ADDITIONAL FILE 1

Incremental prognostic value of early symptom reassessment for later depression remission

This file contains the supplementary methods, Tables S1-S13 and Figures S1-S4 for the accompanying BMC Medicine Research article.

Contents

Supplementary Methods

Source-cohort provenance and analytical eligibility

Study-investigator confirmation and record-level concordance indicate that the reports by Qu et al. (2013) and Fu et al. (2023) analysed selected subsets of the same source database used here; the publications alone do not establish that relationship. Fu et al. selected 124 participants from the manual-acupuncture and electroacupuncture monotherapy groups with observed week-6 outcomes; Qu et al. reported a selected 160-participant, three-centre, three-arm treatment comparison. The present source cohort comprised all 347 supplied randomised records. Analytical eligibility then required week-1 and week-2 attendance for the week-2 landmark and an observed outcome for each outcome-specific complete-case cohort; outcomes were not imputed. Exact parent-protocol eligibility thresholds, recruitment dates and governance identifiers remain subject to reconciliation described in the main manuscript and title page.

The locked analysis input was the 347-record, 589-field SPSS source dataset used by the R pipeline (SHA-256 B04DD404A6947CA27D668572456B55416445ED5B147D72D503194380A64314FF). No other supplied database contributed values to any model, table or figure. This source separation was confirmed during the reproducibility audit.

Scoring and landmark construction

HAMD totals were computed from 17 items and SDS index scores from 20 items after reverse scoring items 2, 5, 6, 11, 12, 14, 16, 17, 18 and 20 and multiplying the raw sum by 1.25. HAMD attendance at a visit required at least 15 recorded HAMD items, while the total score required all 17 items. The main item-14 convention recoded code 3 to zero; the strict sensitivity convention treated code 3 as missing. The week-2 landmark required week-1 and week-2 attendance. Outcome-specific complete cases were then selected without imputing outcomes.

Feature engineering

Absolute change equalled later minus earlier score. Percentage change equalled absolute change divided by the baseline score with the denominator bounded below at 1. Baseline-to-week-1, baseline-to-week-2 and week-1-to-week-2 changes were calculated. Body mass index used weight in kilograms and height in metres after excluding height outside 120-220 cm, weight outside 30-200 kg and BMI outside 12-55 kg/m2. Episode duration was conservatively parsed to months from free text containing years, months, weeks, days or plain numeric values; values outside 0-600 months were missing.

Nested validation

For each outcome, common outer folds were used across staged and benchmark models. The inner loss was binomial log loss. Numeric median imputation, missing indicators, standardisation and categorical design matrices were all learned in training data. The custom ridge logistic model left the intercept unpenalised. Three outer repeats produced three held-out predictions per person, which were averaged before evaluation. The reported inference is conditional on these mean predictions.

Internal-external and transport analyses

Centre IECV omitted centre from the predictor matrix, tuned by leaving out each available training centre in turn and applied a locked pipeline to the held-out centre. Leave-one-arm-out analysis omitted treatment group to avoid encoding an unseen arm. The date-field-ordered holdout used the earliest 70% of MHSTDT values within centre for training and the latest 30% for evaluation, with all preprocessing and tuning confined to the earlier-ordered records. Because the field definition and range require reconciliation with the recruitment log, this is reported as a sensitivity analysis rather than confirmed temporal validation. These designs share a parent trial and are not independent external validation.

Statistical inference

DeLong confidence intervals and paired tests were used for AUROC. Holm correction was applied across nine staged comparisons and, separately, across 15 benchmark comparisons. AP, Brier, log loss and calibration confidence intervals used 500 case-bootstrap samples of fixed mean out-of-fold predictions. Validation intervals used 2000 stratified samples of fixed held-out predictions. DCA used non-remission as the hypothetical action event over thresholds of 0.10-0.70; figure 5 displays 0.20-0.60 for legibility. No confidence bands were calculated, and no prespecified intervention effect was assumed.

Supplementary Table S1. Participant flow, outcome availability and events

Stage/outcome	N	Events	Availability	Comment
Enrolled	347	-	-	Source cohort
Week-1 information available	340	-	98.0%	Predictor-window attendance
Week-2 landmark	338	-	97.4%	Week-1 and week-2 attendance
Week-6 outcome observed within landmark	306	81	90.5% of landmark	26.5% remission
Week-10 outcome observed within landmark	269	94	79.6% of landmark	34.9% remission

Stage/outcome	N	Events	Availability	Comment
Both outcomes observed within landmark	258	47	76.3% of landmark	18.2% two-point remission
Strict item-14: week 6 within landmark	285	81	84.3% of landmark	Code 3 treated as missing
Strict item-14: week 10 within landmark	250	90	74.0% of landmark	Code 3 treated as missing
Strict item-14: both visits within landmark	237	47	70.1% of landmark	Code 3 treated as missing

Note: Outcome events are remission events. Missing outcomes were excluded, not coded as non-remission. Before applying the common week-2 landmark restriction, week-10 outcome scoring yielded 273/95 under the primary item-14 convention and 254/91 under the strict convention; the model cohorts were 269/94 and 250/90, respectively. Week-6 and two-point counts were unchanged by the landmark restriction.

Supplementary Table S2. Predictor definitions and staged-model membership

Predictor	Available	Type	Model membership	Definition
Age	Baseline	Continuous	All	Years
Sex; marital status; ethnicity	Baseline	Categorical	All	One-hot encoded
Centre; randomised treatment group	Baseline	Categorical	All	Known by landmark; modified in transport analyses
Body mass index	Baseline	Continuous	All	kg/m2 after range checks
Episode duration	Baseline	Continuous	All	Parsed months; 69.5% missing
HAMD-17; SDS	Baseline	Continuous	All	Scale totals
Week-1 HAMD-17; SDS	Week 1	Continuous	+week 1; +week 2	Current scores
Baseline-to-week-1 HAMD/SDS change	Week 1	Continuous	+week 1; +week 2	Absolute and percentage
Week-2 HAMD-17; SDS	Week 2	Continuous	+week 2	Current scores
Baseline-to-week-2 HAMD/SDS change	Week 2	Continuous	+week 2	Absolute and percentage
Week-1-to-week-2 HAMD/SDS change	Week 2	Continuous	+week 2	Absolute

Note: No predictors measured after week 2 entered any model. Percentage change used baseline bounded below at 1 as denominator.

Supplementary Table S3. Core continuous-predictor missingness and preprocessing

Variable	Missing	Total	Percent	Rule
age	0	347	0.0%	Training-fold median imputation + indicator
disease_months	241	347	69.5%	Training-fold median imputation + indicator
bmi	5	347	1.4%	Training-fold median imputation + indicator
baseline_hamd	0	347	0.0%	Training-fold median imputation + indicator
baseline_sds	2	347	0.6%	Training-fold median imputation + indicator
hamd_v1	7	347	2.0%	Training-fold median imputation + indicator
sds_v1	7	347	2.0%	Training-fold median imputation + indicator
hamd_v2	9	347	2.6%	Training-fold median imputation + indicator
sds_v2	9	347	2.6%	Training-fold median imputation + indicator

Note: Imputation, scaling and indicator creation were fitted only in training folds. Outcomes were never imputed.

Supplementary Table S4. Full repeated nested cross-validated performance

Outcome	Model	N/events	AUROC	AP	Brier	Log loss	Cal. intercept	Cal. slope
Week-6 remission	Baseline	306/81	0.613 (0.537 to 0.691)	0.384 (0.293 to 0.494)	0.190 (0.166 to 0.212)	0.569 (0.512 to 0.621)	0.019 (-0.266 to 0.244)	0.835 (0.077 to 1.776)
Week-6 remission	+ week 1	306/81	0.741 (0.672 to 0.804)	0.497 (0.400 to 0.614)	0.168 (0.145 to 0.193)	0.512 (0.454 to 0.576)	0.009 (-0.239 to 0.274)	1.026 (0.675 to 1.497)
Week-6 remission	+ week 2	306/81	0.833 (0.781 to 0.877)	0.600 (0.493 to 0.716)	0.145 (0.122 to 0.169)	0.440 (0.383 to 0.502)	0.005 (-0.294 to 0.295)	1.013 (0.768 to 1.369)
Week-10 remission	Baseline	269/94	0.578 (0.498 to 0.658)	0.424 (0.349 to 0.531)	0.225 (0.206 to 0.244)	0.643 (0.603 to 0.682)	-0.026 (-0.285 to 0.240)	0.714 (-0.197 to 1.697)
Week-10 remission	+ week 1	269/94	0.671 (0.600 to 0.743)	0.490 (0.408 to 0.601)	0.213 (0.193 to 0.235)	0.619 (0.573 to 0.670)	-0.026 (-0.272 to 0.227)	0.921 (0.412 to 1.599)
Week-10 remission	+ week 2	269/94	0.742 (0.673 to 0.802)	0.568 (0.477 to 0.680)	0.194 (0.170 to 0.217)	0.576 (0.520 to 0.633)	-0.031 (-0.305 to 0.223)	0.948 (0.615 to 1.429)
Two-point remission	Baseline	258/47	0.563 (0.467 to 0.668)	0.236 (0.169 to 0.352)	0.149 (0.122 to 0.179)	0.479 (0.410 to 0.555)	0.007 (-0.324 to 0.293)	0.308 (-0.649 to 1.414)
Two-point remission	+ week 1	258/47	0.770 (0.693 to 0.845)	0.382 (0.293 to 0.521)	0.135 (0.107 to 0.163)	0.428 (0.356 to 0.507)	-0.002 (-0.379 to 0.309)	0.831 (0.474 to 1.423)
Two-point remission	+ week 2	258/47	0.828 (0.771 to 0.886)	0.434 (0.328 to 0.582)	0.127 (0.101 to 0.153)	0.396 (0.325 to 0.469)	0.001 (-0.423 to 0.350)	0.838 (0.538 to 1.284)

Note: AUROC intervals use DeLong. Other intervals use 500 case-bootstrap samples of fixed mean out-of-fold predictions.

Supplementary Table S5. Performance of same-scale benchmark models

Outcome	Model	N/events	AUROC	AP	Brier	Cal. slope
Week-6 remission	Week-2 HAMD only	306/81	0.830 (0.779 to 0.876)	0.592 (0.467 to 0.713)	0.145 (0.121 to 0.171)	1.022 (0.780 to 1.343)
Week-6 remission	HAMD trajectory	306/81	0.829 (0.781 to 0.875)	0.588 (0.483 to 0.711)	0.146 (0.125 to 0.170)	1.025 (0.804 to 1.368)
Week-6 remission	Non-HAMD	306/81	0.695 (0.632 to 0.762)	0.422 (0.337 to 0.538)	0.182 (0.159 to 0.209)	0.865 (0.480 to 1.384)
Week-6 remission	Full dynamic	306/81	0.833 (0.788 to 0.878)	0.600 (0.501 to 0.703)	0.145 (0.123 to 0.170)	1.013 (0.808 to 1.316)
Week-10 remission	Week-2 HAMD only	269/94	0.709 (0.649 to 0.771)	0.519 (0.427 to 0.616)	0.204 (0.181 to 0.223)	1.014 (0.679 to 1.455)
Week-10 remission	HAMD trajectory	269/94	0.712 (0.650 to 0.778)	0.527 (0.422 to 0.633)	0.203 (0.181 to 0.226)	0.926 (0.602 to 1.370)
Week-10 remission	Non-HAMD	269/94	0.682 (0.615 to 0.751)	0.521 (0.430 to 0.644)	0.209 (0.188 to 0.231)	0.918 (0.499 to 1.566)
Week-10 remission	Full dynamic	269/94	0.742 (0.681 to 0.802)	0.568 (0.481 to 0.688)	0.194 (0.173 to 0.215)	0.948 (0.629 to 1.415)

Outcome	Model	N/events	AUROC	AP	Brier	Cal. slope
Two-point remission	Week-2 HAMD only	258/47	0.788 (0.725 to 0.846)	0.397 (0.291 to 0.537)	0.131 (0.106 to 0.157)	1.016 (0.734 to 1.426)
Two-point remission	HAMD trajectory	258/47	0.790 (0.730 to 0.849)	0.398 (0.281 to 0.550)	0.131 (0.106 to 0.157)	0.988 (0.689 to 1.380)
Two-point remission	Non-HAMD	258/47	0.753 (0.678 to 0.830)	0.369 (0.279 to 0.509)	0.138 (0.109 to 0.164)	1.001 (0.548 to 1.704)
Two-point remission	Full dynamic	258/47	0.828 (0.761 to 0.884)	0.434 (0.336 to 0.592)	0.127 (0.102 to 0.155)	0.838 (0.525 to 1.322)

Note: Models used identical outcome-specific cohorts and fold assignments.

Supplementary Table S6. Paired DeLong comparisons among same-scale benchmark models

Outcome	Comparison	Difference in AUROC (95% CI)	Raw P	Holm-adjusted P
Week-6 remission	HAMD trajectory vs week-2 HAMD	-0.001 (-0.010 to 0.008)	0.7985	1
Week-6 remission	Non-HAMD vs week-2 HAMD	-0.135 (-0.212 to -0.058)	0.000587	0.0082
Week-6 remission	Full vs week-2 HAMD	0.003 (-0.025 to 0.031)	0.839	1
Week-6 remission	Full vs HAMD trajectory	0.004 (-0.022 to 0.031)	0.7617	1
Week-6 remission	Full vs non-HAMD	0.138 (0.078 to 0.198)	5.92e-06	8.87e-05
Week-10 remission	HAMD trajectory vs week-2 HAMD	0.003 (-0.020 to 0.027)	0.7705	1
Week-10 remission	Non-HAMD vs week-2 HAMD	-0.026 (-0.109 to 0.056)	0.5299	1
Week-10 remission	Full vs week-2 HAMD	0.033 (-0.015 to 0.081)	0.178	1
Week-10 remission	Full vs HAMD trajectory	0.030 (-0.014 to 0.073)	0.1822	1
Week-10 remission	Full vs non-HAMD	0.059 (0.010 to 0.108)	0.0174	0.2261
Two-point remission	HAMD trajectory vs week-2 HAMD	0.002 (-0.020 to 0.025)	0.8386	1
Two-point remission	Non-HAMD vs week-2 HAMD	-0.035 (-0.129 to 0.059)	0.4649	1
Two-point remission	Full vs week-2 HAMD	0.040 (-0.013 to 0.094)	0.1392	1
Two-point remission	Full vs HAMD trajectory	0.038 (-0.016 to 0.092)	0.1694	1
Two-point remission	Full vs non-HAMD	0.075 (0.013 to 0.138)	0.018	0.2261

Note: Holm adjustment was applied across all 15 comparisons; confidence intervals were not multiplicity adjusted.

Supplementary Table S7. Detailed leave-one-centre-out internal-external cross-validation

Outcome	Held-out centre/summary	N/events	AUROC	AP	Brier	O:E	Cal. slope
Week-6 remission	1	140/39	0.806 (0.722 to 0.883)	0.618 (0.494 to 0.759)	0.153 (0.122 to 0.185)	0.985 (0.876 to 1.124)	0.771 (0.488 to 1.248)
Week-6 remission	2	64/12	0.806 (0.694 to 0.904)	0.430 (0.289 to 0.700)	0.138 (0.102 to 0.175)	0.825 (0.680 to 1.027)	0.821 (0.440 to 1.563)
Week-6 remission	3	29/8	0.887 (0.750 to 0.988)	0.715 (0.464 to 0.975)	0.146 (0.073 to 0.226)	1.240 (0.894 to 1.833)	0.890 (0.377 to 4.010)
Week-6 remission	4	29/10	0.916 (0.800 to 0.989)	0.874 (0.707 to 0.983)	0.123 (0.073 to 0.182)	1.113 (0.917 to 1.440)	1.510 (0.924 to 4.929)
Week-6 remission	5	44/12	0.927 (0.836 to 0.992)	0.796 (0.580 to 0.979)	0.110 (0.065 to 0.160)	1.028 (0.844 to 1.292)	1.352 (0.698 to 4.821)
Week-6 remission	Micro pooled	306/81	0.839 (0.789 to 0.885)	0.634 (0.544 to 0.730)	0.140 (0.120 to 0.161)	0.997 (0.914 to 1.096)	0.914 (0.697 to 1.229)
Week-6 remission	Macro pooled	NA/NA	0.868 (0.822 to 0.915)	0.687 (0.541 to 0.811)	0.134 (0.120 to 0.147)	1.038 (0.915 to 1.164)	1.069 (0.815 to 1.341)
Week-10 remission	1	117/40	0.722 (0.620 to 0.814)	0.521 (0.413 to 0.666)	0.197 (0.173 to 0.224)	0.942 (0.875 to 1.017)	1.094 (0.546 to 1.897)
Week-10 remission	2	61/20	0.645 (0.488 to 0.798)	0.480 (0.352 to 0.700)	0.227 (0.166 to 0.288)	1.025 (0.857 to 1.257)	0.313 (-0.033 to 0.799)
Week-10 remission	3	24/9	0.844 (0.667 to 0.993)	0.667 (0.472 to 0.989)	0.191 (0.132 to 0.265)	1.131 (0.924 to 1.394)	1.088 (0.186 to 5.063)
Week-10 remission	4	31/9	0.707 (0.465 to 0.909)	0.599 (0.338 to 0.844)	0.188 (0.128 to 0.256)	0.745 (0.631 to 0.915)	0.825 (-0.025 to 2.351)
Week-10 remission	5	36/16	0.950 (0.850 to 1.000)	0.893 (0.720 to 1.000)	0.186 (0.159 to 0.212)	1.372 (1.258 to 1.498)	5.199 (2.671 to 41.456)
Week-10 remission	Micro pooled	269/94	0.720 (0.654 to 0.781)	0.537 (0.465 to 0.630)	0.201 (0.182 to 0.221)	1.004 (0.946 to 1.068)	0.744 (0.463 to 1.119)
Week-10 remission	Macro pooled	NA/NA	0.774 (0.685 to 0.883)	0.632 (0.520 to 0.773)	0.198 (0.188 to 0.213)	1.043 (0.862 to 1.233)	1.704 (0.623 to 3.503)
Two-point remission	1	109/18	0.758 (0.637 to 0.864)	0.340 (0.236 to 0.529)	0.128 (0.106 to 0.151)	0.831 (0.724 to 0.964)	0.833 (0.378 to 1.592)
Two-point remission	2	61/9	0.814 (0.656 to 0.944)	0.413 (0.248 to 0.727)	0.112 (0.075 to 0.154)	0.825 (0.646 to 1.104)	0.706 (0.285 to 1.752)
Two-point remission	3	24/6	0.917 (0.778 to 1.000)	0.683 (0.446 to 1.000)	0.151 (0.087 to 0.242)	1.423 (0.931 to 2.291)	0.782 (0.251 to 28.103)
Two-point remission	4	28/5	0.904 (0.757 to 1.000)	0.725 (0.401 to 1.000)	0.094 (0.056 to 0.135)	0.823 (0.646 to 1.089)	1.844 (1.023 to 7.596)
Two-point remission	5	36/9	0.901 (0.778 to 0.979)	0.780 (0.545 to 0.950)	0.117 (0.063 to 0.175)	1.128 (0.859 to 1.647)	0.981 (0.649 to 2.209)
Two-point remission	Micro pooled	258/47	0.825 (0.766 to 0.879)	0.470 (0.369 to 0.604)	0.121 (0.105 to 0.139)	0.924 (0.831 to 1.032)	0.877 (0.630 to 1.251)
Two-point remission	Macro pooled	NA/NA	0.859 (0.801 to 0.909)	0.588 (0.432 to 0.739)	0.120 (0.105 to 0.137)	1.006 (0.826 to 1.188)	1.029 (0.762 to 1.440)

Note: Centre was omitted as a predictor. Small centre-specific event counts make calibration estimates unstable.

Supplementary Table S8. Detailed leave-one-treatment-arm-out analysis

Outcome	Held-out arm/summary	N/events	AUROC	AP	Brier	O:E	Cal. slope
Week-6 remission	A	55/11	0.907 (0.816 to 0.971)	0.668 (0.447 to 0.908)	0.109 (0.080 to 0.144)	0.730 (0.624 to 0.868)	1.751 (1.046 to 3.990)

Outcome	Held-out arm/summary	N/events	AUROC	AP	Brier	O:E	Cal. slope
Week-6 remission	B	66/14	0.911 (0.815 to 0.982)	0.781 (0.593 to 0.940)	0.103 (0.079 to 0.128)	0.851 (0.754 to 0.970)	2.512 (1.533 to 6.017)
Week-6 remission	C	60/14	0.832 (0.722 to 0.927)	0.537 (0.381 to 0.795)	0.141 (0.102 to 0.183)	0.750 (0.645 to 0.894)	1.115 (0.617 to 2.177)
Week-6 remission	D	58/18	0.768 (0.626 to 0.893)	0.601 (0.429 to 0.796)	0.181 (0.132 to 0.236)	1.078 (0.902 to 1.332)	0.695 (0.262 to 1.650)
Week-6 remission	E	67/24	0.735 (0.608 to 0.846)	0.595 (0.447 to 0.764)	0.236 (0.198 to 0.274)	1.963 (1.613 to 2.453)	0.748 (0.349 to 1.499)
Week-6 remission	Micro pooled	306/81	0.787 (0.734 to 0.837)	0.567 (0.483 to 0.665)	0.155 (0.138 to 0.173)	1.024 (0.951 to 1.106)	0.904 (0.673 to 1.213)
Week-10 remission	A	57/16	0.655 (0.506 to 0.794)	0.438 (0.297 to 0.631)	0.199 (0.177 to 0.221)	0.790 (0.736 to 0.853)	1.068 (-0.105 to 2.427)
Week-10 remission	B	58/18	0.562 (0.385 to 0.732)	0.448 (0.296 to 0.643)	0.212 (0.186 to 0.240)	0.881 (0.810 to 0.955)	0.610 (-0.800 to 1.931)
Week-10 remission	C	56/13	0.764 (0.630 to 0.875)	0.383 (0.286 to 0.553)	0.186 (0.154 to 0.225)	0.596 (0.541 to 0.656)	1.089 (0.393 to 2.453)
Week-10 remission	D	39/15	0.742 (0.558 to 0.900)	0.707 (0.517 to 0.881)	0.208 (0.178 to 0.240)	1.142 (1.050 to 1.260)	1.781 (0.339 to 4.468)
Week-10 remission	E	59/32	0.809 (0.679 to 0.919)	0.777 (0.644 to 0.923)	0.294 (0.271 to 0.316)	2.077 (1.899 to 2.275)	2.536 (1.049 to 6.506)
Week-10 remission	Micro pooled	269/94	0.626 (0.563 to 0.688)	0.441 (0.387 to 0.516)	0.221 (0.209 to 0.233)	1.033 (0.993 to 1.072)	0.695 (0.331 to 1.160)
Two-point remission	A	50/5	0.827 (0.711 to 0.933)	0.268 (0.179 to 0.503)	0.092 (0.075 to 0.110)	0.534 (0.454 to 0.639)	1.393 (0.807 to 3.417)
Two-point remission	B	58/7	0.821 (0.591 to 0.978)	0.569 (0.270 to 0.893)	0.091 (0.075 to 0.108)	0.597 (0.542 to 0.660)	2.774 (0.989 to 8.513)
Two-point remission	C	52/5	0.847 (0.732 to 0.936)	0.281 (0.181 to 0.512)	0.093 (0.076 to 0.114)	0.453 (0.392 to 0.529)	1.522 (0.884 to 3.642)
Two-point remission	D	39/10	0.769 (0.559 to 0.938)	0.546 (0.359 to 0.883)	0.175 (0.151 to 0.201)	1.476 (1.269 to 1.736)	1.308 (0.061 to 4.943)
Two-point remission	E	59/20	0.656 (0.501 to 0.801)	0.485 (0.357 to 0.695)	0.290 (0.261 to 0.318)	4.253 (3.061 to 5.977)	0.472 (0.037 to 1.351)
Two-point remission	Micro pooled	258/47	0.590 (0.519 to 0.658)	0.275 (0.220 to 0.359)	0.150 (0.140 to 0.161)	1.077 (0.998 to 1.161)	0.279 (0.016 to 0.606)

Note: Treatment group was omitted; this is a modified specification and remains within the parent trial.

Supplementary Table S9. Within-centre date-field-ordered holdout sensitivity analysis

Outcome	Validation sample	N/events	AUROC	AP	Brier	O:E	Cal. slope
Week-6 remission	Pooled late cohort	91/22	0.838 (0.754 to 0.915)	0.579 (0.439 to 0.778)	0.142 (0.107 to 0.176)	0.820 (0.717 to 0.949)	0.983 (0.651 to 1.660)
Week-10 remission	Pooled late cohort	78/18	0.694 (0.554 to 0.808)	0.378 (0.263 to 0.521)	0.199 (0.166 to 0.236)	0.571 (0.501 to 0.612)	0.672 (0.127 to 1.490)
Two-point remission	Pooled late cohort	72/10	0.839 (0.741 to 0.908)	0.334 (0.205 to 0.447)	0.122 (0.084 to 0.151)	0.696 (0.469 to 0.754)	0.696 (0.339 to 1.505)

Note: The earliest 70% of MHSTDT values within each centre was used for development and latest 30% for evaluation. The definition and range of MHSTDT must be reconciled with the recruitment log before this analysis is interpreted as temporal validation.

Supplementary Table S10. Outcome-scoring and episode-duration sensitivity analyses

Outcome	Sensitivity	Convention	N/events	AUROC (95% CI if available)	AP	Brier
Week-6 remission	HAMD item-14	Code 3 to zero	306/81	0.839 (0.789 to 0.885)	0.634	0.140
Week-6 remission	HAMD item-14	Code 3 to missing	285/81	0.818 (0.762 to 0.867)	0.599	0.154
Week-10 remission	HAMD item-14	Code 3 to zero	269/94	0.720 (0.654 to 0.781)	0.537	0.201
Week-10 remission	HAMD item-14	Code 3 to missing	250/90	0.726 (0.662 to 0.789)	0.562	0.200
Two-point remission	HAMD item-14	Code 3 to zero	258/47	0.825 (0.766 to 0.879)	0.470	0.121
Two-point remission	HAMD item-14	Code 3 to missing	237/47	0.809 (0.745 to 0.867)	0.490	0.131
Week-6 remission	Episode duration	with duration	306/81	0.836	-	-
Week-6 remission	Episode duration	without duration	306/81	0.837	-	-
Week-10 remission	Episode duration	with duration	269/94	0.736	-	-
Week-10 remission	Episode duration	without duration	269/94	0.752	-	-
Two-point remission	Episode duration	with duration	258/47	0.827	-	-
Two-point remission	Episode duration	without duration	258/47	0.838	-	-

Note: Item-14 estimates are centre-IECV. Episode-duration rows are repeated-CV AUROC sensitivity estimates. The final item-14 convention must be verified from source documents.

Supplementary Table S11. Programmatic output-integrity checks

Check	Status	Detail
DeLong staged comparison count	PASS	9 rows
DeLong same-scale comparison count	PASS	15 rows
DeLong arithmetic	PASS	delta AUROC equals paired AUROC difference
Holm adjustment	PASS	adjusted P values are not smaller than raw P values
Average precision terminology	PASS	metric key is average_precision, not AUPRC
Primary ROC endpoints	PASS	all curves run from (0,0) to (1,1)
Same-scale ROC endpoints	PASS	all curves run from (0,0) to (1,1)
Calibration bin denominators	PASS	grouped calibration bins sum to each analysis cohort

Check	Status	Detail
DCA treat-none reference	PASS	net benefit is zero
DCA treat-all reference	PASS	reference formula reproduced
Clinical-impact monotonicity	PASS	counts decrease with threshold and true events do not exceed high-risk counts

Note: All checks were executed against the statistically revised outputs. These checks verify arithmetic, metric labelling, curve endpoints, calibration-bin denominators and decision-analysis formulas; they are not an independent code review, external validation or substitute for source-document verification.

Supplementary Table S12. Display-level data provenance and reproducibility map

Display	Machine-readable source	Computation	Audit status
Main Table 1	baseline_characteristics.csv	Raw SAV -> scripted baseline summary	Values rebuilt from locked source data
Main Table 2	model_performance.csv; delong_incremental_tests.csv	Mean out-of-fold predictions and paired common-cohort comparisons	Independent rerun matched source CSV
Figure 1	figure1_panel_data.csv; analysis_cohorts.csv	Participant flow and validation-design schematic	Counts reconciled to rerun cohorts
Figure 2	Raw SAV	Observed-case summaries and descriptive spline	Regenerated directly from locked source data
Figure 3	roc_curve_points_all_outcomes.csv; delong_incremental_tests.csv	ROC coordinates and paired AUROC differences	Independent rerun matched source CSV
Figure 4	calibration_curve_points_all_outcomes.csv; validation metric CSVs	Calibration points and validation AUROC	Independent rerun matched source CSV
Figure 5	decision_curve_all_outcomes.csv; clinical_impact_curve_all_outcomes.csv	Net benefit and threshold classifications	Independent rerun matched source CSV
Supplementary Figure S1	pr_curve_points_all_outcomes.csv; model_performance.csv	Precision-recall coordinates, AP and Brier score	Independent rerun matched source CSV
Supplementary Figure S2	same_scale_roc_curve_points_all_outcomes.csv; same_scale_delong_tests.csv	Parsimonious-model ROC and paired comparisons	Independent rerun matched source CSV
Supplementary Figure S3	centre, arm and date-field-ordered validation metric CSVs	Partition-specific held-out performance	Independent rerun matched common source CSVs
Supplementary Figure S4	item14_scoring_sensitivity_summary.csv; figure5_outcome_observability.csv	Scoring-rule sensitivity and observed-outcome proportions	Source-linked; scoring rule remains to be verified

Note: The locked SPSS source file and R scripts were rerun in a separate audit directory. All 31 machine-readable CSV outputs common to the submission pipeline and independent rerun matched exactly within a numerical tolerance of 1e-12; six additional files existed on only one side because they were post-processing summaries or expanded strict-sensitivity outputs. All 18 packaged PDF/TIFF figure files are copied directly from the corresponding source graphics at package generation. This provenance audit verifies reproducibility of the implemented pipeline, not the unresolved clinical meaning of HAMD item-14 code 3, the validity of MHSTDT as enrolment date, or the absence of bias from missing outcomes.

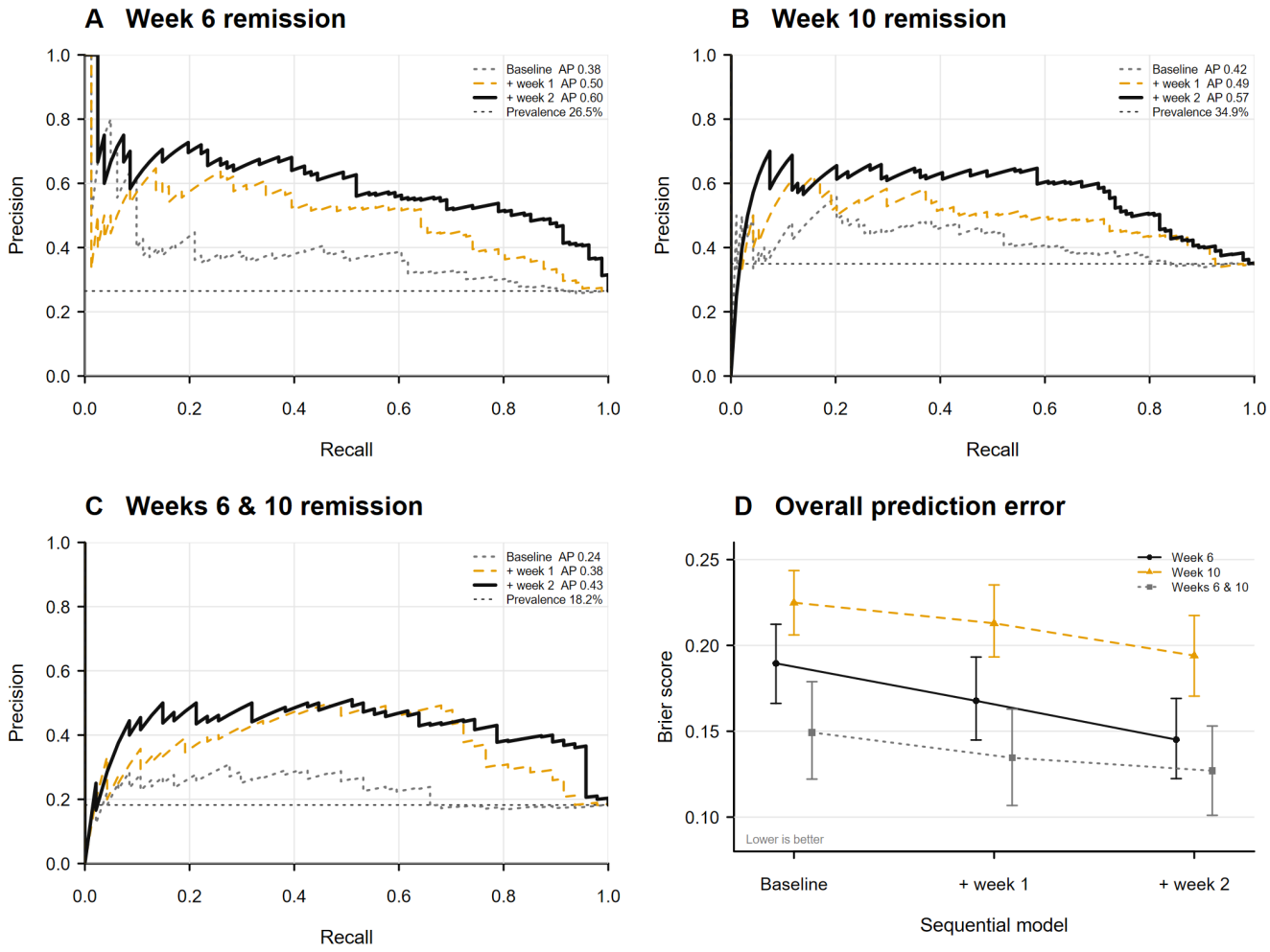
Supplementary Table S13. Paired incremental comparisons for staged models

Outcome	Comparison	Difference in AUROC (95% CI)	Raw P	Holm-adjusted P	Difference in Brier (95% CI)
Week-6 remission	+ week 1 vs baseline	0.128 (0.059 to 0.197)	0.000296	0.0015	-0.022 (-0.035 to -0.010)
Week-6 remission	+ week 2 vs + week 1	0.093 (0.041 to 0.144)	0.000388	0.0016	-0.023 (-0.037 to -0.008)
Week-6 remission	+ week 2 vs baseline	0.220 (0.148 to 0.293)	2.99e-09	2.69e-08	-0.044 (-0.066 to -0.026)
Week-10 remission	+ week 1 vs baseline	0.093 (0.025 to 0.162)	0.0074	0.0222	-0.012 (-0.023 to 0.001)
Week-10 remission	+ week 2 vs + week 1	0.070 (0.019 to 0.122)	0.0078	0.0222	-0.019 (-0.032 to -0.005)
Week-10 remission	+ week 2 vs baseline	0.164 (0.085 to 0.242)	4.29e-05	0.0003	-0.031 (-0.050 to -0.013)
Two-point remission	+ week 1 vs baseline	0.207 (0.102 to 0.311)	0.000105	0.000631	-0.015 (-0.031 to 0.001)
Two-point remission	+ week 2 vs + week 1	0.058 (0.007 to 0.109)	0.0248	0.0248	-0.008 (-0.018 to 0.007)
Two-point remission	+ week 2 vs baseline	0.265 (0.159 to 0.371)	9.07e-07	7.25e-06	-0.022 (-0.039 to -0.002)

Note: AUROC differences and P values are from paired DeLong tests; Holm adjustment was applied across all nine staged comparisons. Brier differences use paired case-bootstrap intervals on fixed out-of-fold predictions. Negative Brier differences favour the newer model. Confidence intervals were not multiplicity adjusted.

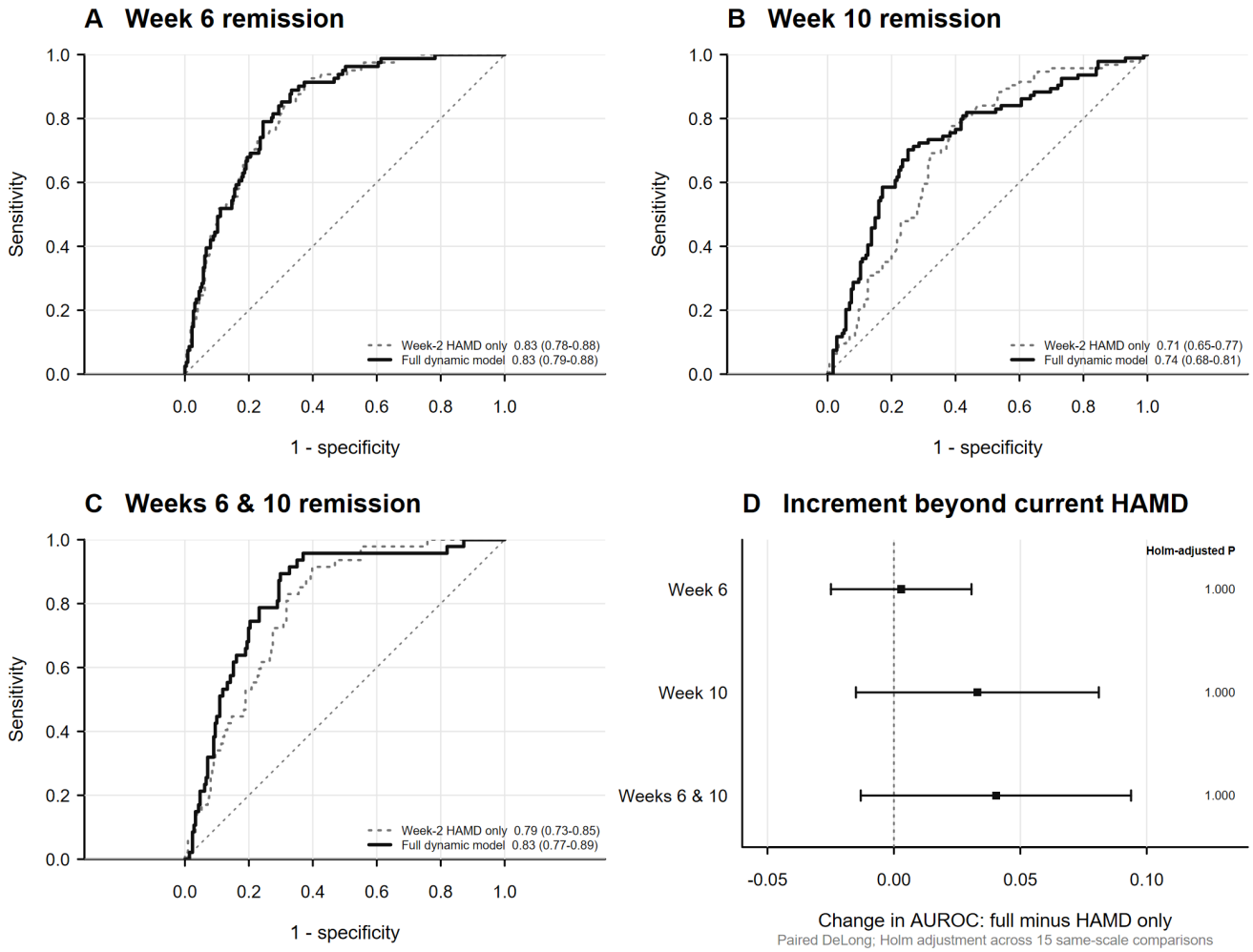
Supplementary Figures

Supplementary Figure S1. Precision-recall curves, average precision and Brier scores.



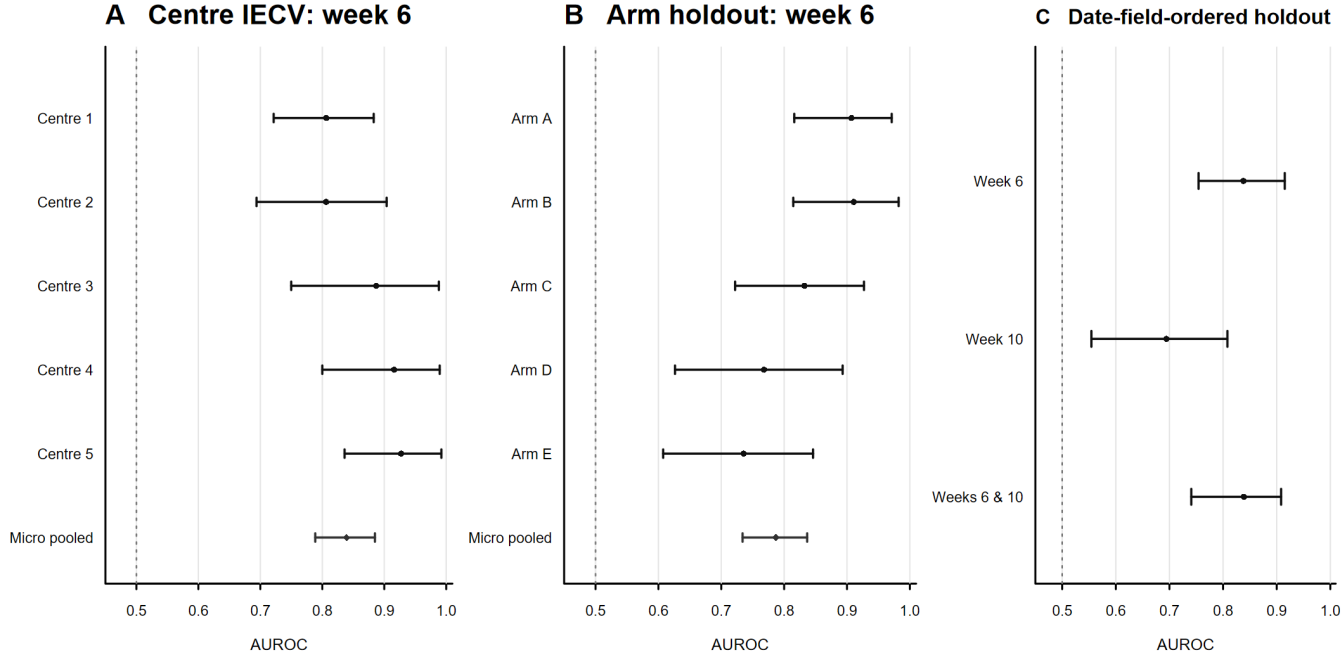
Precision-recall curves and model-level AP and Brier scores are shown for the three staged models. AP, rather than trapezoidal AUPRC, is reported. Confidence intervals are percentile intervals from 500 case-bootstrap samples of fixed out-of-fold predictions.

Supplementary Figure S2. Parsimonious week-2 HAMD benchmark.



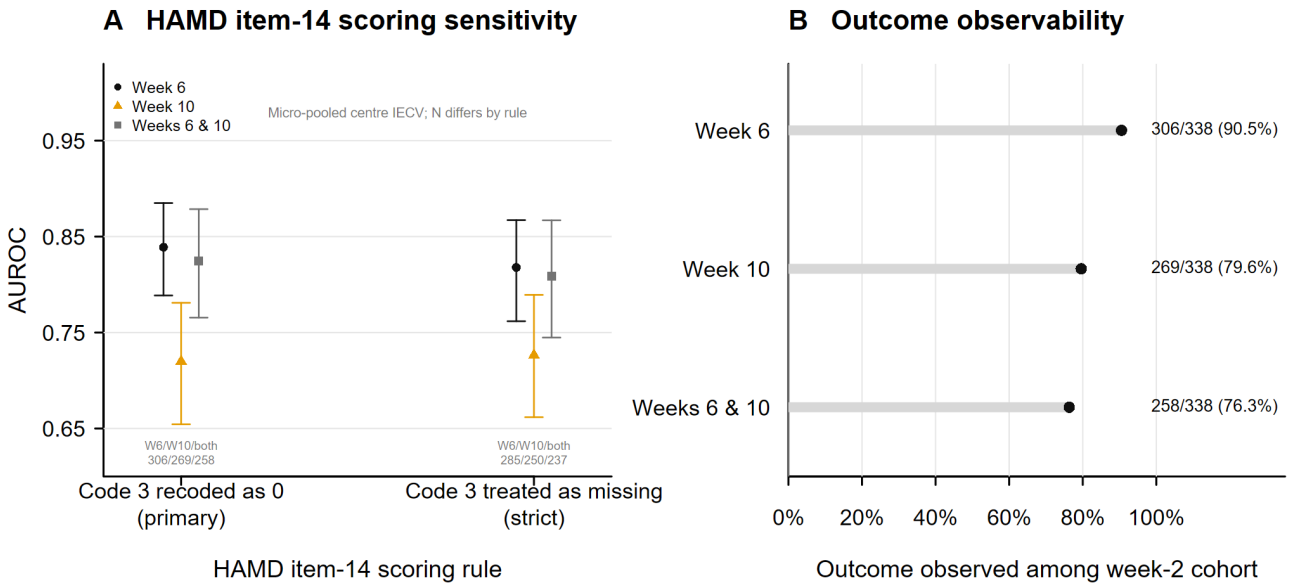
ROC curves compare the week-2 HAMD-only model with the full dynamic model. Paired DeLong differences for full minus week-2 HAMD-only were 0.003, 0.033 and 0.040 for week-6, week-10 and two-point remission, respectively; none was significant after Holm adjustment across 15 same-scale comparisons.

Supplementary Figure S3. Detailed within-trial validation results.



Partition-specific AUROC estimates are shown for held-out centres, treatment arms and the within-centre date-field-ordered holdout. Error bars are conditional bootstrap 95% confidence intervals based on fixed held-out predictions. Small partition event counts produce wide uncertainty. The date field requires verification before this sensitivity analysis can be interpreted as temporal validation.

Supplementary Figure S4. HAMD item-14 scoring sensitivity and outcome observability.



Centre-IECV AUROC is shown under the working convention recoding item-14 code 3 to zero and the strict convention treating code 3 as missing. The strict rule changed cohorts from 306/269/258 to 285/250/237. Outcome availability among 338 week-2 landmark participants was 90.5%, 79.6% and 76.3%.