Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

# Incremental prognostic value of early symptom reassessment for later depression remission: a week-2 landmark prediction study within a five-arm multicentre randomised trial

**Jun-Ming Zhu**
Southern Medical University
**Cai-lin Zhao**
Southern Medical University
**Zhao Yuan**
The Second Affiliated Hospital of Guilin Medical University, Guilin Medical University
**Yong Huang**
Southern Medical University
**Wen-Juan Ji**
Southern Medical University
**Jiani Fu**
Southern Medical University
**Wei Xie**
Southern Medical University
**Jiao Hu**
7341569580qq.com

Southern Medical University

**Research Article**

**Keywords:** major depressive disorder, dynamic prediction, early response, HAMD-17, remission, internal-external cross-validation

**Posted Date:** September 20th, 2026

**DOI:** https://doi.org/10.21203/rs.3.rs-10821681/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

Early symptom change may improve prediction of depression outcomes beyond baseline characteristics, but the incremental value of repeated assessment and the need for complex dynamic models remain uncertain. We quantified the value of week-1 and week-2 symptom information and compared the full model with a contemporaneous symptom-score benchmark.

# Methods

We performed a post hoc secondary prognostic modelling analysis of 347 participants from a five-centre, five-arm randomised depression-treatment trial. Ridge logistic models used baseline information and sequentially added week-1 and week-2 scores and changes on the 17-item Hamilton Depression Rating Scale (HAMD-17) and Self-rating Depression Scale (SDS). Outcomes were HAMD-17 remission (score < = 7) at week 6, week 10 and both visits. Performance was estimated by three-times repeated nested five-fold cross-validation. Incremental area under the receiver operating characteristic curve (AUROC) was assessed with paired DeLong tests and Holm correction. Calibration, average precision, Brier score, decision curves and within-trial transportability analyses were evaluated.

# Results

Outcome cohorts comprised 306 participants (81 remissions) at week 6, 269 (94) at week 10 and 258 (47) at both visits. AUROCs for baseline, week-1 and week-2 models were 0.613, 0.741 and 0.833 at week 6; 0.578, 0.671 and 0.742 at week 10; and 0.563, 0.770 and 0.828 at both visits. All nine staged comparisons remained significant after Holm correction. The full model did not significantly outperform week-2 HAMD-17 alone: AUROC differences were 0.003, 0.033 and 0.040 (all Holm-adjusted P = 1.00). Centre internal-external cross-validation AUROCs were 0.839, 0.720 and 0.825, respectively, while treatment-arm transport was weaker for later outcomes.

# Conclusions

Week-1 and especially week-2 reassessment improved prediction over baseline information. A single week-2 HAMD-17 score captured nearly all week-6 discrimination and much of the discrimination for later outcomes. Independent validation, robust missing-outcome analyses, recalibration and prospective impact evaluation are required before clinical use.

# Trial registration:

The source trial was retrospectively registered with the Chinese Clinical Trial Registry (ChiCTR-TRC-08000278) on 26 December 2008. This post hoc prediction analysis was not prospectively registered.

## Background

Treatment response in major depressive disorder is heterogeneous, and a clinically important proportion of patients do not achieve remission after an initial course of therapy. Baseline characteristics alone have generally provided only modest individual-level prognostic accuracy across trials and settings [1, 2]. Early response offers a different source of information: it reflects the evolving interaction between illness, allocated treatment, adherence and measurement context before the final outcome is known.

Evidence from antidepressant trials has repeatedly linked improvement within the first 2 weeks to later response or remission [3–6]. However, threshold-based early-response rules can have asymmetric predictive value, and substantial proportions of patients classified as early non-responders later respond [6]. Modern prediction studies should therefore quantify probability-based discrimination, calibration and clinical utility; compare sequential models in the same participants; and distinguish internal or internal-external validation from independent external validation [7–13].

Recent depression modelling studies have used cross-trial, multimodal and repeated-measure data [1, 2, 14–18]. Sequential models have shown that prediction can improve as early observations accrue [17], while longitudinal work suggests that the latest symptom measurement may perform as well as more elaborate trajectory summaries [18]. Two questions therefore remain clinically important: how much does each reassessment add beyond information already available, and does a complex dynamic specification outperform the latest score on the same scale used to define remission? A week-2 score on the 17-item Hamilton Depression Rating Scale (HAMD-17) is temporally prior to week-6 and week-10 outcomes and is not leakage, but it may dominate short-horizon prediction through construct continuity.

## Study objective

We performed a secondary prognostic modelling analysis of a five-centre, five-arm randomised trial dataset. The objectives were to quantify the incremental value of week-1 and week-2 symptom information for remission at week 6, week 10 and both visits; compare the full week-2 model with a week-2 HAMD-17-only benchmark; examine transportability across centres and treatment arms; and perform a date-field-ordered holdout sensitivity analysis. We hypothesised that early reassessment would improve prediction over baseline, while treating the added value of the complex model beyond current HAMD severity as an empirical question.

## Methods

# Study design and participants

We conducted a secondary prognostic modelling analysis of individual-participant data from a five-centre, five-arm randomised clinical trial of depression treatment in China. According to confirmation from the original study investigators and record-level concordance, Qu et al. [19] and Fu et al. [20] analysed selected subsets of this same source database rather than independent cohorts. The five-centre report by Fu et al. described recruitment from January 2009 to December 2011 at five medical institutions in Guangdong, China [20].

This prediction analysis was not prospectively registered and no prespecified prediction-model analysis plan was available; it should therefore be interpreted as a post hoc exploratory secondary analysis. Outcomes, the week-2 landmark, staged predictor sets, benchmark comparisons and validation procedures are reported transparently rather than presented as trial-confirmatory analyses. The parent trial registry record is ChiCTR-TRC-08000278, registered retrospectively on 26 December 2008; this secondary prediction analysis was not prospectively registered.

The supplied source cohort comprised 347 participants across five centres (160, 67, 30, 40 and 50 participants) and five groups: medication control, manual acupuncture, manual acupuncture plus medication, electroacupuncture, and electroacupuncture plus medication (68, 70, 69, 70 and 70 participants). The analysis addressed prognosis after treatment assignment and was not designed to estimate causal treatment effects or identify an optimal treatment for an individual.

Published methods for the selected randomised-trial subset described eligible participants as aged 18–60 years, meeting ICD-10 criteria for a depressive episode with bipolar depression excluded, having a baseline HAMD-17 score of at least 17, and providing written informed consent [19]. Reported exclusions were participation in another clinical trial within 4 weeks, current antidepressant treatment without an adequate washout, pregnancy or lactation, serious systemic illness requiring treatment, organic brain disease, and a suicide plan or behaviour [19].

The published eligibility criteria could not be reconciled completely with every supplied record: two ages and 12–15 reconstructed baseline HAMD-17 totals, depending on item-14 handling, fell outside the published thresholds. We therefore treated the supplied 347 records as the source cohort and did not infer protocol eligibility beyond the available documentation.

Published recruitment was reported as January 2009 to December 2011, whereas the supplied source date field MHSTDT (original definition unavailable) ranged from 11 May 2011 to 20 December 2012. Because its definition was unavailable, MHSTDT was not interpreted as a confirmed enrolment date; the ordered holdout using this field was retained only as a source-data sensitivity analysis.

The prediction landmark was week 2. Predictors were restricted to baseline, week 1 and week 2; week-6 and week-10 data were used only to define outcomes. This observation-window/prediction-window separation follows standard longitudinal prediction principles [21]. Of 347 participants, 340 had week-1 information and 338 reached the week-2 landmark. For each outcome, all staged models were evaluated in the same outcome-specific cohort that reached week 2 and had the relevant outcome observed,

preventing cohort composition from masquerading as incremental performance (Fig. 1; Additional file 1: Table S1). This report follows TRIPOD + AI [7].

## Outcomes

The primary outcome was remission at week 6, defined as HAMD-17 total score < = 7. Secondary outcomes were remission at week 10 and remission at both week-6 and week-10 assessments. We use the term two-point remission because observations at two visits do not establish uninterrupted remission between them. Missing outcomes were excluded rather than coded as non-remission.

HAMD-17 totals required all 17 items. Under the working primary convention, item-14 code 3 was interpreted as cannot determine/not applicable and recoded to 0; a sensitivity analysis treated it as missing. The definitive rule must be verified against the original case-report form or scoring manual before submission.

## Candidate predictors and staged models

The baseline model included age, sex, marital status, ethnicity, body mass index, parsed episode duration in months, baseline HAMD-17, baseline Self-rating Depression Scale (SDS), randomised treatment group and centre. The week-1 model added week-1 HAMD-17 and SDS, absolute changes from baseline and percentage changes from baseline. The week-2 model additionally included week-2 HAMD-17 and SDS, their absolute and percentage changes from baseline, and week-1-to-week-2 changes. Negative absolute change denoted improvement. No univariable screening was used. Treatment group was included because allocation was known by week 2; the main model is not a treatment-selection model. Full definitions are provided in Additional file 1: Table S2.

## Preprocessing and missing data

Within each training fold, numeric predictors were median-imputed, centred and scaled; a missingness indicator was added where applicable. Categorical variables were one-hot encoded with explicit missing and unseen levels. The fitted transformations were applied unchanged to validation data. Episode duration was missing for 241/347 (69.5%); other core predictor missingness ranged from 0% to 2.6%. Outcomes were not imputed.

# Model estimation and validation

Each staged model used ridge-penalised logistic regression. The penalty was selected from 0.03, 0.10, 0.30, 1, 3, 10 and 30 by four-fold stratified inner cross-validation minimising log loss. Performance was estimated with a stratified five-fold outer loop repeated three times. Competing models used shared fold assignments. Preprocessing, tuning and fitting were repeated within every outer training fold. The three

held-out probabilities for each participant were averaged to one out-of-fold prediction. The available trial cohort determined sample size; no new a priori sample-size calculation was performed, so event counts and uncertainty are reported for each outcome [9].

## Performance and incremental value

Discrimination was assessed with the area under the receiver operating characteristic curve (AUROC) and average precision (AP), overall accuracy with Brier score and log loss, and calibration with calibration-in-the-large, calibration slope, grouped calibration and a locally smoothed curve, consistent with contemporary model-evaluation guidance [8, 13]. AUROC confidence intervals (CIs) and staged AUROC differences used DeLong's method [10]. Paired DeLong tests compared +week 1 versus baseline, +week 2 versus +week 1, and +week 2 versus baseline; Holm adjustment was applied once across all nine comparisons. AP, Brier, log loss and calibration confidence intervals used 500 case-bootstrap samples of fixed mean out-of-fold predictions. These intervals were conditional on the realised cross-validated predictions and did not fully propagate uncertainty from the complete modelling workflow, including fold allocation, preprocessing, tuning and refitting; they should therefore be interpreted as conditional rather than total uncertainty [22].

## Clinical utility and parsimony

Decision-curve analysis evaluated the hypothetical action of treatment intensification when predicted non-remission risk exceeded thresholds from 0.10 to 0.70 and compared the models with intensify-all and intensify-none strategies [11]. Curves had no resampling confidence bands and were exploratory. To test parsimony, we compared the full week-2 model with a model containing only week-2 HAMD-17, a baseline/week-1/week-2 HAMD trajectory model and a model excluding HAMD-derived predictors. Paired DeLong tests used Holm adjustment across all 15 same-scale comparisons. Incremental value was interpreted jointly across discrimination, calibration, overall error and decision analysis rather than from statistical significance alone [23].

# Within-study transportability and sensitivity analyses

Leave-one-centre-out internal-external cross-validation (IECV) trained on four centres and generated locked predictions in the fifth; centre was omitted as a predictor and tuning was centre based. Leave-one-treatment-arm-out analysis omitted treatment group, retained centre, trained on four arms and evaluated the fifth. A within-centre date-field-ordered sensitivity analysis used the earliest 70% of MHSTDT values at the week-2 landmark for training and the latest 30% for evaluation. Because the definition and range of MHSTDT require reconciliation with the recruitment log, this analysis is not interpreted as confirmed temporal validation. These were all within-trial analyses, not independent external validation. Conditional confidence intervals used 2000 stratified bootstrap samples of fixed

validation predictions. The item-14 sensitivity repeated centre IECV after treating code 3 as missing. Analyses used R 4.4.1 with seed 20260720.

# Patient and public involvement

No patient or public involvement was documented for this secondary prediction analysis.

# Use of AI-assisted tools

OpenAI Codex was used to assist with statistical-programming review, visualisation, language drafting and document preparation. It was not listed as an author and did not make autonomous decisions about study design, analysis, interpretation or submission. The authors independently reviewed and verified all analyses, references and text and accept full responsibility for the work.

## Results

# Cohort and early symptom change

Among 338 participants at the week-2 landmark, 306 had an observed week-6 outcome, 269 had an observed week-10 outcome and 258 had both outcomes. Week-6 remission occurred in 81/306 (26.5%), week-10 remission in 94/269 (34.9%), and two-point remission in 47/258 (18.2%). Baseline and landmark-cohort characteristics are shown in Table 1. Outcome availability was 90.5%, 79.6% and 76.3%, respectively (Additional file 1: Table S1).

Week-6 remission followed a descriptive early-response gradient: 16/139 (11.5%) among participants with < 20% week-2 HAMD reduction, 43/127 (33.9%) with 20%-49% reduction and 22/40 (55.0%) with > = 50% reduction (Fig. 2). These categories were descriptive rather than validated decision thresholds.

Table 1
Characteristics of the enrolled and week-2 landmark cohorts

| Characteristic | Enrolled (N = 347) | Week-2 landmark (N = 338) |
|---|---|---|
| Participants, n | 347 | 338 |
| Age, years | 35.9 (10.5) | 35.9 (10.6) |
| Female sex | 205 (59.2%) | 198 (58.8%) |
| Married | 259 (74.9%) | 251 (74.5%) |
| Han ethnicity | 343 (98.8%) | 334 (98.8%) |
| Body mass index, kg/m2 | 21.8 (4.2) | 21.8 (4.2) |
| Episode duration, months | 11.0 (5.0–24.0) | 12.0 (4.2–24.0) |
| Baseline HAMD-17 | 22.0 (4.9) | 22.0 (4.9) |
| Baseline SDS index | 60.0 (8.6) | 60.0 (8.6) |
| Centre 1 | 160 (46.1%) | 158 (46.7%) |
| Centre 2 | 67 (19.3%) | 66 (19.5%) |
| Centre 3 | 30 (8.6%) | 29 (8.6%) |
| Centre 4 | 40 (11.5%) | 36 (10.7%) |
| Centre 5 | 50 (14.4%) | 49 (14.5%) |
| Treatment arm A: medication control | 68 (19.6%) | 64 (18.9%) |
| Treatment arm B: manual acupuncture | 70 (20.2%) | 69 (20.4%) |
| Treatment arm C: manual acupuncture plus medication | 69 (19.9%) | 68 (20.1%) |
| Treatment arm D: electroacupuncture | 70 (20.2%) | 67 (19.8%) |
| Treatment arm E: electroacupuncture plus medication | 70 (20.2%) | 70 (20.7%) |
| Missing episode duration | 241 (69.5%) | 236 (69.8%) |

**Note**: Values are mean (standard deviation [SD]), median (interquartile range [IQR]), or n (%). Percentages use available data. Episode duration was conservatively parsed from mixed-unit free text and was missing for 69.5% of participants. Ethnicity and treatment labels must be checked against the case-report form before submission. No significance tests compare the cohorts.

# Sequential predictive performance

Discrimination, AP and Brier score improved as early-treatment information was added (Table 2; Fig. 3; Additional file 1: Figure S1). AUROC for baseline, +week 1 and +week 2 was 0.613, 0.741 and 0.833 for

week-6 remission; 0.578, 0.671 and 0.742 for week-10 remission; and 0.563, 0.770 and 0.828 for two-point remission. Week-2 calibration slopes were 1.013, 0.948 and 0.838, respectively, although confidence intervals were wide (Fig. 4).

Table 2
Sequential performance and adjacent incremental value

| Outcome; N/events | Baseline AUROC (95% CI) | Week-1 AUROC (95% CI); adjacent delta; Holm P | Week-2 AUROC (95% CI); adjacent delta; Holm P | Final AP; Brier; calibration slope |
|---|---|---|---|---|
| Week-6 remission<br>N/events 306/81 | 0.613 (0.537 to 0.691) | 0.741 (0.672 to 0.804)<br>Delta 0.128; Holm P = 0.0015 | 0.833 (0.781 to 0.877)<br>Delta 0.093; Holm P = 0.0016 | AP 0.600; Brier 0.145; slope 1.013 |
| Week-10 remission<br>N/events 269/94 | 0.578 (0.498 to 0.658) | 0.671 (0.600 to 0.743)<br>Delta 0.093; Holm P = 0.0222 | 0.742 (0.673 to 0.802)<br>Delta 0.070; Holm P = 0.0222 | AP 0.568; Brier 0.194; slope 0.948 |
| Two-point remission<br>N/events 258/47 | 0.563 (0.467 to 0.668) | 0.770 (0.693 to 0.845)<br>Delta 0.207; Holm P = 0.000631 | 0.828 (0.771 to 0.886)<br>Delta 0.058; Holm P = 0.0248 | AP 0.434; Brier 0.127; slope 0.838 |

**Note**: Each participant contributed one mean out-of-fold probability. AUROC intervals and adjacent differences use paired DeLong methods; Holm adjustment was applied across all nine staged comparisons. Final-model AP, Brier score and calibration slope are reported for the week-2 model. AP=average precision.

# Paired incremental value

Paired DeLong tests supported incremental discrimination at both early landmarks (Table 2 and Additional file 1: Table S13). For week-6 remission, +week 1 versus baseline increased AUROC by 0.128 (95% CI 0.059 to 0.197; Holm P = 0.00148), and +week 2 versus +week 1 added 0.093 (0.041 to 0.144; Holm P = 0.00155). Corresponding increments were 0.093 and 0.070 for week 10, and 0.207 and 0.058 for two-point remission. All nine staged comparisons remained below 0.05 after Holm correction.

# Parsimony, calibration and clinical utility

The full model did not show a statistically demonstrable AUROC gain over week-2 HAMD-17 alone (Additional file 1: Figure S2 and Tables S5-S6). Paired AUROC differences were 0.003 (95% CI -0.025 to 0.031; raw P = 0.839), 0.033 (-0.015 to 0.081; P = 0.178) and 0.040 (-0.013 to 0.094; P = 0.139); all Holm-adjusted P values were 1.00. The HAMD trajectory model likewise offered little observed gain over the current week-2 score. The confidence intervals are not equivalence intervals: they largely exclude an important week-6 gain but remain compatible with modest benefit for the later outcomes. Week-2

HAMD-17 therefore captured nearly all week-6 discrimination and much of the discrimination for later outcomes.

Decision curves suggested higher net benefit for the week-2 model than baseline and default strategies over portions of the evaluated range, most clearly for week-6 non-remission (Fig. 5). Because thresholds were not linked to a prespecified action, the effect of intensification was not estimated and net-benefit uncertainty was not quantified, these results are an exploratory decision-analytic signal rather than demonstrated clinical usefulness or impact.

# Within-study transportability

Centre IECV micro-pooled AUROCs were 0.839 (95% CI 0.789 to 0.885), 0.720 (0.654 to 0.781) and 0.825 (0.766 to 0.879) for week-6, week-10 and two-point remission (Fig. 4; Additional file 1: Tables S7-S9 and Figure S3). Leave-one-treatment-arm-out AUROCs were lower for later outcomes: 0.787, 0.626 and 0.590. In the date-field-ordered holdout sensitivity analysis, AUROCs were 0.838, 0.694 and 0.839, while remission was overpredicted in the records ordered later by MHSTDT, with observed-to-expected ratios of 0.820, 0.571 and 0.696. These figures describe the implemented ordering analysis but should not be interpreted as calendar-time transport until MHSTDT is verified. All analyses used the same trial and do not replace independent external validation.

# Sensitivity analyses

Treating HAMD item-14 code 3 as missing changed centre-IECV AUROC from 0.839 to 0.818 for week-6 remission, from 0.720 to 0.726 for week-10 remission and from 0.825 to 0.809 for two-point remission (Additional file 1: Figure S4 and Table S10). Omitting the highly incomplete episode-duration predictor did not materially change performance. These analyses support directional robustness but do not resolve the source-data definition or informative outcome missingness.

## Discussion

In this week-2 landmark analysis of a five-centre randomised treatment trial, reassessment after 1 and 2 weeks provided additional prognostic information for subsequent HAMD-17-defined remission beyond variables available at baseline. The staged AUROC increments were estimated in identical outcome-specific cohorts and remained statistically detectable after family-wise multiplicity correction. The estimand is nevertheless conditional: prognosis was updated only for participants who remained observed through week 2 and had the relevant later outcome recorded. It is not a baseline prognosis for all randomised participants, and AUROCs across the three outcomes should not be compared as if they arose from one common cohort. Week-6 remission was the only outcome for which discrimination remained comparatively strong across centre, treatment-arm and date-ordered stress tests; later outcomes were more context dependent. The apparent time-related miscalibration remains provisional until the ordering field is verified.

The contribution extends rather than rediscovers the established association between early improvement and later outcome [3–6]. The staged design asks a more specific question: how much does each new clinical measurement add to information already available at the preceding landmark? Early symptom scores may integrate initial treatment exposure, adherence, tolerability, spontaneous fluctuation and measurement context, but the present data cannot separate those mechanisms. Treatment assignment was used only as a known prognostic input after randomisation; the models estimate risk under the observed treatment pathway and neither identify individual treatment effects nor select the treatment that would improve an individual patient's outcome.

Evidence is not uniform regarding the earliest clinically informative reassessment. Furukawa and colleagues reported improved validation performance as week-1 and week-3 information accrued [6], whereas an individual-participant-data meta-analysis found that early non-improvement did not rule out subsequent response and became less informative for more distant outcomes [6]. In GENDEP, the latest symptom measurement generally outperformed more elaborate longitudinal summaries, but clinically useful separation often emerged later than week 2 [18]. By contrast, cross-trial analyses found that week-2 depression severity could generalise at least as well as more complex pretreatment clinical and neuroimaging models [16]. Differences in treatment composition, outcome horizon, scale, case mix and validation design probably explain part of this variation. Accordingly, a high week-2 risk should trigger structured review, not automatic discontinuation or switching.

Week-2 HAMD-17 did not constitute conventional outcome leakage: it preceded week-6 and week-10 assessments, and no post-landmark data entered model development. However, predictor and outcome quantified the same construct on the same scale at neighbouring times, creating strong construct continuity and temporal autocorrelation [21]. This is clinically legitimate for dynamic prognosis but probably contributes to the high short-horizon discrimination. The results therefore concern later HAMD-17-defined status and should not be assumed to generalise to functional recovery, quality of life or patient-reported recovery without separate evaluation. Likewise, two-point remission means remission observed at two visits, not uninterrupted remission between them.

The same-scale benchmark places algorithmic complexity in perspective. Relative to week-2 HAMD-17 alone, the full model increased AUROC by 0.003 for week-6 remission, 0.033 for week-10 remission and 0.040 for two-point remission. All confidence intervals crossed zero, but absence of statistical significance is not formal equivalence. The week-6 interval excluded more than a small gain, whereas the later-outcome intervals remained compatible with gains of approximately 0.08–0.09. The appropriate conclusion is therefore that current symptom severity captured nearly all observed week-6 discrimination and much of the discrimination for later outcomes, while modest later benefit from additional predictors cannot be excluded. Consistent findings from longitudinal and multimodal depression studies suggest that a latest symptom score can be difficult for complex trajectories or high-dimensional features to beat [15, 18]. Any more complex model should demonstrate reproducible improvement in calibration, overall error, net benefit or transportability as well as discrimination [23].

The validation analyses addressed different dimensions of transport and should not be treated as interchangeable external validations. Centre IECV examined location heterogeneity under a shared protocol; it was encouraging for week 6, but some held-out centres had few events. Leave-one-treatment-arm-out analysis evaluated a modified specification because treatment group was omitted, and its micro-pooled AUROC combined predictions from five different training models; weaker later performance cannot be attributed solely to treatment-effect heterogeneity. The date-field-ordered sensitivity analysis preserved ranking for some outcomes but overpredicted remission in records ordered later by MHSTDT. If the field is confirmed as enrolment date, this would illustrate that discrimination can persist while absolute risk becomes clinically misleading; until then it remains a source-data sensitivity result [8, 12, 13]. Recent review evidence likewise shows wide performance variation and few methodologically robust external validations for mental-health treatment prediction models [24, 25].

Decision curves provided an exploratory decision-analytic signal rather than evidence that model-guided care improves outcome. A high predicted probability of non-remission identifies prognostic risk but does not establish that the risk is modifiable, that treatment intensification is effective, or that its benefits exceed adverse effects and resource costs. The analysis also lacked a prespecified user, action, threshold and capacity pathway, and the curves had no uncertainty bands [11, 13]. Measurement-based-care trials indicate that repeated assessment can be beneficial when explicitly connected to protocolised treatment review [26], but that evidence concerns the whole care pathway rather than the predictive value or impact of this model. The next step is therefore prospective comparison of a prespecified model-informed review pathway with structured measurement-based care alone, after independent validation and recalibration.

Finally, consistent with study-investigator confirmation and record-level concordance, Fu et al. [20] and Qu et al. [19] analysed selected subsets of this same 347-participant source database and do not provide replication or external validation. The distinct contribution of the present work is analytical: use of the full supplied source cohort, an explicit week-2 landmark, sequential assessment of incremental information, direct comparison with a contemporaneous same-scale benchmark, three later outcomes and complementary within-trial transport analyses. Claims of novelty should therefore be confined to the prediction question and analytical framework rather than the population. Transparent cross-referencing of overlapping reports and their selection rules is essential.

# Strengths and limitations

Strengths include an explicit prediction landmark, restriction of predictors to information available by that landmark, identical outcome-specific cohorts and fold assignments for staged comparisons, fold-specific preprocessing and tuning, penalised estimation, and one aggregated out-of-fold prediction per participant. The study assessed discrimination, calibration and overall error; used paired comparisons with multiplicity control; challenged the complex model with a same-scale parsimonious benchmark; and examined heterogeneity across centres and treatment arms, with an additional date-field-ordered

sensitivity analysis. These features address common sources of optimistic incremental-value claims [7, 8, 13].

The effective sample size was nevertheless limited, particularly for two-point remission, for which 47 events supported a model containing multiple categorical terms and correlated repeated-score and change variables. Ridge penalisation and nested cross-validation reduce but do not eliminate overfitting or instability, and no development sample-size calculation was undertaken [9]. The comparatively high two-point AUROC should therefore be interpreted alongside its event count, selected complete-follow-up cohort and uncertainty, rather than as evidence that this outcome is intrinsically easier to predict.

Outcomes were analysed as complete cases. Of 338 participants reaching the landmark, 9.5%, 20.4% and 23.7% lacked the week-6, week-10 and two-point outcomes, respectively. If follow-up depended on early improvement, tolerability, treatment arm, centre or unmeasured prognosis, event rates and predictive relationships may be biased; the models currently apply most directly to participants with observed follow-up. The study did not compare retained and lost participants or perform inverse-probability weighting, multiple imputation or departures-from-missing-at-random sensitivity analyses. Episode duration was missing for 69.5%; fold-specific imputation prevented information leakage but could not recreate unrecorded prognostic information.

No independent geographic or contemporary cohort was available. Centre and treatment-arm analyses remained within one parent trial, the treatment-arm analysis used a modified predictor specification, and some centre, arm and date-ordered partitions contained few events. DeLong and bootstrap intervals treated the realised out-of-fold or held-out predictions as fixed and did not fully propagate uncertainty from sample selection, fold allocation, preprocessing, tuning and refitting; the resulting intervals should be interpreted conditionally [22]. The averaged repeated-cross-validation predictions also do not correspond to a single locked model ready for deployment. Generalisability across health systems and current treatment pathways, as well as fairness, clinical utility and impact, remain to be established [24].

Source-data limitations also remain material. The HAMD item-14 code requires confirmation from the original scoring manual or case-report form. Published recruitment dates and eligibility thresholds require reconciliation with the supplied database before submission. The cohort was predominantly Han Chinese and subgroup counts were insufficient for a reliable fairness analysis. Treatment assignment was a main-model prognostic predictor, so the model cannot support causal treatment selection.

# Conclusions

Sequential symptom reassessment at weeks 1 and 2 improved conditional within-trial prediction of later HAMD-17-defined remission beyond baseline information. A single week-2 HAMD-17 score captured nearly all week-6 discrimination and much of the discrimination for later outcomes; the study found no statistically demonstrable incremental AUROC from the full dynamic model, although modest benefit for the later outcomes cannot be excluded. These findings nominate a simple, standardised week-2 reassessment as the first candidate for prospective evaluation, not as a model-guided treatment rule

ready for use. Outcome-missingness sensitivity analyses, source-definition verification, independent external validation, recalibration and a prespecified impact study are required before clinical implementation.

## Abbreviations

AP, average precision; AUROC, area under the receiver operating characteristic curve; CI, confidence interval; DCA, decision-curve analysis; HAMD-17, 17-item Hamilton Depression Rating Scale; IECV, internal-external cross-validation; IQR, interquartile range; MHSTDT, source date field with an unavailable original definition; O:E, observed-to-expected ratio; ROC, receiver operating characteristic; SD, standard deviation; SDS, Self-rating Depression Scale.

## Declarations

# Ethics approval and consent to participate

The study was approved by the Medical Ethics Committee of Southern Medical University Hospital of Integrated Traditional Chinese and Western Medicine (approval number 201105002, approved on 29 May 2011), and all participants provided written informed consent. The study was conducted in accordance with the Declaration of Helsinki. The ethics approval has been independently verified to cover all five centres, five treatment arms, and all 347 included records.

# Consent for publication

Not applicable. This report contains no identifiable individual participant information.

# Availability of data and materials

The participant-level data are not publicly available because the dataset contains clinical-trial data governed by the source-study investigators and applicable institutional and ethics approvals. Deidentified data, the data dictionary and analysis code may be available from the corresponding author on reasonable request, subject to permission from the source-study investigators and applicable governance. The authors must confirm this access process before submission.

# Competing interests

The authors declare that they have no competing interests.

# Funding

This work was supported by the Clinical collaboration between traditional Chinese medicine and Western medicine in the treatment of major and intractable diseases (ZDYN-2024-A-120), the National Natural Science Foundation of China (82374193, 82405071, 82574754), the Guangdong Basic and Applied Basic Research Foundation (2023A1515220030) and the second national famous Traditional Chinese Medicine Inheritance Studio (G724290126).

# Authors' contributions

Concept and design: J-MZ, C-LZ, ZY, HY and JH. Acquisition of data: J-MZ, C-LZ, HY, W-JJ and JF. Analysis and interpretation: J-MZ, ZY, HY and WX. Substantial contributions to development of methodology: J-MZ, C-LZ, ZY and JF. Statistical and computational analysis of data: C-LZ, ZY, HY and JF. Reviewed and performed revision of the manuscript: J-MZ, C-LZ, ZY, HY, JF, WX and JH. Full access to all the data and verification of the underlying data: JF, WX and JH. All authors critically reviewed the manuscript for important intellectual content, revised it, and approved the final version. JH is the guarantor. All authors agree to be accountable for all aspects of the work and have read and approved the final manuscript.

# Acknowledgements

The authors thank all participants who contributed to this study and the clinical staff for their assistance during data collection.

# Prior publication and participant overlap

Study-investigator confirmation and record-level concordance indicate that this source database has been used in prior publications with different selection rules; the publications alone do not establish the dataset relationship. Fu et al. [20] analysed 124 participants selected from the manual-acupuncture and electroacupuncture monotherapy arms who had observed week-6 outcomes, using baseline variables to predict acupuncture efficacy. Qu et al. [19] reported a selected 160-participant, three-centre, three-arm treatment-efficacy analysis. The present study uses the full 347-participant source cohort, applies a week-2 prognostic landmark, evaluates week-6, week-10 and two-point remission, and uses staged incremental-value, parsimony and within-trial transportability analyses. It is a distinct secondary analysis, but not an independent cohort or external validation of either earlier report.

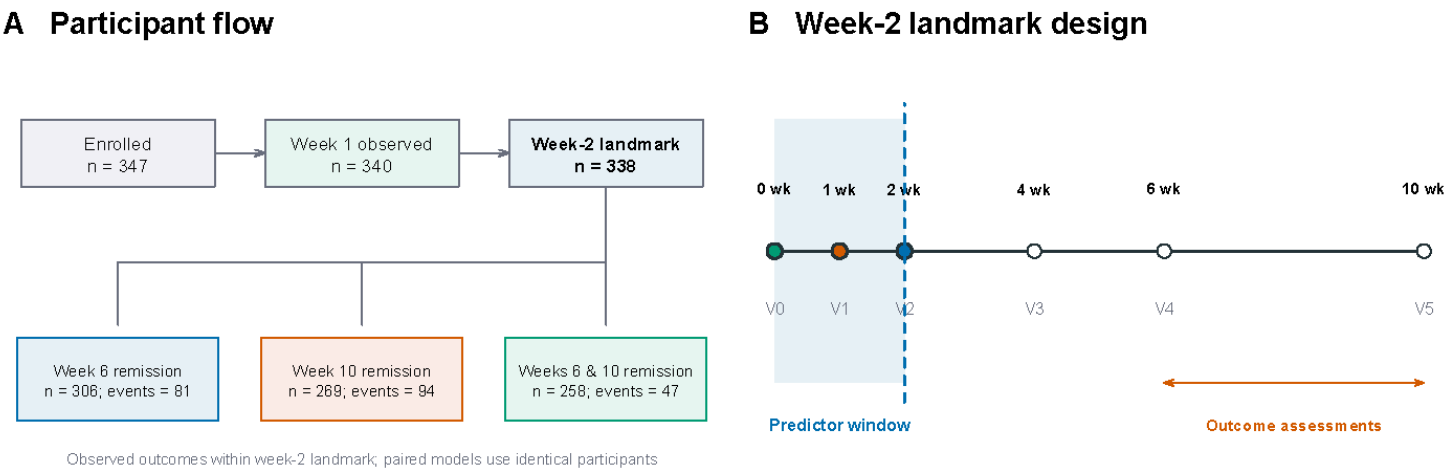
## References

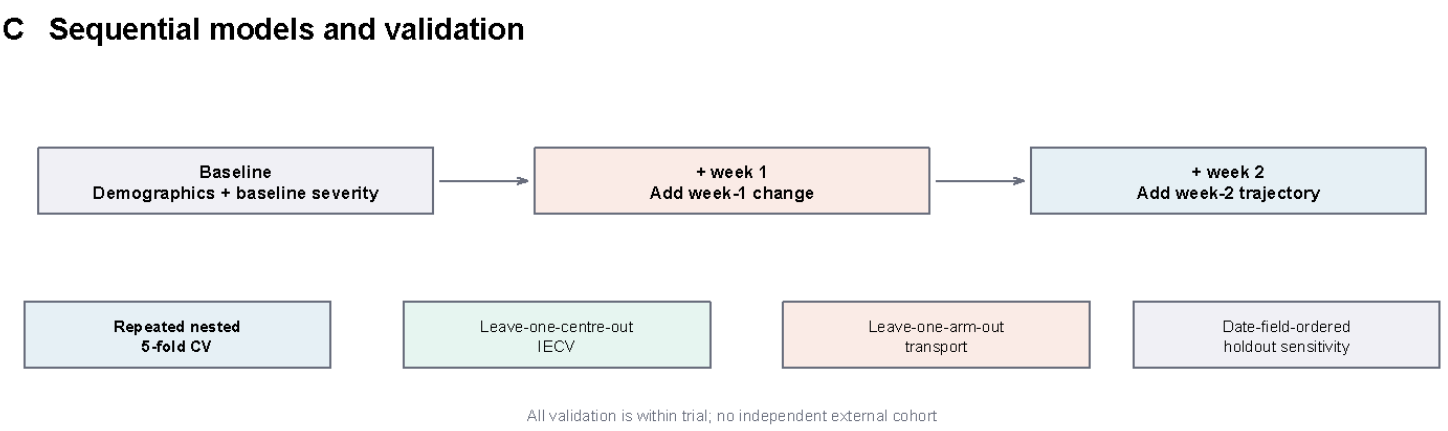
1. Chekroud AM, Zotti RJ, Shehzad Z, Gueorguieva R, Johnson MK, Trivedi MH, et al. Cross-trial prediction of treatment outcome in depression: a machine learning approach. Lancet Psychiatry. 2016;3(3):243–50. 10.1016/S2215-0366(15)00471-X.

2. Moriarty AS, Paton LW, Snell KIE, Archer L, Riley RD, Buckman JEJ, et al. Development and validation of a prognostic model to predict relapse in adults with remitted depression in primary care: secondary analysis of pooled individual participant data from multiple studies. BMJ Ment Health. 2024;27(1):e301226. 10.1136/bmjment-2024-301226.
3. Henkel V, Seemüller F, Obermeier M, Adli M, Bauer M, Mundt C, et al. Does early improvement triggered by antidepressants predict response/remission? Analysis of data from a naturalistic study on a large sample of inpatients with major depression. J Affect Disord. 2009;115(3):439–49. 10.1016/j.jad.2008.10.011.
4. Szegedi A, Jansen WT, van Willigenburg APP, van der Meulen E, Stassen HH, Thase ME. Early improvement in the first 2 weeks as a predictor of treatment outcome in patients with major depressive disorder: a meta-analysis including 6562 patients. J Clin Psychiatry. 2009;70(3):344–53. 10.4088/JCP.07m03780.
5. Hicks PB, Sevilimedu V, Johnson GR, Tal I, Chen P, Davis LL, et al. Predictability of nonremitting depression after first 2 weeks of antidepressant treatment: a VAST-D trial report. Psychiatr Res Clin Pract. 2019;1(2):58–67. 10.1176/appi.prcp.20190003.
6. de Vries YA, Roest AM, Bos EH, Burgerhof JGM, van Loo HM, de Jonge P. Predicting antidepressant response by monitoring early improvement of individual symptoms of depression: individual patient data meta-analysis. Br J Psychiatry. 2019;214(1):4–10. 10.1192/bjp.2018.122.
7. Collins GS, Moons KGM, Dhiman P, Riley RD, Beam AL, Van Calster B, 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.
8. Collins GS, Dhiman P, Ma J, Schlussel MM, Archer L, Van Calster B, et al. Evaluation of clinical prediction models (part 1): from development to external validation. BMJ. 2024;384:e074819. 10.1136/bmj-2023-074819.
9. Riley RD, Snell KIE, Ensor J, Burke DL, Harrell FE Jr, Moons KGM, et al. Minimum sample size for developing a multivariable prediction model: PART II-binary and time-to-event outcomes. Stat Med. 2019;38(7):1276–96. 10.1002/sim.7992.
10. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics. 1988;44(3):837–45. 10.2307/2531595.
11. 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.
12. Van Calster B, McLernon DJ, van Smeden M, Wynants L, Steyerberg EW. Topic Group 'Evaluating diagnostic tests and prediction models' of the STRATOS initiative. Calibration: the Achilles heel of predictive analytics. BMC Med. 2019;17(1):230. 10.1186/s12916-019-1466-7.
13. Riley RD, Archer L, Snell KIE, Ensor J, Dhiman P, Martin GP, et al. Evaluation of clinical prediction models (part 2): how to undertake an external validation study. BMJ. 2024;384:e074820. 10.1136/bmj-2023-074820.

14. Salem H, Huynh T, Topolski N, Mwangi B, Trivedi MH, Soares JC, et al. Temporal multi-step predictive modeling of remission in major depressive disorder using early stage treatment data; STAR*D based machine learning approach. J Affect Disord. 2023;324:286–93. 10.1016/j.jad.2022.12.076.
15. Sajjadian M, Uher R, Ho K, Hassel S, Milev R, Frey BN, et al. Prediction of depression treatment outcome from multimodal data: a CAN-BIND-1 report. Psychol Med. 2023;53(12):5374–84. 10.1017/S0033291722002124.
16. Zhukovsky P, Trivedi MH, Weissman M, Parsey R, Kennedy S, Pizzagalli DA. Generalizability of treatment outcome prediction across antidepressant treatment trials in depression. JAMA Netw Open. 2025;8(3):e251310. 10.1001/jamanetworkopen.2025.1310.
17. Furukawa TA, Kato T, Shinagawa Y, Miki K, Fujita H, Tsujino N, et al. Prediction of remission in pharmacotherapy of untreated major depression: development and validation of multivariable prediction models. Psychol Med. 2019;49(14):2405–13. 10.1017/S0033291718003331.
18. Carr E, Rietschel M, Mors O, Henigsberg N, Aitchison KJ, Maier W, et al. Optimizing the prediction of depression remission: a longitudinal machine learning approach. Am J Med Genet B Neuropsychiatr Genet. 2025;198(3):e33014. 10.1002/ajmg.b.33014.
19. Qu SS, Huang Y, Zhang ZJ, Chen JQ, Lin RY, Wang CQ, et al. A 6-week randomized controlled trial with 4-week follow-up of acupuncture combined with paroxetine in patients with major depressive disorder. J Psychiatr Res. 2013;47(6):726–32. 10.1016/j.jpsychires.2013.02.004.
20. Fu J, Cai X, Huang S, Qiu X, Li Z, Hong H, et al. Predicting acupuncture efficacy for major depressive disorder using baseline clinical variables: a machine learning study. J Psychiatr Res. 2023;168:64–70. 10.1016/j.jpsychires.2023.10.040.
21. Bull LM, Lunt M, Martin GP, Hyrich K, Sergeant JC. Harnessing repeated measurements of predictor variables for clinical risk prediction: a review of existing methods. Diagn Progn Res. 2020;4:9. 10.1186/s41512-020-00078-z.
22. Riley RD, Collins GS, Kirton L, Snell KIE, Ensor J, Whittle R, et al. Uncertainty of risk estimates from clinical prediction models: rationale, challenges, and approaches. BMJ. 2025;388:e080749. 10.1136/bmj-2024-080749.
23. Cook NR. Quantifying the added value of new biomarkers: how and how not. Diagn Progn Res. 2018;2:14. 10.1186/s41512-018-0037-2.
24. Burghoorn DG, Booij SH, Schoevers RA, Riese H. Externally validated clinical prediction models for estimating treatment outcomes for patients with a mood, anxiety or psychotic disorder: systematic review and meta-analysis. BJPsych Open. 2024;10(6):e221. 10.1192/bjo.2024.789.
25. van Dijk DA, Stoel NG, Meijer RJ, Repko RJE, van den Boogaard TM, Ruhé HG, et al. Clinical prediction instruments available for clinicians treating major depressive disorder: a systematic review. J Affect Disord. 2025;382:68–84. 10.1016/j.jad.2025.04.050.
26. Husain MI, Nigah Z, Ansari SUH, Khoso AB, Kiran T, Umer M, et al. Measurement-based care to enhance antidepressant treatment outcomes in major depressive disorder: a randomized clinical trial. JAMA Netw Open. 2025;8(9):e2529427. 10.1001/jamanetworkopen.2025.29427.

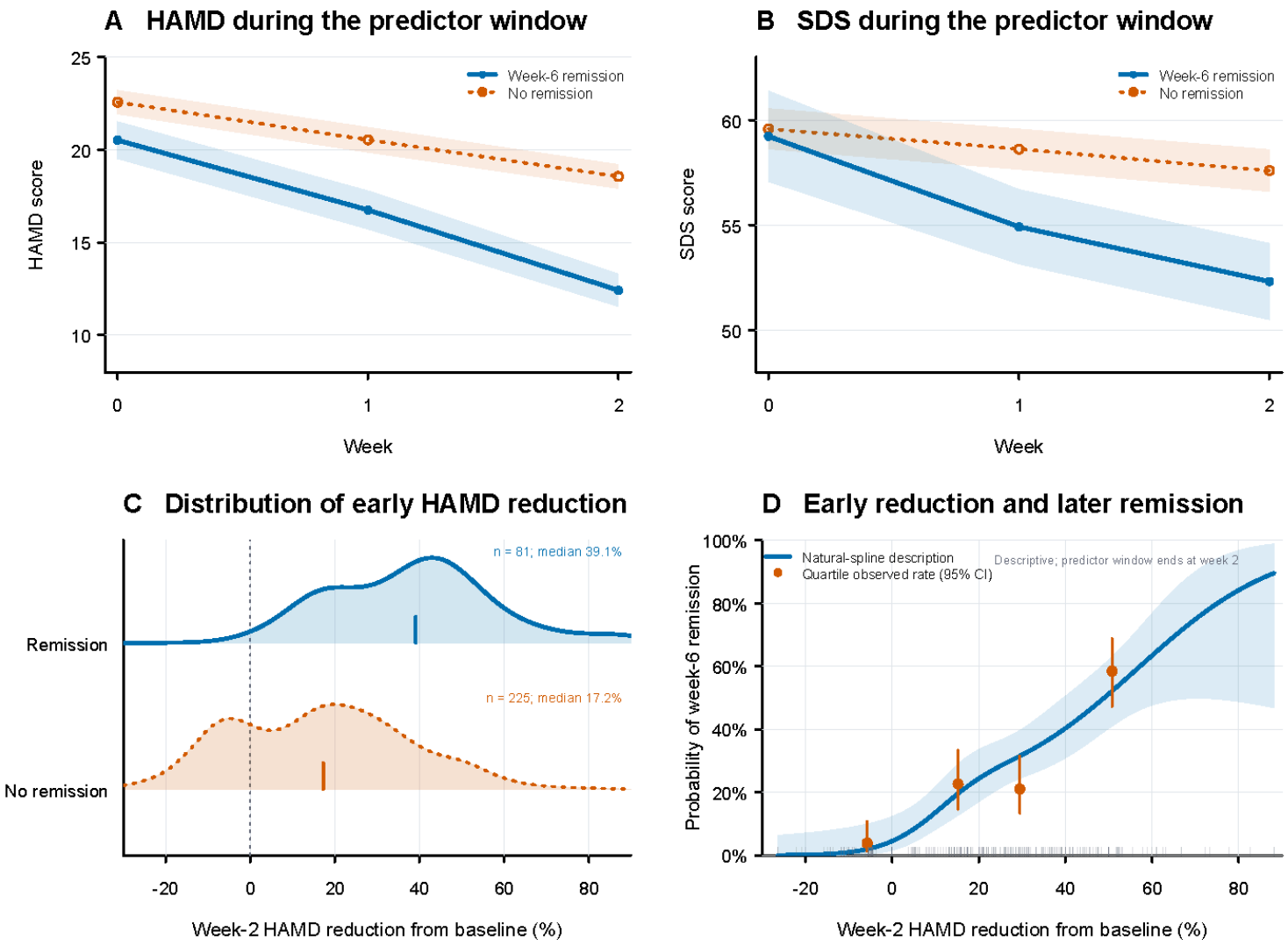
# Figures




## Figure 1

**Study design, week-2 landmark and validation framework.** Participant flow from 347 enrolled participants to 340 with week-1 data, 338 reaching the week-2 landmark, and outcome-specific complete cohorts of 306, 269 and 258 for week-6, week-10 and two-point remission, respectively. The latter counts are observed outcomes within the week-2 landmark cohort, not all raw records with an outcome. Predictor information was restricted to baseline through week 2. Validation comprised repeated nested cross-validation, leave-one-centre-out internal-external cross-validation, leave-one-treatment-arm-out analysis and a date-field-ordered holdout sensitivity analysis. All validation used the same trial and was not independent external validation.




## Figure 2

**Early symptom trajectories and association with week-6 remission.** Panels A and B show observed-case mean HAMD-17 and SDS scores with 95% confidence intervals from baseline to week 2, stratified by week-6 remission. Panel C shows the distribution of week-2 HAMD reduction in participants with and without week-6 remission. Panel D shows a descriptive natural-spline association and quartile-specific observed rates with Wilson 95% confidence intervals. These associations do not establish a causal effect or a validated treatment threshold.

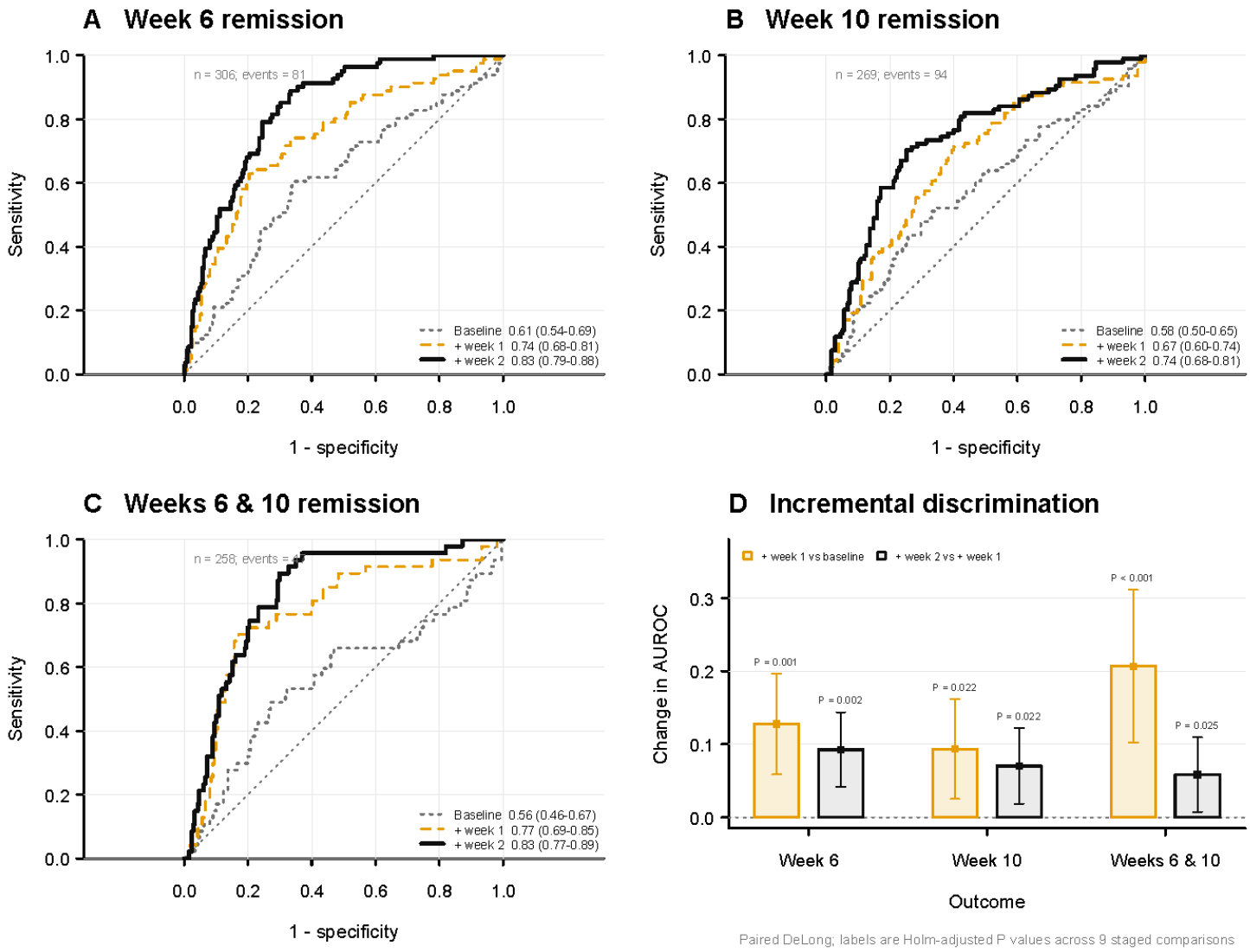
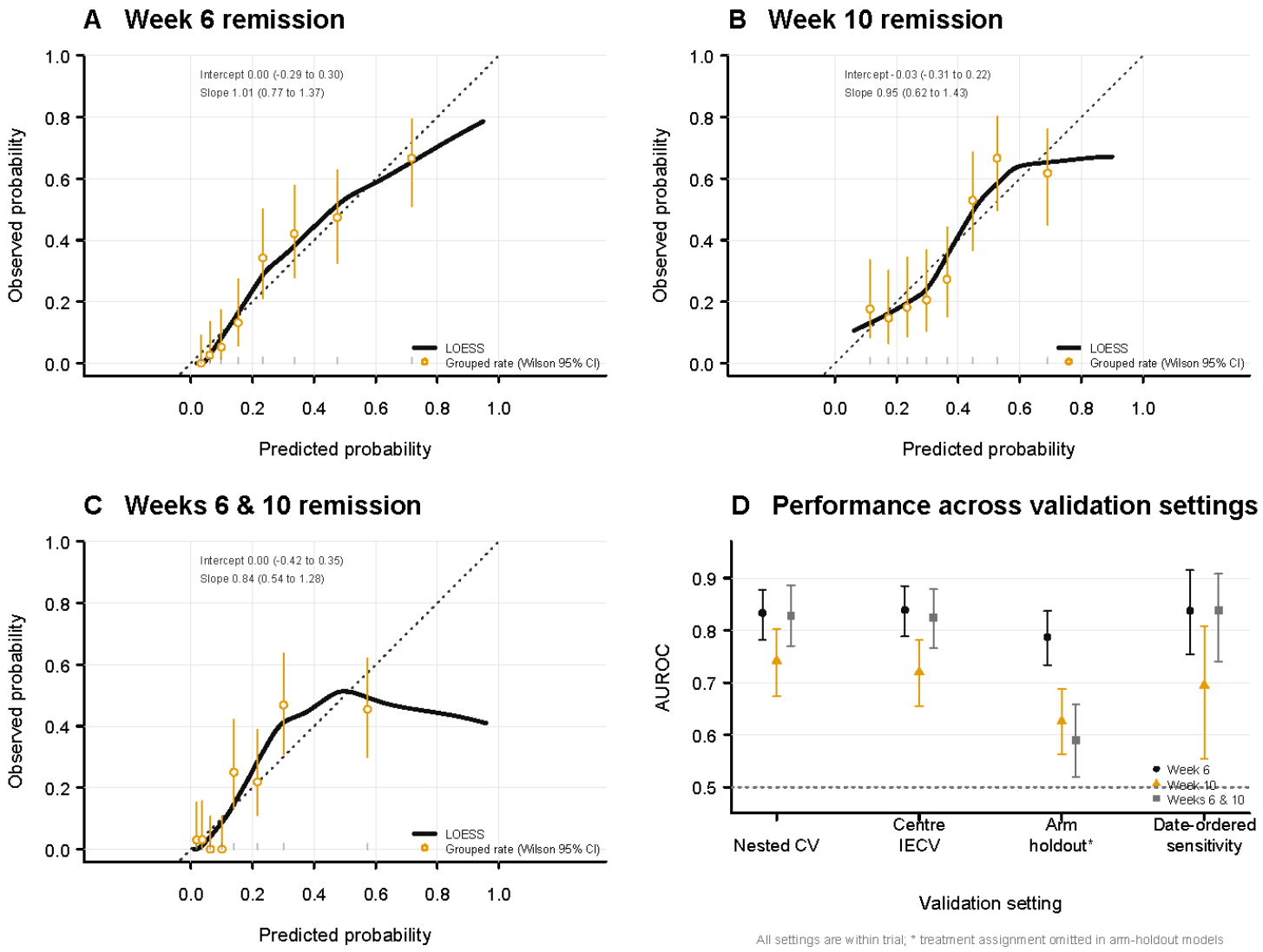



**Figure 3**

**Discrimination and incremental predictive value of sequential models.**Panels A-C show out-of-fold ROC curves for baseline, week-1-updated and week-2-updated models for week-6 remission, week-10 remission and two-point remission. Panel D shows paired DeLong differences in AUROC for +week 1 versus baseline and +week 2 versus +week 1. Holm adjustment was applied across all nine staged comparisons.




## Figure 4

**Calibration and within-trial transportability of the week-2 model.** Panels A-C show out-of-fold calibration for the week-2 model using locally smoothed curves and grouped observed proportions with Wilson 95% confidence intervals. Panel D summarises AUROC across repeated nested cross-validation, centre internal-external cross-validation, leave-one-treatment-arm-out analysis and the date-field-ordered sensitivity analysis. All settings used the same trial and do not constitute independent external validation.

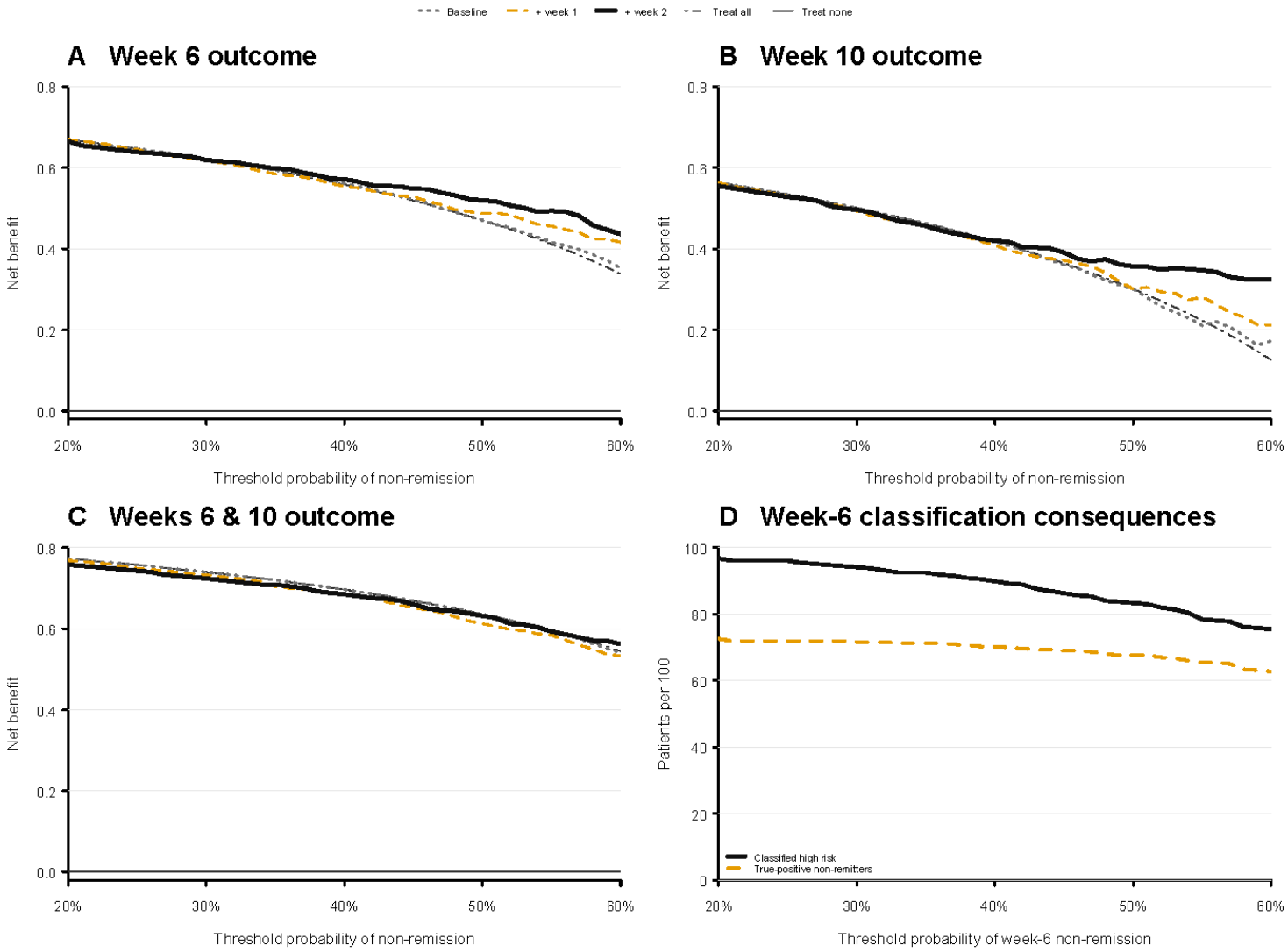



**Figure 5**

**Decision-curve analysis and classification consequences.** Panels A-C show decision curves for predicted non-remission at week 6, week 10 and both visits over displayed thresholds of 20%-60%; the analysis covered 10%-70%. Panel D shows, per 100 participants, those classified as high risk and true-positive non-remitters for week 6. Curves had no resampling confidence bands, thresholds were not linked to a prespecified action, and the analysis is an exploratory decision-analytic signal rather than evidence of clinical impact.

# Supplementary Files

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

- 04Additionalfile1.pdf