

The Prognostic Value of Early Treatment Factors for Cognitive Behavioral Therapy for Insomnia (CBT-I) Outcomes

Yanyu Hu¹, Zhiyong Ni¹, Shiyu Wang¹, Shuchuan Zheng¹, Nana Weng¹, Peng Lai¹, Shujuan Huang¹, Xuan Wei^{1,2}

¹Sleep Medicine Center, Xiamen Xianyue Hospital, Xianyue Hospital Affiliated with Xiamen Medical College, Fujian Psychiatric Center, Fujian Clinical Research Center for Mental Disorders, Xiamen, Fujian, 361012, People's Republic of China; ²The Graduate School of Fujian Medical University, Fuzhou, Fujian, 350000, People's Republic of China

Correspondence: Xuan Wei, Xiamen Xianyue Hospital, Xiamen, Fujian, 361012, People's Republic of China, Tel +86 18950089893, Email sleepxm93@163.com

Background: Cognitive behavioral therapy for insomnia (CBT-I) is the first-line treatment for chronic insomnia, yet a significant proportion of patients fail to achieve favorable outcomes. Baseline-only prognostic models show limited predictive performance, and the added value of early treatment variables remains unclear. This retrospective study aimed to investigate this issue.

Methods: We retrospectively included 1059 patients who completed face-to-face group CBT-I. Post-treatment outcomes were assessed using the Insomnia Severity Index (ISI), and 12-month follow-up outcomes were evaluated using the Pittsburgh Sleep Quality Index (PSQI). The dataset was randomly split into training (70%) and test (30%) sets. Candidate predictors were selected using Least Absolute Shrinkage and Selection Operator, Random Forest, and eXtreme Gradient Boosting. Variables consistently identified across all three algorithms were entered into multivariable logistic regression models. Predictive performance of baseline-only and combined models (baseline plus early treatment variables) was evaluated using the area under the receiver operating characteristic curve (AUC) and Brier Score, with bootstrap resampling used for model comparison.

Results: Unfavorable outcomes were observed in 32.4% of patients at post-treatment (neither ISI < 8 nor a reduction > 7) and in 44.5% at the 12-month follow-up (PSQI > 5). The combined model significantly outperformed the baseline-only model in predicting post-treatment outcomes in the test set (AUC: 0.741 vs 0.630, $p < 0.001$), with week 2 ISI change identified as the strongest predictor. For the 12-month follow-up, the combined model demonstrated a non-significantly higher AUC compared with the baseline-only model (0.742 vs 0.735, $p > 0.05$), with age emerging as the strongest predictor.

Conclusion: Incorporating early treatment variables, particularly Week 2 ISI score reduction, may provide incremental value in predicting short-term CBT-I outcomes. However, the added value of early treatment variables for long-term outcomes appeared limited.

Keywords: cognitive behavioral therapy for insomnia, treatment outcome, prognostic modeling, personalized treatment, early treatment response

Background

Insomnia is a prevalent clinical condition affecting approximately 10–15% of the population.¹ Beyond sleep disturbances, this disorder imposes a multifaceted burden on patients, including reduced quality of life, psychological distress, and impaired physical functioning. Cognitive behavioral therapy for insomnia (CBT-I) is a structured multicomponent intervention that targets the maladaptive behaviors and cognitions which perpetuate insomnia. Clinical guidelines consistently designate CBT-I as the first-line treatment for insomnia.^{2,3} Substantial evidence from meta-analyses demonstrates the significant efficacy of CBT-I in alleviating insomnia symptoms and improving sleep quality.^{3,4} Moreover, these benefits are durable, with long-term effects maintained after treatment discontinuation.^{5,6}

However, despite its established efficacy, approximately 40% of patients fail to achieve a clinically meaningful response to standard face-to-face CBT-I.⁷ This heterogeneity in treatment response highlights the critical need for

predictive tools to identify individuals at risk of suboptimal outcomes prior to or early in the treatment course. Accurate prognostic models could facilitate personalized treatment strategies, such as treatment intensification or early adaptation, thereby improving overall clinical effectiveness.

Several studies have attempted to predict CBT-I outcomes using baseline demographic, clinical, and sleep-related variables.⁸ For short-term outcomes, baseline insomnia severity has been identified as an important predictor in most prognostic models.^{9–11} Additionally, some models found that sleep-related beliefs and baseline anxiety and depression symptoms may also contribute to predicting treatment outcomes.^{9,12} For long-term outcomes, baseline insomnia severity and age have also been reported as predictors in some studies.^{13,14} However, the clinical utility of existing baseline-only prognostic models remains limited, with reported discrimination frequently below acceptable thresholds (area under the receiver operating characteristic curve [AUC] < 0.70 in test samples).¹¹ These findings suggest that pretreatment characteristics alone may be insufficient to capture the dynamic processes that unfold during CBT-I and influence treatment response.

In recent years, there has been a growing recognition that personalized care should not be limited to pretreatment matching, but should also include dynamic treatment adjustments based on ongoing response indicators.¹⁵ First, these early treatment indicators may help identify a critical clinical decision window, during which timely adjustments to the intervention can be taken. Second, their inclusion may contribute to the predictive performance of the models. In psychotherapy research, early symptom change has consistently been shown to predict treatment outcomes across a range of disorders, including depression, anxiety, and personality disorders.^{16–18} Early improvement may reflect patients' positive characteristics, such as higher therapeutic engagement and better behavioral adherence, which are associated with favorable outcomes. Moreover, early response might promote neurocognitive adaptation, including reduced disorder-related anxiety and enhanced self-efficacy, thereby establishing a positive feedback loop that sustains treatment success.

Within the context of CBT-I, however, the prognostic value of early treatment variables remains largely unexplored. It is unclear whether integrating early treatment factors with baseline characteristics improves outcome prediction, and which early indicators are most informative. Moreover, little is known about whether early treatment variables differentially predict short-term versus long-term outcomes following CBT-I.

In the present study, we sought to address these gaps by developing prognostic models for post-treatment and 12-month follow-up outcomes of CBT-I that integrate both baseline and early treatment variables. By comparing combined models incorporating early treatment variables with baseline-only models, we aimed to evaluate the added prognostic value of early treatment factors and to identify key predictors emerging during the early phase of CBT-I.

Method

Participants

In this retrospective study, we reviewed clinical records of patients who received CBT-I at the Sleep Medicine Center of Xiamen Xianyue Hospital between September 2015 and May 2022. A total of 1274 patients who underwent face-to-face, group-based CBT-I were initially identified from the database. Inclusion criteria were: (1) diagnosis of chronic insomnia disorder based on the International Classification of Sleep Disorders, third edition criteria (ICSD-3); (2) a baseline Insomnia Severity Index (ISI) score ≥ 8 ; (3) age ≥ 18 . Exclusion criteria were: (1) non-attendance of the final CBT-I session; (2) missing ISI data at post-treatment. A detailed participant flowchart is presented in [Figure 1](#).

Procedure of CBT-I

All patients underwent a formal psychiatric evaluation and were diagnosed with insomnia according to the ICSD-3 criteria by a psychiatrist or sleep physician prior to receiving the therapy. Standard face-to-face group CBT-I was delivered by therapists with specialized training in behavioral sleep medicine. The main components of therapy include sleep restriction therapy (SRT), stimulus control therapy (SCT), cognitive therapy, sleep hygiene, relaxation therapy.⁷ Treatment consisted of five weekly sessions (180 min each) with between four and seven participants per group. Baseline assessments were conducted one week prior to the treatment using standardized questionnaires, sleep diaries, and

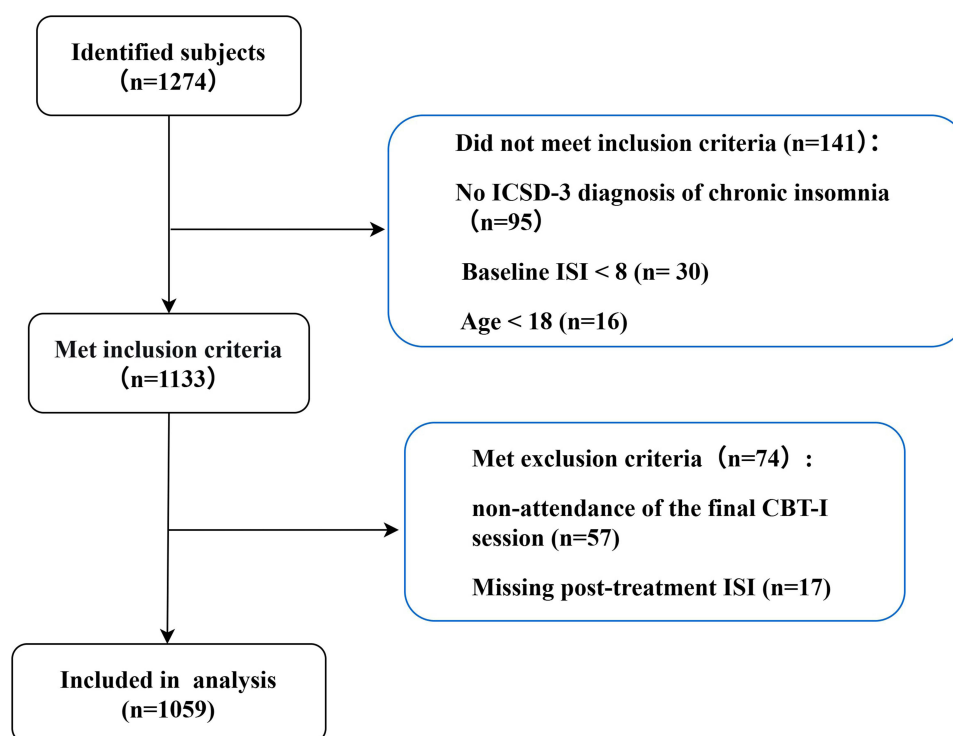


Figure 1 Flowchart of the study participants.

actigraphy. SCT and SRT were started in week 1, while cognitive therapy started in week 3. Detailed CBT-I procedure as delivered in our sleep center is described in [Supplementary Table S1](#).

Data Collection

All data were obtained from clinical treatment records. Variables were collected from two time points: (1) Baseline variables, including demographics, clinical characteristics, psychometric measures and sleep parameters assessed before treatment; and (2) Early treatment variables, which included psychometric measures and sleep parameters (see Psychometric Measures for details).

Early treatment variables were defined as those assessed at Week 2. This time point was selected to balance predictive accuracy with the availability of sufficient remaining treatment sessions for potential clinical adaptation. Week 2 falls within the first half of the 5-session treatment protocol, consistent with prior psychotherapy research indicating that early symptom change during the first half of treatment provides clinically meaningful prognostic information.¹⁷

In total, 45 candidate predictors were included in the analysis, comprising 24 baseline variables and 21 early treatment variables. A complete list of all candidate predictors is provided in [Table 1](#). The pool of these candidate variables was selected based on extensive literature review and clinical expertise.

Demographics and Clinical Characteristics

As shown in [Table 1](#), candidate predictors for demographic and clinical variables included sex, age, educational attainment, employment status, marital status, body mass index (BMI), duration of insomnia, tea or coffee consumption, and pre-treatment medication use. Employment status was categorized as “Employed” or “Other”. The “Other” category mainly comprised participants who were unemployed, retired, students, or homemakers. Pre-treatment medication use was thoroughly documented, with a focus on two drug classes: (a) hypnotics, primarily including benzodiazepines (BZDs), non-benzodiazepine receptor agonists (non-BZDs), sedating antidepressants, sedating antipsychotics, and melatonin receptor agonists, among others; and (b) non-sedating antidepressants/anxiolytics, primarily including selective serotonin reuptake Inhibitors (SSRIs), serotonin–norepinephrine reuptake Inhibitors (SNRIs), and 5-HT1A receptor agonists, among others.

Table 1 Comparison of candidate variables between the two post-treatment outcome groups

	Missing (%)	Total (n = 1059)	Favorable Outcome (n = 716)	Unfavorable Outcome (n = 343)	p-value
Baseline variables					
Sex, n (%)	0%				0.067
Female		694 (66)	483 (67)	211 (62)	
Male		365 (34)	233 (33)	132 (38)	
Age, M (P25, P75)	0.1%	41 (33, 50)	40 (33, 49)	43 (34, 52)	0.002
Educational attainment, n (%)	0.8%				0.595
Illiterate or Primary school		68 (6)	42 (5)	26 (7)	
Junior high school		192 (17)	130 (17)	62 (17)	
Senior high school		449 (39)	305 (39)	144 (39)	
College or above		444 (38)	309 (39)	135 (37)	
employment status, n (%)	0.8%				0.001
Employed		666 (63)	474 (67)	192 (56)	
Other		391 (37)	234 (33)	151 (44)	
Married, n (%)	0.4%	875 (83)	587 (82)	288 (84)	0.598
BMI, M (P25, P75)	7.1%	21.5 (19.7, 23.3)	21.5 (19.7, 23.4)	21.4 (19.7, 23.3)	0.930
Duration of insomnia (month), M (P25, P75)	7.5%	36 (12, 120)	36 (12, 120)	36 (12, 120)	0.803
Tea/coffee Consumption, n (%)	0%	245 (23)	179 (25)	66 (19)	0.045
Hypnotic use, n (%)	0%	727 (69)	475 (66)	252 (74)	0.021
Antidepressant/ anxiolytic use, n (%)	0%	238 (23)	160 (22)	78 (23)	0.938
Baseline psychometric measures					
ISI, M (P25, P75)	0%	19 (15, 22)	20 (16, 23)	17 (15, 20)	< 0.001
BDI, M (P25, P75)	0.7%	13 (8, 19)	13 (8, 19)	13 (8, 19)	0.846
BAI, M (P25, P75)	1.0%	29 (25, 35)	29 (25, 36)	29 (25, 34)	0.360
FFS, M (P25, P75)	0.6%	13 (9, 17)	13 (9, 17)	13 (9, 16)	0.835
DBAS, M (P25, P75)	14.4%	84 (76, 90)	84 (76, 90)	84 (76, 91)	0.746
Baseline sleep diary					
TST (min), M (P25, P75)	1.9%	369.3 (308.4, 417.1)	375.7 (318.6, 420.4)	357.1 (294.3, 406.0)	0.002
SE (%), M (P25, P75)	1.8%	75.5 (64.8, 83.6)	76.4 (66.4, 84.2)	73.0 (6, 82.1)	0.002
SOL (min), M (P25, P75)	1.5%	42.0 (27.1, 65.4)	42.1 (27.9, 64.3)	41.6 (27.1, 65.7)	0.741
WASO (min), M (P25, P75)	2.2%	22.0 (6.9, 49.0)	20.0 (5.7, 43.0)	29.0 (9.5, 52.0)	0.006
TIB (min), M (P25, P75)	1.6%	494.2 (461.4, 531.2)	495.0 (461.9, 533.2)	493.3 (461.1, 526.0)	0.331
Baseline actigraphy					
TST (min), M (P25, P75)	2.7%	414.4 (384.8, 450.1)	413.7 (386.0, 450.1)	415.9 (382.2, 449.3)	0.835
SE (%), M (P25, P75)	2.9%	84.0 (80.0, 87.6)	83.8 (80.0, 87.6)	84.3 (80.0, 87.8)	0.705
SOL (min), M (P25, P75)	2.7%	14.2 (8.6, 24.2)	14.3 (8.6, 25.0)	13.9 (8.1, 22.7)	0.407
TIB (min), M (P25, P75)	2.8%	498.8 (466.9, 532.1)	498.6 (467.9, 534.2)	499.4 (464.3, 528.5)	0.608
Early treatment variables					
Week 2 psychometric measures					
ISI, M (P25, P75)	0.8%	13 (9, 16)	12 (9, 15)	14 (11, 18)	< 0.001
ISI change, M (P25, P75)	0.8%	-6 (-9, -2)	-7 (-10, -4)	-3 (-5, 0)	< 0.001

(Continued)

Table 1 (Continued).

	Missing (%)	Total (n = 1059)	Favorable Outcome (n = 716)	Unfavorable Outcome (n = 343)	p-value
Week 2 sleep diary					
TST (min), M (P25, P75)	1.3%	323.6 (284.3, 360.9)	327.9 (289.3, 363.7)	312.1 (271.4, 353.0)	<0.001
SOL (min), M (P25,P75)	1.4%	27.1 (17.3, 42.1)	27.1 (17.1, 40.0)	28.6 (18.6, 47.1)	0.016
SE (%), M (P25, P75)	1.5%	83.5 (75.3, 89.5)	84.9 (77.1, 89.9)	80.5 (72.6, 88.0)	<0.001
WASO (min), M (P25, P75)	1.9%	13.0 (3.1, 29.0)	11.0 (2.9, 27.8)	16.0 (4.1, 34.0)	0.002
TIB (min), M (P25, P75)	1.4%	390.3 (361.0, 422.2)	391.6 (363.0, 423.6)	390.0 (360.0, 421.1)	0.271
Week 2 actigraphy					
TST (min), M (P25, P75)	2.1%	343.7 (314.7, 370.9)	344.4 (317.0, 371.4)	340.3 (311.4, 368.9)	0.130
SE (%), M (P25, P75)	2.2%	86.6 (82.6, 89.9)	86.9 (82.8, 90.1)	86.2 (82.1, 89.4)	0.063
SOL (min), M (P25, P75)	2.1%	6.7 (3.9, 12.1)	6.6 (3.9, 11.7)	6.9 (4.1, 13.1)	0.285
TIB (min), M (P25, P75)	2.1%	397.9 (370.6, 427.8)	398.1 (370.7, 427.4)	395.7 (369.3, 427.9)	0.481
Week 2 sleep diary change					
TST change, M (P25, P75)	3.1%	-44.4 (-85.4, -3.7)	-43.0 (-87.1, -4.6)	-44.6 (-82.4, -2.9)	0.955
SE change, M (P25, P75)	3.3%	6.9 (0.9, 15.3)	7.0 (1.7, 15.2)	6.5 (-0.4, 15.9)	0.235
SOL change, M (P25, P75)	2.9%	-12.9 (-31.4, -1.4)	-14.3 (-32.3, -2.9)	-10.0 (-27.9, 0.8)	0.024
WASO change, M (P25, P75)	3.4%	-6.4 (-23.6, 1.8)	-5.7 (-22.0, 1.3)	-9.0 (-28.0, 2.8)	0.497
TIB change, M (P25, P75)	3.0%	-102.3 (-135.9, -70.1)	-101.4 (-136.0, -70.1)	-104.3 (-135.0, -69.3)	0.887
Week2 actigraphy change					
TST change, M (P25, P75)	4.8%	-71.0 (-100.3, -43.7)	-69.9 (-98.1, -41.9)	-75.6 (-108.1, -45.5)	0.052
SE change, M (P25, P75)	4.9%	2.3 (-0.3, 5.5)	2.5 (0.0, 5.7)	1.6 (-0.8, 4.5)	0.004
SOL change, M (P25, P75)	4.7%	-6.2 (-15.4, -0.9)	-6.6 (-16.1, -1.3)	-5.8 (-13.6, -0.4)	0.159
TIB change, Mean \pm SD	4.7%	-99.94 \pm 54.05	-99.34 \pm 53.38	-101.25 \pm 55.53	0.606
Week2 TIB difference (min), M (P25, P75)	2.1%	28.7 (11.7, 58.1)	28.0 (12.1, 55.3)	29.4 (10.2, 61.9)	0.806

Abbreviations: M (P25, P75), median (25th percentile, 75th percentile); BMI, body mass index; BDI, Beck Depression Inventory; BAI, Beck Anxiety Inventory; FFS, Flinders Fatigue Scale; ISI, Insomnia Severity Index; DBAS, Dysfunctional Beliefs and Attitudes about Sleep Scale; TST, total sleep time; SE, sleep efficiency; SOL, sleep onset latency; WASO: wake after sleep onset; TIB, time in bed.

Psychometric Measures

Baseline Psychometric measures included ISI, Beck Depression Inventory (BDI), Beck Anxiety Inventory (BAI), Flinders Fatigue Scale (FFS), and Dysfunctional Beliefs and Attitudes about Sleep Scale (DBAS). Psychometric measures at week 2 included only the ISI.

The ISI is a 7-item self-report questionnaire that assesses the severity and impact of insomnia symptoms over the past two weeks; scores range from 0 to 28, with higher scores indicating greater severity.¹⁹ The BDI and BAI are both 21-item questionnaires designed to assess the intensity of depressive and anxiety symptoms over the past week, respectively.^{20,21} The FFS (7 items) measures insomnia-related daytime fatigue over the past two weeks,²² and the DBAS (30 items) evaluates sleep-related beliefs and attitudes.²³

Sleep Parameters

Subjective sleep pattern was assessed using a Chinese-language sleep diary completed daily by participants. The following five items from the sleep diary were used to derive the subjective sleep parameters: (1) "What time did you turn off the lights?" (lights-off time); (2) "How many minutes did it take you to fall asleep?" (sleep onset latency, SOL); (3) "What was your total sleep time last night?" (total sleep time, TST); (4) "What time did you finally get out of bed?"

(get-up time); (5) “What was the total time you spent awake during the night, from first falling asleep until the final awakening?” (wake after sleep onset, WASO). From these items, time in bed (TIB) was calculated as the interval between lights-off time and get-up time, and sleep efficiency (SE) was calculated as $(TST/TIB) \times 100\%$. Weekly mean values were then computed for SOL, TST, WASO, TIB, and SE at baseline and week 2.

Objective sleep parameters were assessed using actigraphy. Portable devices provide a practical approach for evaluating sleep disorders,²⁴ and previous studies have supported the validity of actigraphy for sleep measurement.²⁵ Weekly averages of SOL, TST, TIB, and SE were obtained from actigraphy. Changes in sleep parameters were calculated as the difference between Week 2 and baseline values (Week 2 minus baseline).

Adherence

Following the method described by Riedel et al,²⁶ we quantified adherence to SRT using the weekly TIB difference (calculated as mean actigraphy-measured TIB minus prescribed TIB). A smaller difference indicated better adherence. Adherence calculated at week 2 was also included as early treatment variable.

Outcome Measures

Post-treatment outcomes were assessed using the ISI. Treatment response was defined as a reduction of > 7 points from baseline, and remission was defined as a post-treatment ISI score < 8 .¹⁹ Patients who failed to meet either criterion were classified as having an unfavorable post-treatment outcome.

Long-term outcomes were assessed at the 12-month follow-up using the Pittsburgh Sleep Quality Index (PSQI), which evaluates sleep quality and disturbances over the past month.²⁷ A global PSQI score > 5 was categorized as an unfavorable outcome.

Statistical Analysis

Clinical characteristics were compared between patients with favorable and unfavorable outcomes at post-treatment. Continuous data following a normal distribution were presented as the mean \pm standard deviation (SD); otherwise, they were expressed as the median with interquartile range, shown as M (P25, P75). Comparisons were conducted using Wilcoxon rank-sum tests for continuous variables and χ^2 -tests for categorical variables. All analyses were performed using R (version 4.4.2). A two-tailed p-value < 0.05 was considered statistically significant.

Prognostic Model Development

The dataset was randomly divided into a training set (70%) for model development and a test set (30%) for validation. Prior to analysis, missing values were imputed using the *missForest* R package. The proportion of missing data for each variable is reported in Table 1. Missing data pattern analysis showed that most missing values were isolated to single variables (200 cases, 58.5%), with no systematic co-occurrence between variables.

Following a validated hybrid strategy,²⁸ machine learning methods were used for feature selection only, while multivariable logistic regression was employed for final model construction to maintain interpretability and clinical applicability.

Feature Selection

Three machine learning algorithms—Least Absolute Shrinkage and Selection Operator (LASSO) regression with 10-fold cross-validation (10-fold CV), Random Forest (RF), and eXtreme Gradient Boosting (XGBoost)—were applied for preliminary feature ranking. Variables were ranked according to their importance in RF and XGBoost models, and by the absolute value of regression coefficients in LASSO. A consensus ranking strategy was then used to identify stable predictors, defined as variables consistently appearing among the top N ranked features across all three algorithms.

The optimal value of N was determined using a sensitivity analysis. Specifically, for each candidate value of N, multivariable logistic regression models were constructed using the corresponding top N consensus features, and predictive performance was evaluated using repeated 10-fold CV. The optimal feature set size was defined as the smallest N at which the mean area under the receiver operating characteristic curve (AUC) reached a plateau, indicating

no meaningful improvement with increasing feature number. The results of the sensitivity analysis are presented in [Supplementary Figure S1](#).

Multicollinearity among selected features was assessed using variance inflation factors (VIF), and no significant collinearity was observed. All feature selection procedures were performed exclusively in the training set.

Model Construction and Evaluation

Separate multivariable logistic regression models were constructed for post-treatment and 12-month follow-up outcomes using the selected predictors. To evaluate the added predictive value of early treatment variables, two models were developed for each outcome: a baseline model including only baseline variables, and a combined model additionally incorporating early treatment variables. For clinical interpretability, final models were visualized using nomograms.

Model discrimination was evaluated using the AUC, and calibration was assessed using the Brier Score (BS) and calibration plots. Model performance was evaluated using 10-fold cross-validation with 10 repeats for internal validation, and an independent hold-out test set was subsequently used to assess the final model performance. To formally compare the performance of baseline-only and combined models, bootstrap resampling with 1000 iterations was used to estimate the distribution of AUC differences in both training and test sets.

Given the potential for bias due to loss at the 12-month follow-up, we conducted a sensitivity analysis assess the robustness of our findings for the long-term outcome model. We performed best-case and worst-case scenario imputations for missing outcomes and re-evaluated the model's performance.

Results

General Characteristics

Of the 1274 patients initially identified, 1133 met the inclusion criteria. After excluding 74 patients who met the exclusion criteria, a total of 1059 patients were included in the final analysis. The 12-months follow-up completion rate was 75.4% ($n = 799$).

At post-treatment, 343 patients (32.4%) were classified as having an unfavorable outcome (neither response nor remission). At the 12-month follow-up, 355 patients (44.5%) had an unfavorable outcome ($PSQI > 5$). Of the 1059 patients included in the final analysis, 1006 (95.0%) completed all five treatment sessions, 49 (4.6%) attended four sessions, and 4 (0.4%) attended fewer than four sessions. There were no significant differences in treatment completion rates across outcome groups (all $p > 0.05$).

Comparison of candidate variables between the two post-treatment outcome groups is shown in [Table 1](#). Compared with patients who achieved a favorable outcome, those with an unfavorable outcome were older, less likely to be employed, and had higher rates of hypnotic use and lower rates of tea/coffee consumption (all $p < 0.05$). No significant between-group differences were observed in sex, educational attainment, marital status, duration of insomnia, BMI and rates of antidepressant/anxiolytic use (all $p > 0.05$).

Regarding baseline psychometric and sleep characteristics, the unfavorable outcome group had lower baseline ISI scores than the favorable outcome group (median, 17 vs 20; $p < 0.001$). No significant differences were observed between the two groups on the BDI, BAI, FFS, or DBAS (all $p > 0.05$). Sleep diary data indicated poorer subjective sleep at baseline in the unfavorable group, characterized by shorter TST, longer WASO, and lower SE (all $p < 0.05$). In contrast, no significant between-group differences were observed in objective actigraphy-derived sleep parameters (all $p > 0.05$).

At Week 2, compared to patients with favorable outcomes, those with unfavorable outcomes exhibited a markedly smaller reduction in ISI score from baseline (median change: -3 vs -7 , $p < 0.001$), and had higher absolute ISI scores at the Week 2 assessment (median: 14 vs 12, $p < 0.001$). Sleep diary data at Week 2 consistently indicated poorer subjective sleep in the unfavorable outcome group, characterized by longer SOL and WASO, shorter TST, and lower SE (all $p < 0.05$). Reductions in SOL from baseline to Week 2 were significantly smaller in the unfavorable group (median change: -10.0 vs -14.3 ; $p = 0.024$). In contrast, actigraphy-derived sleep parameters at Week 2 showed minimal between-group differences; only the change in SE differed significantly, with a lower change observed in the unfavorable outcome group (median change: 1.6% vs 2.5%; $p = 0.004$). The TIB difference at Week 2 did not differ significantly between groups.

In exploratory analyses comparing sleep diary and actigraphy parameters, actigraphy yielded longer TST, shorter SOL, and higher SE than sleep diary at baseline and Week 2 (see [Supplementary Table S2](#)).

Feature Selection and Model Construction For Post-Treatment Outcome

Sensitivity analysis for feature set size determination (Supplementary Figure S1) showed that the AUC plateaued when consensus features were selected as those common to the top 8 features of LASSO, Random Forest, and XGBoost for the baseline-only model and top 12 features for the combined model. Using this threshold, six consensus features were identified for the baseline-only model (Baseline ISI score; Baseline FFS score; Baseline sleep diary SE; Age; Baseline sleep diary TST; Duration of insomnia) (Figure 2A), and four for the combined model (Week 2 ISI change; Week 2 TIB difference; Week 2 actigraphy TST change; Baseline ISI score) (Figure 2B).

Using the selected features, multivariable logistic regression models were constructed and presented as nomograms. The nomogram for the baseline-only model (Figure 3A) indicated that baseline ISI score had the widest point range among the predictors, identifying it as the most influential contributor to outcome prediction. However, in the nomogram for the combined model (Figure 3B), the top three predictors ranked by contribution were all early-treatment variables. Week 2 ISI change emerged as the strongest predictor with the widest point range, whereas Week 2 TIB difference and Week 2 actigraphy TST change exhibited substantially narrower point ranges.

In the training set, the baseline-only model achieved a mean CV-AUC of 0.683 (SD = 0.065), whereas the combined model yielded a mean CV-AUC of 0.752 (SD = 0.060). In the test set, the baseline-only model achieved an AUC of 0.630 (95% CI: 0.562–0.698) and a Brier score of 0.207 (95% CI: 0.184–0.228), while the combined model achieved an AUC of 0.741 (95% CI: 0.683–0.798) and a Brier score of 0.180 (95% CI: 0.159–0.202). Figure 4A compares the ROC curves of the two models in the test set.

Bootstrap resampling indicated that the combined model outperformed the baseline-only model, with a difference in mean CV-AUC of 0.061 (95% CI: 0.012–0.107; p = 0.013) in the training set and a difference in AUC of 0.111 (95% CI: 0.045–0.182; p < 0.001) in the test set.

Calibration curves for the two models are shown in Figure 4B. The combined model tended to show better calibration than the baseline-only model, with closer alignment with the ideal diagonal line across the entire predicted probability range.

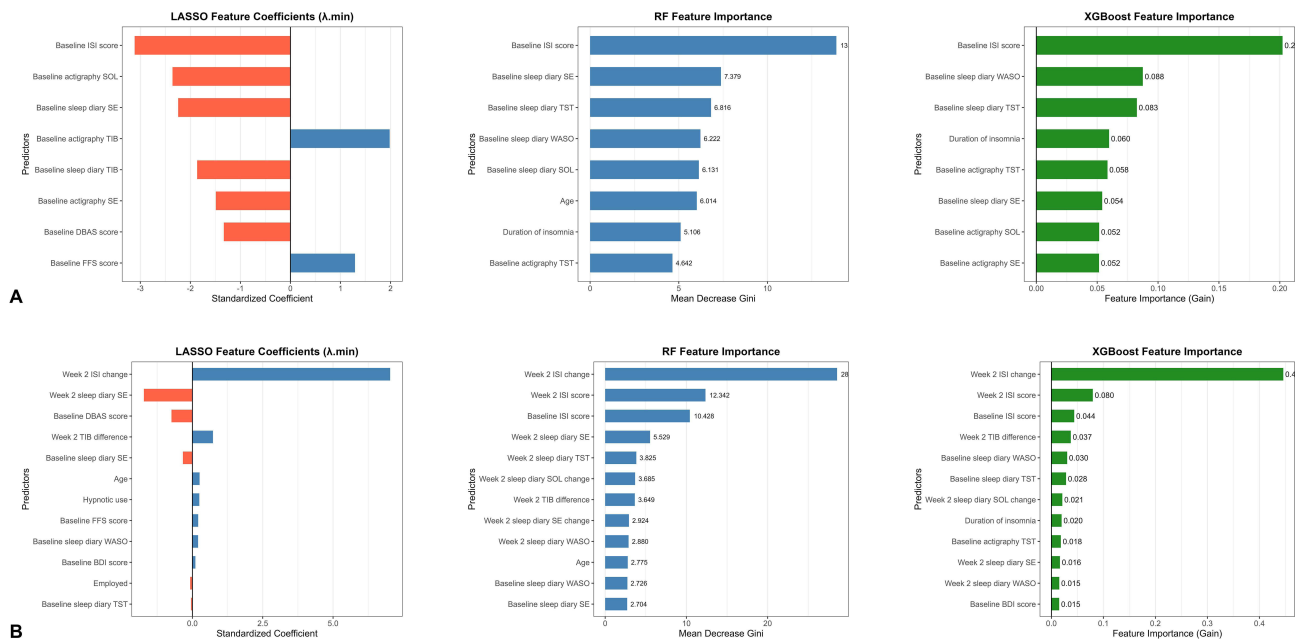


Figure 2 Top important features selected by machine learning algorithms for post-treatment outcome models. (A) Feature selection for the baseline-only model. Panel shows the top 8 important features selected from 24 baseline variables. (B) Feature selection for the combined model. Panel shows the top 12 important features selected from 45 variables. Features are ranked by decreasing importance (or coefficient magnitude).

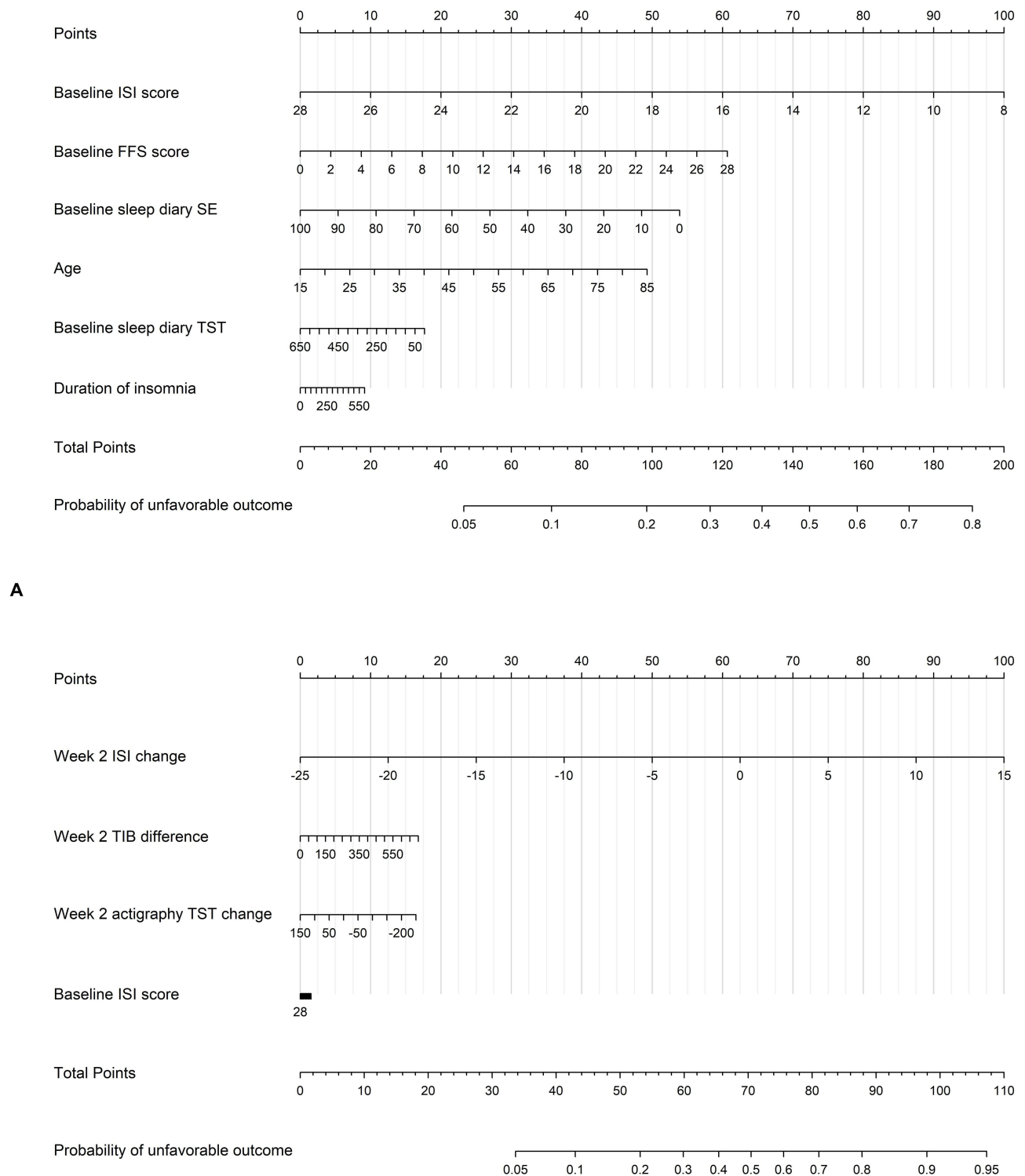


Figure 3 Nomograms for predicting the probability of post-treatment unfavorable outcome. **(A)** Nomogram derived from the baseline-only model. **(B)** Nomogram derived from the combined model. To estimate risk, sum the points for each predictor; then project the total points downward to the probability axis.

Notes: For graphical purposes, the nomogram displays age starting from 15 and insomnia duration from 0. However, the actual study population ranges were age ≥ 18 years and insomnia duration ≥ 3 months. Risk predictions should only be made using values within these actual study ranges.

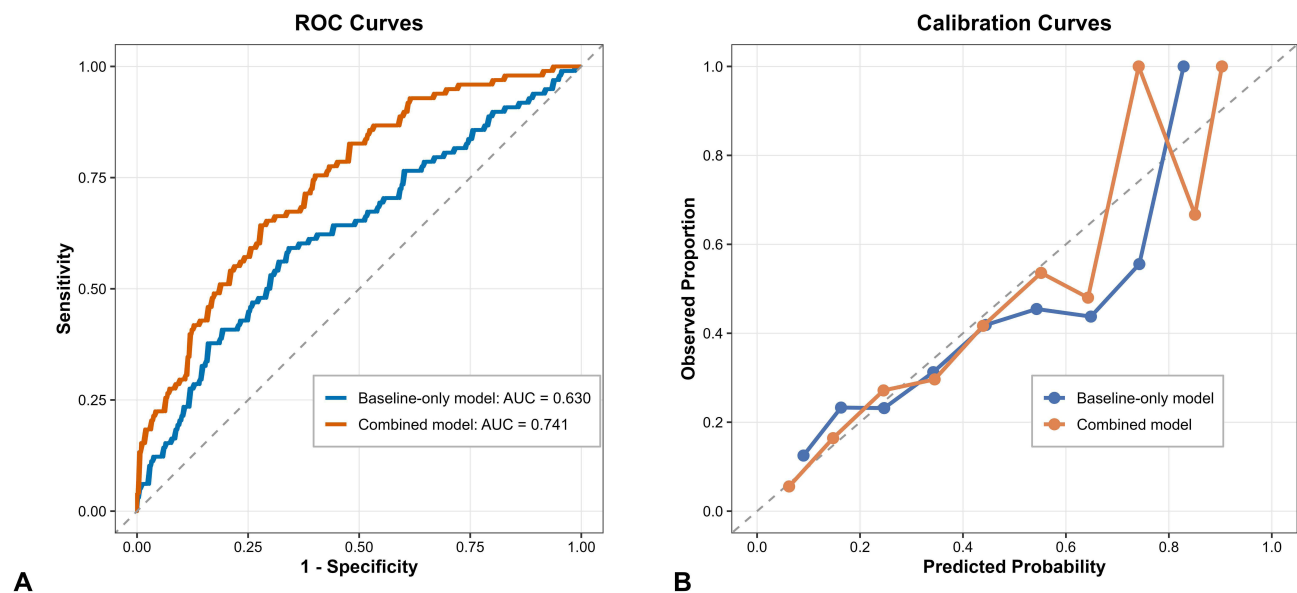


Figure 4 Comparison of ROC curves and Calibration curves for two post-treatment models in the test set. **(A)** ROC curves: The combined model (red curve) demonstrated superior discriminative ability to the baseline-only model (blue curve), with an AUC of 0.741 versus 0.630. The difference between the two AUCs was statistically significant ($p < 0.001$). **(B)** Calibration curves: The combined model (red curve) shows better alignment with the ideal diagonal line than the Baseline-only model (blue curve), indicating better calibration.

Abbreviations: ROC, Receiver Operating Characteristic curve; AUC, Area Under the Curve.

For 12-Month Follow-Up Outcome

Sensitivity analysis further showed that the AUC plateaued when consensus features were selected as those common to the top 5 features of LASSO, Random Forest, and XGBoost for the baseline-only model and top 11 features for the combined model. Using this threshold, four consensus features were identified for the baseline-only model (Age; Duration of insomnia; Baseline ISI score; Baseline sleep diary SE) (Figure 5A), and five for the combined model (Age; Duration of insomnia; Week 2 sleep diary SE; Baseline ISI score; Baseline sleep diary SE) (Figure 5B).

Nomograms for both the baseline-only (Figure 6A) and combined models (Figure 6B) consistently indicated age as the most influential predictor, followed by duration of insomnia.

The baseline-only model achieved a mean CV-AUC of 0.739 (SD = 0.061) in the training set, an AUC of 0.735 (95% CI: 0.664–0.792) (Figure 7A), and a Brier score of 0.210 (95% CI: 0.187–0.232) in the test set. In comparison, the combined model yielded slightly higher performance, with a mean CV-AUC of 0.743 (SD = 0.061) in the training set, and an AUC of 0.742 (95% CI: 0.683–0.807) and a Brier score of 0.204 (95% CI: 0.180–0.226) in the test set. However, bootstrap resampling analysis indicated that the improvement in discriminatory performance was not statistically significant (all $p > 0.05$).

Calibration curves for the two models are shown in Figure 7B. Both models demonstrated good and similar calibration performance, with their curves closely following the ideal diagonal line across most of the predicted probability range.

Consistent with the primary analysis, the sensitivity analysis for missing 12-month follow-up data yielded similar results. When all lost-to-follow-up patients were alternatively assumed to have unfavorable or favorable outcomes, the combined model consistently demonstrated slightly higher AUCs than the baseline-only model (0.693 vs 0.684 and 0.688 vs 0.685). Nevertheless, bootstrap analyses confirmed that these differences did not reach statistical significance under either assumption (all $p > 0.05$).

Discussion

This study suggests that incorporating early treatment variables can meaningfully improve the prediction of post-treatment outcomes following CBT-I, beyond what can be achieved using baseline characteristics alone. In particular,

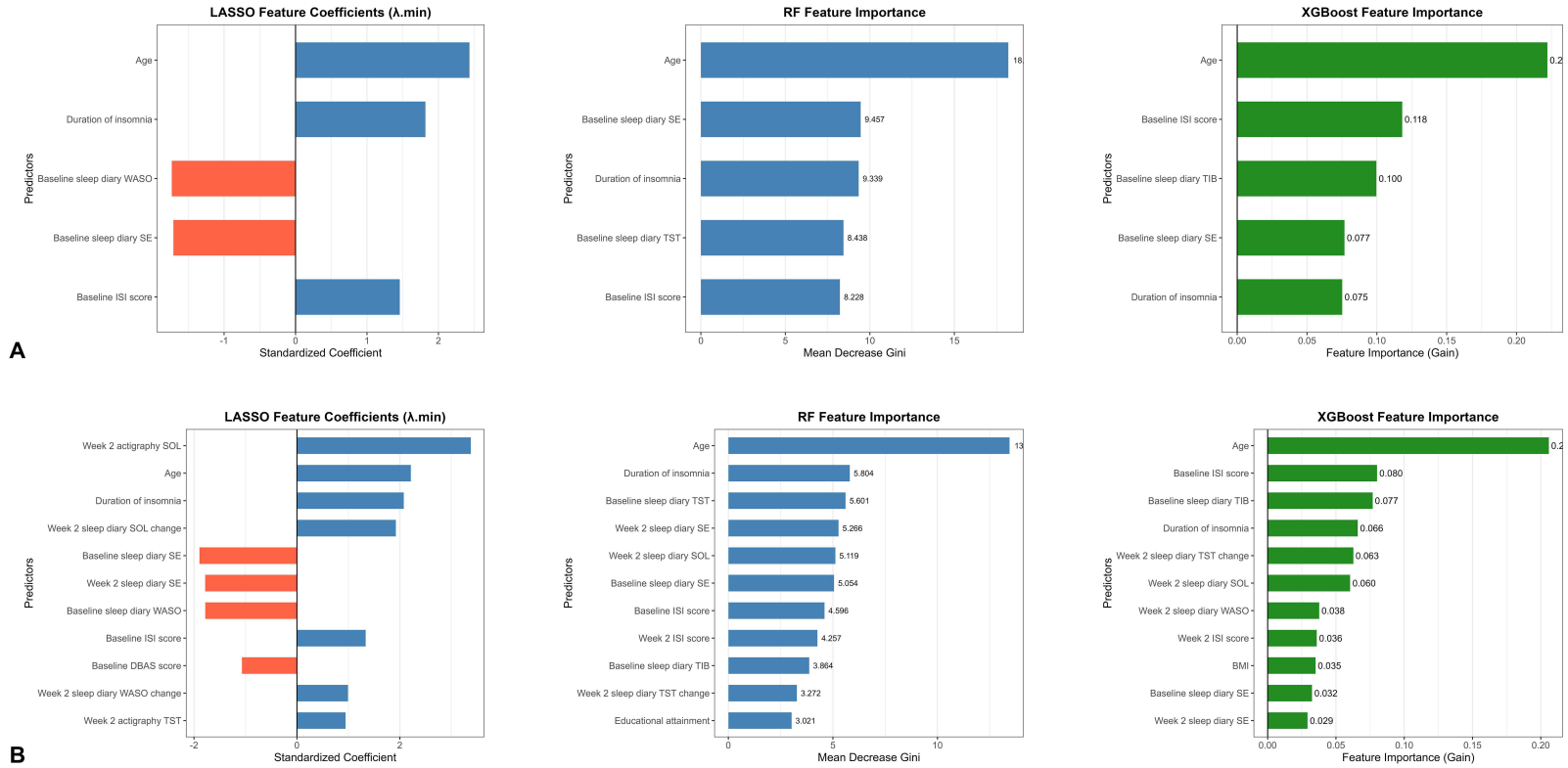
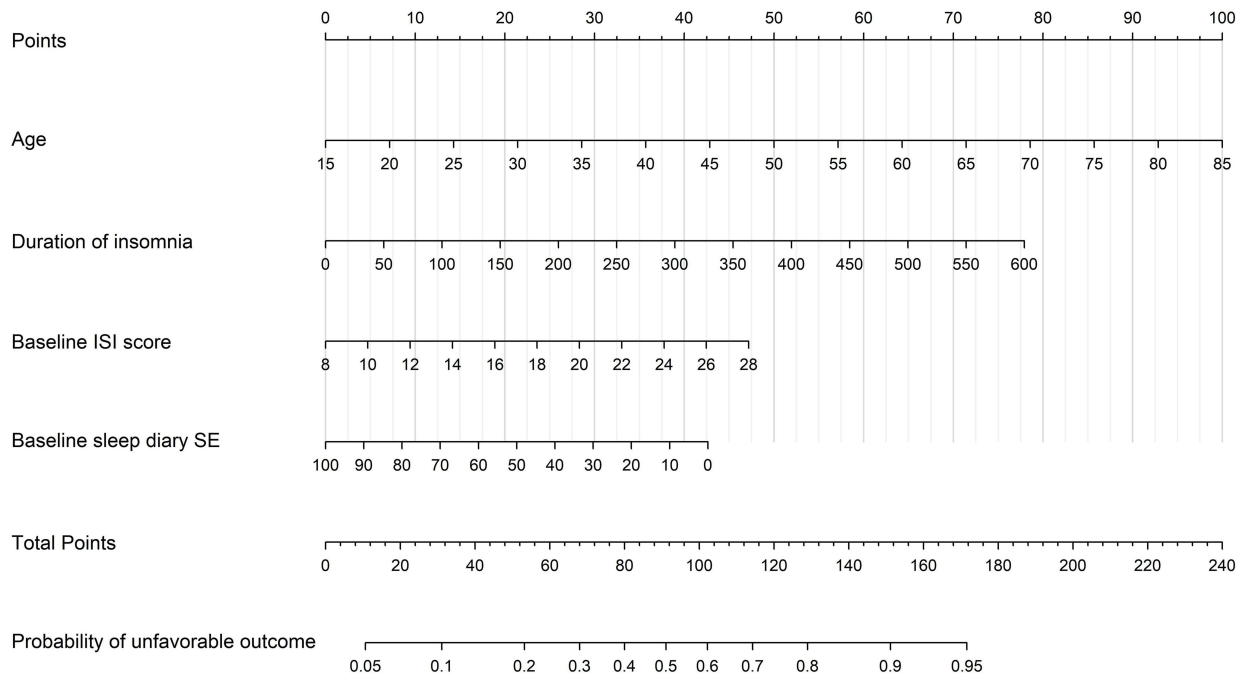
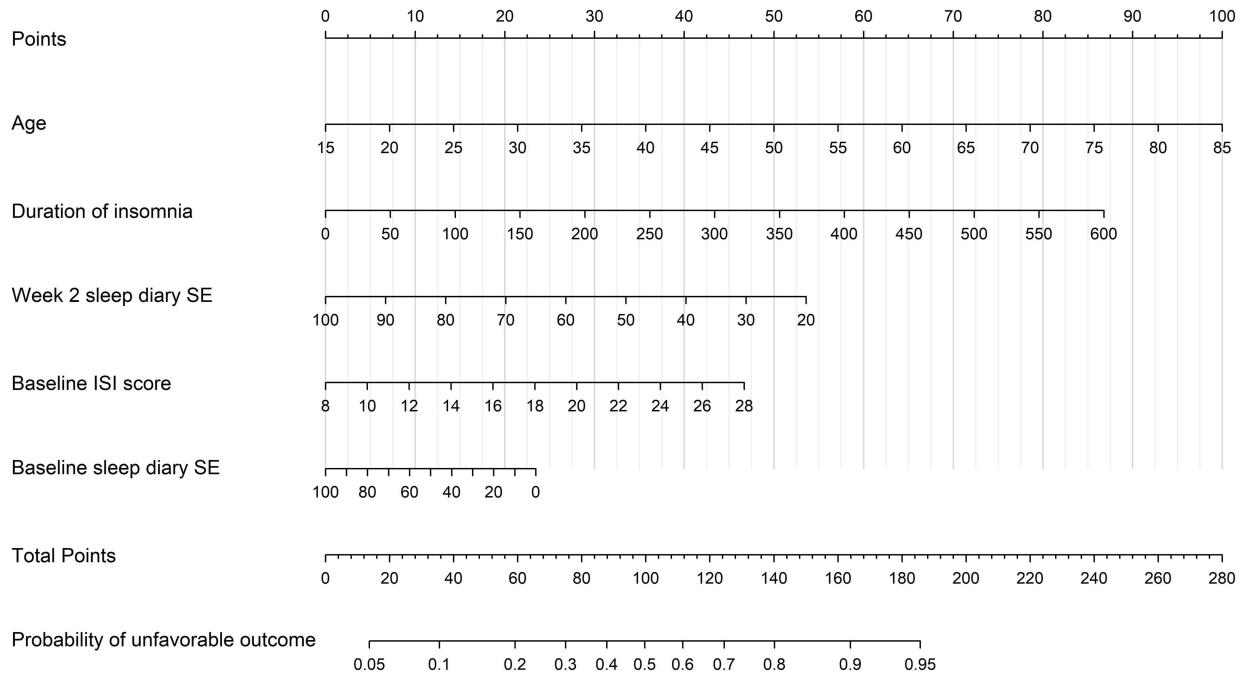


Figure 5 Top important features selected by machine learning algorithms for 12-month follow-up outcome models. **(A)** Feature selection for the baseline-only model. Panel shows the top 5 important features selected from 24 baseline variables; **(B)** Feature selection for the combined model. Panel shows the top 11 important features selected from 45 variables.



A



B

Figure 6 Nomograms for predicting the probability of 12-month follow-up unfavorable outcome. **(A)** Nomogram derived from the baseline-only model; **(B)** Nomogram derived from the combined model.

Notes: For graphical purposes, the nomogram displays age starting from 15 and insomnia duration from 0. However, the actual study population ranges were age ≥ 18 years and insomnia duration ≥ 3 months. Risk predictions should only be made using values within these actual study ranges.

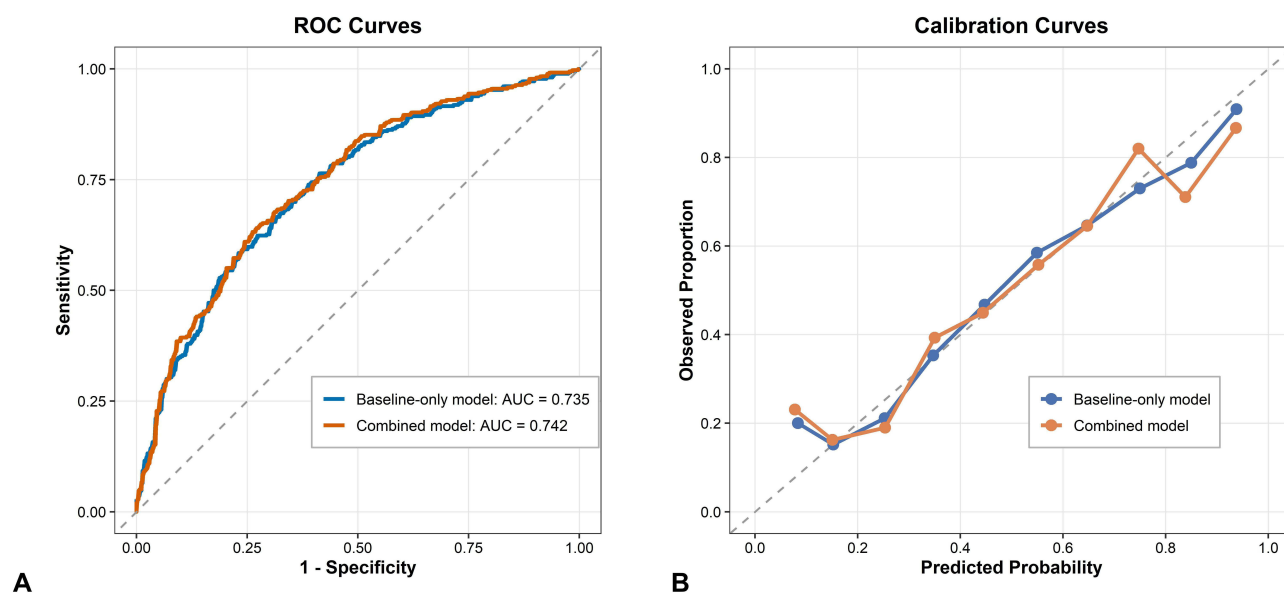


Figure 7 Comparison of ROC curves and Calibration curves for two 12-month follow-up models in the test set. **(A)** ROC curves: The combined model (red curve) demonstrated slightly improved discriminative ability compared with the baseline-only model (blue curve), with an AUC of 0.742 versus 0.735. This improvement was not statistically significant ($p > 0.05$); **(B)** Calibration curves. Both models showed good and similar calibration, each closely tracking the ideal diagonal line across most of the predicted probability range.

Abbreviations: ROC, Receiver Operating Characteristic curve; AUC, Area Under the Curve.

reduction in insomnia severity during the early treatment phase emerged as the most informative predictor of post-treatment outcomes. However, the added predictive information of early treatment variables for 12-month follow-up outcomes appears limited and did not reach statistical significance, suggesting that the prognostic value of early treatment factors differs between short-term and long-term outcomes.

Regarding post-treatment outcome, most prior studies have consistently reported baseline insomnia severity as the most important predictor of CBT-I outcome.^{8–11} This is partially reflected in our baseline-only model for the post-treatment outcome, where a higher baseline ISI score was associated with a lower risk of an unfavorable outcome. However, the predictive importance of Week 2 ISI change outweighed that of the baseline ISI score in the combined model, suggesting that early treatment variables may carry more prognostic information for short-term treatment outcomes than baseline variables.

We found that a greater reduction in the ISI score at Week 2 was associated with a reduced probability of an unfavorable outcome. This pattern is broadly consistent with studies on psychological treatments for other disorders, which suggest that early symptom improvement may indicate better outcomes.^{16–18} These early changes may be associated with underlying processes such as therapeutic engagement, behavioral adherence, and neurocognitive adaptation. Within the specific context of CBT-I, it is possible that early symptom improvement may correspond to greater responsiveness to the behavioral components delivered in the initial treatment phase. Moreover, because the ISI captures subjective sleep quality, its improvement might be detectable as early as week 2, potentially earlier than that of some other sleep parameters. This could partly explain why ISI was the most informative predictor among all early treatment variables at week 2.

Week 2 TIB difference and Week 2 actigraphy TST change were also identified as the important early treatment predictor for post-treatment outcomes. Within the multivariable logistic regression model, higher Week 2 TIB difference (indicating lower adherence to SRT) and greater reduction in Week 2 actigraphy TST were associated with a higher probability of an unfavorable outcome. Previous studies have shown that adherence to behavioral strategies influences CBT-I outcomes.²⁹ Our study further found that adherence at an early stage was associated with treatment outcomes. However, on the nomogram, the narrower point ranges of Week 2 actigraphy TST change and TIB difference suggest that

these two factors provide only modest incremental predictive value beyond the strong signal captured by Week 2 ISI change.

In the 12-month follow-up outcome models, baseline predictors showed significantly greater predictive importance than early treatment predictors. Specifically, older age and longer insomnia duration were the two factors that were most strongly associated with an increased risk of unfavorable outcomes. This suggests that the long-term efficacy of CBT-I may be associated with patients' pre-existing, relatively stable characteristics. In older adults, age-related changes in sleep architecture may limit the extent to which insomnia is fully reversible.³⁰ Similarly, a longer duration of insomnia may correspond to more deeply entrenched perpetuating factors, which might limit sustained treatment response even after completion of a standard CBT-I program.

Notably, even after incorporating early treatment variables, model performance for predicting 12-month outcomes did not improve significantly. First, early treatment predictors may be more vulnerable to confounding, attenuating their signal-to-noise ratio for long-term prediction and thereby biasing estimates toward the null. Second, measurement limitations should also be considered. Third, improvements achieved during the early phase of treatment (eg, behavioral adjustments) may not be consistently maintained after treatment completion. Variability in post-treatment adherence to CBT-I skills and environmental influences are plausible but unmeasured moderators of long-term outcomes.

In summary, this study demonstrated the short-term efficacy of CBT-I and the durability of treatment gains at 12-month follow-up, which is consistent with prior evidence.^{5,6} Nevertheless, a significant proportion of patients still fail to achieve favorable outcomes. To better identify these individuals, we developed prognostic models to examine the added prognostic value of early treatment variables. Our results suggest that the predictive utility of early treatment variables may be time-horizon dependent: they are more informative for short-term outcomes, whereas long-term outcomes are more strongly influenced by stable baseline characteristics.

In this study, we used the ISI for post-treatment assessment and the PSQI for follow-up evaluation. Both questionnaires are widely used outcome measures in CBT-I research.³ The version of the PSQI employed in this study assesses insomnia symptoms over the past month, making it more suitable for long-term follow-up than for immediate post-treatment evaluation. Moreover, since the primary aim of this study was to compare the predictive performance of the combined and baseline-only models for each outcome, the use of different instruments for post-treatment and follow-up assessments is unlikely to substantially affect the comparative conclusions.

This study has several limitations. First, the retrospective nature of the data did not allow for the inclusion of all potential predictors, which might have impacted the model's predictive accuracy. Thus, future prospective studies are warranted to validate these findings. Second, missing data may have introduced bias. To mitigate this, we imputed variables with missing values using the MissForest algorithm. Third, loss to follow-up at 12-month is a potential source of bias. Although sensitivity analyses that imputed extreme outcomes for missing data consistently supported the robustness of our primary finding, the possibility of non-random attrition cannot be fully ruled out. Fourth, the consensus-based feature selection strategy is relatively stringent. However, it can help control model complexity and reduce overfitting risk from excessive features. This allows for a more accurate assessment of the true predictive gain of early treatment features. Fifth, the risk of overfitting cannot be entirely ruled out, even though we performed internal validation using repeated cross-validation and an independent hold-out set. Therefore, external validation in a distinct cohort is required to confirm the model's generalizability.

Conclusions

This study suggests that incorporating early treatment variables may provide incremental value over baseline information alone in predicting CBT-I short-term outcomes. Week 2 ISI score reduction appeared to be the most informative predictor of post-treatment outcome. Nevertheless, 12-month follow-up outcomes may be more closely related to baseline characteristics, and the added value of early treatment variables for predicting long-term outcomes appeared limited. These findings are preliminary and warrant further validation.

Use of Generative AI Tools

The authors used ChatGPT (GPT-4) as an AI-assisted language tool to improve grammar and readability. The authors take full responsibility for the content, analysis, and interpretation of the data.

Data Sharing Statement

The data that support the findings of this study are not publicly available due to ethical and privacy restrictions. Data may be made available from the corresponding author upon reasonable request, subject to approval by the institutional ethics committee.

Ethics Approval and Consent to Participate

The study was conducted in full accordance with the ethical standards of the Declaration of Helsinki and its subsequent revisions. Ethical approval for this study was obtained from the Institutional Review Board of Xiamen Xianyue Hospital (Approval No. 2022-KY-033). The Board also waived the requirement for individual informed consent due to the fully anonymized and non-traceable nature of the data.

Author Contributions

YH; Conceptualization, Funding acquisition; Data curation, Writing—original draft; Writing—review and editing. XW; Conceptualization, Supervision, Writing—review and editing. ZN; Data curation, Writing—review and editing. SW; Data curation, Writing—review and editing. NW; Data curation, Writing—review and editing. SZ; Software, Writing—review and editing. PL; Software, Writing—review and editing. SH; Software, Writing—review and editing. All authors gave final approval of the version to be published; have agreed on the journal to which the article has been submitted; and agree to be accountable for all aspects of the work.

Funding

This work was supported by the Medical and Health Guidance Project of Xiamen, China [grant number 3502Z20224ZD1313].

Disclosure

Dr Yanyu Hu reports grants from the Medical and Health Guidance Project of Xiamen, China, during the conduct of the study. All other authors declare that they have no competing interests relating to this work.

References

1. Morin CM, Benca R. Chronic insomnia. *Lancet*. 2012;379(9821):1129–1141. doi:10.1016/S0140-6736(11)60750-2
2. Riemann D, Baglioni C, Bassetti C, et al. European guideline for the diagnosis and treatment of insomnia. *J Sleep Res*. 2017;26(6):675–700. doi:10.1111/jsr.12594
3. Trauer JM, Qian MY, Doyle JS, Rajaratnam SMW, Cunnington D. Cognitive behavioral therapy for chronic insomnia: a systematic review and meta-analysis. *Ann Intern Med*. 2015;163(3):191–204. doi:10.7326/M14-2841
4. van Straten A, van der Zweerde T, Kleiboer A, et al. Cognitive and behavioral therapies in the treatment of insomnia: a meta-analysis. *Sleep Med Rev*. 2018;38:3–16. doi:10.1016/j.smrv.2017.02.001
5. Perlis ML, Posner D, Riemann D, et al. Insomnia. *Lancet*. 2022;400(10357):1047–1060. doi:10.1016/S0140-6736(22)00879-0
6. Furukawa Y, Sakata M, Furukawa TA, Efthimiou O, Perlis M. Initial treatment choices for long-term remission of chronic insomnia disorder in adults: a systematic review and network meta-analysis. *Psychiatry Clin Neurosci*. 2024;78(11):646–653. doi:10.1111/pcn.13730
7. Muench A, Vargas I, Posner D, Perlis ML. Standard cognitive behavioral therapy for insomnia (CBT-I): when is the adaptation of CBT-I warranted? *Adapt Cognitive Behav Ther Insomnia*. 2022;3–24.
8. Holler E, Du Y, Barboi C, Owora A. Prognostic models for predicting insomnia treatment outcomes: a systematic review. *J Psychiatr Res*. 2024;170:147–157. doi:10.1016/j.jpsychires.2023.12.017
9. Blom K, Hentati Isacson N, Forsell E, et al. An investigation and replication of sleep-related cognitions, acceptance and behaviours as predictors of short- and long-term outcome in cognitive behavioural therapy for insomnia. *J Sleep Res*. 2021;30(5):e13376. doi:10.1111/jsr.13376
10. Forsell E, Jernelöv S, Blom K, Kaldo V. Clinically sufficient classification accuracy and key predictors of treatment failure in a randomized controlled trial of internet-delivered cognitive behavior therapy for insomnia. *Internet Interv*. 2022;29:100554. doi:10.1016/j.invent.2022.100554
11. Gabbay FH, Wynn GH, Georg MW, et al. Toward personalized care for insomnia in the US Army: a machine learning model to predict response to cognitive behavioral therapy for insomnia. *J Clin Sleep Med*. 2024;20(6):921–931. doi:10.5664/jcs.m.11026

12. Pruiksma KE, Hale WJ, Mintz J, et al. Predictors of cognitive behavioral therapy for insomnia (CBTi) outcomes in active-duty US army personnel. *Behav Ther.* 2020;51(4):522–534. doi:10.1016/j.beth.2020.02.001
13. Espie CA, Inglis SJ, Harvey L. Predicting clinically significant response to cognitive behavior therapy for chronic insomnia in general medical practice: analysis of outcome data at 12 months posttreatment. *J Consult Clin Psychol.* 2001;69(1):58–66. doi:10.1037/0022-006X.69.1.58
14. Jansson-Fröjmark M, Linton SJ. The role of sleep-related beliefs to improvement in early cognitive behavioral therapy for insomnia. *Cogn Behav Ther.* 2008;37(1):5–13. doi:10.1080/16506070801907013
15. Forsell E, Jernelöv S, Blom K, et al. Proof of concept for an adaptive treatment strategy to prevent failures in internet-delivered CBT: a single-blind randomized clinical trial with insomnia patients. *Am J Psychiatry.* 2019;176(4):315–323. doi:10.1176/appi.ajp.2018.18060699
16. Lutz W, Stulz N, Köck K. Patterns of early change and their relationship to outcome and follow-up among patients with major depressive disorders. *J Affect Disord.* 2009;118(1–3):60–68. doi:10.1016/j.jad.2009.01.019
17. Schibbye P, Ghaderi A, Ljotsson B, et al. Using early change to predict outcome in cognitive behaviour therapy: exploring timeframe, calculation method, and differences of disorder-specific versus general measures. *PLoS One.* 2014;9(6):e100614. doi:10.1371/journal.pone.0100614
18. Lutz W, Hofmann SG, Rubel J, et al. Patterns of early change and their relationship to outcome and early treatment termination in patients with panic disorder. *J Consult Clin Psychol.* 2014;82(2):287–297. doi:10.1037/a0035535
19. Morin CM, Belleville G, Belanger L, Ivers H. The insomnia severity index: psychometric indicators to detect insomnia cases and evaluate treatment response. *Sleep.* 2011;34(5):601–608. doi:10.1093/sleep/34.5.601
20. Richter P, Werner J, Heerlein A, Kraus A, Sauer H. On the validity of the beck depression inventory. A review. *Psychopathology.* 1998;31(3):160–168. doi:10.1159/000066239
21. Beck AT, Epstein N, Brown G, Steer RA. An inventory for measuring clinical anxiety: psychometric properties. *J Consult Clin Psychol.* 1988;56(6):893–897. doi:10.1037/0022-006X.56.6.893
22. Gradisar M, Lack L, Richards H, et al. The flinders fatigue scale: preliminary psychometric properties and clinical sensitivity of a new scale for measuring daytime fatigue associated with insomnia. *J Clin Sleep Med.* 2007;3(7):722–728. doi:10.5664/jcsm.27030
23. Espie CA, Inglis SJ, Harvey L, Tessier S. Insomniacs' attributions. psychometric properties of the dysfunctional beliefs and attitudes about sleep scale and the sleep disturbance questionnaire. *J Psychosom Res.* 2000;48(2):141–148. doi:10.1016/S0022-3999(99)00090-2
24. Cid-Verdejo R, Chávez Fariás C, Martínez-Pozas O, et al. Instrumental assessment of sleep bruxism: a systematic review and meta-analysis. *Sleep Med Rev.* 2024;74:101906. doi:10.1016/j.smr.2024.101906
25. Smith MT, McCrae CS, Cheung J, et al. Use of actigraphy for the evaluation of sleep disorders and circadian rhythm sleep-wake disorders: an American academy of sleep medicine clinical practice guideline. *J Clin Sleep Med.* 2018;14(7):1231–1237. doi:10.5664/jcsm.7230
26. Riedel BW, Lichstein KL. Strategies for evaluating adherence to sleep restriction treatment for insomnia. *Behav Res Ther.* 2001;39(2):201–212. doi:10.1016/S0005-7967(00)00002-4
27. Buysse DJ, Reynolds CFR, Monk TH, Berman SR, Kupfer DJ. The Pittsburgh sleep quality index: a new instrument for psychiatric practice and research. *Psychiatry Res.* 1989;28(2):193–213. doi:10.1016/0165-1781(89)90047-4
28. Deng Y, Ma Y, Fu J, et al. Combinatorial use of machine learning and logistic regression for predicting carotid plaque risk among 5.4 million adults with fatty liver disease receiving health check-ups: population-based cross-sectional study. *JMIR Public Health Surveill.* 2023;9:e47095. doi:10.2196/47095
29. Dong L, Soehner AM, Bélanger L, Morin CM, Harvey AG. Treatment agreement, adherence, and outcome in cognitive behavioral treatments for insomnia. *J Consult Clin Psychol.* 2018;86(3):294–299. doi:10.1037/ccp0000269
30. Cohen ZL, Eigenberger PM, Sharkey KM, Conroy ML, Wilkins KM. Insomnia and other sleep disorders in older adults. *Psychiatr Clin North Am.* 2022;45(4):717–734. doi:10.1016/j.psc.2022.07.002

Nature and Science of Sleep

Publish your work in this journal

Nature and Science of Sleep is an international, peer-reviewed, open access journal covering all aspects of sleep science and sleep medicine, including the neurophysiology and functions of sleep, the genetics of sleep, sleep and society, biological rhythms, dreaming, sleep disorders and therapy, and strategies to optimize healthy sleep. The manuscript management system is completely online and includes a very quick and fair peer-review system, which is all easy to use. Visit <http://www.dovepress.com/testimonials.php> to read real quotes from published authors.

Submit your manuscript here: <https://www.dovepress.com/nature-and-science-of-sleep-journal>

Dovepress
Taylor & Francis Group