiety status; they do not predict future onset, establish causation, or replace structured clinical assessment. Their use should remain research oriented until performance, fairness, thresholds, workflow consequences, and clinical impact are evaluated in independent prospective settings with clinically assessed outcomes.

Abbreviations

AAP

American Academy of Pediatrics

ACE

adverse childhood experience

ADHD

attention-deficit/hyperactivity disorder

AIC

Akaike information criterion

ASD

autism spectrum disorder

AUC

area under the receiver-operating-characteristic curve

CAHMI

Child and Adolescent Health Measurement Initiative

CI

confidence interval

DCA

decision-curve analysis

HHID

household identifier

LASSO

least absolute shrinkage and selection operator

MICE

multiple imputation by chained equations

NPV

negative predictive value

NSCH

National Survey of Children's Health

PPV

positive predictive value

PSU
primary sampling unit
SHAP
Shapley additive explanations
STROBE
Strengthening the Reporting of Observational Studies in Epidemiology
TRIPOD + AI
Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis plus Artificial Intelligence.

Declarations

>

Ethics approval and consent to participate

This study was a secondary analysis of the deidentified public-use 2023–2024 National Survey of Children's Health combined topical dataset obtained from the Child and Adolescent Health Measurement Initiative Data Resource Center for Child and Adolescent Health. The analysis involved no direct interaction with participants and no access to identifiable private information. Institutional review board approval and additional informed consent were therefore not required. Trial registration: Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Funding

No funding was received for this work.

Author Contribution

CQ conceived the study, developed the methodology, conducted the statistical analyses, interpreted the results, and drafted the manuscript. YS and QC contributed to the clinical interpretation of the findings and critically revised the manuscript. JX and YF contributed to the literature review, interpretation of the results, and critical revision of the manuscript. GC and MW contributed to data curation, preparation of tables and figures, and critical revision of the manuscript. CY supervised the study, contributed to the

study design and interpretation of the findings, and critically revised the manuscript. All authors read and approved the final manuscript.

Acknowledgements

Not applicable.

Data Availability

The 2023–2024 combined NSCH dataset and codebook used in this study are available from the CAHMI Data Resource Center under its data-use conditions [19]. Official annual NSCH public-use microdata and technical documentation are available from the US Census Bureau [20]. Additional file 1 provides variable definitions, audit summaries, exact model coefficients, tuning results, performance estimates, and sensitivity-analysis outputs.

References

1. Wolraich ML, Hagan JF Jr, Allan C, et al. Clinical practice guideline for the diagnosis, evaluation, and treatment of attention-deficit/hyperactivity disorder in children and adolescents. *Pediatrics*. 2019;144(4):e20192528. 10.1542/peds.2019-2528.
2. Danielson ML, Claussen AH, Bitsko RH, et al. ADHD prevalence among US children and adolescents in 2022: diagnosis, severity, co-occurring disorders, and treatment. *J Clin Child Adolesc Psychol*. 2024;53(3):343–60. PMID:38778436.
3. Njardvik U, Wergeland GJ, Riise EN, Hannesdottir DK, Ost LG. Psychiatric comorbidity in children and adolescents with ADHD: a systematic review and meta-analysis. *Clin Psychol Rev*. 2025;118:102571. 10.1016/j.cpr.2025.102571.
4. Zhang Y, Liao W, Rao Y, Gao W, Yang R. Effects of ADHD and ADHD medications on depression and anxiety in children and adolescents: a systematic review and meta-analysis. *J Psychiatr Res*. 2025;181:623–39. 10.1016/j.jpsychires.2024.12.022.
5. Collins GS, Moons KGM, Dhiman P, et al. TRIPOD + AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. *BMJ*. 2024;385:e078378. 10.1136/bmj-2023-078378.
6. von Elm E, Altman DG, Egger M, Pocock SJ, Gøtzsche PC, Vandenbroucke JP, STROBE Initiative. The Strengthening of Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. *Lancet*. 2007;370(9596):1453–7. 10.1016/S0140-6736(07)61602-X.
7. 2023. US Census Bureau, National Survey of Children's Health Methodology Report, Published. January 31, 2025. <https://www2.census.gov/programs-surveys/nsch/technical-documentation/methodology/2023-NSCH-Methodology-Report.pdf>. Accessed 10 Aug 2026.

8. US Census Bureau. 2024 National Survey of Children's Health Methodology Report. Published December 1, 2025. <https://www2.census.gov/programs-surveys/nsch/technical-documentation/methodology/2024-NSCH-Methodology-Report.pdf>. Accessed 10 Aug 2026.
9. US Census Bureau. 2023 National Survey of Children's Health questionnaires. <https://www.census.gov/programs-surveys/nsch/technical-documentation/questionnaires/2023.html>. Accessed 2 Sep 2026.
10. US Census Bureau. 2024 National Survey of Children's Health questionnaires. <https://www.census.gov/programs-surveys/nsch/technical-documentation/questionnaires/2024.html>. Accessed 2 Sep 2026.
11. van Buuren S, Groothuis-Oudshoorn K. mice: multivariate imputation by chained equations in R. *J Stat Softw.* 2011;45(3):1–67. 10.18637/jss.v045.i03.
12. Lumley T. Analysis of complex survey samples. *J Stat Softw.* 2004;9(8):1–19. 10.18637/jss.v009.i08.
13. Friedman J, Hastie T, Tibshirani R. Regularization paths for generalized linear models via coordinate descent. *J Stat Softw.* 2010;33(1):1–22. 10.18637/jss.v033.i01.
14. Wright MN, Ziegler A. ranger: a fast implementation of random forests for high dimensional data in C++ and R. *J Stat Softw.* 2017;77(1):1–17. 10.18637/jss.v077.i01.
15. Chen T, Guestrin C. XGBoost: a scalable tree boosting system. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. Association for Computing Machinery; 2016:785–794. 10.1145/2939672.2939785
16. Van Calster B, McLernon DJ, van Smeden M, Wynants L, Steyerberg EW. Calibration: the Achilles heel of predictive analytics. *BMC Med.* 2019;17:230. 10.1186/s12916-019-1466-7.
17. Vickers AJ, Elkin EB. Decision curve analysis: a novel method for evaluating prediction models. *Med Decis Mak.* 2006;26(6):565–74. 10.1177/0272989X06295361.
18. Lundberg SM, Lee SI. A unified approach to interpreting model predictions. *Adv Neural Inf Process Syst.* 2017;30:4765–74.
19. Child and Adolescent Health Measurement Initiative. DRC Indicator Dataset & Codebook Instructions: 2023–2024 NSCH (2 years combined). <https://www.childhealthdata.org/help/dataset/2023-2024-combined-nsch-dataset--codebook-instruction>. Accessed 10 Aug 2026.
20. US Census Bureau. National Survey of Children's Health datasets. <https://www.census.gov/programs-surveys/nsch/data/datasets.html>. Accessed 10 Aug 2026.

Figures

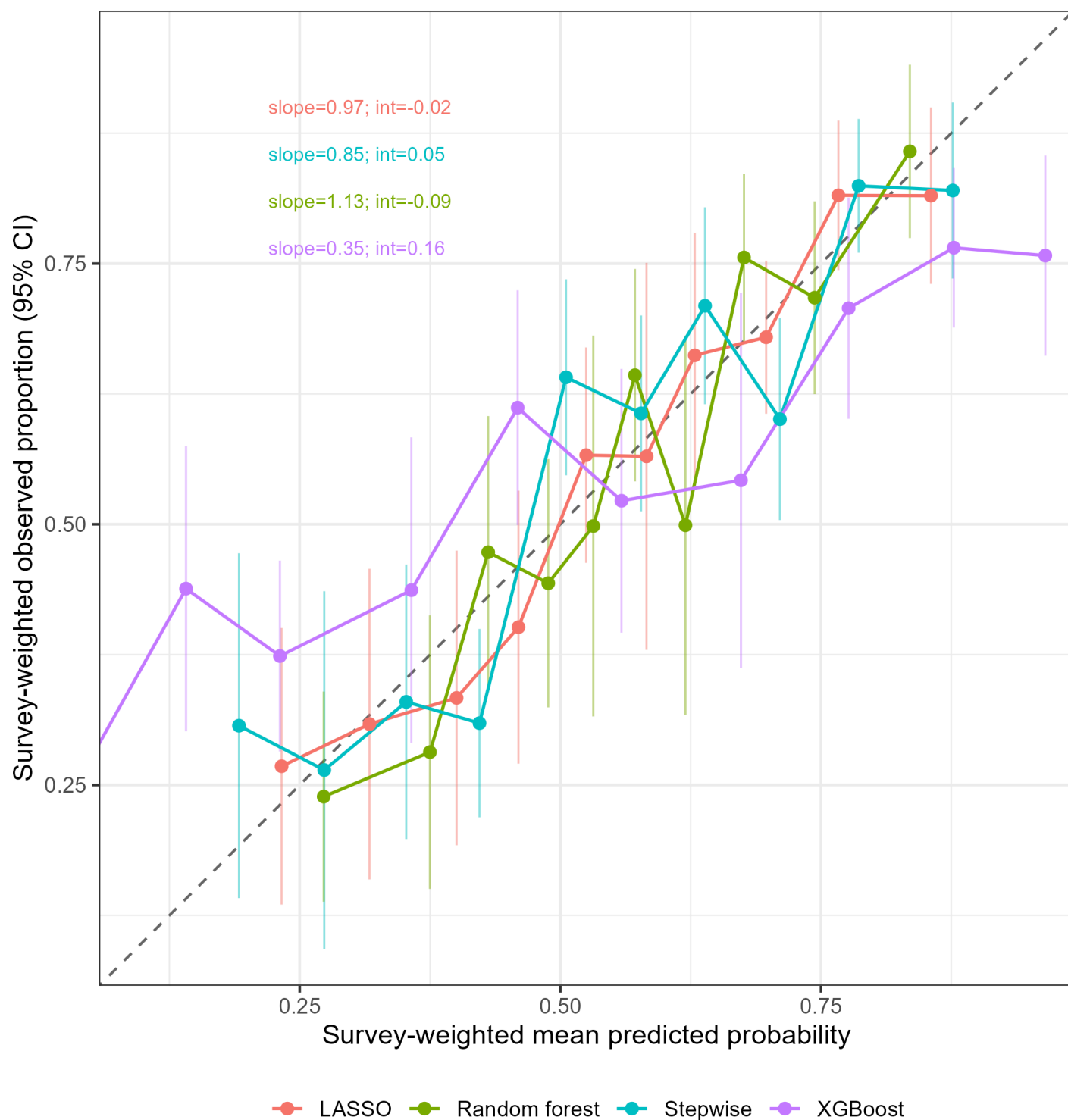


Figure 1

Survey-weighted calibration in the independent 2024 annual-sample evaluation.

Each model was divided into 10 nonempty, model-specific survey-weighted prediction-decile bins. Bin-level mean predicted probability and observed prevalence were estimated using svymean under the complex survey design; observed 95% Wald CIs were calculated as ± 1.96 SE and truncated to [0,1]. The

diagonal denotes ideal calibration. Calibration intercepts and slopes were estimated by survey-weighted logistic regression of outcome on predicted logit.

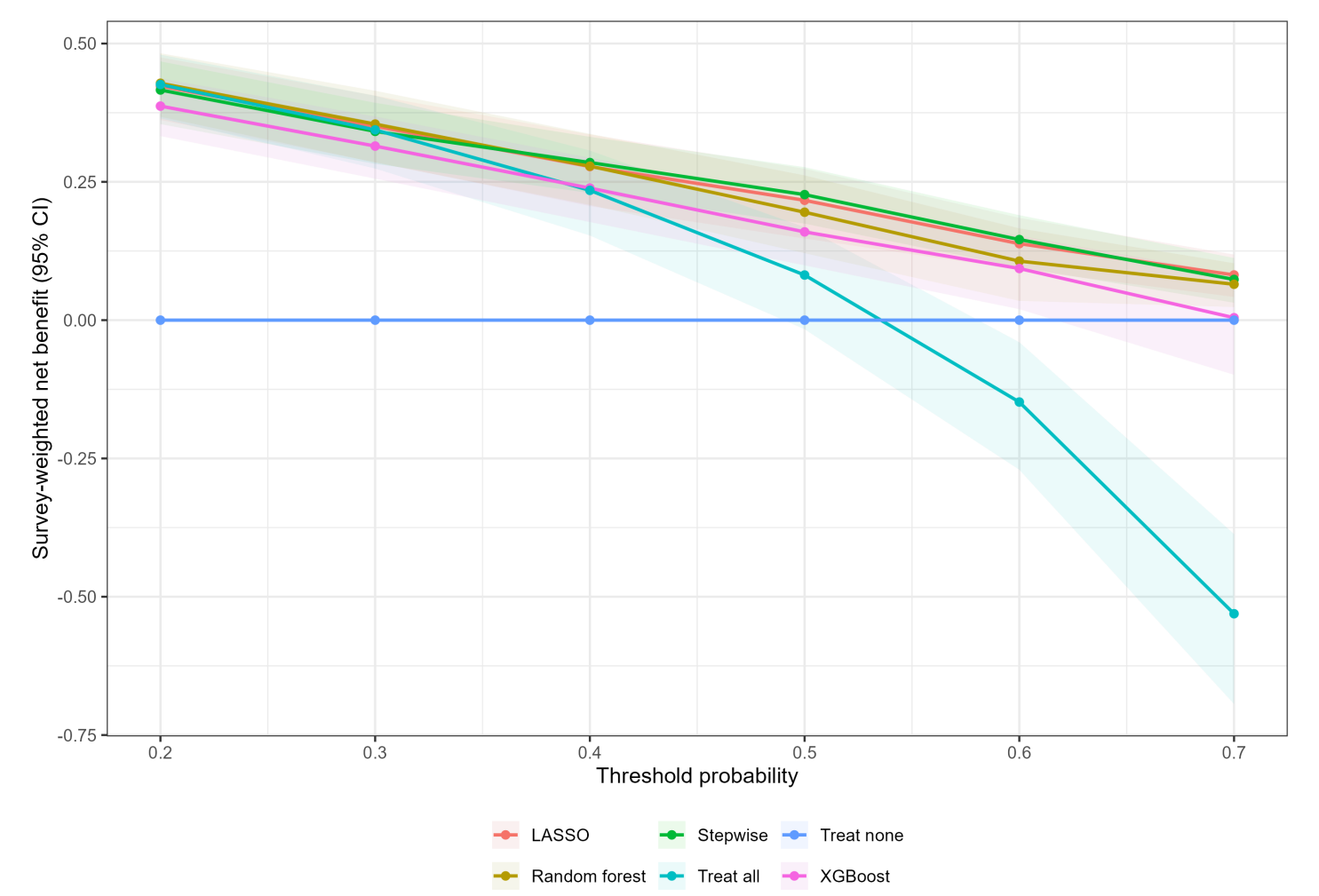


Figure 2

Survey-weighted decision curves in the independent 2024 annual-sample evaluation.

Net benefit is shown at prespecified exploratory threshold probabilities from 0.20 to 0.70 for the four locked models and treat-all and treat-none strategies. The treat action denotes hypothetical high-risk classification; thresholds were not stakeholder derived or clinically validated. The curves do not establish clinical utility or patient benefit. Shaded bands are conditional survey-design 95% percentile intervals from 500 fixed-prediction HHID-cluster bootstrap replicates.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [Additionalfile1.xlsx](#)

- [Additionalfile2SupplementaryFigures.pdf](#)
- [Additionalfile3TRIPODAIChecklist.docx](#)
- [Additionalfile4STROBECrosswalk.docx](#)