ACCSCIENCE PUBLISHING

Artificial Intelligence in Health

ORIGINAL RESEARCH ARTICLE

Knowledge-constrained neural additive modeling for hypoglycemia risk stratification in older adults with type 2 diabetes

Supplementary Files

This supplementary file provides robustness and sensitivity analyses that support the manuscript's main findings. Table S1 reports the ablation comparison of knowledge-constrained neural additive modeling (CNAM) variants. Table S2 summarizes subgroup precision–recall-area under the curve results. Table S3 reports calibration-metric sensitivity under sigmoid and isotonic mapping. Table S4 presents seed sensitivity analyses. Table S5 summarizes variable-level missingness in the raw analysis layer. Table S6 compares median imputation and multiple imputation by chained equations sensitivity for the evaluated baseline models. Table S7 reports the sensitivity analysis for excluding predictors with >20% missingness. Table S8 compares CNAM and explainable boosting machine shape stability, and Table S9 reports monotonicity-violation analysis for the constrained features. Figures S1–S4 provide the corresponding visual summaries. These tables and figures support the robustness, sensitivity, and interpretability analyses reported in Sections 3.4–3.5 of the main manuscript.

Note S1 and Algorithm S1 describe the outcome definition and calibration pipeline. Tables S10–S13 and Figures S5 and S6 provide additional calibration, operating-point, and decision-curve sensitivity analyses.

Note S1. Outcome definition and severity level limitation

The outcome in this study was the hypoglycemia history in a cross-sectional outpatient survey. It is an ascertainment-dependent endpoint rather than a prospectively monitored incident hypoglycemia outcome. Recorded history may reflect actual event occurrence, patient recall, glucose monitoring behavior, symptom recognition, and clinical documentation practice. In older adults, nocturnal hypoglycemia and impaired symptom awareness may further contribute to under-ascertainment.

The available dataset did not systematically distinguish Level 1, Level 2, and Level 3 hypoglycemia. We therefore used one composite recorded-history endpoint based on the available 3.9 mmol/L clinical alert threshold. This definition was feasible for the present retrospective cross-sectional analysis, but it may combine clinically heterogeneous events. Accordingly, the learned shape functions and calibrated probabilities should be interpreted as associations with recorded hypoglycemia history in this cohort, not as severity-specific estimates of all underlying hypoglycemic events.

Algorithm S1. Pooled development out-of-fold calibration and locked-test application

Input: Full cohort D; fixed locked-test split; five development outer folds; model-training procedure; sigmoid calibration method; development threshold-selection rules.

Output: Calibrated locked-test probabilities and locked-test operating-point evaluation without using locked-test outcomes for model development.

Procedure:

(i) Split D into development set D_dev and locked test set D_test using the prespecified stratified hold-out split.
(ii) For each outer fold $k = 1,...,5$ do.
(iii) Fit the model using only the corresponding outer-training fold.
(iv) Store raw held-out predictions for development participants in the outer validation fold.
(v) Generate raw predictions for D_test using the kth outer-fold model.
(vi) End for.
(vii) Pool all held-out development predictions into one complete development out-of-fold (OOF) prediction stack.
(viii) Fit one sigmoid calibration model on the pooled development OOF stack and corresponding development labels.
(ix) Average the five locked-test raw prediction vectors to obtain the cross-validation-ensemble raw probability.
(x) Apply the single sigmoid calibrator from Line 8 unchanged to the averaged locked-test raw probabilities.

(xi) Select operating thresholds only from development-calibrated probabilities and apply them unchanged to D_test.
(xii) Evaluate D_test once for final discrimination, calibration, threshold-dependent, and decision-curve metrics.
(xiii) Do not use locked-test outcomes for model fitting, hyperparameter selection, calibration fitting, threshold selection, or model selection.

Note: A single pooled development OOF calibrator was used; fold-specific calibration functions were not fitted and averaged.

Table S1. Ablation study of CNAM variants on the locked test set

Variant	CV ROC-AUC (mean)	CV PR-AUC (mean)	Test ROC-AUC	Test PR-AUC	Test Brier score	Test sensitivity (%)	Test specificity (%)
NAM	0.7015	0.2769	0.7312	0.4084	0.2039	74.36	65.77
CNAM (smooth)	0.7024	0.2836	0.7324	0.4080	0.2054	74.36	63.51
CNAM (smooth + mono)	0.7045	0.2885	0.7387	0.3876	0.2040	79.49	59.91

Abbreviations: AUC: Area under the curve; CNAM: knowledge-constrained neural additive modeling; CV: Cross-validation; NAM: Neural additive modeling; PR: Precision–recall; ROC: Receiver operating characteristic.

Table S2. Subgroup PR-AUC (calibrated probabilities) for CNAM and EBM on the locked test set

Subgroup type	Subgroup level	*n*	Positive cases	CNAM PR-AUC (95% CI)	EBM PR-AUC (95% CI)
Age group	60–69	154	23	0.3602 (0.2005–0.5434)	0.3678 (0.1954–0.5501)
Age group	70–79	74	12	0.5336 (0.2883–0.7789)	0.5738 (0.3117–0.8018)
Age group	≥80	33	4	0.5164 (0.1250–1.0000)	0.3958 (0.1369–0.9211)
Insulin use	No	140	13	0.2116 (0.0731–0.4121)	0.2115 (0.0685–0.4343)
Insulin use	Yes	121	26	0.5074 (0.3306–0.6815)	0.4967 (0.3283–0.6961)
Secretagogue use	No	176	26	0.4021 (0.2393–0.6032)	0.4488 (0.2786–0.6216)
Secretagogue use	Yes	85	13	0.4534 (0.2131–0.7339)	0.4074 (0.2060–0.6872)
Sex	Female	128	19	0.3440 (0.1883–0.5764)	0.3265 (0.1809–0.5374)
Sex	Male	133	20	0.4877 (0.2833–0.6916)	0.5465 (0.3131–0.7408)

Note: Results for small subgroups, particularly the ≥80 years subgroup, should be interpreted cautiously due to the limited number of positive cases.
Abbreviations: AUC: Area under the curve; CI: Confidence interval; CNAM: knowledge-constrained neural additive modeling; EBM: Explainable boosting machine; PR: Precision–recall.

Table S3. Calibration sensitivity analysis on the locked test set, using calibration models fitted on development out-of-fold predictions

Model	Calibration method	Fit source	Brier score	ECE	Calibration intercept	Calibration slope
Logistic regression	Sigmoid	Dev OOF	0.1145	0.0363	1.0493	1.6518
Logistic regression	isotonic	Dev OOF	0.1159	0.0362	0.3207	1.2002
Tree	Sigmoid	Dev OOF	0.1189	0.0573	1.0971	1.6914
Tree	Isotonic	Dev OOF	0.1186	0.0502	-1.2984	0.2466
EBM	Sigmoid	Dev OOF	0.1140	0.0371	0.9192	1.5784
EBM	Isotonic	Dev OOF	0.1163	0.0324	0.2873	1.1943
CNAM	Sigmoid	Dev OOF	0.1136	0.0189	0.7529	1.4759
CNAM	Isotonic	Dev OOF	0.1139	0.0273	0.4481	1.2965

Abbreviations: CNAM: knowledge-constrained neural additive modeling; Dev: Development; EBM: Explainable boosting machine; ECE: Expected calibration error; OOF: Out-of-fold.

Table S4. Sensitivity of locked-test CNAM performance metrics to random seed under the fixed data split

Seed	ROC-AUC	PR-AUC	Brier score	ECE	Sensitivity (%)	Specificity (%)	Threshold
42	0.7345	0.4197	0.1141	0.0296	82.05	54.50	0.1363
52	0.7341	0.3849	0.1144	0.0236	79.49	54.50	0.1338
62	0.7384	0.4026	0.1145	0.0240	66.67	68.02	0.1646
72	0.7390	0.4093	0.1137	0.0230	76.92	60.36	0.1464
82	0.7345	0.3906	0.1142	0.0276	71.79	64.41	0.1555
Mean	0.7361	0.4014	0.1142	0.0256	75.38	60.36	0.1473
SD	0.0024	0.0140	0.0003	0.0029	6.18	5.99	0.0129

Note: Thresholds were selected using the Youden criterion for all seeds.
Abbreviations: AUC: Area under the curve; CNAM: knowledge-constrained neural additive modeling; ECE: Expected calibration error; PR: Precision–recall; ROC: Receiver operating characteristic; SD: Standard deviation.

Table S5. Missingness summary by variable in the raw analysis layer used for the review-driven sensitivity analyses

Variable	Type	Missing count	Missing %	Primary handling method
BMI	Numeric	132	5.07	Fold-wise median imputation; sensitivity: MICE
HbA1c (%)	Numeric	99	3.80	Fold-wise median imputation; sensitivity: MICE
Age at diagnosis	Numeric	45	1.73	Fold-wise median imputation; sensitivity: MICE
Age	Numeric	0	0.00	Fold-wise median imputation; sensitivity: MICE
Diabetes duration	Numeric	0	0.00	Fold-wise median imputation; sensitivity: MICE
All modeled categorical predictors	Categorical	0	0.00	Primary "Missing" category + one-hot encoding

Notes: Raw missingness was observed only for BMI, HbA1c(%), and age at diagnosis; the paper-facing train/test cross-validation scheme package is complete after preprocessing.
Abbreviations: BMI: Body mass index; HbA1c: Glycated hemoglobin; MICE: Multiple imputation by chained equations.

Table S6. Median-imputation versus MICE sensitivity on the locked test set for the evaluated baseline models

Model	Imputation	ROC-AUC	PR-AUC	Brier score	ECE
Logistic regression	Median	0.7414	0.4220	0.1138	0.0334
Logistic regression	MICE	0.7454	0.4190	0.1142	0.0360
Tree	Median	0.7092	0.3424	0.1198	0.0684
Tree	MICE	0.7129	0.3291	0.1197	0.0497
EBM	Median	0.7227	0.4125	0.1140	0.0294
EBM	MICE	0.7244	0.4037	0.1140	0.0288

Note: MICE produced only modest performance differences relative to median imputation across the evaluated baseline models.
Abbreviations: AUC: Area under the curve; EBM: Explainable boosting machine; ECE: Expected calibration error; MICE: Multiple imputation by chained equations; PR: Precision–recall; ROC: Receiver operating characteristic.

Table S7. Sensitivity analysis after excluding participants with more than 20% missing predictors

Model	*n* total	*n* excluded (>20%)	*n* retained	Retained prevalence	ROC-AUC	PR-AUC	Brier	ECE
Logistic regression	2,603	0	2,603	0.1506	0.7300	0.3991	0.1145	0.0363
Tree	2,603	0	2,603	0.1506	0.7167	0.3411	0.1189	0.0573
EBM	2,603	0	2,603	0.1506	0.7323	0.4002	0.1140	0.0371
CNAM	2,603	0	2,603	0.1506	0.7350	0.4127	0.1136	0.0189

Notes: No participants exceeded the >20% missing-predictor threshold under the current 22-predictor set; the retained cohort was therefore unchanged by construction.
Abbreviations: AUC: Area under the curve; CNAM: knowledge-constrained neural additive modeling; EBM: Explainable boosting machine; ECE: Expected calibration error; PR: Precision–recall; ROC: Receiver operating characteristic.

Table S8. Shape-stability comparison between CNAM and EBM, summarized by mean pairwise Pearson correlation across outer-fold shape curves interpolated on a common quantile grid

Feature	CNAM mean pairwise correlation	EBM mean pairwise correlation	Delta (CNAM-EBM)
Age	0.7700	0.8993	−0.1293
Age at diagnosis	0.9449	0.8353	+0.1096
BMI	0.6801	0.6819	−0.0019
Diabetes duration	0.9001	0.8639	+0.0361
HbA1c(%)	0.9751	0.9480	+0.0271

Notes: Shape stability was feature-dependent rather than uniformly superior for one model across all variables; Fold-wise curves were aligned on a common quantile grid before pairwise correlation was computed.
Abbreviations: BMI: Body mass index; CNAM: knowledge-constrained neural additive modeling; EBM: Explainable boosting machine; HbA1c: Glycated hemoglobin.

Table S9. Monotonicity-violation analysis for the evaluated constrained features

Variant	Feature	Direction	Violation rate	Mean violation rate
NAM	Age	Increase	0.0612	0.1497
NAM	Age at diagnosis	Decrease	0.2449	0.1497
NAM	Diabetes duration	Increase	0.1429	0.1497
CNAM (smooth)	Age	Increase	0.1224	0.1769
CNAM (smooth)	Age at diagnosis	Decrease	0.3061	0.1769
CNAM (smooth)	Diabetes duration	Increase	0.1020	0.1769
CNAM (smooth + mono)	Age	Increase	0.0000	0.0000
CNAM (smooth + mono)	Age at diagnosis	Decrease	0.0000	0.0000
CNAM (smooth + mono)	Diabetes duration	Increase	0.0000	0.0000

Notes: The zero-violation rate of the monotonicity-constrained variant is expected from the constrained parameterization. It supports directionally plausible shape functions but should not be interpreted as an independent empirical performance gain.
Abbreviations: CNAM: knowledge-constrained neural additive modeling; NAM: Neural additive modeling.

Table S10. Locked test calibration metrics with bootstrap confidence intervals

Model	Probability type	Brier score (95% CI)	ECE (95% CI)	Calibration intercept (95% CI)	Calibration slope (95% CI)
Logistic regression	Raw	0.1128 (0.0860–0.1391)	0.0296 (0.0222–0.0778)	0.4411 (−0.3259–1.2564)	1.2721 (0.8332–1.8347)
Logistic regression	Calibrated	0.1145 (0.0889–0.1432)	0.0363 (0.0223–0.0820)	1.0493 (−0.0143–2.2085)	1.6518 (1.0802–2.4050)
CNAM	Raw	0.2085 (0.1922–0.2251)	0.3050 (0.2645–0.3457)	−1.7552 (−2.2132–−1.4237)	1.3863 (0.8533–2.0035)
CNAM	Calibrated	0.1136 (0.0858–0.1385)	0.0189 (0.0151–0.0677)	0.7529 (−0.2007–1.7412)	1.4759 (0.9164–2.1495)

Notes: Expected calibration error was computed using 10 equal-width bins in the primary analysis; CIs were estimated using 1,000 nonparametric bootstrap resamples; Bootstrap samples containing only one class were excluded when calibration intercept and slope were estimated.
Abbreviations: CI: Confidence interval; CNAM: knowledge-constrained neural additive modeling; ECE: Expected calibration error.

Table S11. Sensitivity of expected calibration error to binning strategy

Model	Probability type	Bins	Binning strategy	ECE
Logistic regression	Raw	10	Equal width	0.0296
Logistic regression	Raw	10	Quantile equal mass	0.0446
Logistic regression	Raw	15	Equal width	0.0246
Logistic regression	Raw	15	Quantile equal mass	0.0584
Logistic regression	Calibrated	10	Equal width	0.0363
Logistic regression	Calibrated	10	Quantile equal mass	0.0488
Logistic regression	Calibrated	15	Equal width	0.0341
Logistic regression	Calibrated	15	Quantile equal mass	0.0651
CNAM	Raw	10	Equal width	0.3050
CNAM	Raw	10	Quantile equal mass	0.3021
CNAM	Raw	15	Equal width	0.3050
CNAM	Raw	15	Quantile equal mass	0.3021
CNAM	Calibrated	10	Equal width	0.0189
CNAM	Calibrated	10	Quantile equal mass	0.0361
CNAM	Calibrated	15	Equal width	0.0360
CNAM	Calibrated	15	Quantile equal mass	0.0566

Notes: The primary expected calibration error used 10 equal-width bins; Equal-mass binning and 15-bin settings are reported to show how ECE estimates vary with binning choices on the locked test set.
Abbreviations: CNAM: knowledge-constrained neural additive modeling; ECE: Expected calibration error.

Table S12. CNAM-locked test performance at alternative calibrated operating points.

Operating point	Threshold	Sensitivity	Specificity	PPV	NPV	TP/FP/TN/FN	Flagged proportion
Existing calibrated Youden	0.1301	0.8205	0.5405	0.2388	0.9449	32/102/120/7	0.5134
Existing calibrated sensitivity ≥ 0.80	0.1182	0.8462	0.4955	0.2276	0.9483	33/112/110/6	0.5556
Dev OOF calibrated sensitivity ≥ 0.90	0.0888	0.9231	0.2973	0.1875	0.9565	36/156/66/3	0.7356
Fixed 0.10	0.1000	0.8718	0.3964	0.2024	0.9462	34/134/88/5	0.6437
Fixed 0.15	0.1500	0.7692	0.6171	0.2609	0.9384	30/85/137/9	0.4406
Fixed 0.20	0.2000	0.4359	0.8288	0.3091	0.8932	17/38/184/22	0.2107

Notes: The sensitivity-oriented threshold targeting 90% sensitivity was selected using development out-of-fold calibrated probabilities and then applied unchanged to the locked test set; Fixed probability thresholds are illustrative operating points, not clinical recommendations.
Abbreviations: CNAM: knowledge-constrained neural additive modeling; FN: False negative; FP: False positive; NPV: Negative predictive value; OOF: Out-of-fold; PPV: Positive predictive value; TN: True negative; TP: True positive.

Table S13. Net benefit differences between CNAM and logistic regression at selected threshold probabilities

Threshold probability	CNAM net benefit	Logistic regression net benefit	Difference	95% CI
0.10	0.0732	0.0775	−0.0043	−0.0209–0.0102
0.15	0.0575	0.0376	0.0198	0.0032–0.0390
0.20	0.0287	0.0278	0.0010	−0.0125–0.0153
0.30	0.0197	0.0241	−0.0044	−0.0186–0.0104

Notes: Net benefit differences were computed as CNAM minus logistic regression using calibrated locked-test probabilities; CIs were estimated with 1,000 bootstrap resamples; Positive values favor CNAM.
Abbreviations: CI: Confidence interval; CNAM: knowledge-constrained neural additive modeling.

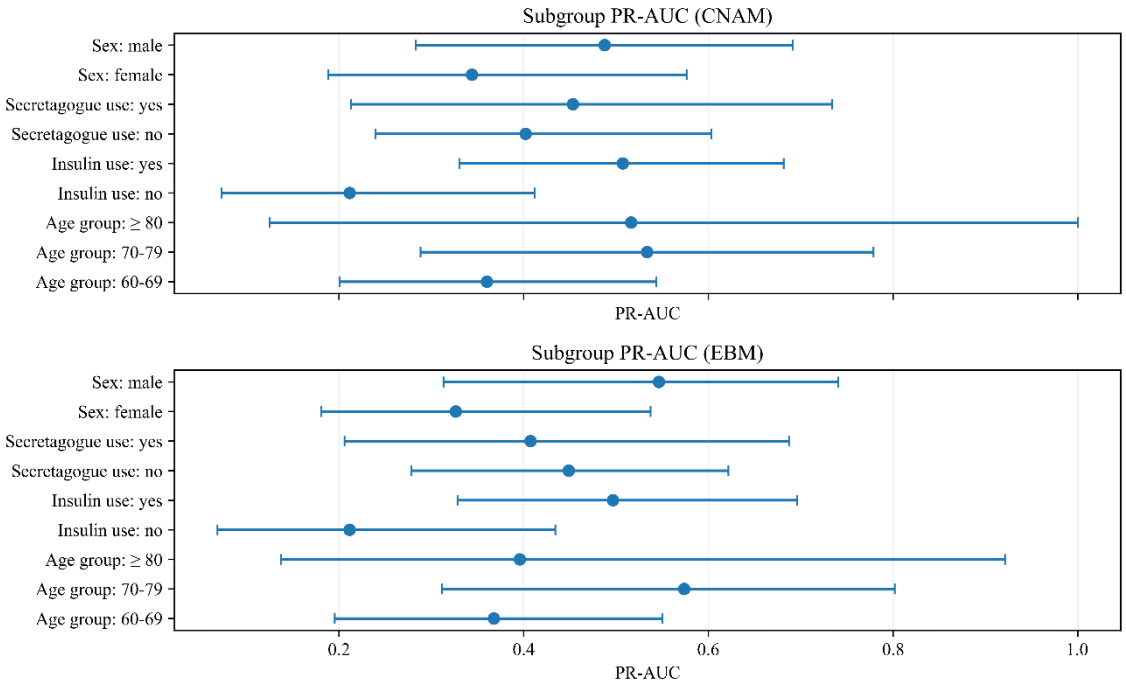


Figure S1. Subgroup PR-AUC forest plots for CNAM and EBM on the locked test set
Abbreviations: AUC: Area under the curve; CNAM: knowledge-constrained neural additive modeling; EBM: Explainable boosting machine; PR: Precision–recall.

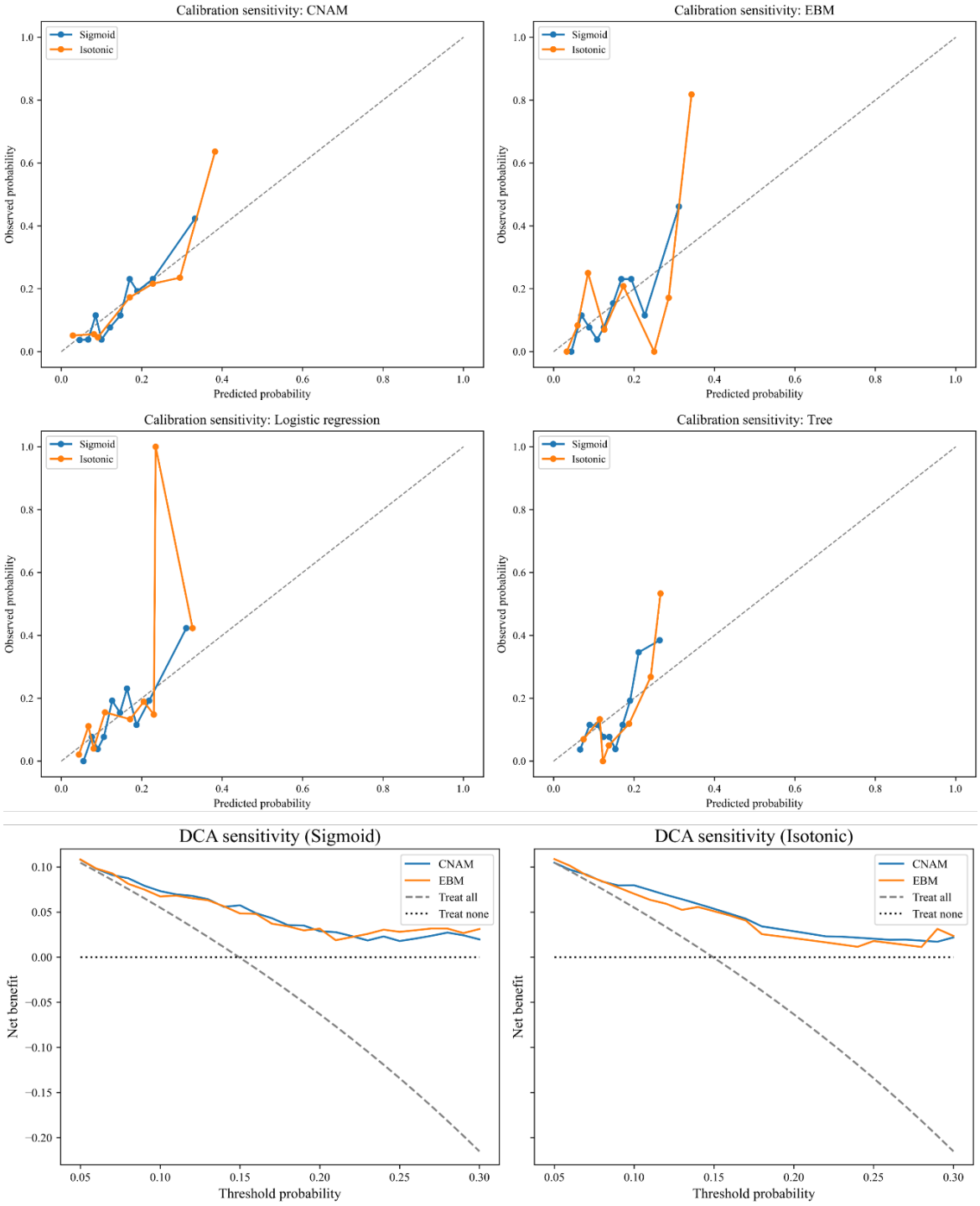
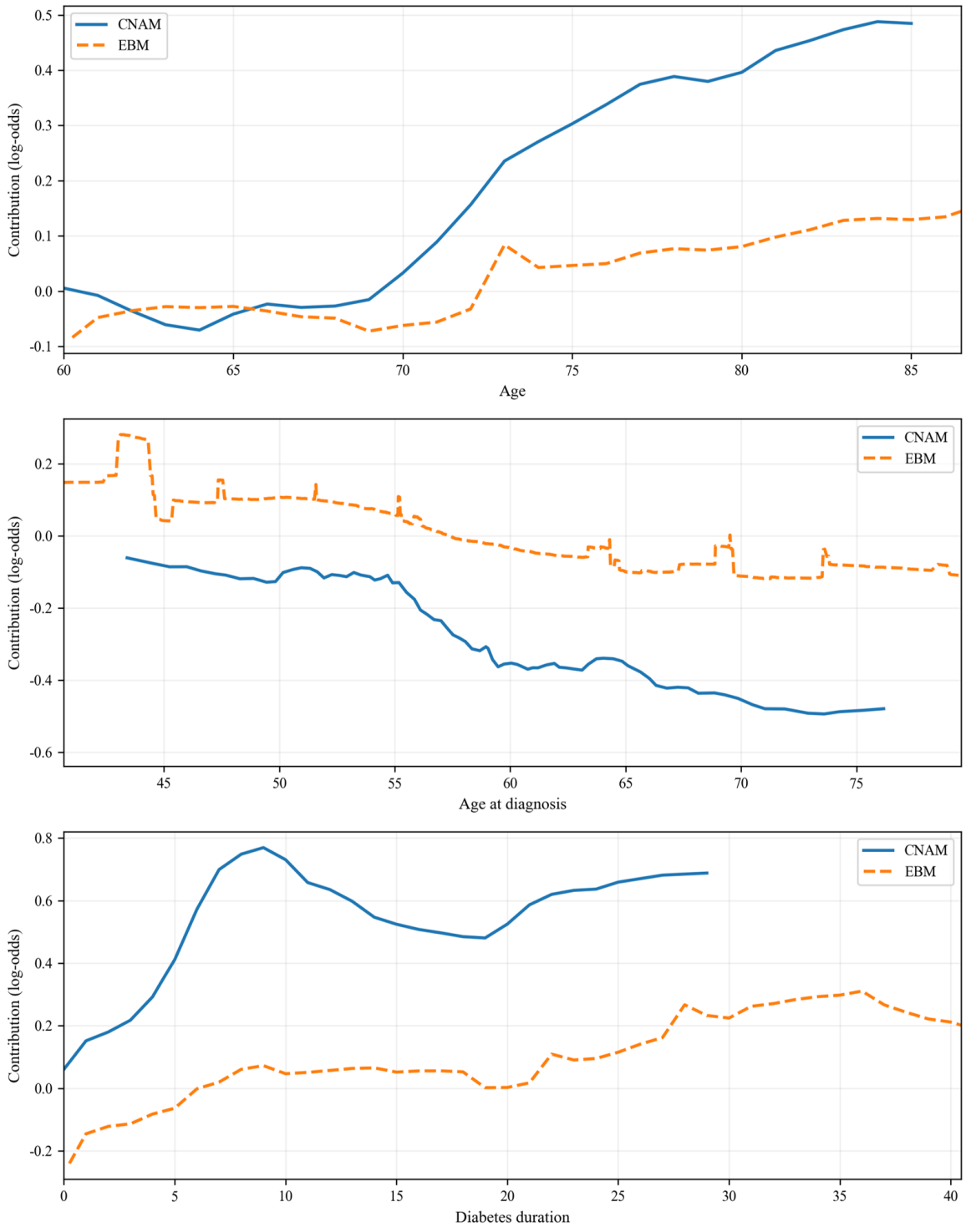


Figure S2. Calibration-metric and decision-curve sensitivity analyses. (A) Calibration-metric summary comparing sigmoid and isotonic calibration for each model, including Brier score, expected calibration error, calibration intercept, and calibration slope. Decision curve analysis under different calibration methods: (B) sigmoid; (C) isotonic.

Abbreviations: CNAM: knowledge-constrained neural additive modeling; DCA: Decision curve analysis; EBM: Explainable boosting machine.


CNAM
EBM
Contribution (log-odds)
Age
Age at diagnosis
Diabetes duration

(*Cont'd...*)

Figure S3. (***Continued***)

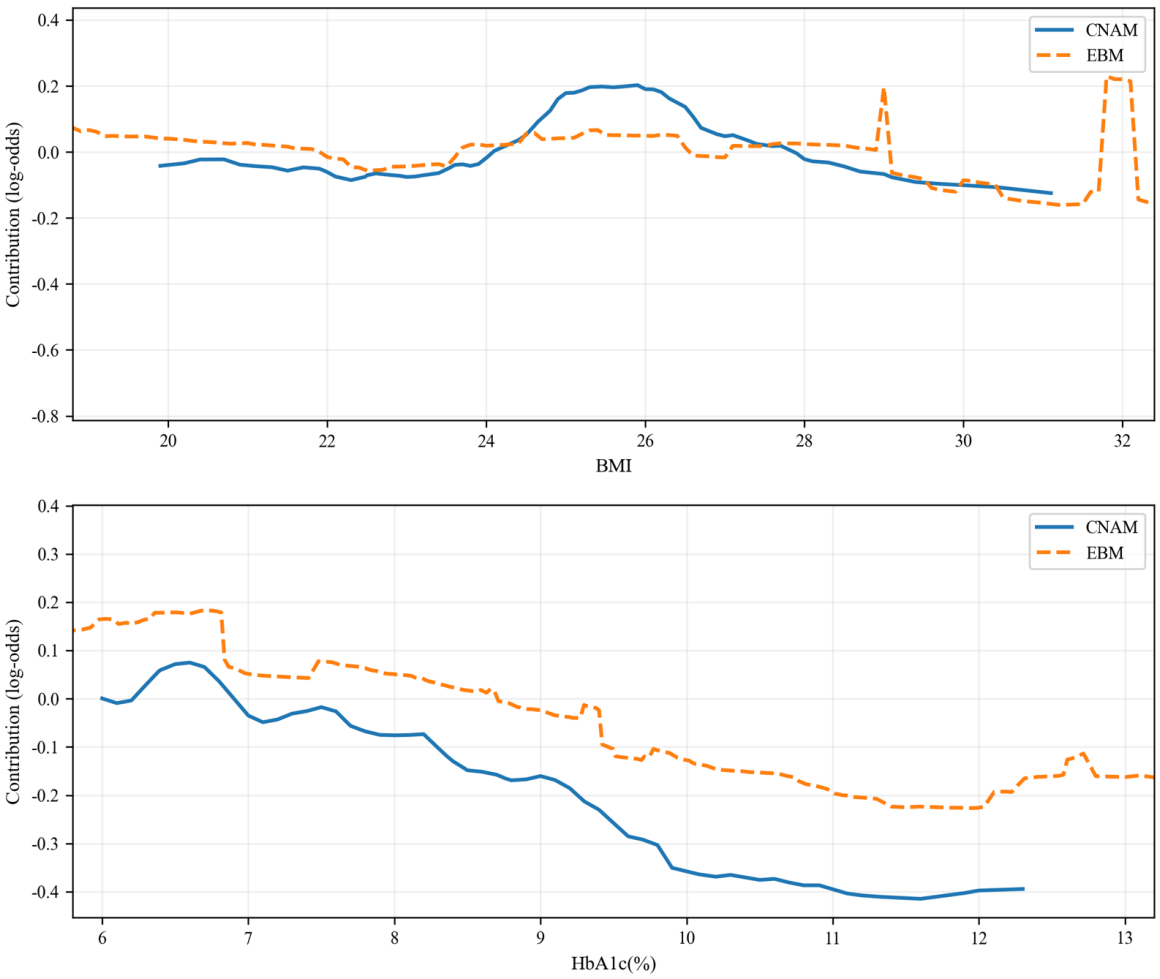


Figure S3. Comparison of global shape functions between CNAM and EBM for key numerical predictors. Panels are split across two pages for readability. Abbreviations: BMI: Body mass index; CNAM: knowledge-constrained neural additive modeling; EBM: Explainable boosting machine; HbA1c: Glycated hemoglobin.

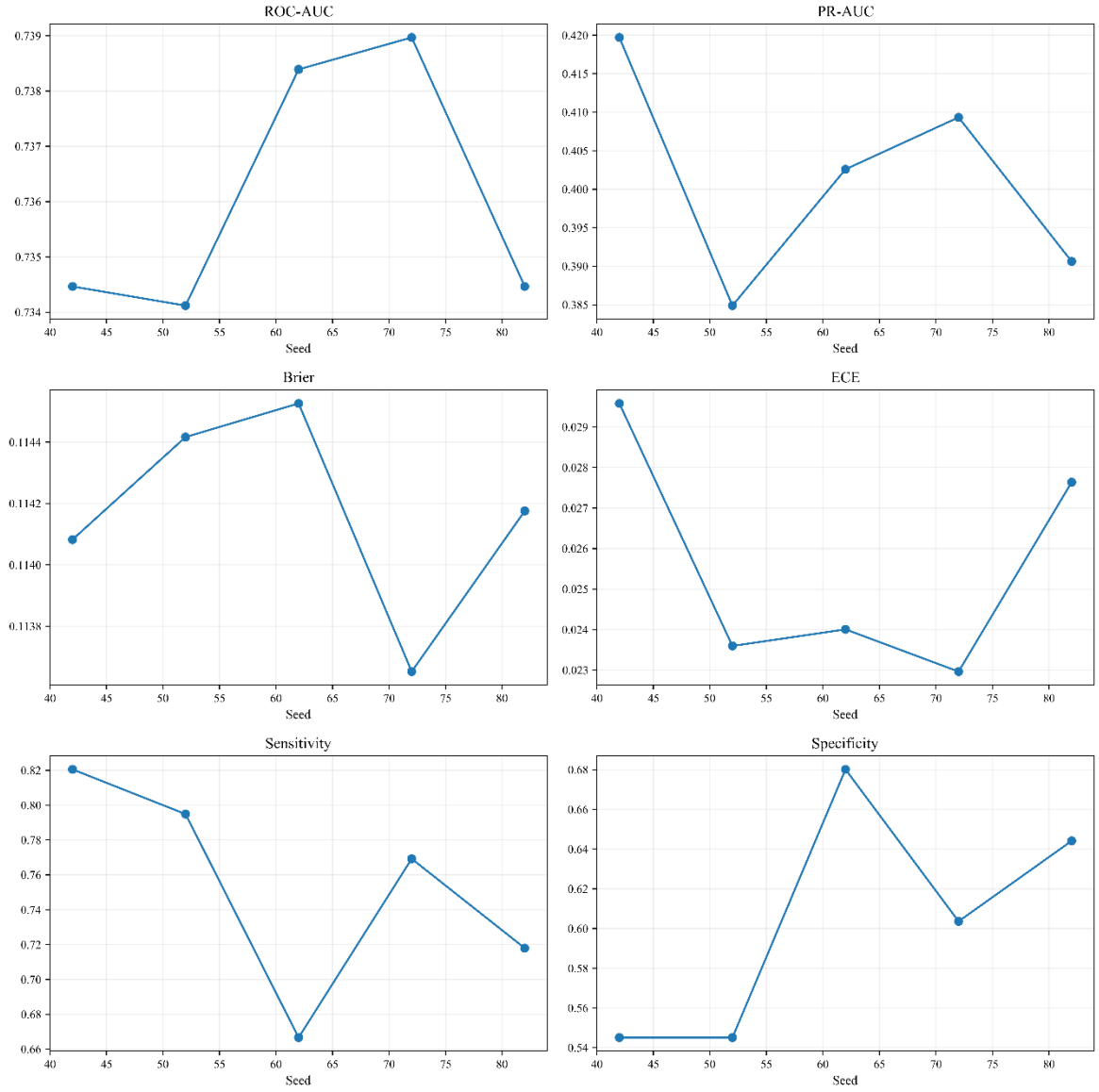


Figure S4. Sensitivity of locked-test knowledge-constrained neural additive modeling performance metrics to random seed under the fixed data split. Abbreviations: AUC: Area under the curve; ECE: Expected calibration error; PR: Precision–recall; ROC: Receiver operating characteristic.

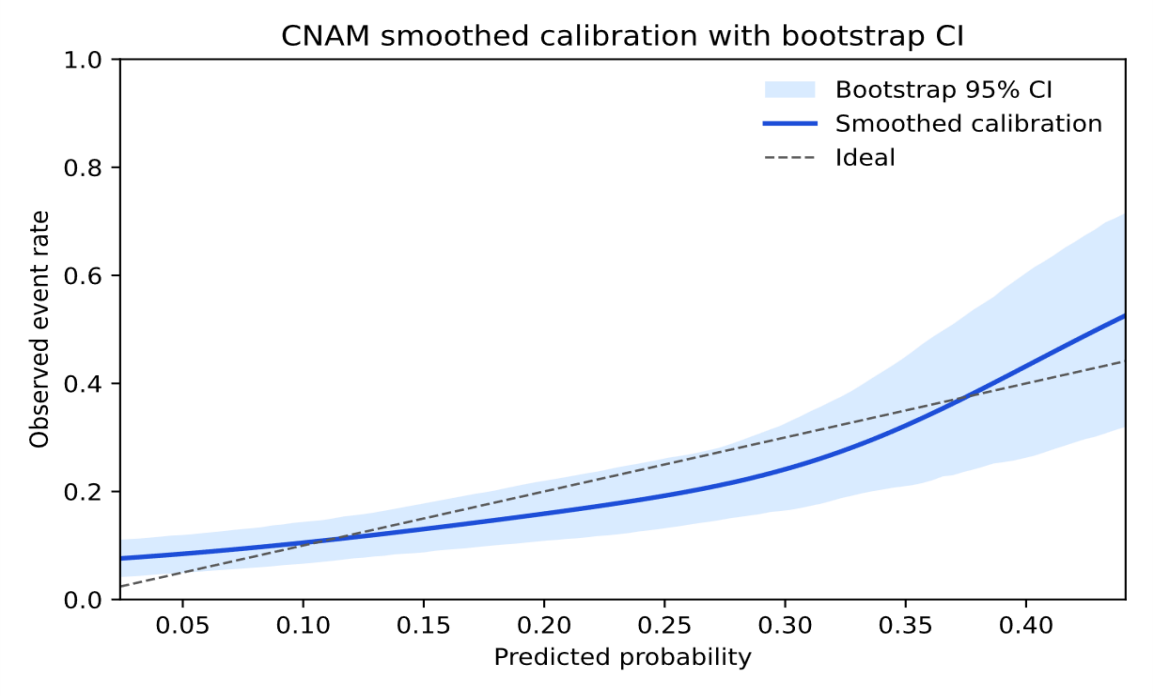


Figure S5. Smoothed CNAM calibration curve with bootstrap confidence bands. Smoothed calibration curve for calibrated CNAM probabilities on the locked test set. The curve was estimated with a Gaussian kernel smoother over the observed predicted-probability range, and pointwise confidence bands were estimated by nonparametric bootstrap resampling. This figure complements the binned calibration plot in the main manuscript.
Abbreviations: CI: Confidence interval; CNAM: knowledge-constrained neural additive modeling.

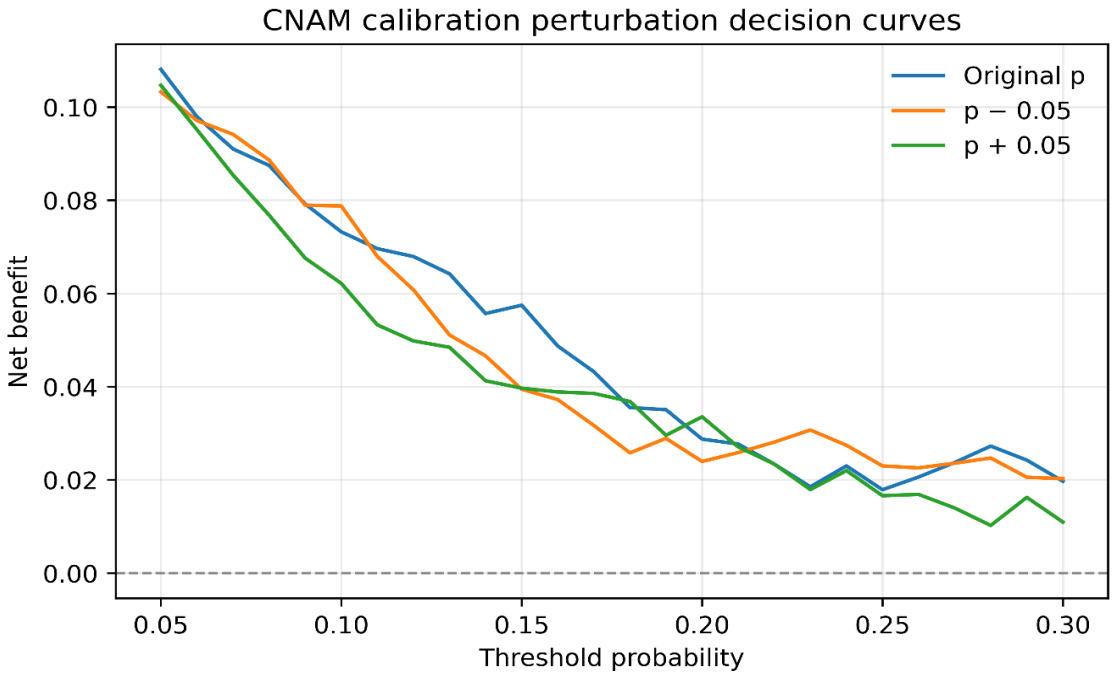


Figure S6. Calibration perturbation sensitivity analysis for CNAM decision curves. Decision curve sensitivity under calibration perturbation for CNAM. The original calibrated CNAM probabilities were compared with shifted versions by −0.05 and +0.05, and with clipped versions in [0, 1]. This analysis tests the sensitivity of net benefit to simple probability-scale perturbations and does not involve model refitting or recalibration.
Abbreviations: BMI: Body mass index; CNAM: knowledge-constrained neural additive modeling; HbA1c: Glycated hemoglobin.