PSILOGOS

Research Article

open-access

When Transportability Matters More Than Complexity: Multicentre Calibration of a High-Frequency rTMS Response Score in an In-Silico Population Study

Barbara H. Stanley

ABSTRACT

Prognostic models for repetitive transcranial magnetic stimulation (rTMS) are commonly judged by discrimination within the population from which they were estimated. Clinical deployment instead requires accurate probabilities after the diagnostic composition, symptom severity, treatment resistance, and anatomical characteristics of the treated population have changed. We examined whether a site-aware penalized recalibration procedure improves transport of a published high-frequency left dorsolateral prefrontal cortex response score and whether added complexity is justified relative to a centre-specific intercept correction. Eight prespecified in-silico populations contained 48,000 records with systematic differences in sex composition, bipolar-depression prevalence, age, Montgomery–Åsberg Depression Rating Scale score, treatment-resistance stage, dorsolateral prefrontal cortical thickness, and scalp-to-cortex distance. Outcomes were sampled from a nonlinear probability function anchored to published coefficients but altered by centre and case mix. Four approaches were compared: the unchanged transported score, centre-specific intercept correction, pooled logistic recalibration, and partially pooled transport calibration (PPTC). Four hundred records per centre were reserved for updating and 5,600 per centre for held-out evaluation. Probability accuracy, calibration, discrimination, and decision-curve utility were quantified. Paired differences were evaluated by 600 centre-stratified bootstrap samples. Among 44,800 held-out records, response prevalence was 45.24%. The unchanged score attained an area under the receiver-operating-characteristic curve of 0.790 and a Brier score of 0.1856, yet its largest centre-level absolute calibration error was 9.90 percentage points. Centre-specific intercept correction reduced that maximum error to 2.46 points, improved the Brier score to 0.1812, and increased net benefit at a 0.35 treatment threshold from 0.2498 to 0.2579. Corresponding paired differences were -0.00434 (95% bootstrap interval -0.00496 to -0.00377) for the Brier score and $+0.00809$ (0.00626 to 0.00997) for net benefit. PPTC reduced the maximum centre error to 5.61 points and improved net benefit to 0.2567, but it did not outperform intercept correction; its largest subgroup error was 2.92 points. Pooled recalibration offered negligible improvement. The research question was whether a site-aware multivariable update warrants its complexity when an rTMS response model crosses populations. Under the prespecified case-mix changes, it did not. A local intercept correction was the most reliable approach because it addressed prevalence drift without destabilizing individual risk contrasts. High discrimination alone concealed clinically important centre-level probability errors; local calibration should therefore precede threshold-based use of transported rTMS response scores.

Keywords: repetitive transcranial magnetic stimulation; treatment-resistant depression; probability calibration; transportability; decision curve; structural magnetic resonance imaging; in-silico study

Department of Psychiatry, Vagelos College of Physicians and Surgeons, Columbia University, New York, USA
bhs2@cumc.columbia.edu

– **Published:** 11 October 2024

https://doi.org/DFdfD

1. INTRODUCTION

Treatment-resistant depression imposes prolonged disability, recurrent service use, and elevated mortality risk despite successive pharmacological and psychotherapeutic interventions. Repetitive transcranial magnetic stimulation has become an important non-invasive treatment for patients whose depressive episode has not responded adequately to medication. Contemporary clinical guidance supports high-frequency stimulation of the left dorsolateral prefrontal cortex (DLPFC), among other evidence-based protocols, because it offers antidepressant benefit without the systemic adverse-effect burden of many drug combinations.[1,2] Randomized evidence and network syntheses confirm that rTMS is effective for acute major depressive episodes, while large clinical registries show that meaningful improvement also occurs in routine practice.[3–5] The central implementation problem is therefore no longer whether rTMS can work, but how to estimate an individual patient's chance of benefit accurately enough to support treatment selection.

Observed response varies considerably. Diagnostic polarity, number of failed treatments, symptom burden, age, medication exposure, stimulation dose, and targeting method can all alter the apparent treatment effect or the population in which it is measured.[6,7] Naturalistic investigations have linked response to clinical characteristics, although findings for age, sex, medication, and bipolar diagnosis are not consistent across services.[8–10] The bipolar subgroup is especially consequential. Meta-analytic evidence supports active rTMS for bipolar depression, but the number of dedicated trials is limited, stimulation protocols differ, and participants with bipolar illness are often embedded within broader depressive cohorts.[11–13] A probability model trained in a centre with one diagnostic mixture can consequently be miscalibrated when applied in a centre with another mixture, even if its rank ordering of patients remains acceptable.

Structural magnetic resonance imaging has been investigated as a practical complement to clinical predictors. DLPFC thickness, anterior cingulate morphology, hippocampal and amygdala volumes, scalp-to-cortex distance, and distributed gray-matter organization have each been proposed as correlates of response.[14–16] Anatomical T1-weighted images are attractive because they are widely available and can be processed with established cortical parcellation procedures.[17] However, a statistically associated region is not automatically a transportable predictor. Parcellation error, scanner characteristics, medication-related morphology, targeting variability, and centre-specific case mix can change both the predictor distribution and its relation to outcome. Studies of functional and structural connectivity likewise indicate that response-relevant information may be distributed across networks rather than confined to a single cortical region.[18–20] These considerations make transport evaluation a distinct scientific task rather than an optional final check.

Most biomarker studies emphasize association tests or the area under the receiver-operating-characteristic curve (AUC). Discrimination answers whether patients who respond tend to receive higher scores than patients who do not; it does not answer whether a predicted probability of 0.60 corresponds to a 60% response frequency. This distinction matters whenever a service uses a probability threshold to determine whether rTMS should be offered, delayed, or discussed alongside alternatives. A model may preserve its AUC after transport because the ordering of patients survives, while systematically overestimating response in a low-prevalence centre and underestimating it in a high-prevalence centre. Probability calibration has therefore been described as a central weakness of clinical prediction research.[21] Reporting guidance similarly requires separate examination of discrimination and calibration, particularly during external validation.[22]

Several updating strategies are possible. An unchanged model preserves the derivation coefficients and requires no local outcomes, but it assumes that outcome prevalence and predictor effects are stable. Intercept correction retains all relative effects while shifting the probability scale to the local event rate. Pooled logistic recalibration estimates a common intercept and slope from several centres, which can correct global overfitting but may average away meaningful centre differences. A larger site-aware model can include patient characteristics and centre indicators under shrinkage, allowing partial adaptation without estimating an unconstrained model separately in every centre. Added flexibility is not automatically advantageous. When the local update sample is modest, a complex procedure can trade lower centre bias for unstable subgroup probabilities or reduced discrimination. The appropriate question is not which method appears most sophisticated, but which produces probabilities that remain useful at clinically relevant thresholds.

Decision-curve analysis links probability estimates to action by weighting false-positive treatment recommendations against true-positive recommendations at a stated threshold.[23] It is particularly suitable for rTMS because treatment choice reflects time, cost, access, alternative therapies, and patient preference rather than a purely statistical classification. A threshold of 0.35, for example, encodes a different clinical trade-off from a threshold of 0.20; the preferred prediction method may change across that range. Calibration, discrimination, and net benefit must consequently be considered together. A method with a slightly improved Brier score but no gain in decision utility may offer little practical value, whereas a simple correction that reduces severe centre-level probability error may be important even if the pooled AUC barely changes.

The present study asks a research question that differs from biomarker discovery: when a high-frequency left-DLPFC rTMS response score is transported across ser-

vices with deliberately different clinical and anatomical compositions, does a partially pooled, site-aware update produce more reliable and useful probabilities than simpler recalibration? We constructed eight controlled populations, reserved local observations for updating, and evaluated four methods in held-out records. The design was intended to expose the difference between apparently satisfactory pooled performance and clinically important local error. We prespecified that a useful update should reduce centre-level calibration error and Brier loss without sacrificing decision-curve utility or generating disproportionate error in sex-by-diagnosis groups. This emphasis shifts the scientific target from identifying another correlate of response to determining how an existing multivariable signal should be carried across populations.

2. MATERIALS AND METHODS

2.1. Study design

We performed a reproducible in-silico multicentre cohort study. The estimand was the performance of a transported probability model among patients eligible for 20-Hz left-DLPFC rTMS after a local update sample had been observed. The primary comparison was centre-specific intercept correction versus PPTC in records excluded from model updating. The unchanged score and pooled logistic recalibration served as additional comparators. All construction rules, random-number seeds, analysis code, and figure code were fixed in the accompanying project.

The study did not simulate a randomized treatment contrast. Every record represented a treatment-eligible individual receiving the same rTMS course, and the binary outcome represented at least 50% reduction in Montgomery–Åsberg Depression Rating Scale (MADRS) score at follow-up. The analysis therefore concerns prognostic transport among treated patients. It cannot estimate the treatment effect relative to sham stimulation, medication, electroconvulsive therapy, or another rTMS protocol.

Clinical marginals and the coefficient vector were transcribed from a naturalistic cohort in which 131 records were screened, 95 patients contributed complete clinical outcomes, and 90 contributed MRI analyses after acquisition or parcellation exclusions. The reported treatment consisted of ten sessions over two weeks, 20-Hz stimulation at 110% of resting motor threshold, 80 two-second trains per session, a 10-second inter-train interval, and 3,200 pulses per session. Response at day 45 occurred in 38 of 95 patients. The multivariable coefficient vector contained sex, depressive diagnosis, age, treatment-resistance stage, illness duration, entry MADRS score, DLPFC thickness, and minimum brain–scalp distance; the published AUC was 0.786 and Nagelkerke R^2 was 0.317.[24] This citation is the sole provenance reference for the numerical values taken from the attached article.

The transported log odds for person i were

$$\eta_i = 0.440149 - 1.454878M_i - 1.380683U_i + 0.023513A_i \quad (1)$$
$$- 0.690661R_i - 0.117649D_i + 2.361551T_i - 0.053008S_i, \quad (2)$$

where M denotes male sex, U unipolar depression, A age in years, R treatment-resistance stage, D the MADRS score at cohort entry, T DLPFC thickness in millimetres, and S minimum brain–scalp distance in millimetres. The illness-duration coefficient was zero at the precision reported and was retained in the generator without contributing to η_i. The unchanged transported probability was $\hat{p}_i = \text{logit}^{-1}(\eta_i)$.

This equation was used as a transport object, not as evidence that each coefficient is causal. In particular, the sex and diagnostic terms may encode differences in referral, concomitant treatment, symptom expression, or follow-up as well as biological variation. Retaining the coefficient signs and scales allowed the experiment to examine transport failure without inventing a second clinical effect estimate.

2.2. Multicentre population construction

Eight populations of 6,000 records were generated with the prespecified distributions shown in Table 1. The total population was 48,000. Age followed a truncated Gaussian distribution with centre means from 48 to 58 years and a common standard deviation of 11.2 years. MADRS score followed a truncated Gaussian distribution with centre means from 25.5 to 32.5 and a common standard deviation of 5.5. DLPFC thickness and brain–scalp distance were similarly truncated to physiologically credible ranges. Treatment-resistance stage was rounded to an integer from 2 to 5. Illness duration followed a right-skewed log-normal distribution restricted to 12–600 months. Sex and diagnosis were Bernoulli variables with centre-specific probabilities.

The composition spans centres with relatively low symptom burden and short brain–scalp distance as well as centres with greater treatment resistance, greater symptom burden, and longer distance. It also breaks the correlation between bipolar prevalence and female representation: Centre A has low values for both, Centre E has high bipolar prevalence but a lower proportion of women, and Centre H has high values for both. This deliberate structure permits diagnostic and sex-related calibration error to be separated from a simple prevalence gradient.

2.3. Outcome mechanism

The response probability was generated from the transported log odds plus centre effects and prespecified departures from linear additivity:

$$\text{logit}(p_i) = 0.82\eta_i + \delta_{c[i]} - 0.018\frac{(D_i - 28)^2}{5.5} + 0.34F_iB_i \quad (3)$$
$$- 0.055(R_i - 2)\frac{(S_i - 18)}{3.3} + 0.20\frac{(T_i - 2.52)}{0.13}B_i, \quad (4)$$

where $\delta_{c[i]}$ is the centre shift, F denotes female sex, and B denotes bipolar depression. Centre shifts from A to H

Table 1. Prespecified centre composition.

Centre	Women (%)	Bipolar (%)	Age	Resistance	MADRS	Thickness	Distance
A	40	10	48	2.20	25.5	2.62	14.5
B	49	18	51	2.45	27.0	2.57	16.5
C	57	25	52	2.55	28.3	2.52	18.0
D	66	33	55	2.75	30.0	2.48	20.0
E	46	40	58	3.10	32.5	2.43	22.5
F	61	15	50	2.35	26.5	2.59	15.5
G	53	29	54	2.90	31.0	2.45	21.5
H	70	36	56	2.65	29.0	2.50	19.0

Values for age, resistance, MADRS, thickness, and distance are distribution means. Thickness and distance are in millimetres.

were 0.55, 0.30, 0.10, -0.10, -0.55, 0.40, -0.35, and -0.20. Binary outcomes were sampled from $\mathrm{Bernoulli}(p_i)$ using seed 20260814.

The multiplicative factor of 0.82 creates moderate coefficient attenuation after transport. The quadratic MADRS term represents decreasing linear adequacy at symptom extremes; the sex-by-diagnosis and anatomy-by-diagnosis terms create clinically interpretable effect modification; and the resistance-by-distance term increases difficulty when pharmacological refractoriness and a less favourable stimulation geometry coincide. These perturbations were not labelled as patient-level biological effects. They are controlled challenges used to determine whether an updating method can recover probability accuracy when the original linear specification is imperfect.

2.4. Evaluation samples

Four hundred records were sampled without replacement from each centre for model updating, yielding 3,200 update records. The remaining 5,600 records per centre, or 44,800 in total, were held out until the models were fixed. The same held-out outcomes were used for paired comparison of all four approaches. No record used for tuning the penalization parameter contributed to evaluation.

The choice of 400 update records per centre reflects a service-scale audit rather than a full refit. It supplies local outcome prevalence with useful precision while remaining small relative to the number of candidate effects in a site-specific multivariable model. Sample-size considerations for external validation emphasize the number of events, the precision of calibration estimates, and the intended range of predicted probabilities rather than a single universal rule.[25] Each held-out centre included 5,600 records so that graphical calibration and subgroup comparisons would not be dominated by Monte Carlo noise.

2.5. Prediction approaches

The *fixed score* applied $\mathrm{logit}^{-1}(\eta_i)$ without updating. It represents direct transport of the reported coefficients.

The *local intercept* approach estimated one additive constant a_c for each centre by solving

$$\frac{1}{n_c}\sum_{i=1}^{n_c}\mathrm{logit}^{-1}(\eta_i + a_c) = \bar{y}_c \tag{5}$$

in the 400 update records from centre c. The slope and all predictor contrasts remained unchanged. This operation aligns the mean predicted probability with the observed local response proportion and is therefore targeted at prevalence drift.

The *pooled recalibration* approach fitted $\mathrm{logit}(p_i) = a + b\eta_i$ to all 3,200 update records. It can correct a common intercept shift and uniform coefficient attenuation but cannot represent centre-specific prevalence.

The new *partially pooled transport calibration* approach fitted an L_2-penalized logistic model containing the transported log odds, sex, diagnosis, their interaction, age, resistance stage, MADRS score, DLPFC thickness, brain–scalp distance, and one-hot centre terms. Continuous predictors were standardized within the update sample. The regularization constant was selected from nine values between 0.01 and 100 by five-fold cross-validated log loss. Penalization shrinks centre and patient-level adjustments toward a shared solution, allowing heterogeneity while limiting the variance expected from eight unconstrained local models. The selected inverse penalty was 0.01, indicating strong shrinkage.

These four approaches answer progressively different transport hypotheses. The fixed score assumes complete stability; the local intercept assumes only prevalence drift; pooled recalibration assumes a common probability-scale distortion; and PPTC permits centre and patient-characteristic adjustment. Their comparison reveals whether added flexibility addresses the type of shift that actually determines decision quality.

2.6. Performance measures and statistical analysis

Overall probability accuracy was quantified with the Brier score,

$$\mathrm{Brier} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{p}_i)^2, \tag{6}$$

for which lower values are preferable. Calibration was summarized by the calibration intercept and slope from logistic regression of outcome on predicted log odds, an equal-frequency ten-bin expected calibration error, and the absolute difference between mean prediction and observed response within each centre. A calibration intercept of zero and slope of one indicate agreement on the logistic scale. Discrimination was quantified by AUC.

Clinical utility was assessed with net benefit,

$$\mathrm{NB}(t) = \frac{\mathrm{TP}(t)}{n} - \frac{\mathrm{FP}(t)}{n}\frac{t}{1-t}, \tag{7}$$

where t is the probability threshold, TP is the number of true-positive treatment recommendations, and FP is the number of false-positive recommendations. The expression converts false-positive recommendations into the same units as true positives using the odds at the chosen threshold. Curves were evaluated from 0.10 to 0.60, with numerical summaries at 0.25, 0.35, and 0.45.

Group calibration was examined for men with unipolar depression, women with unipolar depression, men with bipolar depression, and women with bipolar depression. These groups were selected before outcome generation because the transported score contains strong sex and diagnostic coefficients and the outcome mechanism contains their interaction. The largest absolute group error was used as a guard against acceptable pooled results masking concentrated miscalibration.

Paired differences from the fixed score were estimated in 600 bootstrap samples, resampling individuals with replacement separately within each centre. The 2.5th and 97.5th percentiles formed 95% intervals. Analyses used Python 3.12, NumPy, pandas, SciPy, scikit-learn, and Matplotlib. All reported values were read directly from machine-generated comma-separated files to avoid manual transcription errors.

3. RESULTS

3.1. Population and response distribution

The constructed population contained 48,000 records, of which 3,200 were assigned to updating and 44,800 to held-out evaluation. Response prevalence in the held-out population was 45.24%, but centre prevalence ranged from 19.73% in Centre E to 63.04% in Centre A. The median transported score also varied substantially because symptom severity, resistance, diagnosis, sex composition, thickness, and distance differed by centre.

As shown in Figure 1, the transported probability distributions preserved wide within-centre variation, yet the observed response proportions did not consistently align with their medians. The most visible discrepancies occurred in Centres A and E: the unchanged score underpredicted the high response frequency in A and overpredicted the low response frequency in E. This reversal is important because it cannot be resolved by a single pooled intercept change; a correction that helps one centre can worsen the other.

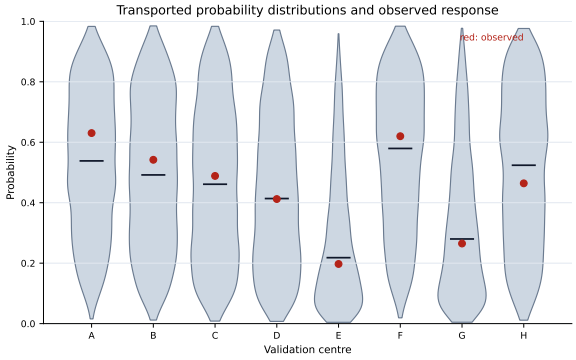

Figure 1. Transported probabilities and centre response.

3.2. Overall probability accuracy and discrimination

The unchanged score achieved an AUC of 0.790, a Brier score of 0.1856, a calibration intercept of -0.001, and a calibration slope of 0.953 (Table 2). Its near-zero pooled intercept concealed local differences in opposing directions. Centre-specific intercept correction produced the lowest Brier score, 0.1812, and the highest AUC, 0.801. Although intercept correction does not alter ranking within a centre, its centre-dependent shifts changed the ordering of records across the pooled population, accounting for the modest AUC increase.

The equal-frequency expected calibration error favoured the fixed score because errors of opposite sign cancelled in the pooled population and the binning procedure did not respect centre. This result illustrates why a single pooled calibration summary is insufficient for transport evaluation. PPTC improved the Brier score to 0.1823 and AUC to 0.798, but its calibration slope of 1.090 indicated mild compression of the updated probabilities. Pooled recalibration changed neither rank discrimination nor probability loss in a clinically meaningful way.

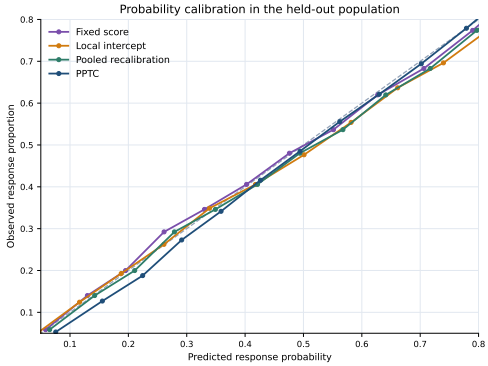

Figure 2. Calibration in held-out records.

The pooled reliability curves in Figure 2 appear close to the identity line for all approaches. PPTC showed underprediction in the lower-middle probability range and slight overprediction at the upper end, consistent with its slope above one. Local intercept correction tracked the identity line closely despite a non-zero pooled calibration intercept because the eight locally updated probabil-

Table 2. Held-out performance.

Approach	Brier	ECE	AUC	Cal. int.	Cal. slope	NB .25	NB .35	NB .45
Fixed score	0.1856	0.0106	0.7901	−0.001	0.953	0.3085	0.2498	0.1947
Local intercept	**0.1812**	0.0168	**0.8012**	−0.078	0.922	**0.3137**	**0.2579**	**0.2036**
Pooled recalibration	0.1857	0.0141	0.7901	−0.069	0.972	0.3086	0.2505	0.1945
PPTC	0.1823	0.0156	0.7984	−0.060	1.090	0.3122	0.2567	0.2007

ity scales overlap unevenly when combined. The graph reinforces that pooled visual agreement should be interpreted together with centre-stratified evidence.

3.3. Centre-level calibration

Absolute centre errors for the unchanged score were 9.90, 4.63, 1.97, 2.18, 8.68, 6.55, 6.55, and 4.87 percentage points from A through H. Pooled recalibration reduced error in Centres A–C but increased error in D, E, G, and H, reaching 9.98 points in Centre E. A common recalibration line was unable to accommodate both high- and low-prevalence services.

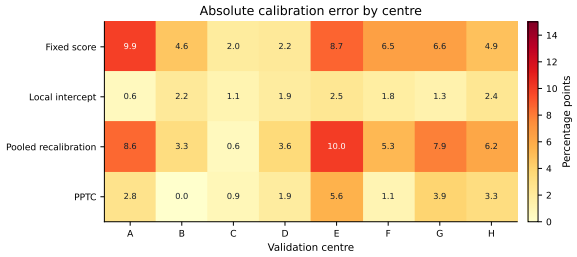

Figure 3. Centre-level calibration error.

The centre-resolved map in Figure 3 identifies local intercept correction as the only approach that kept every absolute error below 2.5 percentage points. Its largest error was 2.46 points in Centre E and its smallest was 0.63 points in Centre A. PPTC reduced error relative to the fixed score in every centre, but residual error remained 5.61 points in Centre E, 3.93 points in Centre G, and 3.34 points in Centre H. The new method therefore achieved partial adaptation without matching the directness of a prevalence correction.

The bootstrap difference in maximum centre error was -7.04 percentage points for local intercept correction, with a 95% interval from -8.48 to -5.75 points. PPTC reduced the maximum by -4.33 points (-5.91 to -3.07). Pooled recalibration differed from the fixed score by only $+0.07$ points, and its interval included improvements and deterioration (-1.33 to $+1.30$). These paired results exclude random outcome sampling as a plausible explanation for the superiority of the local correction within the constructed populations.

3.4. Decision-curve utility

Net benefit at thresholds of 0.25, 0.35, and 0.45 was highest for the local intercept approach. At 0.35, the fixed score yielded 0.2498, local intercept correction yielded 0.2579, pooled recalibration yielded 0.2505, and PPTC yielded 0.2567. Relative to the fixed score, the paired improvement was 0.00809 (95% interval 0.00626–0.00997) for local correction and 0.00689 (0.00500–0.00909) for PPTC. The pooled difference was 0.00076 (-0.00023 to 0.00178).

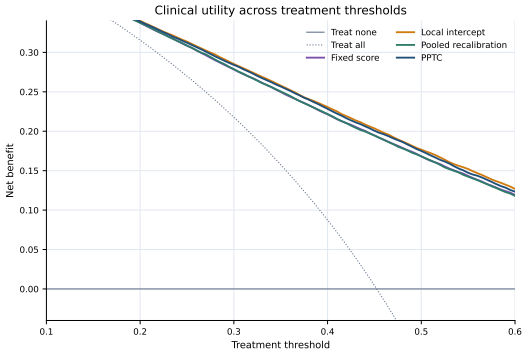

Figure 4. Decision curves.

Across the threshold interval shown in Figure 4, local intercept correction remained at or near the upper envelope of model-based strategies. The absolute separation is modest because all methods share the same dominant transported signal, but the advantage persists through the clinically plausible middle of the range. Treating every patient becomes inferior to probability-guided selection as the threshold rises, while treating none has zero net benefit by definition. The decision curves therefore support the calibration findings: centre prevalence contains actionable information that a pooled update fails to preserve.

3.5. Sex-by-diagnosis calibration

The four prespecified groups included 14,976 men with unipolar depression, 18,207 women with unipolar depression, 5,122 men with bipolar depression, and 6,495 women with bipolar depression. Observed response proportions were 29.78%, 50.49%, 40.69%, and 69.78%, respectively. The unchanged score underpredicted men with unipolar depression by 4.48 points and overpredicted women with bipolar depression by 3.36 points.

Group results in Figure 5 reveal a trade-off not visible in pooled accuracy. Local intercept correction limited the largest absolute error to 2.76 points, occurring among women with unipolar depression. Pooled recalibration reached 5.00 points among men with bipolar depression. PPTC limited errors to 2.92 points among men with unipolar depression, 2.69 points among men with bipolar depression, 1.03 points among women with unipolar

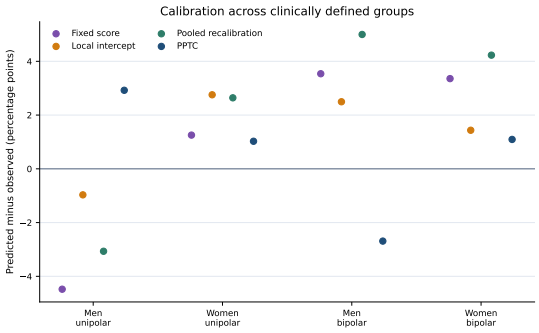

Figure 5. Calibration by sex and diagnosis.

depression, and 1.09 points among women with bipolar depression. Explicit modelling of sex, diagnosis, and their interaction improved balance across groups, although its largest error remained slightly greater than that of local intercept correction.

The subgroup finding explains why PPTC did not overtake the simpler correction despite lower centre errors than the fixed score. Strong penalization stabilized the site and patient-level terms, but the finite update samples did not identify centre shifts as directly as a prevalence correction. A local intercept does not attempt to relearn clinical effects; it preserves the transported contrasts while correcting the most precisely estimable local quantity, the response prevalence.

4. DISCUSSION

This study asked whether a site-aware penalized update justifies its complexity when a high-frequency rTMS response score is applied to populations with different clinical and anatomical compositions. The answer was no under the prespecified conditions. PPTC improved Brier loss, centre calibration, AUC, and net benefit relative to the unchanged score, demonstrating that partial pooling can recover some transport error. Nevertheless, a centre-specific intercept correction produced lower probability loss, smaller maximum centre error, greater decision utility, and better worst-group calibration. Pooled logistic recalibration was largely ineffective because a single intercept and slope averaged centre shifts that occurred in opposite directions.

The most consequential observation is the contrast between pooled and local calibration. The fixed score had a pooled calibration intercept close to zero, a slope close to one, and an AUC of 0.790. Those values could be presented as successful external performance. Yet the same score missed response prevalence by almost ten percentage points in one centre and nearly nine points in another. The apparent pooled agreement arose because underprediction in high-response centres offset overprediction in low-response centres. This cancellation is a mathematical property of aggregation, not evidence that probabilities are clinically reliable. Calibration assessment must therefore reflect the unit at which decisions are made. If referrals and treatment thresholds are set by a local service, centre-level calibration is the relevant quantity.

The superiority of intercept correction is consistent with the design of the transport challenge. Centre prevalence varied markedly, whereas much of the individual ordering encoded by the transported coefficients remained useful. A local intercept uses one degree of freedom to target precisely that change. It cannot repair a reversed predictor effect or a severe nonlinear relation, but it estimates a common local shift with much less variance than a multivariable update. The 400-record update sample was adequate for estimating prevalence but comparatively sparse for learning site indicators, six continuous effects, and correlated clinical-anatomical structure. Strong penalization protected PPTC from extreme variance, as shown by the selected inverse penalty of 0.01, but it also attenuated adjustments and contributed to the calibration slope above one.

These findings have direct implications for rTMS biomarker research. Investigations frequently seek increasingly rich predictor sets, including cortical thickness, gray-matter volume, functional connectivity, diffusion measures, electric-field estimates, and symptom dimensions.[26,27] Multivariate neuroimaging can improve prediction in selected samples, but complexity raises requirements for harmonization, missing-data handling, scanner stability, and external sample size. A more complex model is useful only if it improves decisions in the population where it will be deployed. Before adding a new imaging feature, investigators should determine whether local probability error is already resolved by an intercept update. If it is, the imaging feature must demonstrate incremental value beyond that corrected reference, not merely beyond an unchanged score.

The present results do not diminish the biological importance of prefrontal anatomy. DLPFC and anterior cingulate morphology have plausible relations to the networks engaged by stimulation, and cortical depth influences the electric field delivered to neural tissue.[28,29] Studies have reported associations between cortical structure and response to rTMS, electroconvulsive therapy, transcranial direct current stimulation, antidepressant medication, and psychotherapy.[30–32] Longitudinal work also suggests that neuromodulation may be accompanied by structural plasticity.[33] The transport issue is different: even a genuine biological association may yield inaccurate absolute probabilities when diagnosis, treatment resistance, symptom burden, coil placement, or concomitant medication differs between centres.

Diagnostic composition deserves particular attention. Bipolar and unipolar depressive episodes share substantial symptom burden but may differ in clinical history, network physiology, treatment exposure, and response to stimulation.[34,35] Published naturalistic comparisons have not agreed on whether bipolar diagnosis predicts a more favourable response to high-frequency left-DLPFC treatment.[8,10] Frequency, intensity, pulse count, course

length, and eligibility criteria vary across studies, making a diagnostic coefficient especially vulnerable to transport. In the constructed populations, bipolar prevalence ranged from 10% to 40%, and diagnostic status interacted with sex and cortical thickness in the outcome mechanism. The resulting subgroup analysis showed that a flexible update could improve one group while worsening another. This finding supports routine reporting of group calibration rather than reliance on an overall interaction p-value.

Sex-related effects require similar caution. Meta-analytic observations have linked a greater proportion of women to larger antidepressant effects in some rTMS trials, but study-level associations cannot establish an individual causal effect.[36,37] Sex may correlate with age, menopausal status, illness course, medication, motor threshold, cortical geometry, or referral patterns. Encoding a large sex coefficient in a clinical tool can therefore transfer poorly when those correlated factors change. The safest response is not to delete the predictor automatically, because that can also impair calibration, but to evaluate sex-specific probability agreement and clinical consequences in each intended population.

Treatment resistance and symptom burden also contribute to case-mix shift. Lower degrees of prior treatment failure have repeatedly been associated with better rTMS outcomes, although definitions of resistance are heterogeneous.[7,38] Entry MADRS score affects both the available range for a 50% reduction and the clinical interpretation of response. A model transported from a service treating moderately severe episodes to one treating more severe and refractory episodes may overstate benefit even if the direction of every coefficient remains correct. Centre E combined the greatest mean resistance, highest mean MADRS score, shortest mean cortical thickness, and longest mean brain–scalp distance; it consequently exhibited the lowest response prevalence and the largest error after pooled recalibration. This combination illustrates why one-variable subgroup checks cannot replace multivariable transport assessment.

Concomitant medication is another source of unmodelled heterogeneity. Benzodiazepines, anticonvulsants, antidepressants, antipsychotics, and lithium may influence cortical excitability or mark differences in illness severity. Naturalistic reports disagree about the extent to which benzodiazepine use attenuates rTMS response.[39,40] Medication was not explicitly generated here because the published coefficient vector did not include drug terms. In a prospective validation, medication classes, doses, and changes during treatment should be recorded and assessed as transport variables rather than introduced after model failure is observed.

The decision-curve findings translate statistical improvement into treatment-selection consequences. At a threshold of 0.35, intercept correction increased net benefit by approximately 0.008 relative to the fixed score. Interpreted per 100 treated candidates, this is equivalent to about 0.8 additional true-positive recommendations without increasing false positives, after weighting false positives by the threshold odds. The number is not a universal clinical effect; it depends on the chosen threshold and population. Its importance lies in demonstrating that calibration repair can improve decisions without discovering a new biomarker. Decision thresholds should be selected with patients, clinicians, resource constraints, and alternatives in mind rather than chosen to maximize an internal classification statistic.

Pooled recalibration performed poorly because it estimated a common correction across heterogeneous centres. In Centre A, the score needed to move upward, whereas in Centre E it needed to move downward. The pooled model settled near the aggregate response rate and therefore left both errors largely intact. This behaviour cautions against combining institutions during validation without modelling institution. Large sample size does not compensate for estimating the wrong target. A dataset with tens of thousands of records can give an extremely precise estimate of an average correction that is inappropriate for every participating service.

PPTC occupied an intermediate position. Its site indicators moved probabilities in the correct direction, reducing the maximum centre error by 4.33 percentage points relative to the unchanged score. It also improved net benefit and pooled AUC, while explicit sex-by-diagnosis modelling kept every prespecified group error below three points. The result still illustrates bias–variance allocation: penalized multivariable updating borrows strength across centres and spreads a limited update sample across several effects, whereas an intercept correction directs all available information toward local prevalence. A larger local sample, hierarchical priors designed for sparse centres, or a calibration constraint could change the ranking. Those alternatives require separate evaluation rather than retrospective selection based on the present outcomes.

Several features strengthen the study. The research question, population shifts, update sizes, nonlinear departures, and evaluation measures were specified in executable code. Every method was compared on the same 44,800 held-out records, and uncertainty intervals used paired, centre-stratified resampling. The eight centres were designed to prevent one demographic variable from serving as a proxy for all others. Probability accuracy, discrimination, centre calibration, group calibration, and decision utility were examined together. Finally, the results include an unfavourable finding for the new method, reducing the risk that novelty was equated with superiority.

Important limitations define the scope of inference. First, the populations are in silico and contain no newly treated patients. Numerical control provides a clean transport experiment but cannot reproduce the full dependence structure of clinical variables, scanner effects, comorbidity, medication, attrition, coil positioning, or clinician

behaviour. The results should guide a prospective validation plan, not a treatment decision. Second, the generator was anchored to one reported coefficient vector. Other protocols, including 10-Hz stimulation, intermittent theta-burst stimulation, bilateral stimulation, longer courses, or individualized functional targets, may produce different transport behaviour. Comparative effectiveness can depend on frequency and target.[41,42]

Third, the outcome mechanism included explicit nonlinear and interaction terms selected for interpretability rather than estimated from patient records. The numerical magnitude of every performance difference is conditional on those choices. The qualitative conclusion is therefore strongest for settings dominated by prevalence drift with preserved individual ranking. If predictor effects reverse, intercept correction will be insufficient. Fourth, the analysis used a binary 50% MADRS reduction. Continuous symptom change, remission, durability, functioning, adverse effects, and patient-valued outcomes may yield different conclusions. Binary response also discards information near the cutoff.

Fifth, centre-specific intercept correction requires local outcomes collected with the same definition and follow-up timing as the transported model. Differences in assessment schedule or missing follow-up can create apparent prevalence drift that is actually measurement drift. Sixth, the study did not model a sham-treated population and cannot distinguish prognostic response from treatment-specific benefit. A predictor of improvement among treated patients may also predict spontaneous recovery, regression to the mean, or nonspecific care effects. Prospective decision tools should ideally estimate individualized treatment effects when a suitable comparator exists.

The next empirical test should be a locked, multicentre external validation in which the unchanged score, local intercept correction, and PPTC are specified before outcomes are examined. Recruitment should span services with different diagnostic composition, treatment resistance, medication use, scanners, and targeting practices. Each centre should report calibration-in-the-large, calibration slope, flexible calibration curves, Brier score, AUC, decision curves, and prespecified subgroup calibration. Model updating should occur only after the unchanged score has been evaluated, and all update sample sizes should be justified by the desired precision of local probability estimates. This sequence would determine whether the present negative comparison for added complexity persists in treated patients.

5. CONCLUSION

The paper's research question was whether a site-aware penalized update provides enough transport benefit to justify its complexity when a high-frequency left-DLPFC rTMS response score crosses populations with different clinical and anatomical compositions. In this controlled multicentre study, PPTC improved on unchanged transport but did not surpass a centre-specific intercept correction. The local correction reduced the largest centre error from 9.90 to 2.46 percentage points, lowered Brier loss, increased net benefit across the main treatment-threshold range, and limited error in sex-by-diagnosis groups. Pooled recalibration failed because opposite centre shifts cancelled in aggregate. The appropriate operational conclusion is specific: before using an rTMS response probability to guide treatment thresholds, each service should measure local outcome prevalence and test an intercept update; additional multivariable adaptation should be adopted only if it demonstrates incremental centre and subgroup benefit in held-out local records. s.

FUNDING

No external funding was used for this computational study.

COMPETING INTERESTS

The authors declare no competing interests.

BIBLIOGRAFIA / REFERENCES

1. Lefaucheur J-P, Aleman A, Baeken C, et al. Evidence-based guidelines on the therapeutic use of repetitive transcranial magnetic stimulation: an update (2014–2018). *Clinical Neurophysiology*. 2020;131:474–528. doi:10.1016/j.clinph.2019.11.002.
2. Perera T, George MS, Grammer G, Janicak PG, Pascual-Leone A, Wirecki TS. The Clinical TMS Society consensus review and treatment recommendations for TMS therapy for major depressive disorder. *Brain Stimulation*. 2016;9:336–346. doi:10.1016/j.brs.2016.03.010.
3. Sehatzadeh S, Daskalakis ZJ, Yap B, et al. Unilateral and bilateral repetitive transcranial magnetic stimulation for treatment-resistant depression: a meta-analysis of randomized controlled trials over two decades. *Journal of Psychiatry and Neuroscience*. 2019;44:151–163. doi:10.1503/jpn.180056.
4. Brunoni AR, Chaimani A, Moffa AH, et al. Repetitive transcranial magnetic stimulation for the acute treatment of major depressive episodes: a systematic review with network meta-analysis. *JAMA Psychiatry*. 2017;74:143–152. doi:10.1001/jamapsychiatry.2016.3644.
5. Sackeim HA, Aaronson ST, Carpenter LL, et al. Clinical outcomes in a large registry of patients with major depressive disorder treated with transcranial magnetic stimulation. *Journal of Affective Disorders*. 2020;277:65–74. doi:10.1016/j.jad.2020.08.005.
6. Kar SK. Predictors of response to repetitive transcranial magnetic stimulation in depression: a review of recent updates. *Clinical Psychopharmacology and Neuroscience*. 2019;17:25–33. doi:10.9758/cpn.2019.17.1.25.
7. Garnaat SL, Fukuda AM, Yuan S, Carpenter LL. Identification of clinical features and biomarkers

that may inform a personalized approach to rTMS for depression. *Personalized Medicine in Psychiatry*. 2019;17–18:4–16. doi:10.1016/j.pmip.2019.09.001.

8. Rostami R, Kazemi R, Nitsche MA, Gholipour F, Salehinejad MA. Clinical and demographic predictors of response to rTMS treatment in unipolar and bipolar depressive disorders. *Clinical Neurophysiology*. 2017;128:1961–1970. doi:10.1016/j.clinph.2017.07.395.
9. Lacroix A, Calvet B, Laplace B, et al. Predictors of clinical response after rTMS treatment of patients suffering from drug-resistant depression. *Translational Psychiatry*. 2021;11:587. doi:10.1038/s41398-021-01555-9.
10. Yang YB, Chan P, Rayani K, McGirr A. Comparative effectiveness of repetitive transcranial magnetic stimulation in unipolar and bipolar depression. *Canadian Journal of Psychiatry*. 2021;66:313–315. doi:10.1177/0706743720950938.
11. Nguyen TD, Hieronymus F, Lorentzen R, McGirr A, Østergaard SD. The efficacy of repetitive transcranial magnetic stimulation for bipolar depression: a systematic review and meta-analysis. *Journal of Affective Disorders*. 2021;279:250–255. doi:10.1016/j.jad.2020.10.013.
12. McGirr A, Karmani S, Arsappa R, et al. Clinical efficacy and safety of repetitive transcranial magnetic stimulation in acute bipolar depression. *World Psychiatry*. 2016;15:85–86. doi:10.1002/wps.20300.
13. Mutz J, Edgcumbe DR, Brunoni AR, Fu CHY. Efficacy and acceptability of non-invasive brain stimulation for adult unipolar and bipolar depression: a systematic review and meta-analysis of randomised sham-controlled trials. *Neuroscience and Biobehavioral Reviews*. 2018;92:291–303. doi:10.1016/j.neubiorev.2018.05.015.
14. Boes AD, Uitermarkt BD, Albazron FM, et al. Rostral anterior cingulate cortex is a structural correlate of repetitive TMS treatment response in depression. *Brain Stimulation*. 2018;11:575–581. doi:10.1016/j.brs.2018.01.029.
15. Baeken C, van Beek V, Vanderhasselt M-A, Duprat R, Klooster D. Cortical thickness in the right anterior cingulate cortex relates to clinical response to left prefrontal accelerated intermittent theta-burst stimulation. *Neuromodulation*. 2021;24:938–949. doi:10.1111/ner.13380.
16. Seeberg I, Kjaerstad HL, Miskowiak KW. Neural and behavioral predictors of treatment efficacy on mood symptoms and cognition in mood disorders: a systematic review. *Frontiers in Psychiatry*. 2018;9:337. doi:10.3389/fpsyt.2018.00337.
17. Desikan RS, Ségonne F, Fischl B, et al. An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest. *NeuroImage*. 2006;31:968–980. doi:10.1016/j.neuroimage.2006.01.021.
18. Cash RFH, Cocchi L, Anderson R, et al. A multivariate neuroimaging biomarker of individual outcome to transcranial magnetic stimulation in depression. *Human Brain Mapping*. 2019;40:4618–4629. doi:10.1002/hbm.24725.
19. Ge R, Downar J, Blumberger DM, Daskalakis ZJ, Lam RW, Vila-Rodriguez F. Structural network integrity of the central executive network is associated with the therapeutic effect of rTMS in treatment-resistant depression. *Progress in Neuro-Psychopharmacology and Biological Psychiatry*. 2019;92:217–225. doi:10.1016/j.pnpbp.2019.01.012.
20. Hamilton JP, Chen MC, Gotlib IH. Neural systems approaches to understanding major depressive disorder: an intrinsic functional organization perspective. *Neurobiology of Disease*. 2013;52:4–11. doi:10.1016/j.nbd.2012.01.015.
21. Van Calster B, McLernon DJ, van Smeden M, Wynants L, Steyerberg EW. Calibration: the Achilles heel of predictive analytics. *BMC Medicine*. 2019;17:230. doi:10.1186/s12916-019-1466-7.
22. Collins GS, Reitsma JB, Altman DG, Moons KGM. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis: the TRIPOD statement. *Annals of Internal Medicine*. 2015;162:55–63. doi:10.7326/M14-0697.
23. Vickers AJ, Elkin EB. Decision curve analysis: a novel method for evaluating prediction models. *Medical Decision Making*. 2006;26:565–574. doi:10.1177/0272989X06295361.
24. Harika-Germaneau G, Wassouf I, Le Tutour T, et al. Baseline clinical and neuroimaging biomarkers of treatment response to high-frequency rTMS over the left DLPFC for resistant depression. *Frontiers in Psychiatry*. 2022;13:894473. doi:10.3389/fpsyt.2022.894473.
25. Riley RD, Debray TPA, Collins GS, et al. Minimum sample size for external validation of a clinical prediction model with a binary outcome. *Statistics in Medicine*. 2021;40:4230–4251. doi:10.1002/sim.9025.
26. Cash RFH, Weigand A, Zalesky A, et al. Using brain imaging to improve spatial targeting of transcranial magnetic stimulation for depression. *Biological Psychiatry*. 2021;90:689–700. doi:10.1016/j.biopsych.2020.05.033.
27. Miron J-P, Jodoin VD, Lespérance P, Blumberger DM. Repetitive transcranial magnetic stimulation for major depressive disorder: basic principles and future directions. *Therapeutic Advances in Psychopharmacology*. 2021;11:20451253211042696. doi:10.1177/20451253211042696.
28. Lee EG, Duffy W, Hadimani RL, et al. Investigational effect of brain–scalp distance on the efficacy of transcranial magnetic stimulation treatment in depression. *IEEE Transactions on Magnetics*. 2016;52:1–4. doi:10.1109/TMAG.2015.2514158.
29. Davis NJ. Variance in cortical depth across the brain surface: implications for transcranial stimulation of the brain. *European Journal of Neuroscience*. 2021;53:996–1007. doi:10.1111/ejn.14957.
30. Filmer HL, Ehrhardt SE, Shaw TB, Mattingley JB, Dux PE. The efficacy of transcranial direct current stimulation to prefrontal areas is related to underlying

cortical morphology. *NeuroImage*. 2019;196:41–48. doi:10.1016/j.neuroimage.2019.04.026.
31. Redlich R, Opel N, Grotegerd D, et al. Prediction of individual response to electroconvulsive therapy via machine learning on structural magnetic resonance imaging data. *JAMA Psychiatry*. 2016;73:557–564. doi:10.1001/jamapsychiatry.2016.0316.
32. Costafreda SG, Khanna A, Mourao-Miranda J, Fu CHY. Neural correlates of sad faces predict clinical remission to cognitive behavioural therapy in depression. *NeuroReport*. 2009;20:637–641. doi:10.1097/WNR.0b013e3283294159.
33. Dalhuisen I, Ackermans E, Martens L, et al. Longitudinal effects of rTMS on neuroplasticity in chronic treatment-resistant depression. *European Archives of Psychiatry and Clinical Neuroscience*. 2021;271:39–47. doi:10.1007/s00406-020-01135-w.
34. Han K-M, De Berardis D, Fornaro M, Kim Y-K. Differentiating between bipolar and unipolar depression in functional and structural MRI studies. *Progress in Neuro-Psychopharmacology and Biological Psychiatry*. 2019;91:20–27. doi:10.1016/j.pnpbp.2018.03.022.
35. Gold AK, Ornelas AC, Cirillo P, et al. Clinical applications of transcranial magnetic stimulation in bipolar disorder. *Brain and Behavior*. 2019;9:e01419. doi:10.1002/brb3.1419.
36. Kedzior KK, Azorina V, Reitz SK. More female patients and fewer stimuli per session are associated with the short-term antidepressant properties of rTMS: a meta-analysis of 54 sham-controlled studies. *Neuropsychiatric Disease and Treatment*. 2014;10:727–756. doi:10.2147/NDT.S58405.
37. Kedzior KK, Reitz SK. Short-term efficacy of repetitive transcranial magnetic stimulation in depression: reanalysis of data from meta-analyses up to 2010. *BMC Psychology*. 2014;2:39. doi:10.1186/s40359-014-0039-y.
38. Brakemeier E-L, Wilbertz G, Rodax S, et al. Patterns of response to repetitive transcranial magnetic stimulation in major depression: replication study in drug-free patients. *Journal of Affective Disorders*. 2008;108:59–70. doi:10.1016/j.jad.2007.09.007.
39. Hunter AM, Minzenberg MJ, Cook IA, et al. Concomitant medication use and clinical outcome of rTMS treatment of major depressive disorder. *Brain and Behavior*. 2019;9:e01275. doi:10.1002/brb3.1275.
40. Deppe M, Abdelnaim M, Hebel T, et al. Concomitant lorazepam use and antidepressive efficacy of repetitive transcranial magnetic stimulation in a naturalistic setting. *European Archives of Psychiatry and Clinical Neuroscience*. 2021;271:61–67. doi:10.1007/s00406-020-01160-9.
41. Cohen RB, Brunoni AR, Boggio PS, Fregni F. Clinical predictors associated with duration of repetitive transcranial magnetic stimulation treatment for remission in bipolar depression. *Journal of Nervous and Mental Disease*. 2010;198:679–681. doi:10.1097/NMD.0b013e3181ef2175.
42. Fitzgerald PB, Hoy KE, Anderson RJ, Daskalakis ZJ. A study of the pattern of response to rTMS treatment in depression. *Depression and Anxiety*. 2016;33:746–753. doi:10.1002/da.22503.