PSILOGOS

Research Article

open-access

Does a Lesion–Symptom Probability Cascade Improve Early Post-Stroke Depression Stratification? An Aggregate-Calibrated Digital-Cohort Study

Wolfgang Wagner

ABSTRACT

Early post-stroke depression is clinically heterogeneous, and an admission-time estimate should respect both lesion anatomy and the ordered dependence of depressive symptoms. Whether an explicit lesion-to-central-symptom-to-peripheral-symptom cascade adds predictive information beyond direct lesion-aware regression is unknown. A digital cohort of 17,220 profiles was generated from prespecified demographic, vascular, neurological, anatomical, and 13-symptom constraints. Six deep-structure indicators represented the right and left basal ganglia, right and left internal capsules, right external capsule, and right insula. A lesion–symptom probability cascade first estimated depressed mood, psychiatric anxiety, and loss of interest; estimated the remaining ten symptoms from admission variables and out-of-fold central-symptom probabilities; and then estimated the probability of a Hamilton Depression Rating Scale total above 7. Clinical logistic regression, lesion-augmented logistic regression, extra trees, and histogram boosting were evaluated with the same repeated outer validation splits. Discrimination, average precision, Brier loss, calibration intercept and slope, operating characteristics, prevalence shifts, probability attenuation, and subgroup consistency were examined. Lesion-augmented logistic regression attained an area under the receiver-operating curve (AUROC) of 0.6912 and average precision of 0.7317. The cascade attained an AUROC of 0.6911, average precision of 0.7316, Brier loss of 0.2196, calibration intercept of -0.001, and calibration slope of 0.980. Its AUROC therefore differed from lesion-augmented logistic regression by -0.0001. Right basal-ganglia involvement produced the largest absolute risk difference (36.4 percentage points, bootstrap 95% interval 34.7–38.4). Cascade discrimination remained between 0.677 and 0.697 across prespecified subgroups and was stable when outcome prevalence ranged from 0.35 to 0.65, whereas probability attenuation predictably distorted calibration slopes. Under aggregate-calibrated conditions, modeling a symptom cascade did not improve discrimination beyond direct lesion-aware regression. Deep-structure anatomy accounted for the measurable gain over clinical variables, while the cascade supplied an interpretable decomposition and well-calibrated probabilities. Prospective external validation with individual imaging and symptom trajectories is required before clinical use.

Keywords: post-stroke depression; lesion anatomy; synthetic health data; stacked probability model; calibration; clinical prediction

Department of Psychiatry and Psychotherapy, Medical University of Vienna, Vienna, Austria
wolfgang.wagner99@gmail.com

– **Published:** 20 November 2024

1. INTRODUCTION

Stroke remains a leading cause of death and disability, but survival statistics alone do not describe the neuropsychiatric burden carried by patients and families.[1] Depression after stroke is among the most consequential of these burdens. Pooled estimates place its frequency near one third of survivors, although estimates change with assessment instrument, time since stroke, care setting, and eligibility rules.[2,3] Depressive symptoms are associated with poorer functional recovery, lower treatment participation, recurrent vascular events, diminished quality of life, and excess mortality.[4,5] The first weeks after stroke are especially important because neurological deficits, sleep disruption, uncertainty, separation from ordinary social roles, and direct injury to mood-related circuits converge before long-term adaptation has occurred. A screening decision made during this period must therefore distinguish a clinically meaningful depressive state from symptoms that may accompany acute illness while remaining sensitive enough not to miss patients who need structured follow-up.

The determinants of post-stroke depression span several levels. Demographic and psychiatric history, stroke severity, disability, cognitive impairment, social support, lesion volume, and lesion location have all been implicated.[6–8] Their effects are not interchangeable. A stressful life event can alter affect without focal neuroanatomical damage; a capsular lesion may simultaneously increase motor disability and disrupt subcortical connections; cognitive impairment can reduce the reliability of self-report; and aphasia can exclude high-risk patients from conventional assessments. Social isolation and reduced participation also have durable effects that cannot be reduced to lesion coordinates.[9] A clinically credible prediction method should consequently preserve the distinction between admission-time neurological information and the depressive manifestations that emerge during hospitalization.

The use of a single depression sum score introduces a further difficulty. Two individuals with the same total can have markedly different symptom compositions, severities, functional implications, and safety needs. Psychiatric network research addresses this heterogeneity by treating symptoms as interacting elements rather than exchangeable indicators of one hidden disorder.[10,11] In non-stroke depression, depressed mood, anhedonia, anxiety, guilt, sleep disturbance, and fatigue have shown different patterns of association with risk factors and illness course.[12,13] Networks can also change over time, making a central symptom at discharge less influential during community recovery.[14] Centrality estimates, however, are descriptive summaries of a fitted dependence structure; they do not by themselves establish temporal or causal direction, and their stability must be evaluated.[15,16] These cautions are particularly relevant when behavioral measurements are collected at one occasion.

Measurement timing changes the meaning of the target. A two-week assessment captures an early depressive phenotype during neurological stabilization, initiation of secondary prevention, and adaptation to disability. It is neither an admission diagnosis nor a chronic prognosis. Predictors measured after the two-week outcome would contaminate an admission-time estimate even if they were clinically plausible. Symptom scores collected at the target time pose the clearest risk: they are excellent descriptors of the outcome but unavailable when the prediction is supposed to be made. A valid staged model must therefore learn symptom tendencies from earlier neurological and anatomical variables, pass only predicted probabilities forward, and reserve observed symptoms for training targets. This temporal rule is more restrictive than a conventional cross-sectional symptom network, but it corresponds to the decision faced by a stroke team planning enhanced observation or scheduled psychiatric assessment.

The central–peripheral distinction also needs restraint. Central symptoms can be defined statistically, clinically, or temporally, and those definitions need not agree. A symptom with many partial correlations may not be the first symptom to appear; a clinically urgent symptom such as suicidality may be uncommon and therefore not central; and a highly predictable symptom may carry little additional information once disability is measured. The present ordering treats depressed mood, psychiatric anxiety, and loss of interest as intermediate targets because they connect affective, anxious, and motivational domains. It does not claim that the remaining symptoms are secondary in importance or treatment priority. This separation allows the numerical experiment to test information flow while preserving the distinct clinical status of each symptom.

Screening practice adds another layer. A useful estimate should support a proportionate action, such as repeat interview, referral, collateral history, or closer monitoring, rather than replacing professional assessment. False negatives may delay recognition of severe mood disturbance, whereas false positives consume scarce psychiatric resources and may medicalize transient distress. The appropriate threshold therefore depends on the action and setting. A model intended for broad ward screening may accept lower specificity than one that directly triggers medication. These consequences motivate simultaneous reporting of probability calibration, sensitivity, specificity, and prevalence-dependent precision instead of presenting one discrimination statistic as a complete measure of usefulness.

Lesion studies add anatomical specificity but have produced heterogeneous findings. Earlier location studies were limited by small samples, coarse hemispheric classifications, variable depression definitions, and systematic exclusion of patients with language impairment. Contemporary work has emphasized circuits linking prefrontal, limbic, striatal, and thalamic territories rather than a simple left–right rule.[17,18] Lesion-

network approaches show that anatomically distant injuries can be functionally connected to a common circuit,[19,20] while focal-lesion evidence has been used to derive a distributed depression circuit.[21] Voxelwise lesion–behavior mapping increases spatial resolution but remains sensitive to lesion coverage, sample size, multiple-comparison control, and correlations among neighboring voxels.[22,23] Atlas labeling is likewise an interpretive layer rather than direct proof of a biological mechanism.[24] These considerations favor anatomically constrained prediction accompanied by uncertainty and robustness analyses.

Clinical prediction research presents its own methodological hazards. High apparent accuracy can arise from predictor leakage, repeated use of one test set, optimistic tuning, inadequate events per parameter, or failure to assess calibration. Transparent reporting recommendations therefore require a clear target population, predictor timing, outcome definition, validation design, and full performance description.[26] Risk-of-bias assessment further emphasizes participant selection, predictor measurement, outcome measurement, and statistical analysis.[27] Discrimination alone is insufficient: a model may rank patients correctly but systematically overestimate or underestimate their probabilities.[28] Calibration intercept and slope, Brier loss, and decision-relevant operating characteristics complement the AUROC and provide information that rank measures cannot.[29,30]

Digital cohorts offer a controlled setting in which a new computational method can be tested without claiming access to unavailable individual records. Synthetic health data can preserve selected univariate and multivariate properties, support reproducible method development, and reduce disclosure risks, but its utility depends on how generation constraints and intended analyses are defined.[31,32] Validation studies have shown that synthetic clinical data can reproduce some analytical conclusions while also warning that distributional similarity does not guarantee equal subgroup behavior or transportability.[33,34] A digital cohort is consequently appropriate for a numerical experiment about method behavior, but it cannot establish clinical effectiveness.

The present study asks a specific question: *when admission variables and deep-structure lesion descriptors are available, does an explicitly ordered lesion-to-central-symptom-to-peripheral-symptom probability cascade improve stratification of suprathreshold depressive symptoms at two weeks beyond a direct lesion-augmented model?* The analysis uses a cohort twenty times larger than the calibration population to reduce Monte Carlo noise while retaining reported marginal characteristics. The original network estimation and lesion localization methods are retained as scientific premises: depressive symptoms remain distinct variables, central symptoms are modeled separately, and lesion descriptors are anatomically localized. The added method is a leakage-controlled stacked cascade evaluated against linear and nonlinear comparators, with repeated validation, calibration analysis, anatomical risk differences, prevalence shifts, probability attenuation, and subgroup auditing. The principal objective is not to manufacture a favorable result but to determine whether the cascade contributes information that is absent from a flat anatomical predictor set.

2. METHODS

This was a computational study using digitally generated patient profiles. The target estimand was the out-of-sample probability of a Hamilton Depression Rating Scale (HDRS) total above 7 at two weeks after a first stroke, conditional on variables available during the initial admission. The digital records contain no names, dates, image files, institutions, or direct copies of patient-level observations. The calculation therefore evaluates statistical behavior under stated constraints rather than treatment efficacy, diagnostic safety, or deployment readiness.

The sample size was fixed at 17,220, equal to twenty replications of an 861-person calibration population. Increasing the digital cohort reduced random variation in the generation step and supplied sufficient positive and negative outcomes for repeated validation. It did not create new clinical evidence. The analysis protocol fixed the random seed at 20260814 before generation, defined all predictors and outcome rules in code, and stored fold-level estimates and out-of-fold probabilities.

2.1. Aggregate calibration and digital cohort generation

Demographic and clinical constraints comprised age, sex, education, stressful life events, smoking, alcohol use, diabetes, hypertension, hyperlipidemia, coronary disease, stroke type, National Institutes of Health Stroke Scale (NIHSS), Barthel Index, modified Rankin Scale, Mini-Mental State Examination (MMSE), lesion laterality, lesion volume, imaging delay, 13 depressive symptoms, HDRS total, and the proportion with HDRS above 7. The numerical targets were taken once from a three-center cohort of first-ever stroke patients assessed two weeks after onset.[40] Binary variables were assigned to achieve their target counts after rounding. Continuous variables were sampled from bounded normal, gamma, or log-normal distributions selected according to their scale and skew. Disability and cognition were coupled to age, education, stroke severity, and lesion volume so that the resulting profiles were clinically ordered rather than independently assembled.

The achieved clinical summaries in table 1 closely matched age, education, stroke severity, lesion volume, HDRS total, and all binary proportions. Barthel and MMSE variability was lower after values were constrained to their valid ranges and coupled to NIHSS. This difference is important because the digital cohort should not be mistaken for a record-level facsimile. The experi-

Table 1. Clinical calibration

Characteristic	Calibration value	Digital cohort
Age, years	57.98 (11.22)	57.84 (11.16)
Male sex, %	76.4	76.4
Education, years	9.40 (4.15)	9.52 (3.98)
Stressful life event, %	5.2	5.2
Current smoking, %	54.4	54.4
Alcohol use, %	49.2	49.2
Diabetes, %	22.8	22.8
Hypertension, %	57.7	57.7
Hyperlipidemia, %	18.7	18.7
Coronary disease, %	6.9	6.9
Ischemic stroke, %	90.0	90.0
NIHSS	3.85 (3.44)	3.85 (3.43)
Barthel Index	76.40 (27.95)	80.58 (18.88)
Modified Rankin Scale	2.26 (1.42)	2.20 (1.27)
MMSE	24.64 (4.74)	24.89 (3.33)
Lesion volume, cm^3	18.08 (32.69)	17.93 (27.66)
Imaging delay, days	3.06 (2.16)	3.43 (1.86)
HDRS total	9.06 (6.86)	9.37 (7.10)
HDRS $>$ 7, %	52.7	52.7

Values are mean (standard deviation) unless stated otherwise. Bounded dependence equations produced modest differences in some continuous summaries. HDRS, Hamilton Depression Rating Scale; MMSE, Mini-Mental State Examination; NIHSS, National Institutes of Health Stroke Scale.

ment retains the intended ordering among neurological severity, disability, cognition, and lesion burden while documenting where bounded generation narrows dispersion.

Lesion laterality was assigned as left hemispheric, right hemispheric, bilateral, or infratentorial. Six deep-structure indicators were then generated conditionally on laterality and log lesion volume: right and left basal ganglia, right and left internal capsules, right external capsule, and right insula. Conditional probabilities were higher for anatomically compatible laterality, bilateral lesions, and larger volumes. The right basal-ganglia and capsular coefficients were larger than their left-sided counterparts because the calibration evidence concentrated more associated voxels in right deep structures. These coefficients are explicit generative assumptions; they are not patient-specific anatomical reconstructions.

The realistic axial image in figure 1 was created only to orient readers to the anatomical level of the basal ganglia and internal capsules. It is a synthetic image, contains no participant information, and was not processed by the analysis. The absence of overlays and labels separates anatomical orientation from quantitative evidence.

Thirteen symptom liabilities were generated as correlated ordinal variables. Depressed mood, psychiatric anxiety, and loss of interest were designated central symptoms. Their liabilities depended on a shared affective term, neurological disability, psychosocial load, vascular burden, deep-structure lesion indicators, and random variation. The remaining symptoms depended on the

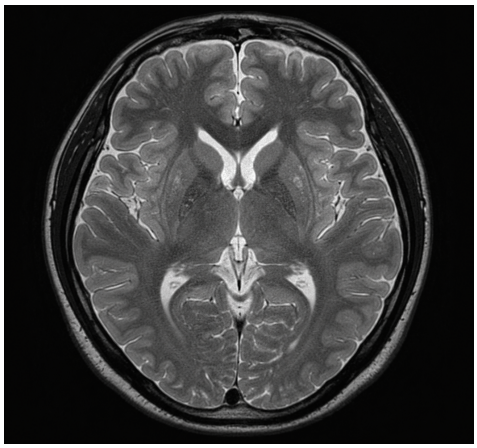

Figure 1. Synthetic axial MRI reference.

standardized central-symptom liability plus symptom-specific neurological or psychosocial terms. Ordinal thresholds were chosen to reproduce each nonzero prevalence; insomnia used a 0–6 range, gastrointestinal and general somatic symptoms used 0–2, and the other items used 0–4.

Exact agreement in nonzero symptom prevalence is visible in table 2, whereas the mean positive severity is a property of the ordinal generation rule. The total HDRS outcome was produced from the ordered symptom burden, deep-lesion load, disability, psychosocial load, and residual variation. The 52.7% highest liabilities were assigned HDRS above 7; integer totals were then drawn within the clinically compatible 0–7 and 8–45 ranges. This construction makes the outcome dependent on the symptom system without using the observed symptom scores as admission-time predictors.

The categorical agreement plot in figure 2 provides a direct audit of the generated counts. Constraint markers and digital-cohort markers are nearly superimposed for all displayed variables, confirming that later performance differences cannot be attributed to accidental prevalence drift in the generation step.

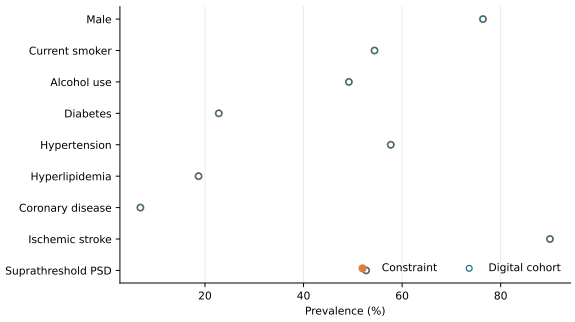

Figure 2. Agreement of cohort constraints.

2.2. Lesion–symptom probability cascade

Let X_i denote the admission clinical vector for digital patient i, R_i the six deep-structure lesion indicators, C_{ij} the three central symptoms, P_{ik} the ten peripheral symp-

Table 2. Symptom calibration

Symptom	Calibration prevalence (%)	Digital prevalence (%)	Digital score, mean (SD)
Depressed mood	55.9	55.9	1.40 (1.50)
Guilt	51.8	51.8	1.30 (1.49)
Suicidality	10.8	10.8	0.27 (0.86)
Insomnia	57.1	57.1	2.00 (2.16)
Loss of interest	52.3	52.3	1.31 (1.49)
Retardation	42.5	42.5	1.06 (1.43)
Agitation	39.1	39.1	0.98 (1.41)
Psychiatric anxiety	56.8	56.8	1.42 (1.50)
Somatic anxiety	36.0	36.0	0.90 (1.37)
Gastrointestinal somatic	25.6	25.6	0.38 (0.70)
General somatic	32.4	32.4	0.49 (0.76)
Hypochondriasis	30.2	30.2	0.76 (1.30)
Weight loss	19.9	19.9	0.50 (1.12)

toms, and Y_i the indicator of HDRS above 7. The first stage estimated each central symptom separately:

$$\operatorname{logit}\{\Pr(C_{ij}=1 \mid X_i, R_i)\} = \alpha_j + X_i^\top \beta_j + R_i^\top \gamma_j, \qquad j = 1,2,3. \tag{1}$$

Equation (1) preserves symptom specificity: a right basal-ganglia indicator may carry a different coefficient for anxiety than for loss of interest. L2 regularization limits unstable coefficients when lesion indicators are correlated with laterality and volume. Probabilities for the training records were generated strictly out of fold with three inner folds.

The second stage estimated each peripheral symptom from the same admission variables and the three out-of-fold central probabilities:

$$\operatorname{logit}\{\Pr(P_{ik}=1 \mid X_i, R_i, \widehat{\boldsymbol{c}}_i)\} = \delta_k + X_i^\top \theta_k + R_i^\top \phi_k + \widehat{\boldsymbol{c}}_i^\top \lambda_k, \qquad k = 1,\ldots,10. \tag{2}$$

The probability vector in equation (2) represents uncertainty about central activation rather than observed two-week symptoms. This distinction prevents outcome-time measurements from entering an admission-time classifier. Peripheral training probabilities were again produced out of fold. Final models for each symptom were fitted on the complete outer-training portion and applied to the untouched outer-test portion.

The third stage combined admission variables, lesion indicators, 13 predicted symptom probabilities, the mean predicted central burden, the mean predicted peripheral burden, and dispersion across the three central probabilities:

$$\operatorname{logit}\{\Pr(Y_i=1)\} = \eta_0 + Z_i^\top \eta, \quad Z_i = [X_i, R_i, \widehat{\boldsymbol{c}}_i, \widehat{\boldsymbol{p}}_i, \bar{c}_i, \bar{p}_i, s(c_i)]\,. \tag{3}$$

The construction in equation (3) is termed the lesion–symptom probability cascade. It differs from a symptom network: there are no estimated contemporaneous partial-correlation edges, and the stage order is imposed for prediction. It also differs from ordinary stacked classification because the intermediate tasks correspond to clinically named symptoms and a prespecified central–peripheral ordering. The method remains interpretable at the level of symptom probabilities, but its final coefficients should not be interpreted as causal effects.

2.3. Comparator models

Four comparators were fitted on each outer-training portion. Clinical logistic regression used demographics, vascular factors, stroke type, NIHSS, Barthel Index, modified Rankin Scale, MMSE, lesion volume, imaging delay, and laterality. Lesion-augmented logistic regression added the six deep-structure indicators. Extra trees supplied a randomized nonlinear comparator with 220 trees, balanced class weights, a minimum leaf size of 18, and 75% feature sampling. Histogram gradient boosting used 180 iterations, learning rate 0.045, at most 15 leaves, minimum leaf size 35, and L2 regularization 1.5. All hyperparameters were fixed before outer validation. Random forests and boosting can capture interactions but often require careful probability assessment in clinical applications.[35,36] The logistic comparators supply a transparent reference against which additional complexity can be judged.

2.4. Validation and statistical analysis

The five methods used identical repeated stratified outer splits: five folds repeated twice, giving ten evaluations with approximately 80% development and 20% test records per fold. Every symptom-stage prediction for cascade training was generated by three-fold inner cross-fitting. Standardization was learned only from the relevant training portion. This separation is essential because stacked models can otherwise learn from fitted values that are systematically more optimistic than test-time predictions. Repeated cross-validation provides empirical variation across splits, although these fold ranges

are not a substitute for confidence intervals from independent clinical samples.

Discrimination was measured with AUROC. Average precision was included because it responds to the proportion of positive outcomes and summarizes the precision–recall relation. Overall probability error was measured with Brier loss. Calibration intercept and slope were obtained by logistic recalibration of the binary outcome on the logit of predicted probability. An ideal intercept is zero and an ideal slope is one. Sensitivity and specificity were calculated at the Youden threshold within each test fold. These measures were selected before analysis to prevent selective reporting.

Anatomical risk differences were calculated as the absolute difference in suprathreshold prevalence between profiles with and without each deep-structure indicator. Six hundred bootstrap samples supplied percentile intervals. These differences describe the digital cohort and include both direct and indirect generative dependencies; they are not clinical effect estimates.

Robustness was examined in two ways. First, test records were resampled to outcome prevalences of 0.35, 0.45, 0.527, 0.60, and 0.65. Second, predicted logits were multiplied by 0.85 and 0.70 to represent probability attenuation without changing rank order. Sex, age below or above 60 years, ischemic versus hemorrhagic stroke, and NIHSS at most 4 versus above 4 were audited separately. The code was implemented in Python with NumPy, pandas, SciPy, scikit-learn, Matplotlib, and seaborn. All generated records, predictions, fold results, figures, and the fixed-seed script accompany the manuscript.

2.5. Ethics and reporting

No human participant was recruited, contacted, or reidentified. No individual clinical record or participant image was available to this analysis. Institutional review approval and individual consent were therefore not applicable to the computational experiment. Reporting follows the relevant principles of TRIPOD for predictor timing, validation, and complete performance description, while explicitly distinguishing digital evaluation from external clinical validation.[26] Clinical use is outside the scope of the study.

3. RESULTS AND DISCUSSION

3.1. Cohort behavior and symptom organization

The digital cohort contained 17,220 profiles, of which 9,076 had an HDRS total above 7. Mean age was 57.84 years, 76.4% were male, 90.0% had ischemic stroke, and mean NIHSS was 3.85. The mean HDRS total was 9.37 (standard deviation 7.10). These values place the numerical experiment in a relatively mild-stroke, early-hospitalization population. That restriction matters: associations learned here should not be projected to severe aphasia, prolonged impaired consciousness, or rehabilitation populations with a different symptom composition.

Deep-lesion indicators were associated most strongly with the three central symptoms in the generated cohort. The lesion–symptom portrait in figure 3 shows the largest prevalence differences for right basal-ganglia involvement, followed by right internal capsule and right external capsule. Depressed mood, psychiatric anxiety, and loss of interest occupy the high-intensity columns for these regions. Peripheral symptoms show smaller but broadly positive differences because their liabilities partly depend on central activation.

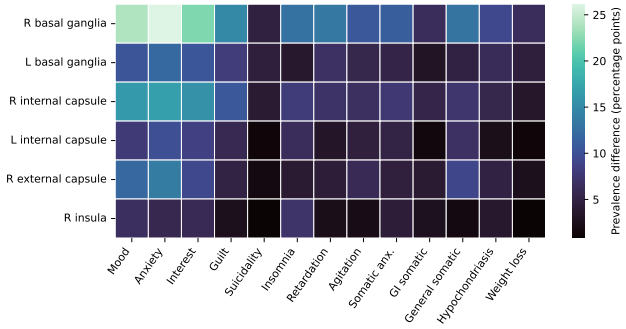

Figure 3. Deep-lesion symptom differences.

This pattern is not evidence that the colored anatomical cells independently cause their corresponding symptoms. It reflects the prespecified anatomical compatibility rules, shared disability pathways, and central-to-peripheral dependence. Its value is methodological: it establishes a test environment in which the cascade has a genuine ordered signal to exploit. If the cascade cannot improve out-of-sample stratification in this favorable environment, the burden of proof for using it in clinical data becomes higher.

3.2. Model discrimination and precision

The clinical logistic model attained a mean AUROC of 0.6066 (empirical 95% fold range 0.5954–0.6159) and average precision of 0.6332. Adding the six deep-structure indicators raised AUROC to 0.6912 and average precision to 0.7317. Extra trees attained an AUROC of 0.6790, while histogram boosting reached 0.6867. The lesion–symptom cascade reached 0.6911. Mean receiver-operating curves in figure 4 show a clear separation between the clinical-only model and all lesion-aware models, but near-complete overlap between the cascade and lesion-augmented logistic regression.

The direct answer to the study question is therefore negative. The cascade did not add material rank information beyond deep-lesion indicators included directly in logistic regression. The AUROC difference was -0.0001, and the average-precision difference was -0.00002. The ordered symptom probabilities largely recoded information already available through lesion anatomy, severity, disability, cognition, and laterality. This redundancy is plausible because every intermediate symptom model

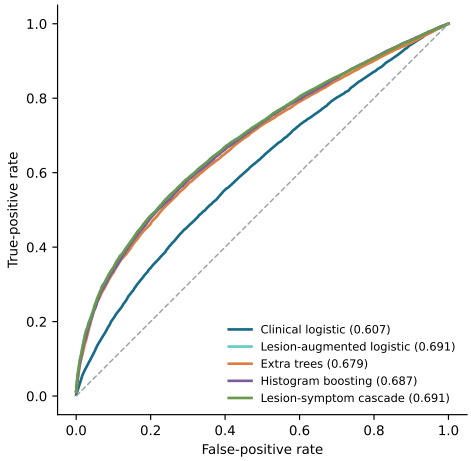

Figure 4. Receiver-operating characteristics.

used the same admission predictor set. A cascade can reorganize a signal without increasing the information content of that signal.

The nonlinear methods did not exceed the lesion-augmented logistic model. Extra trees lost approximately 0.012 AUROC, and histogram boosting lost approximately 0.004. Their lower performance suggests that the calibrated relationships were adequately represented by regularized additive terms and that flexible interactions mainly fitted split-specific variation. This finding aligns with broader clinical prediction experience: complexity is beneficial only when stable nonlinear structure exists and is measured with sufficient sample diversity. It also reinforces the need to compare any new method with a well-specified regression model rather than with an intentionally weak reference.

The results in table 3 show that the main improvement came from anatomical representation, not from model layering. The reduction in Brier loss from 0.240 to approximately 0.220 is consistent with the increase in discrimination. Cascade and lesion-augmented regression were practically indistinguishable across all four principal measures, so selecting between them should depend on the intended use. The direct model is simpler and easier to transport; the cascade is useful only when symptom-specific intermediate probabilities are themselves clinically relevant and prospectively validated.

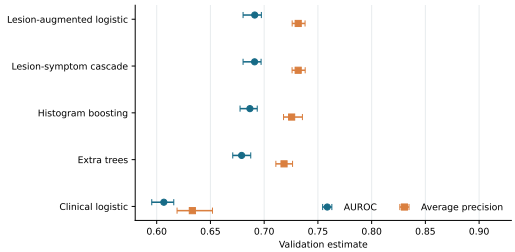

Figure 5. Repeated-validation estimates.

Fold ranges in figure 5 reinforce this conclusion. The two leading models have overlapping intervals for both AUROC and average precision. The plot also prevents a visually exaggerated interpretation of a fourth-decimal difference. Model comparison should focus on clinically meaningful separation, probability quality, and transportability, not on the identity of the numerically highest mean.

3.3. Calibration and operating characteristics

The cascade calibration intercept was -0.001 (fold range -0.015 to 0.013), and its calibration slope was 0.980 (0.936–1.023). The mean curve in figure 6 followed the identity line over the range where most predictions occurred. Thin fold curves show modest variation around the mean, particularly at the extremes, where fewer records contribute.

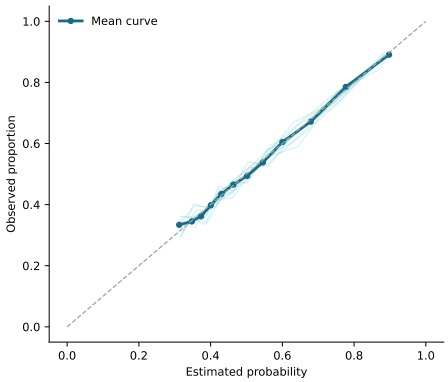

Figure 6. Calibration of cascade probabilities.

Near-ideal calibration is expected in a large digital cohort when development and test records arise from the same generation mechanism. It nevertheless verifies that cross-fitting did not introduce systematic overconfidence. The result should not be interpreted as evidence of calibration in another hospital. Calibration is sensitive to outcome frequency, eligibility rules, measurement practice, and the relationships among predictors. External recalibration would probably be required even if discrimination transported.

At the fold-specific Youden thresholds, cascade sensitivity averaged 0.541 and specificity 0.753. Lesion-augmented logistic regression produced sensitivity 0.550 and specificity 0.745. These operating points illustrate a tradeoff rather than a clinical recommendation. A stroke service prioritizing sensitivity for safety screening would select a lower threshold and accept more false-positive assessments. Decision thresholds should be tied to the consequences of missed depression, psychiatric referral capacity, and the availability of repeat assessment.[37,38]

3.4. Anatomical risk differences

Right basal-ganglia involvement was associated with an absolute increase of 36.4 percentage points in suprathreshold prevalence (bootstrap 95% interval 34.7–38.4). Right internal capsule followed at 23.7 points (21.6–

Table 3. Repeated-validation performance

Model	AUROC	Average precision	Brier loss	Calibration slope
Clinical logistic	0.607	0.633	0.240	0.962
Lesion-augmented logistic	**0.691**	**0.732**	**0.220**	0.984
Extra trees	0.679	0.718	0.224	0.897
Histogram boosting	0.687	0.726	0.221	1.057
Lesion–symptom cascade	0.691	0.732	0.220	0.980

25.5), right external capsule at 16.8 (13.9–19.7), left internal capsule at 16.4 (14.0–18.8), left basal ganglia at 14.6 (11.9–17.1), and right insula at 9.7 (6.5–13.1). The ordered display in figure 7 shows a right-deep-structure gradient rather than a uniform hemispheric effect.

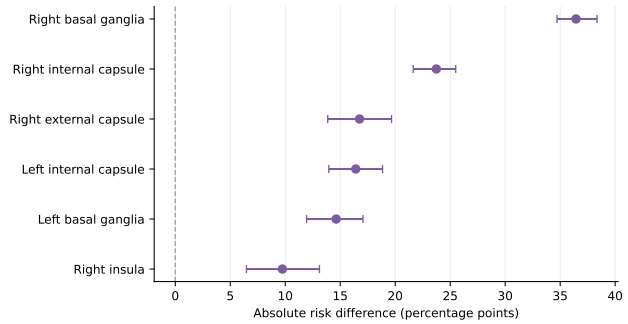

Figure 7. Regional absolute risk differences.

The anatomical hierarchy is compatible with models of mood regulation involving corticostriatal and thalamocortical pathways. Basal ganglia participate in reward, action selection, motivation, and affective control; the internal capsule carries fibers connecting frontal and subcortical territories; and the external capsule contains association fibers that may influence distributed networks.[17,18] Disconnection studies further show that behavioral consequences can extend beyond visibly damaged tissue.[25] Even so, the present risk differences are properties of the generation equations. They justify the model experiment but cannot settle debates about lesion laterality or regional causation.

The magnitude of the right basal-ganglia difference also explains why a flat lesion-aware regression performed so well relative to the cascade. When one admission indicator strongly shifts the outcome and the intermediate symptoms, both model types receive essentially the same dominant signal. A cascade becomes more likely to add value when intermediate measurements introduce new information, such as early repeated symptom observations, ecological momentary assessment, functional connectivity, or longitudinal change. None of those inputs were available by design.

3.5. Subgroup consistency

Cascade AUROC was 0.687 among women and 0.693 among men; 0.687 below age 60 and 0.697 at age 60 or older; 0.691 for ischemic and 0.695 for hemorrhagic stroke; and 0.677 for NIHSS at most 4 versus 0.681 above 4. These estimates appear in figure 8. Average precision was higher in the NIHSS-above-4 group because outcome prevalence was 0.600 in that group, compared with 0.484 among lower-severity profiles.

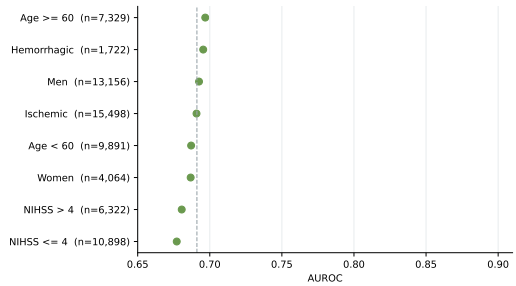

Figure 8. Subgroup discrimination.

The narrow AUROC span suggests that the generation and model-fitting procedures did not produce a large rank disparity across the audited groups. This is not a fairness guarantee. Sex and stroke type were assigned under simple aggregate constraints, and subgroup-specific clinical mechanisms were not calibrated from individual observations. The audit demonstrates what the model does in this digital population; it does not establish equitable performance across real clinical groups. Independent cohorts must examine missingness, language impairment, access to imaging, assessment reliability, and threshold consequences.

3.6. Distribution shifts and probability attenuation

Changing outcome prevalence from 0.35 to 0.65 left AUROC between 0.686 and 0.697, as expected for a rank measure under outcome resampling. Average precision rose from 0.581 to 0.815 because positive predictive value depends on prevalence. The robustness plot in figure 9 makes the divergence visible: the AUROC line is nearly horizontal while the precision line follows the changing positive proportion.

Calibration intercept moved from -0.727 at prevalence 0.35 to 0.506 at prevalence 0.65. This behavior is clinically important. A model transported to a service with fewer depressed patients would systematically overestimate absolute probability unless its intercept were adjusted. Multiplying logits by 0.85 and 0.70 left AUROC and average precision unchanged but increased calibration

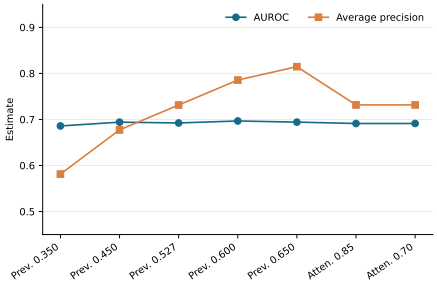

Figure 9. Robustness under distribution shifts.

slopes to 1.154 and 1.400, respectively. Rank preservation therefore does not imply probability preservation. Any deployment study that reports only AUROC could miss a substantial distortion in estimated risk.

3.7. Methodological interpretation

The probability cascade has three defensible features. First, it enforces temporal discipline through out-of-fold intermediate predictions. Second, it provides named symptom probabilities that can be inspected separately. Third, it permits different admission predictors to influence different symptoms before aggregation. These features make it more clinically legible than a single opaque nonlinear classifier. Interpretability, however, is useful only when the intermediate estimates are accurate and actionable. The digital experiment shows that interpretability does not automatically translate into higher discrimination.

The negative comparison is informative for method selection. A layered architecture adds fitting steps, regularization choices, computational cost, and opportunities for error. If its final probability does not improve materially, a direct lesion-aware model is preferable for the narrow task of admission-time stratification. The cascade may still be retained as an auxiliary analysis when clinicians need symptom-specific probabilities, but those probabilities would require their own calibration and validation. Model complexity should be justified by added clinical information, not by novelty alone.

The result also separates representation from prediction. The lesion–symptom portrait is visually rich because it describes thirteen outcomes across six anatomical indicators, yet a final binary classifier can compress most of that structure into a few lesion coefficients. Rich descriptive organization is not proof of incremental predictive value. To contribute new rank information, an intermediate symptom estimate must vary between patients who otherwise have similar clinical and anatomical predictors and must retain a stable association with the final outcome. In this experiment, intermediate probabilities were deterministic functions of the same variables already offered to lesion-augmented regression. Their near-zero incremental value is therefore a coherent statistical outcome rather than a failure of implementation.

Clinical interpretation should likewise avoid treating the right basal ganglia as a solitary depression center. The large digital risk difference summarizes multiple linked paths: direct affective liability, motor and motivational consequences, lesion volume, laterality, and propagation through central symptoms. Real strokes also damage fibers of passage and connected territories, making regional indicators correlated rather than isolated. A patient with right striatal injury may develop apathy, slowed behavior, fatigue, or impaired reward learning that overlaps with depression rating items. Careful assessment must distinguish these manifestations while recognizing that they can coexist. Anatomical information may raise vigilance, but the diagnosis remains behavioral and contextual.

The modest overall AUROC is also clinically revealing. A value near 0.69 means that many positive and negative profiles remain inseparable from admission variables alone. This ceiling is unsurprising because mood after stroke is shaped by post-admission experiences, family support, complications, sleep, pain, communication, rehabilitation progress, and pre-existing vulnerability that were absent from the predictor set. Enlarging the digital sample reduces numerical uncertainty but cannot compensate for omitted information. Improving future models will require better timed measurements and richer content, not simply more elaborate algorithms applied to the same variables.

Calibration findings suggest a practical evaluation sequence for future work. Investigators should first verify predictor timing and prevent symptom leakage; next compare a direct regularized model with any layered or nonlinear alternative; then examine calibration over the clinically occupied probability range; and finally evaluate threshold consequences in independent sites. When prevalence changes, intercept adjustment may restore overall calibration without altering discrimination. When the calibration slope changes, coefficient shrinkage or more substantial model updating may be necessary. This sequence keeps the analysis tied to patient-care decisions and discourages selection of a model solely because its AUROC is marginally larger.

Synthetic health-data research reaches a similar conclusion about utility. Agreement of marginal distributions is necessary but insufficient; the important test is whether conclusions remain stable for the intended analysis.[31,34] Here, categorical constraints were exact and continuous summaries were close, yet the joint dependence structure was deliberately authored. Results therefore answer a conditional question: what happens when deep lesions affect central symptoms and those symptoms propagate to peripheral manifestations according to the stated equations? They do not estimate what would happen under unknown patient-level dependencies.

Several limitations delimit interpretation. The digital cohort excludes unmeasured biology, medication, inflammatory markers, prior psychiatric vulnerability beyond the exclusion premise, detailed social environment, rehabilitation exposure, and time-varying recovery. Apha-

sia and severe consciousness disturbance are not represented, despite their importance for selection bias. Lesions are binary regional indicators rather than three-dimensional masks, tract disconnections, or functional-connectivity disturbances. Symptom direction is imposed by construction and cannot validate causal ordering. The HDRS threshold is a screening definition, not a structured psychiatric diagnosis. Finally, repeated internal validation samples one generation mechanism; only external patient-level validation can assess transport.

These limitations also define the next empirical test. A prospective multisite cohort should record admission lesion masks, tract disconnection, NIHSS domains, cognition, disability, prior psychiatric history, social support, and repeated symptom-level HDRS or another validated scale. Central and peripheral predictions should be trained only on earlier observations, with site-level external validation and prespecified handling of aphasia. Sample-size planning should address the number of outcome events, predictor parameters, and anticipated calibration precision.[39] The direct lesion-aware model and the cascade should then be compared on discrimination, calibration, net benefit, symptom-specific accuracy, subgroup performance, and model updating needs.

4. CONCLUSION

The study asked whether an admission-time lesion-to-symptom probability cascade improves identification of suprathreshold depressive symptoms at two weeks beyond direct lesion-aware regression. Under an aggregate-calibrated digital population with explicit central-to-peripheral symptom dependence, it did not. The cascade AUROC was 0.6911 compared with 0.6912 for lesion-augmented logistic regression, and average precision was effectively identical. Deep-structure lesion descriptors, especially right basal-ganglia and right internal-capsule involvement, accounted for the improvement over clinical variables alone. The cascade nevertheless produced well-calibrated probabilities and a transparent decomposition into symptom-specific risks. For the narrow task of final risk stratification, the simpler lesion-augmented model is therefore the more economical choice. The cascade becomes scientifically worthwhile only if prospective longitudinal symptoms or richer connectivity measures contribute information not already carried by admission anatomy and disability.

BIBLIOGRAFIA / REFERENCES

1. GBD 2019 Diseases and Injuries Collaborators. Global burden of 369 diseases and injuries in 204 countries and territories, 1990–2019: a systematic analysis for the Global Burden of Disease Study 2019. *Lancet*. 2020;396:1204–1222. https://doi.org/10.1016/S0140-6736(20)30925-9.
2. M. L. Hackett and K. Pickles. Part I: frequency of depression after stroke: an updated systematic review and meta-analysis of observational studies. *Int. J. Stroke*. 2014;9:1017–1025. https://doi.org/10.1111/ijs.12357.
3. L. Ayerbe, S. Ayis, C. D. Wolfe, and A. G. Rudd. Natural history, predictors and outcomes of depression after stroke: systematic review and meta-analysis. *Br. J. Psychiatry*. 2013;202:14–21. https://doi.org/10.1192/bjp.bp.111.107664.
4. A. Towfighi, B. Ovbiagele, N. El Husseini, et al. Poststroke depression: a scientific statement for healthcare professionals from the American Heart Association/American Stroke Association. *Stroke*. 2017;48:e30–e43. https://doi.org/10.1161/STR.0000000000000113.
5. W. Cai, C. Mueller, Y.-J. Li, W.-D. Shen, and R. Stewart. Post stroke depression and risk of stroke recurrence and mortality: a systematic review and meta-analysis. *Ageing Res. Rev.*. 2019;50:102–109. https://doi.org/10.1016/j.arr.2019.01.013.
6. R. G. Robinson and R. E. Jorge. Post-stroke depression: a review. *Am. J. Psychiatry*. 2016;173:221–231. https://doi.org/10.1176/appi.ajp.2015.15030363.
7. R. F. Villa, F. Ferrari, and A. Moretti. Post-stroke depression: mechanisms and pharmacological treatment. *Pharmacol. Ther.*. 2018;184:131–144. https://doi.org/10.1016/j.pharmthera.2017.11.005.
8. Y. Shi, Y. Yang, J. Zhang, Y. Chen, and L. Wang. Risk factors for post-stroke depression: a meta-analysis. *Front. Aging Neurosci.*. 2017;9:218. https://doi.org/10.3389/fnagi.2017.00218.
9. S. Northcott, B. Moss, K. Harrison, and K. Hilari. A systematic review of the impact of stroke on social support and social networks: associated factors and patterns of change. *Clin. Rehabil.*. 2016;30:811–831. https://doi.org/10.1177/0269215515602136.
10. D. Borsboom and A. O. J. Cramer. Network analysis: an integrative approach to the structure of psychopathology. *Annu. Rev. Clin. Psychol.*. 2013;9:91–121. https://doi.org/10.1146/annurev-clinpsy-050212-185608.
11. E. I. Fried and A. O. J. Cramer. Moving forward: challenges and directions for psychopathological network theory and methodology. *Perspect. Psychol. Sci.*. 2017;12:999–1020. https://doi.org/10.1177/1745691617705892.
12. E. I. Fried, R. M. Nesse, K. Zivin, C. Guille, and S. Sen. Depression is more than the sum score of its parts: individual DSM symptoms have different risk factors. *Psychol. Med.*. 2014;44:2067–2076. https://doi.org/10.1017/S0033291713002900.
13. C. van Borkulo, L. Boschloo, D. Borsboom, B. W. J. H. Penninx, L. J. Waldorp, and R. A. Schoevers. Association of symptom network structure with the course of depression. *JAMA Psychiatry*. 2015;72:1219–1226. https://doi.org/10.1001/jamapsychiatry.2015.2079.
14. S. A. Ashaie, J. Hung, C. J. Funkhouser, S. A. Shankman, and L. R. Cherney. Depression over time in persons with stroke: a network analysis approach. *J. Affect. Disord. Rep.*. 2021;4:100131. https://doi.

org/10.1016/j.jadr.2021.100131.
15. S. Epskamp, D. Borsboom, and E. I. Fried. Estimating psychological networks and their accuracy: a tutorial paper. *Behav. Res. Methods*. 2018;50:195–212. https://doi.org/10.3758/s13428-017-0862-1.
16. D. J. Robinaugh, A. J. Millner, and R. J. McNally. Identifying highly influential nodes in the complicated grief network. *J. Abnorm. Psychol.*. 2016;125:747–757. https://doi.org/10.1037/abn0000181.
17. W. D. Taylor, H. J. Aizenstein, and G. S. Alexopoulos. The vascular depression hypothesis: mechanisms linking vascular disease with depression. *Mol. Psychiatry*. 2013;18:963–974. https://doi.org/10.1038/mp.2013.20.
18. J. L. Lanciego, N. Luquin, and J. A. Obeso. Functional neuroanatomy of the basal ganglia. *Cold Spring Harb. Perspect. Med.*. 2012;2:a009621. https://doi.org/10.1101/cshperspect.a009621.
19. M. D. Fox. Mapping symptoms to brain networks with the human connectome. *N. Engl. J. Med.*. 2018;379:2237–2245. https://doi.org/10.1056/NEJMra1706158.
20. A. D. Boes, S. Prasad, H. Liu, et al. Network localization of neurological symptoms from focal brain lesions. *Brain*. 2015;138:3061–3075. https://doi.org/10.1093/brain/awv228.
21. J. L. Padmanabhan, D. Cooke, J. Joutsa, et al. A human depression circuit derived from focal brain lesions. *Biol. Psychiatry*. 2019;86:749–758. https://doi.org/10.1016/j.biopsych.2019.05.007.
22. C. Sperber and H.-O. Karnath. On the validity of lesion-behaviour mapping methods. *Neuropsychologia*. 2018;115:17–24. https://doi.org/10.1016/j.neuropsychologia.2017.08.026.
23. C. Rorden, H.-O. Karnath, and L. Bonilha. Improving lesion-symptom mapping. *J. Cogn. Neurosci.*. 2007;19:1081–1088. https://doi.org/10.1162/jocn.2007.19.7.1081.
24. E. T. Rolls, C.-C. Huang, C.-P. Lin, J. Feng, and M. Joliot. Automated anatomical labelling atlas 3. *NeuroImage*. 2020;206:116189. https://doi.org/10.1016/j.neuroimage.2019.116189.
25. A. Salvalaggio, M. De Filippo De Grazia, M. Zorzi, M. Thiebaut de Schotten, and M. Corbetta. Post-stroke deficit prediction from lesion and indirect structural and functional disconnection. *Brain*. 2020;143:2173–2188. https://doi.org/10.1093/brain/awaa156.
26. G. S. Collins, J. B. Reitsma, D. G. Altman, and K. G. M. Moons. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. *BMJ*. 2015;350:g7594. https://doi.org/10.1136/bmj.g7594.
27. R. F. Wolff, K. G. M. Moons, R. D. Riley, et al. PROBAST: a tool to assess the risk of bias and applicability of prediction model studies. *Ann. Intern. Med.*. 2019;170:51–58. https://doi.org/10.7326/M18-1376.
28. B. Van Calster, D. J. McLernon, M. van Smeden, L. Wynants, and E. W. Steyerberg. Calibration: the Achilles heel of predictive analytics. *BMC Med.*. 2019;17:230. https://doi.org/10.1186/s12916-019-1466-7.
29. B. Van Calster, L. Wynants, J. F. M. Verbeek, et al. Reporting and interpreting decision curve analysis: a guide for investigators. *Eur. Urol.*. 2018;74:796–804. https://doi.org/10.1016/j.eururo.2018.08.038.
30. B. Van Calster, D. Nieboer, Y. Vergouwe, B. De Cock, M. J. Pencina, and E. W. Steyerberg. A calibration hierarchy for risk models was defined: from utopia to empirical data. *J. Clin. Epidemiol.*. 2016;74:167–176. https://doi.org/10.1016/j.jclinepi.2015.12.005.
31. A. Goncalves, P. Ray, B. Soper, J. Stevens, L. Coyle, and A. P. Sales. Generation and evaluation of synthetic patient data. *BMC Med. Res. Methodol.*. 2020;20:108. https://doi.org/10.1186/s12874-020-00977-1.
32. D. Kaur, M. Sobiesk, S. Patil, et al. Application of Bayesian networks to generate synthetic health data. *J. Am. Med. Inform. Assoc.*. 2021;28:801–811. https://doi.org/10.1093/jamia/ocaa303.
33. J. Chen, D. Chun, M. Patel, E. Chiang, and J. James. The validity of synthetic clinical data: a validation study of a leading synthetic data generator (Synthea) using clinical quality measures. *BMC Med. Inform. Decis. Mak.*. 2019;19:44. https://doi.org/10.1186/s12911-019-0793-0.
34. Z. Azizi, C. Zheng, L. Mosquera, L. Pilote, and K. El Emam. Can synthetic data be a proxy for real clinical trial data? A validation study. *BMJ Open*. 2021;11:e043497. https://doi.org/10.1136/bmjopen-2020-043497.
35. L. Breiman. Random forests. *Mach. Learn.*. 2001;45:5–32. https://doi.org/10.1023/A:1010933404324.
36. J. H. Friedman. Greedy function approximation: a gradient boosting machine. *Ann. Stat.*. 2001;29:1189–1232. https://doi.org/10.1214/aos/1013203451.
37. A. J. Vickers and E. B. Elkin. Decision curve analysis: a novel method for evaluating prediction models. *Med. Decis. Making*. 2006;26:565–574. https://doi.org/10.1177/0272989X06295361.
38. A. J. Vickers, B. Van Calster, and E. W. Steyerberg. Net benefit approaches to the evaluation of prediction models, molecular markers, and diagnostic tests. *BMJ*. 2016;352:i6. https://doi.org/10.1136/bmj.i6.
39. R. D. Riley, J. Ensor, K. I. E. Snell, et al. Calculating the sample size required for developing a clinical prediction model. *BMJ*. 2020;368:m441. https://doi.org/10.1136/bmj.m441.
40. C. Pan, G. Li, W. Sun, et al. Psychopathological network for early-onset post-stroke depression symptoms. *BMC Psychiatry*. 2023;23:114. https://doi.org/10.1186/s12888-023-04606-1.