Scientia Psychiatrica
Vol. 7, No. 3, 2026, pp. 160–170 · e-ISSN 2715-9736 journal homepage: scientiapsychiatrica.com

ORIGINAL RESEARCH ARTICLE

Transdiagnostic Brain Age Gap in Schizophrenia and Bipolar Disorder: A Leakage-Controlled OpenNeuro Analysis

Aline Hafidzah[1*], Rashidah Unaib Al-Zayid[2], Cindy Susanti[3]
[1]*Department of Radiology, Phlox Institute, Palembang, Indonesia*
[2]*Department of Radiology, Family Medicine Academy, Riyadh, Saudi Arabia*
[3]*Department of Internal Medicine, Provita Hospital, Jayapura, Indonesia*

ARTICLE INFO

Article history
Received: May 18, 2026
Revised: June 28, 2026
Accepted: July 15, 2026
Available online: July 25, 2026
Published: July 30, 2026

Keywords
Bipolar disorder
Brain age gap
FreeSurfer
Normative modeling
OpenNeuro
Schizophrenia

***Corresponding author**
Aline Hafidzah
E-mail: aline.hafidzah@phlox.or.id
ORCID: 0009-0004-1115-5785

DOI
10.37275/scipsy.v7i3.218

ABSTRACT

Background. Brain age gap (BAG), predicted minus chronological brain age, is often elevated in psychotic and mood disorders. Its magnitude depends on the reference sample, age range, scanner, image quality, and age-bias correction.

Objective. To estimate transdiagnostic BAG differences in schizophrenia and bipolar disorder with a control-only model, nested validation, and adjustment for scanner and ghost artifact.

Methods. This secondary cross-sectional study used the University of California, Los Angeles (UCLA) Consortium for Neuropsychiatric Phenomics dataset (OpenNeuro ds000030): 125 controls, 49 participants with bipolar disorder, and 50 with schizophrenia. Ridge regression used 68 Desikan–Killiany cortical-thickness and 37 intracranial-volume-normalized FreeSurfer measures. Models were trained only in controls within repeated 10-fold validation, with hyperparameters and age-bias correction estimated in training data. Group effects came from heteroskedasticity-consistent linear models adjusted for age, gender, scanner, and ghost artifact.

Results. In controls, raw out-of-fold mean absolute error was 6.48 years (95% bootstrap confidence interval [CI] 5.83–7.11), r = 0.562, and R^2 = 0.277. Adjusted BAG differed from controls by 0.55 years in bipolar disorder (95% CI −0.50 to 1.59; p = 0.304; false-discovery-rate q = 0.304) and by 0.63 years in schizophrenia (95% CI −0.51 to 1.77; p = 0.276; q = 0.304). The schizophrenia–bipolar contrast was 0.08 years (95% CI −1.16 to 1.33; p = 0.895). Sensitivity analyses gave similar estimates.

Conclusion. Both patient groups showed small positive estimates with wide overlap, and the confidence intervals spanned zero. The findings support a reproducible open-data workflow, not individual diagnosis or a causal claim of accelerated aging.

Copyright: © 2026 The Author(s). Published by HM Publisher in collaboration with CMHC Research Center on behalf of Scientia Psychiatrica. The authors retain copyright of their work.

Access licence: This is an open access article distributed under the terms of the Creative Commons Attribution–NonCommercial–ShareAlike 4.0 International Licence (CC BY-NC-SA 4.0), which permits non-commercial use, sharing, adaptation and reproduction in any medium, provided the original author and source are credited and any derivative work is distributed under the same licence.

1. Introduction

Brain-age modeling predicts chronological age from multivariate neuroimaging features. The difference between predicted and observed age is the brain age gap (BAG). Large multisite benchmarks and uncertainty-aware models show that this quantity can be estimated reproducibly. Its calibration, however, stays conditional on the training population, the acquisition environment, and the modeling architecture.[1,2] BAG is therefore a model-dependent summary of anatomical resemblance. It is not a direct measurement of biological aging, neurodegeneration, or longitudinal tissue change. Performance metrics need equally careful reading, because age range, sample size, correction strategy, feature representation, and regression algorithm can all shift mean absolute error, correlation, and downstream associations materially.[3–7]

Schizophrenia is a useful test case. Neurodevelopmental variation, illness exposure, treatment, physical health, and technical imaging factors may all contribute to an older-appearing structural pattern. Recent primary studies report advanced BAG across psychotic psychopathology, describe the structural underpinnings of advanced apparent age, and find altered multimodal trajectories in younger patients. Others document changes after early medication and associations spanning psychosis and psychopathy.[8–12] These findings support a distributed transdiagnostic signal. They do not show that BAG can classify an individual, or that it identifies when the underlying deviation arose.

Evidence in bipolar disorder, and across the schizophrenia–bipolar spectrum, is more heterogeneous. Brain age has been related to cognition in first-episode bipolar disorder, to rumination in schizophrenia, to explainable regional morphometry, and to multimodal trajectories specific to schizophrenia.[13–16] Larger and newer investigations add deep learning, psychosis biotypes, international bipolar-disorder consortia, alternative model specifications, and direct schizophrenia–bipolar comparisons. They demonstrate that the estimated difference varies with diagnostic definition, age, medication, scanner, and model selection.[17–21] Overlap within each diagnosis matters as much as the group mean. Longitudinal and clinically anchored studies show why interpretation should go beyond a single case-control coefficient. BAG has been examined as a predictor of early treatment response, in relation to illness duration and antipsychotic exposure, among young adults with psychotic experiences, after exercise in schizophrenia-spectrum disorders, and across first-episode mania treated with lithium or quetiapine.[22–27] Multimodal ensemble modeling has also tied age-prediction signals to schizophrenia symptoms.[28] Disease stage, treatment, and individual heterogeneity can thus shape both the size and the meaning of the gap.

Analytical design can matter just as much. Multidimensional psychiatric models show diagnosis-dependent developmental trajectories, and external benchmarking reveals sensitivity to scanner variance and prediction bias.[29,30] Head motion and related image-quality variation can systematically alter predicted age and reduce reliability.[31,32] Regression towards the mean typically makes younger participants look older and older participants look younger. A correction estimated from the very observations being evaluated can then yield optimistic downstream metrics. A defensible analysis therefore separates prediction performance from group inference. It trains the normative model only in controls, keeps every data-dependent step inside the training folds, and accounts for measured technical imbalances.

The novelty of this study is the joint application of four safeguards within one reproducible transdiagnostic analysis of OpenNeuro ds000030: an identical control-only model for schizophrenia and bipolar disorder, nested hyperparameter selection, fold-specific age-bias correction estimated without evaluated observations, and explicit adjustment for scanner and the dataset-provided ghost-artifact label, with raw prediction metrics retained. The primary aim was to estimate adjusted age-debiased BAG differences between bipolar disorder and controls and between schizophrenia and controls. Secondary objectives were to quantify raw out-of-fold prediction performance and to compare schizophrenia directly with bipolar disorder. We also tested robustness after excluding ghost-labeled scans and after using raw BAG, and described the morphometric features contributing most consistently to the ridge models.

2. Methods

2.1. Study design and setting

This was a secondary cross-sectional analysis of the UCLA Consortium for Neuropsychiatric Phenomics (CNP) LA5c study, distributed as OpenNeuro dataset ds000030. The source study was designed as a phenome-wide investigation of neural, cognitive, and behavioral function in healthy adults and diagnostic groups.[33] Release R1.0.5 was used because it contains the participant table and the public morphometric derivatives needed for a lightweight, auditable reconstruction. Reporting followed the Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement.[34] No new participants were recruited, contacted, scanned, or phenotyped.

The unit of analysis was the individual participant. The analysis was explanatory and estimative, not diagnostic. BAG was treated as a continuous outcome, and the primary estimands were adjusted mean differences, in years, between each patient group and controls. The pipeline was specified so that information from the test observations, the patient diagnoses, or the group-comparison results could not influence the fitted age-prediction model.

2.2. Participants

The OpenNeuro participant table contains 272 records: 130 controls, 49 participants with bipolar disorder, 50 with schizophrenia, and 43 with attention-deficit/hyperactivity disorder, all aged 21–50 years. Eligibility required a diagnosis coded as CONTROL, BIPOLAR, or SCHZ, a T1-weighted scan indicator, and the three FreeSurfer statistics files used for feature extraction (aseg.stats, lh.aparc.stats, and rh.aparc.stats). The attention-deficit/hyperactivity disorder group fell outside the predefined transdiagnostic comparison and was excluded. All eligible records were included. No sampling or post hoc case selection took place.

FreeSurfer derivatives were available for 265 participants in the release. After the diagnostic and T1-weighted criteria were applied, the analytic sample comprised 224 unique participants: 125 controls, 49 with bipolar disorder, and 50 with schizophrenia. The sample was modest. The 125-person normative group in particular was treated as a precision constraint and not as a matter for retrospective power calculations. Confidence intervals, out-of-fold performance, regularization, and sensitivity analyses characterized the information that the available sample could support.

2.3. Variables and instruments

Data acquisition and provenance. The source T1-weighted images were acquired on two Siemens Trio 3-T scanners at UCLA. The distributed data were deidentified, and the anatomical images had already been processed with FreeSurfer using the Desikan–Killiany cortical parcellation framework.[35] Cortical reconstruction was not rerun. The analysis instead downloaded participants.tsv and the three statistics files for each participant from the public OpenNeuro

Amazon Web Services object store. This approach saved storage and transfer while keeping the morphometric information that the predefined model required.

In all, 796 successfully retrieved files occupied 6,154,477 bytes. Each file was hashed with SHA-256, and the download manifest recorded its uniform resource locator, size, checksum, and acquisition status. Twenty-one expected derivative-file requests returned no file. They corresponded to seven participants without the full three-file set, and these records were excluded before diagnostic filtering. The dataset is released under Creative Commons Zero. Raw files were kept unchanged, and every transformation was applied in derived analysis files.

Morphometric variables. Cortical predictors were mean thickness values from the 34 Desikan–Killiany regions in each hemisphere, which gave 68 features. Subcortical and ventricular predictors were volumes from aseg.stats divided by estimated total intracranial volume, to reduce between-person variation attributable to head size. Vessel labels, the optic chiasm, and fields with absent or negligible variance were excluded. The retained matrix held 37 normalized volumetric measures and 68 cortical-thickness measures, 105 predictors in total.

The outcome predicted by the normative model was chronological age in years. For participant i, raw BAG was calculated as $BAG_i = \hat{y}_i - y_i$, where $\hat{y}_i$ is predicted brain age and y_i is chronological age. Age-debiased BAG was calculated as $BAG_i^{debiased} = BAG_i - (\hat{\alpha}_{-i} + \hat{\beta}_{-i} \times y_i)$, where the correction parameters were estimated only from out-of-fold predictions in the corresponding control training set. Diagnosis was represented by control, bipolar disorder, and schizophrenia. Recorded gender, scanner serial number, and ghost status were taken directly from participants.tsv. Ghost status was treated as a binary technical-quality indicator and not as a clinical characteristic or a comprehensive substitute for image-quality assessment.

2.4. Data collection procedure

Normative modeling and nested validation. Ridge regression was chosen because it can stabilize prediction when correlated morphometric features are numerous relative to the training sample. The model was fitted exclusively in controls. In each of 10 repetitions, controls were split into 10 outer folds, balanced by age quintile, recorded gender, and scanner. In every outer iteration, feature-scaling parameters were estimated from the outer training controls alone. Five-fold inner cross-validation within that training set then selected the ridge penalty by mean absolute error, using the one-standard-error rule. Neither patient data nor the held-out control fold contributed to scaling, penalty selection, or coefficient estimation. The separation was kept deliberately. Leakage across preprocessing or validation stages can materially inflate neuroimaging prediction performance.[36]

Each control received one held-out prediction per repetition, and the 10 predictions were averaged at participant level. Each patient was predicted by all 100 outer-training models, and those predictions were averaged into an ensemble estimate. The asymmetry is intentional. Controls need held-out predictions to avoid self-influence. Patients were never used to train any model, so the complete control-model ensemble could evaluate them. A deterministic seed (20260831) governed fold assignment.

Age-bias correction and prediction performance. Within each outer training set, inner out-of-fold predictions were used to regress raw BAG on chronological age. The resulting intercept and slope were then applied to the outer held-out controls and to all patients predicted by that outer model. This kept bias correction independent of the participant being evaluated. It also avoided learning correction parameters from patient data. Participant-level age-debiased BAG was the average of the corrected held-out or ensemble predictions. Recent empirical work confirms that the choice and placement of bias correction materially affect the residual age dependence of BAG.[4,37]

Prediction performance was evaluated only in controls, with raw, uncorrected held-out predicted age. Predefined metrics were mean absolute error (MAE), root mean squared error (RMSE), the Pearson correlation between predicted and chronological age, and the coefficient of determination (R^2). Ninety-five percent confidence intervals for MAE, RMSE, and Pearson correlation came from 5,000 participant-level bootstrap resamples. Corrected predictions were not used to claim better age-prediction accuracy, because chronological age enters the correction and can mechanically improve apparent performance.

2.5. Statistical analysis

Continuous variables were summarized by mean and standard deviation, because their group distributions were regular enough for descriptive comparison. Categorical variables were reported as counts and percentages. Age was compared across diagnostic groups by one-way analysis of variance, and recorded gender, scanner, and ghost status by Pearson chi-square tests. These bivariate tests described sample composition. They did not guide covariate selection. Age, recorded gender, scanner, and ghost status were retained a priori, since each could influence morphometry, prediction error, or diagnostic-group comparisons.

The primary model was an ordinary linear regression of age-debiased BAG on diagnosis, chronological age, recorded gender, scanner, and ghost status, with controls as the diagnostic reference. Heteroskedasticity-consistent HC3 standard errors were used. The Benjamini–Hochberg procedure adjusted the two predefined patient-versus-control coefficients. Effect estimates are given in BAG years with 95% confidence intervals, unadjusted p values, and false-discovery-rate q values. The schizophrenia-versus-bipolar contrast was computed from the coefficient covariance matrix. All tests were two-sided.

Two sensitivity analyses were predefined. The adjusted model was first repeated after excluding all ghost-

labeled scans. The primary covariate model was then fitted with raw BAG in place of age-debiased BAG. Model diagnostics covered residual distribution, heteroskedasticity, collinearity, and influential observations. Feature contribution was described as the median absolute standardized ridge coefficient across the 100 outer models. Ridge coefficients are conditional on correlated predictors. Their ranking was therefore read as descriptive model contribution, not as a regional causal effect or an independent biomarker.

Reproducibility. Acquisition, feature extraction, model fitting, validation, statistical analysis, and figure generation were all scripted. The analysis archive holds the download manifest, derived analytic dataset, subject- and model-level predictions, regression tables, bootstrap performance results, diagnostics, feature summaries, figures, random seed, and software-session information. This structure supports distributed reconstruction and site-aware validation, priorities also shown in decentralized brain-age estimation research.[38] Independent reconstruction from the public files reproduced the reported sample flow, together with the direction, magnitude, and inference of the primary and sensitivity results.

2.6. Ethical approval

The CMHC Ethics Committee Indonesia granted ethical approval for the secondary analysis (No. 2026/0252). The source UCLA CNP study obtained written informed consent and institutional review board approval, as documented in the original data descriptor.[33] The present study used only public, deidentified data. No reidentification was attempted.

3. Results

3.1. Participant flow and characteristics

The source release included 272 participants, of whom 265 had the relevant FreeSurfer derivatives. After restriction to controls, bipolar disorder, and schizophrenia with T1-weighted data and the complete set of required statistics files, 224 participants entered the analysis. There were no duplicate participant identifiers. The final 105-feature matrix had no missing values. The analytic cohort contained 125 controls, 49 participants with bipolar disorder, and 50 with schizophrenia; the complete demographic and technical profile is given in Table 1.

Table 1. Characteristics of the analytic sample.

Characteristic	Control (n = 125)	Bipolar disorder (n = 49)	Schizophrenia (n = 50)	p value
Age, years, mean (SD)	31.53 (8.80)	35.29 (9.03)	36.46 (8.88)	0.001
Female-coded gender, n (%)	59 (47.2)	21 (42.9)	12 (24.0)	0.018
Scanner 35426, n (%)	23 (18.4)	23 (46.9)	25 (50.0)	< 0.001
Ghost artifact, n (%)	15 (12.0)	8 (16.3)	20 (40.0)	< 0.001
Age-debiased BAG, years, mean (SD)	0.01 (2.67)	0.79 (2.96)	0.89 (3.11)	–

BAG, brain age gap; SD, standard deviation. Age was compared by one-way analysis of variance; categorical variables were compared by Pearson chi-square tests. BAG p values are presented in Table 3 after covariate adjustment.

Mean age was 31.53 years in controls, 35.29 years in bipolar disorder, and 36.46 years in schizophrenia (p = 0.001). The proportion coded female was 47.2%, 42.9%, and 24.0%, respectively (p = 0.018). Scanner 35426 contributed 18.4% of controls, 46.9% of the bipolar group, and 50.0% of the schizophrenia group (p < 0.001). Ghost artifact was labeled in 12.0% of controls, 16.3% of the bipolar group, and 40.0% of the schizophrenia group (p < 0.001). These imbalances supported keeping all technical and demographic covariates.

3.2. Raw age-prediction performance

In the 125 controls, repeated held-out prediction gave an MAE of 6.48 years (95% bootstrap confidence interval [CI] 5.83–7.11) and an RMSE of 7.45 years (95% CI 6.78–8.11). The Pearson correlation between chronological and predicted age was 0.562 (95% CI 0.440–0.663), and R^2 was 0.277. Table 2 gives these performance measures.

Table 2. Raw out-of-fold age-prediction performance in 125 controls.

Metric	Estimate	95% bootstrap CI
Mean absolute error	6.48 years	5.83 to 7.11
Root mean squared error	7.45 years	6.78 to 8.11
Pearson correlation	0.562	0.440 to 0.663
R^2	0.277	Not estimated

CI, confidence interval; MAE, mean absolute error; RMSE, root mean squared error. Metrics were calculated from raw predicted age before age-bias correction; bootstrap intervals used 5,000 participant-level resamples.

Figure 1 shows the chronological-versus-predicted pattern. It displayed the expected regression towards the sample mean: younger adults tended to receive older predictions, and adults near the upper end of the 21–50-year range tended to receive younger ones. These raw metrics and the restricted age span support exploratory group-level estimation with explicit uncertainty. They do not support individual age estimation or clinical classification.

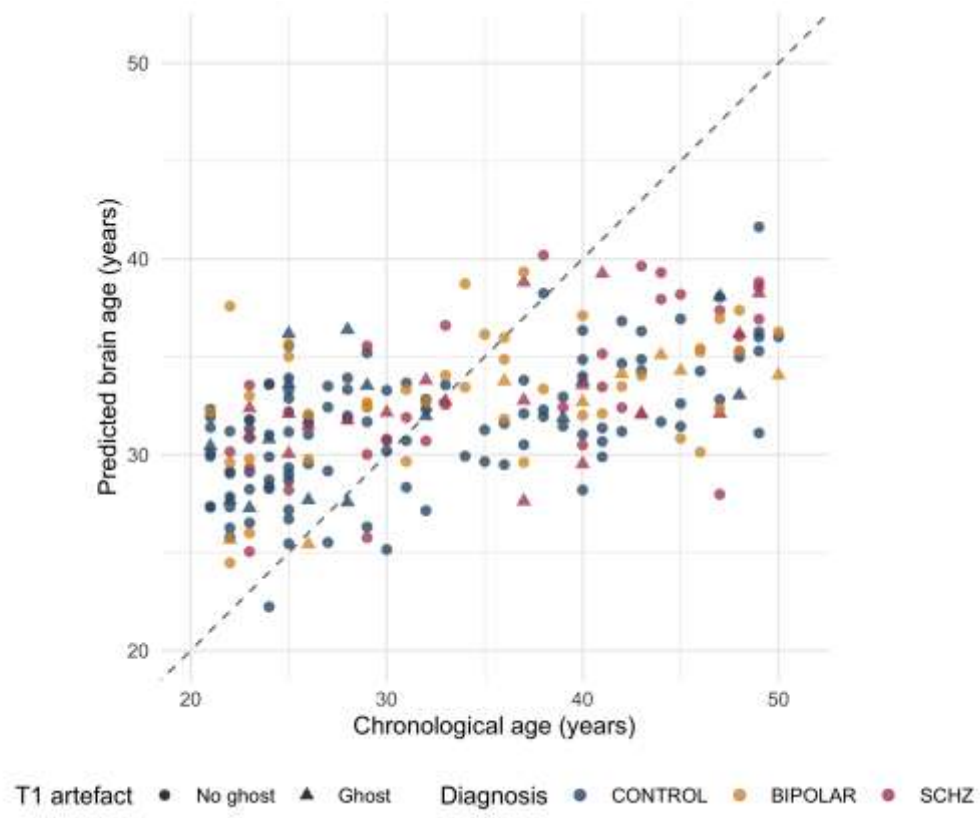


Figure 1. Chronological age versus raw predicted brain age. Control predictions were held out; patient predictions were averaged across 100 control-trained models. The dashed line indicates perfect prediction. Triangles identify scans labeled with ghost artifact. CONTROL, controls; BIPOLAR, bipolar disorder; SCHZ, schizophrenia.

3.3. Brain age gap distributions and primary comparisons

Mean age-debiased BAG was 0.01 years (standard deviation [SD] 2.67) in controls, 0.79 years (SD 2.96) in bipolar disorder, and 0.89 years (SD 3.11) in schizophrenia. Figure 2 shows the complete distributions. Both patient groups were shifted modestly in the positive direction, but individual values overlapped extensively across all three groups.

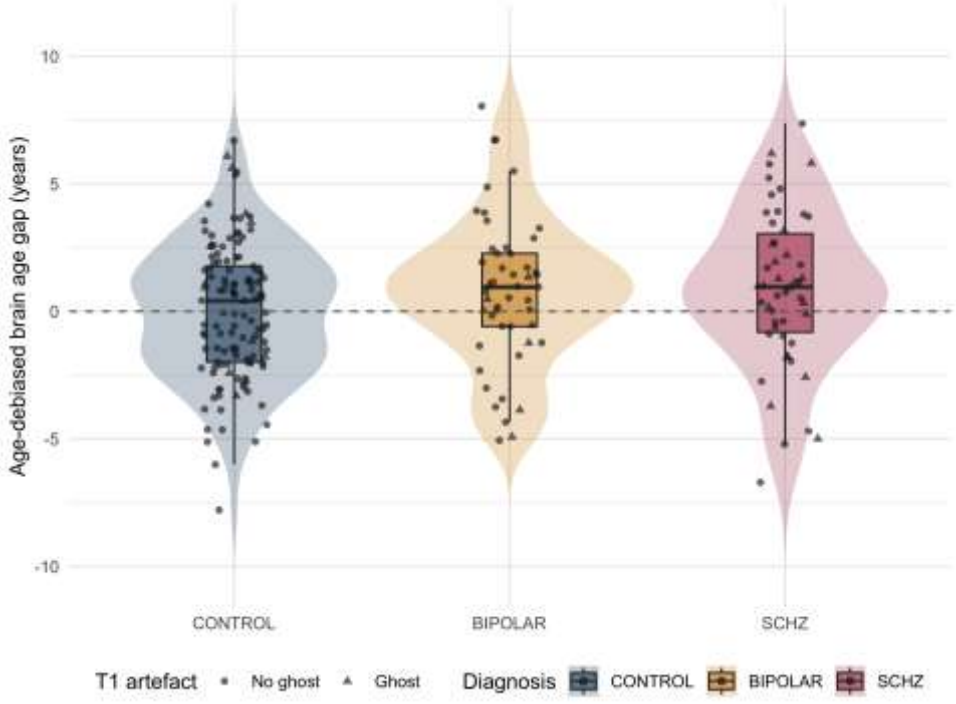


Figure 2. Distribution of age-debiased brain age gap by diagnosis. Violin shapes show density, boxes show the median and interquartile range, and points represent individual participants. Triangles identify scans labeled with ghost artifact. The dashed line marks zero. CONTROL, controls; BIPOLAR, bipolar disorder; SCHZ, schizophrenia.

After adjustment for age, recorded gender, scanner, and ghost status, bipolar disorder was associated with a 0.55-year higher BAG than controls (95% CI −0.50 to 1.59; p = 0.304; q = 0.304). Schizophrenia was associated with a 0.63-year higher BAG than controls (95% CI −0.51 to 1.77; p = 0.276; q = 0.304). The adjusted schizophrenia-versus-bipolar difference was 0.08 years (95% CI −1.16 to 1.33; p = 0.895). Table 3 gives the full coefficient set, and Figure 3 visualizes the two patient-versus-control estimates with their 95% CIs.

Table 3. Adjusted diagnostic-group differences and sensitivity analyses.

Analysis and contrast	Difference, years	95% CI	p value	q value
Primary: bipolar disorder vs control	0.55	−0.50 to 1.59	0.304	0.304
Primary: schizophrenia vs control	0.63	−0.51 to 1.77	0.276	0.304
Primary: schizophrenia vs bipolar disorder	0.08	−1.16 to 1.33	0.895	–
No-ghost subset: bipolar disorder vs control	0.94	−0.19 to 2.08	0.103	–
No-ghost subset: schizophrenia vs control	0.68	−0.80 to 2.16	0.364	–
Raw BAG: bipolar disorder vs control	0.56	−0.48 to 1.61	0.287	–
Raw BAG: schizophrenia vs control	0.65	−0.48 to 1.78	0.257	–

BAG, brain age gap; CI, confidence interval; HC3, heteroskedasticity-consistent standard errors; q, Benjamini–Hochberg false-discovery-rate q value. Primary and raw-BAG models were adjusted for age, recorded gender, scanner, and ghost status with HC3 standard errors. The no-ghost model was adjusted for age, recorded gender, and scanner. Benjamini–Hochberg q values were applied only to the two predefined primary patient-versus-control comparisons.

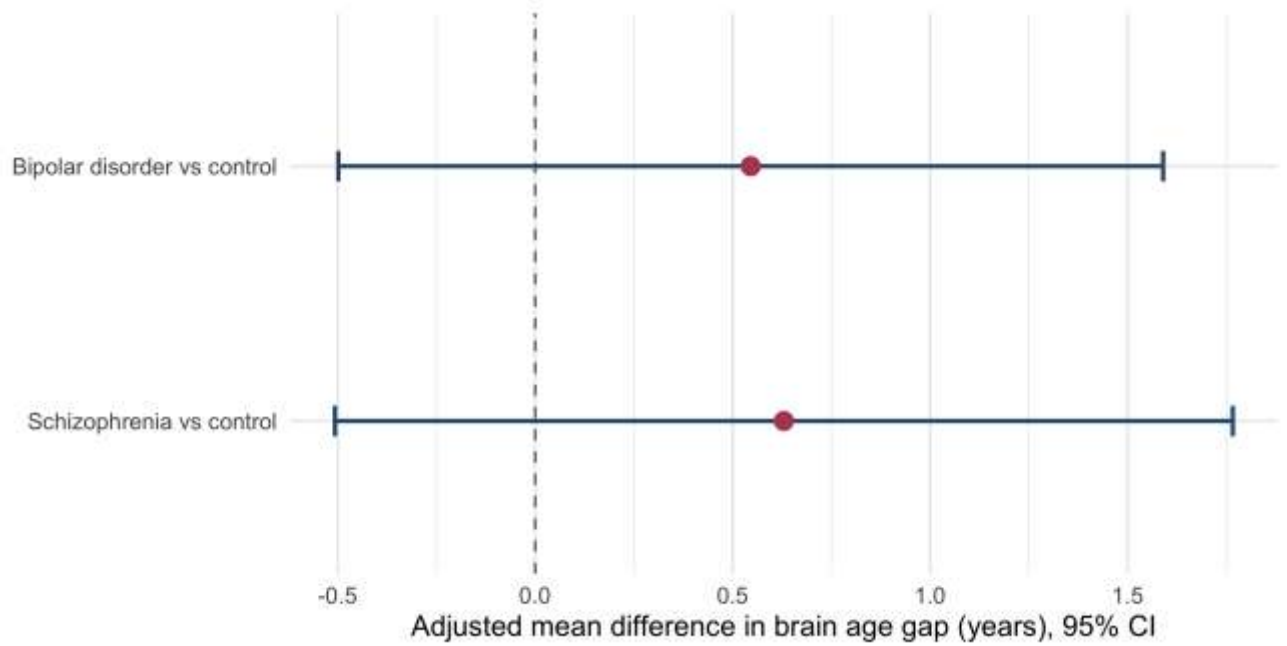


Figure 3. Adjusted differences in age-debiased brain age gap relative to controls. Points are regression coefficients and lines are 95% confidence intervals (CI) from models adjusted for age, recorded gender, scanner, and ghost status using HC3 standard errors. The dashed line marks zero. HC3, heteroskedasticity-consistent standard errors.

Both primary confidence intervals crossed zero. Neither comparison met the false-discovery-rate threshold. The intervals are compatible with no difference and with positive differences of up to approximately 1.6 years for bipolar disorder and 1.8 years for schizophrenia. They support a small, imprecisely estimated positive direction, not biological equivalence among groups.

3.4. Sensitivity analyses and model diagnostics

Excluding 43 ghost-labeled scans left 181 participants, 30 of them from the original 50 with schizophrenia. In this subset, the adjusted difference was 0.94 years for bipolar disorder versus controls (95% CI −0.19 to 2.08; p = 0.103) and 0.68 years for schizophrenia versus controls (95% CI −0.80 to 2.16; p = 0.364). Point estimates stayed positive. Precision fell because the exclusion was concentrated in schizophrenia. Table 3 gives these estimates.

When raw BAG replaced age-debiased BAG, adjusted estimates were 0.56 years for bipolar disorder (95% CI −0.48 to 1.61; p = 0.287) and 0.65 years for schizophrenia (95% CI −0.48 to 1.78; p = 0.257), also given in Table 3. Model diagnostics are shown in Table 4: maximum variance inflation factor 1.38, maximum Cook's distance 0.0478, Breusch–Pagan p value 0.962, and Shapiro–Wilk p value 0.896. Seventeen observations exceeded the 4/n threshold for Cook's distance, and HC3 inference was kept as the predefined conservative approach.

Table 4. Diagnostic assessment of the primary adjusted model.

Diagnostic	Result	Interpretation
Observations / residual df	224 / 217	Complete analytic cohort
Adjusted R^2	0.009	Group/covariate model fit
Maximum variance inflation factor	1.377	Low collinearity
Maximum Cook's distance	0.0478	No dominant observation
Cook's distance > 4/n	17 observations	HC3 inference retained
Breusch–Pagan test	p = 0.962	No evidence of heteroskedasticity
Shapiro–Wilk test	p = 0.896	Residual normality not rejected

df, degrees of freedom; HC3, heteroskedasticity-consistent covariance estimator; BAG, brain age gap. Diagnostics refer to the primary model of age-debiased BAG adjusted for diagnosis, chronological age, recorded gender, scanner, and ghost status.

3.5. Morphometric contributions

Across the 100 ridge models, the largest median absolute standardized coefficients included left pars opercularis thickness, white-matter hypointensity volume, bilateral caudate volume, right caudal middle frontal thickness, third-ventricle volume, and bilateral lateral-ventricle volume. Figure 4 shows the 15 leading features, their relative coefficient magnitudes, and the mean coefficient directions. Most leading cortical-thickness and caudate coefficients were negative with age, whereas ventricular and white-matter-hypointensity coefficients were positive. These rankings describe the age-prediction model. They were not tested as diagnosis-specific regional effects.

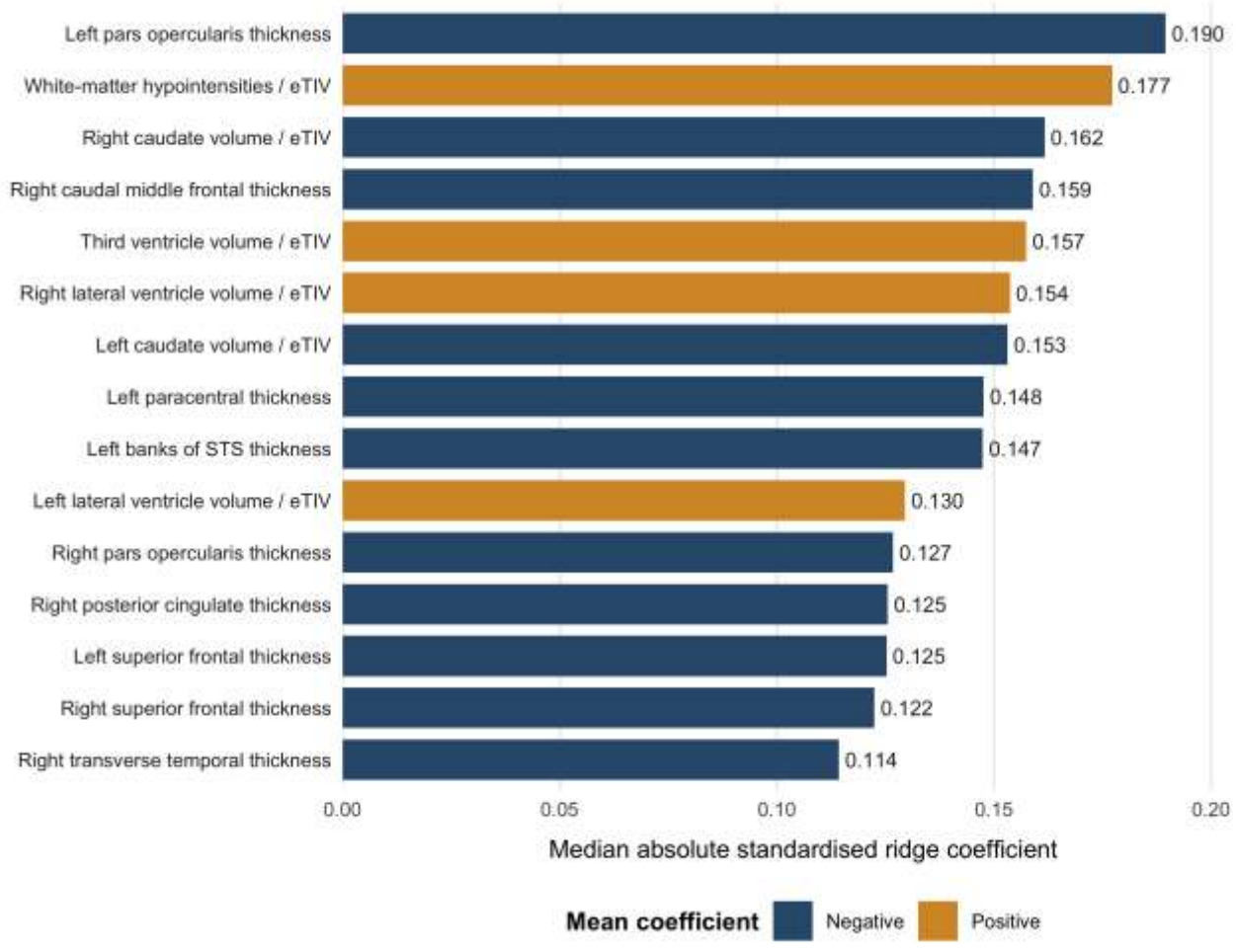


Figure 4. Leading morphometric contributions to the control-trained ridge age model. Bars show the median absolute standardized coefficient across 100 outer models; color indicates the direction of the mean coefficient. Rankings are descriptive contributions within a correlated predictor set and are not diagnosis-specific regional effects. eTIV, estimated total intracranial volume; STS, superior temporal sulcus.

4. Discussion

4.1. Principal findings

This leakage-controlled transdiagnostic analysis found small positive BAG estimates in both bipolar disorder and schizophrenia. The adjusted differences were imprecise, however, and did not differ statistically from zero. The schizophrenia–bipolar contrast was essentially null. Individual BAG values overlapped heavily, which implies that BAG as estimated here would not separate diagnoses at the participant level (no classification analysis was performed). The central result is therefore not that the two disorders share

equivalent brain aging, and it is not that earlier positive studies are contradicted. It is that a modest, narrow-age, single-dataset control model yields considerably smaller and less precise group differences once scanner and ghost artifact are explicitly handled.

The pipeline operated as specified. Every control prediction was held out. Patients never influenced training, hyperparameter selection was nested, and age-bias correction came only from training-control predictions. Raw performance was reported apart from corrected group inference. File-level provenance, repeated modeling, ensemble predictions, HC3 standard errors, and concordant sensitivity analyses gave an auditable estimate of what this dataset supports. Transparency matters most for null or inconclusive findings. It limits the chance that a visually persuasive result stems from optimistic validation or a technical imbalance.

4.2. Comparison with the existing literature

The adjusted differences of 0.55 years for bipolar disorder and 0.63 years for schizophrenia are smaller than many recent primary-study estimates, although the broader literature is not uniform. Transdiagnostic and schizophrenia-focused studies report positive average gaps, yet they also show variation by illness stage, modality, age distribution, and reference model.[8–18] In bipolar disorder, international and single-site studies find that average BAG is sensitive to model selection, medication, subtype, and age-by-diagnosis patterns.[19–21] The present model learned from 125 controls aged 21–50 years. Its restricted reference distribution and moderate raw accuracy limit both the range and the calibration of its predictions.

Recent longitudinal evidence offers several possible explanations for divergent magnitudes. Early treatment response, disease duration, antipsychotic exposure, psychotic experiences, exercise, and treatment during first-episode mania have each been linked to BAG or to its change over time.[22–27] Multimodal symptom analyses likewise suggest that a global gap may combine distinct anatomical and clinical contributions.[28] The small averages in ds000030 may therefore reflect a mixture of subgroups with different developmental histories, illness stages, treatment exposures, and technical characteristics. They need not signal an absence of neuroanatomical variation.

This reading agrees with multidimensional psychiatric trajectories and newer psychosis-biotype findings.[18,29] One adjusted diagnostic coefficient can hide marked within-group heterogeneity. In a modest open cohort, transparent estimation and uncertainty serve better than retrospective classification claims.

4.3. Interpretation and mechanism

Meaning of the effect size and uncertainty. Statistical non-significance should be kept apart from practical magnitude. The point estimates, 0.55 and 0.63 years, are small beside both the 21–50-year sampling range and the 6.48-year MAE. They are also far smaller than the within-group SD of BAG, which was approximately 3 years. Even if a population mean of similar size were confirmed, the overlap would rule out individual diagnosis from BAG alone. Conversely, the upper confidence limits allow positive group differences approaching 1.6–1.8 years. The present study therefore cannot establish equivalence, and it cannot exclude a small-to-moderate effect.

The distinction between prediction error and group-effect uncertainty is crucial. MAE describes how far individual control predictions typically fall from chronological age. It does not directly determine whether a group mean can be estimated. Repeated ensemble averaging can stabilize participant-level predictions, and covariate adjustment can estimate a small mean shift. Even so, the limited normative sample and the uncertainty of model fitting lower confidence in calibration outside this cohort. The analysis supports estimation with explicit intervals, not a binary declaration that BAG is present or absent.

Age bias, validation, and technical confounding. The raw prediction plot displayed regression towards the mean, a common property of brain-age models. Empirical comparisons show that performance metrics depend on the age distribution, and that the correction strategy can alter the apparent relation considerably between BAG and chronological age.[3–5,37] The correction here served only to construct a downstream association outcome, and its parameters were learned without the participant being evaluated. The raw-BAG sensitivity analysis agreed with the primary one, which indicates that the main conclusion was not an artifact of residualization.

Nested validation matters. Feature scaling, nuisance handling, model selection, or related observations crossing folds can change BAG and inflate apparent performance.[30,36] In this analysis, scaling, penalty selection, and age-bias estimation stayed within training data. Patient labels and group results were never used to choose the ridge penalty. These safeguards do not create external validity. They do make the internal estimate interpretable.

Scanner and ghost artifact were not minor covariates. Scanner 35426 was used for approximately half of each patient group but for fewer than one fifth of controls, and ghost artifact was more than three times as common in schizophrenia as in controls. Motion studies show that image degradation can systematically alter brain-age estimates and their reliability.[31,32] Adjustment and exclusion analyses addressed the measured components of this problem. Both are limited. The positive point estimates persisted. The ghost-free schizophrenia analysis, however, lost precision because 40% of that group was removed. Linear adjustment cannot remove every acquisition or segmentation effect, and a binary ghost label cannot replace image-level quality control or harmonization.

Biological interpretation. BAG condenses a distributed anatomical pattern into a single scalar. It can place an individual against a normative age trajectory, but the word “age” should not be read literally as a clock. A cross-sectional positive BAG may arise from early

neurodevelopmental differences, cumulative illness exposure, medication, cardiometabolic health, substance use, sleep, socioeconomic conditions, or combinations of these. It cannot tell accelerated longitudinal tissue change apart from a stable offset set before illness onset. Transdiagnostic and longitudinal findings therefore favor the phrase "older-appearing structural pattern" over a causal diagnosis of accelerated aging.[8,22–27]

The regional coefficients fit known age-related cortical thinning and ventricular or white-matter changes. They do not localize a disease mechanism. Ridge regression spreads weight across correlated predictors, so a large coefficient can reflect unique information, compensation for related features, scaling, or sample-specific covariance. The same regions can contribute to accurate age prediction without showing a diagnosis-specific abnormality. Mechanistic claims at the level of single regions would need prespecified regional deviation scores, uncertainty estimates, replication, and clinical association analyses that go beyond the global BAG model.

4.4. Implications for practice and policy

Clinical use. No clinical decision should rest on the present BAG values. Individual utility would require a large representative lifespan reference, external calibration across scanners and populations, test-retest reliability, fairness assessment, and evidence of incremental value beyond age, symptoms, cognition, illness duration, medication, and conventional MRI measures. A defined intended use, such as prognosis or treatment stratification, would also be needed in place of retrospective diagnostic separation. Recent bipolar and schizophrenia findings show why transportability and clinical context must come before implementation.[19–23]

Implications for research. The immediate next step is external application of a locked model developed on a much larger control sample that spans adulthood. Normative-model features can improve individual deviation mapping relative to raw measures, and empirical first-episode psychosis work shows that extreme deviations capture clinically relevant cortical heterogeneity.[39,40] External evaluation should report calibration-in-the-large, calibration slope, MAE by age stratum, residual BAG-age association, scanner-stratified estimates, harmonization sensitivity analyses, and fairness across adequately represented demographic groups.

Future transdiagnostic work should move past a single global group average. Recent studies have used normative modeling to assess pharmacological effects in schizophrenia, to map molecularly enriched functional circuits across diagnoses, to characterize dopaminergic deviations in psychosis, and to provide reusable morphometric centile platforms.[41–44] Such approaches should keep careful feature construction, because global adjustment can obscure regional abnormalities across the schizophrenia spectrum.[45] Replication should separate first-episode, chronic, and treatment-resistant groups. It should test age-by-diagnosis and medication interactions, link deviation patterns with cognition and functional outcomes, and use longitudinal scans to tell stable neurodevelopmental offsets from within-person change.

4.5. Strengths and limitations

The study has several strengths. It used openly licensed participant-level data and kept file-level provenance. The normative model was trained only in controls, and every control estimate was held out. Hyperparameter tuning, scaling, and age-bias correction were nested. One procedure was applied to both diagnoses, which allows a direct transdiagnostic comparison. Known scanner and ghost imbalances were modeled, HC3 uncertainty was reported, false-discovery-rate correction was limited to the two predefined primary comparisons, and conclusions were checked with raw BAG and a ghost-free subset. The full analysis can be rerun from a small public derivative set. It needs no inaccessible raw-image infrastructure.

The limitations define the scope of inference. First, the 21–50-year age range is too narrow to estimate an adult lifespan trajectory, and predictive metrics depend strongly on the age distribution. Second, 125 controls are few against 105 correlated features; ridge regularization reduces variance but cannot replace a large normative cohort. Third, there was no independent external test site and no pretrained lifespan model. Calibration and transportability therefore remain unknown. Fourth, scanner and ghost labels were highly imbalanced across diagnoses. Adjustment reduces measured confounding, but residual acquisition and processing effects may remain. Fifth, the public derivatives lacked the image-level quality metrics needed for comprehensive surface-reconstruction review in this analysis.

Sixth, medication, illness duration, symptom severity, bipolar subtype, cardiometabolic status, substance exposure, and cognitive function were not entered into the group models. Their omission prevents mechanistic analysis and may leave clinical confounding. Seventh, the cross-sectional design cannot estimate within-person aging rate or temporal ordering. Eighth, participant-level ensemble BAG served as the outcome in the group model, and uncertainty from fitting the age-prediction model was not propagated into the HC3 confidence interval. Ninth, the ghost-free analysis changed the composition of the schizophrenia group, so it is a robustness check and not a purified estimate of the same target population. Finally, null-centered intervals do not demonstrate equivalence. Feature rankings are descriptive. They are not regionally inferential.

5. Conclusion

A control-only, nested, and technically adjusted analysis of OpenNeuro ds000030 found small positive brain age gap estimates in schizophrenia and bipolar disorder. Overlap was wide. The confidence intervals spanned zero. The near-zero direct contrast gave no evidence

that schizophrenia had a larger gap than bipolar disorder in this cohort. Raw-BAG and ghost-exclusion analyses agreed with the direction of the primary conclusion, whereas the restricted age range, the modest normative sample, and the scanner-quality imbalance limited precision and transportability. The study provides an auditable open-data workflow and an internal estimate for this dataset, whose calibration outside this cohort is unknown. It does not support individual diagnosis. Nor does it support a causal claim of accelerated aging. External lifespan models, richer clinical phenotyping, multisite calibration, and longitudinal validation are required before brain age gap can be considered clinically useful.

Declarations

Use of Generative Artificial Intelligence. None. The authors declare that no generative artificial intelligence tool was used in the conception, research, analysis, drafting, revision or preparation of this manuscript, and no AI tool is listed as an author.

Author Contributions. A.H.: conceptualization, methodology, software, formal analysis, data curation, visualization, writing – original draft. R.U.A.-Z.: validation, investigation, writing – review & editing. C.S.: resources, supervision, project administration, writing – review & editing. All authors have read and approved the final version of the manuscript and agree to be accountable for all aspects of the work.

Funding. This work received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Conflict of Interest. The authors declare that they have no competing interests, financial or non-financial, in relation to this work.

Ethical Approval. Ethical approval for the secondary analysis was granted by the CMHC Ethics Committee Indonesia (No. 2026/0252). The source UCLA CNP study obtained institutional review board approval as documented in the original data descriptor.

Informed Consent. Written informed consent was obtained from participants in the original UCLA CNP study. No new participant contact or data collection occurred in this secondary analysis of public, deidentified data.

Data Availability. The source dataset is publicly available through OpenNeuro as ds000030. This analysis used release R1.0.5 and its public FreeSurfer derivatives.

Acknowledgements. The authors acknowledge the UCLA Consortium for Neuropsychiatric Phenomics investigators and participants and the OpenNeuro platform for making the dataset and derivatives publicly available.

References

1. Dufumier B, Grigis A, Victor J, et al. OpenBHB: a large-scale multi-site brain MRI data-set for age prediction and debiasing. Neuroimage. 2022;263:119637. doi:10.1016/j.neuroimage.2022.119637
2. Hahn T, Ernsting J, Winter NR, et al. An uncertainty-aware, shareable, and transparent neural network architecture for brain-age modeling. Sci Adv. 2022;8(1):eabg9471. doi:10.1126/sciadv.abg9471
3. de Lange AG, Anatürk M, Rokicki J, et al. Mind the gap: performance metric evaluation in brain-age prediction. Hum Brain Mapp. 2022;43(10):3113–29. doi:10.1002/hbm.25837
4. Zhang B, Zhang S, Feng J, et al. Age-level bias correction in brain age prediction. Neuroimage Clin. 2023;37:103319. doi:10.1016/j.nicl.2023.103319
5. More S, Antonopoulos G, Hoffstaedter F, et al. Brain-age prediction: a systematic comparison of machine learning workflows. Neuroimage. 2023;270:119947. doi:10.1016/j.neuroimage.2023.119947
6. Beheshti I, Ganaie MA, Paliwal V, et al. Predicting brain age using machine learning algorithms: a comprehensive evaluation. IEEE J Biomed Health Inform. 2022;26(4):1432–40. doi:10.1109/JBHI.2021.3083187
7. Wood DA, Kafiabadi S, Busaidi AA, et al. Accurate brain-age models for routine clinical MRI examinations. Neuroimage. 2022;249:118871. doi:10.1016/j.neuroimage.2022.118871
8. Demro C, Shen C, Hendrickson TJ, et al. Advanced brain-age in psychotic psychopathology: evidence for transdiagnostic neurodevelopmental origins. Front Aging Neurosci. 2022;14:872867. doi:10.3389/fnagi.2022.872867
9. Chen CL, Hwang TJ, Tung YH, et al. Detection of advanced brain aging in schizophrenia and its structural underpinning by using normative brain age metrics. Neuroimage Clin. 2022;34:103003. doi:10.1016/j.nicl.2022.103003
10. Huang J, Ke P, Chen X, et al. Multimodal magnetic resonance imaging reveals aberrant brain age trajectory during youth in schizophrenia patients. Front Aging Neurosci. 2022;14:823502. doi:10.3389/fnagi.2022.823502
11. Xi YB, Wu XS, Cui LB, et al. Neuroimaging-based brain-age prediction of first-episode schizophrenia and the alteration of brain age after early medication. Br J Psychiatry. 2022;220(6):339–46. doi:10.1192/bjp.2021.169
12. Tesli N, Bell C, Hjell G, et al. The age of violence: mapping brain age in psychosis and psychopathy. Neuroimage Clin. 2022;36:103181. doi:10.1016/j.nicl.2022.103181
13. Chakrabarty T, Frangou S, Torres IJ, et al. Brain age and cognitive functioning in first-episode bipolar disorder. Psychol Med. 2023;53(11):5127–35. doi:10.1017/S0033291722002136
14. Abram SV, Roach BJ, Hua JPY, et al. Advanced brain age correlates with greater rumination and less mindfulness in schizophrenia. Neuroimage Clin. 2023;37:103301. doi:10.1016/j.nicl.2022.103301
15. Ballester PL, Suh JS, Ho NCW, et al. Gray matter volume drives the brain age gap in schizophrenia: a SHAP study. Schizophrenia (Heidelb). 2023;9(1):3. doi:10.1038/s41537-022-00330-z
16. Zhu JD, Wu YF, Tsai SJ, et al. Investigating brain aging trajectory deviations in different brain regions of individuals with schizophrenia using multimodal magnetic resonance imaging and brain-age prediction: a multicenter study. Transl Psychiatry. 2023;13(1):82. doi:10.1038/s41398-023-02379-5
17. Kim WS, Heo DW, Maeng J, et al. Deep learning-based brain age prediction in patients with schizophrenia spectrum disorders. Schizophr Bull. 2024;50(4):804–14. doi:10.1093/schbul/sbad167
18. Promet L, Meda SA, Alliey-Rodriguez N, et al. Brain age disparities in psychosis across DSM diagnoses and B-SNIP biotypes. Schizophr Bull. 2026;52(1):sbaf022. doi:10.1093/schbul/sbaf022

19. Ng HX, Abé C, Alda M, et al. Brain aging in bipolar disorder using a neuroimaging and machine learning-derived metric: findings from the ENIGMA BD working group. J Affect Disord. 2026;403:121234.
doi:10.1016/j.jad.2026.121234
20. Topolski N, Barsotti EJ, Boscutti A, et al. Brain age in bipolar disorder: impact of model selection and clinical factors. Biol Psychiatry Cogn Neurosci Neuroimaging. 2026 [Epub ahead of print].
doi:10.1016/j.bpsc.2026.01.012
21. Weerasekera A, Zhou S, Wang C, et al. Deep learning-based structural brain age estimation in bipolar disorder and schizophrenia: a single-site pilot study. Hum Brain Mapp. 2026;47(3):e70479. doi:10.1002/hbm.70479
22. Fan L, Zhang Z, Ma X, et al. Brain age gap as a predictor of early treatment response and functional outcomes in first-episode schizophrenia: a longitudinal study. Can J Psychiatry. 2025;70(3):240–50.
doi:10.1177/07067437241293981
23. Roig-Herrero A, San-José-Revuelta LM, Navarro-González R, et al. Effects of disease duration and antipsychotics on brain age in schizophrenia. Schizophr Res. 2026;287:82–90. doi:10.1016/j.schres.2025.11.008
24. Navarro-González R, Luque-Laguna P, de Luis-García R, et al. Increased brain-age gap in young adults with psychotic experiences. Biol Psychiatry Glob Open Sci. 2026;6(2):100643. doi:10.1016/j.bpsgos.2025.100643
25. Yilmaz D, Papiol S, Keeser D, et al. Brain age gap reduction following exercise mirrors clinical improvements in schizophrenia spectrum disorders. Neuroimage Clin. 2025;48:103881. doi:10.1016/j.nicl.2025.103881
26. Yu K, Ge R, Yu Y, et al. Individual-level brain phenotypes in first-episode mania: normative modelling of brain morphometry and brainAGE. BJPsych Open. 2025;11(3):e95. doi:10.1192/bjo.2025.28
27. Han LKM, Dehestani N, Suo C, et al. Longitudinal brain age in first-episode mania youth treated with lithium or quetiapine. Eur Neuropsychopharmacol. 2025;95:40–8. doi:10.1016/j.euroneuro.2025.03.013
28. Lee J, Kim SY, Park HW, et al. A multimodal ensemble stacking model improves brain age prediction and reveals associations with schizophrenia symptoms. Front Psychiatry. 2025;16:1600479. doi:10.3389/fpsyt.2025.1600479
29. Niu X, Taylor A, Shinohara RT, et al. Multidimensional brain-age prediction reveals altered brain developmental trajectory in psychiatric disorders. Cereb Cortex. 2022;32(22):5036–49. doi:10.1093/cercor/bhab530
30. Jirsaraie RJ, Kaufmann T, Bashyam V, et al. Benchmarking the generalizability of brain age models: challenges posed by scanner variance and prediction bias. Hum Brain Mapp. 2023;44(3):1118–28. doi:10.1002/hbm.26144
31. Vakli P, Weiss B, Rozmann D, et al. The effect of head motion on brain age prediction using deep convolutional neural networks. Neuroimage. 2024;294:120646. doi:10.1016/j.neuroimage.2024.120646
32. Hanson JL, Adkins DJ, Bacas E, et al. Examining the reliability of brain age algorithms under varying degrees of participant motion. Brain Inform. 2024;11(1):9. doi:10.1186/s40708-024-00223-0
33. Poldrack RA, Congdon E, Triplett W, et al. A phenome-wide examination of neural and cognitive function. Sci Data. 2016;3:160110. doi:10.1038/sdata.2016.110
34. von Elm E, Altman DG, Egger M, et al. The strengthening the reporting of observational studies in epidemiology (STROBE) statement: guidelines for reporting observational studies. PLoS Med. 2007;4(10):e296. doi:10.1371/journal.pmed.0040296
35. 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(3):968–80.
doi:10.1016/j.neuroimage.2006.01.021
36. Rosenblatt M, Tejavibulya L, Jiang R, et al. Data leakage inflates prediction performance in connectome-based machine learning models. Nat Commun. 2024;15(1):1829. doi:10.1038/s41467-024-46150-w
37. Wang H, Treder MS, Marshall D, et al. A skewed loss function for correcting predictive bias in brain age prediction. IEEE Trans Med Imaging. 2023;42(6):1577–89. doi:10.1109/TMI.2022.3231730
38. Basodi S, Raja R, Ray B, et al. Decentralized brain age estimation using MRI data. Neuroinformatics. 2022;20(4):981–90. doi:10.1007/s12021-022-09570-x
39. Rutherford S, Barkema P, Tso IF, et al. Evidence for embracing normative modeling. eLife. 2023;12:e85082. doi:10.7554/eLife.85082
40. Worker A, Berthert P, Lawrence AJ, et al. Extreme deviations from the normative model reveal cortical heterogeneity and associations with negative symptom severity in first-episode psychosis from the OPTiMiSE and GAP studies. Transl Psychiatry. 2023;13(1):373. doi:10.1038/s41398-023-02661-6
41. Lin X, Huo Y, Wang Q, et al. Using normative modeling to assess pharmacological treatment effect on brain state in patients with schizophrenia. Cereb Cortex. 2024;34(2):bhae003. doi:10.1093/cercor/bhae003
42. Lawn T, Giacomel A, Martins D, et al. Normative modelling of molecular-based functional circuits captures clinical heterogeneity transdiagnostically in psychiatric patients. Commun Biol. 2024;7(1):689. doi:10.1038/s42003-024-06391-3
43. Giacomel A, Martins D, Nordio G, et al. Investigating dopaminergic abnormalities in schizophrenia and first-episode psychosis with normative modelling and multisite molecular neuroimaging. Mol Psychiatry. 2025;30(8):3533–50. doi:10.1038/s41380-025-02938-w
44. Little B, Alyas N, Surtees A, et al. Brain morphology normative modelling platform for abnormality and centile estimation: brain MoNoCle. Imaging Neurosci (Camb). 2025;3:imag_a_00438.
doi:10.1162/imag_a_00438
45. Hua JPY, Fryer SL, Stuart B, et al. Adjustment of regional cortical thickness measures for global cortical thickness obscures deficits across the schizophrenia spectrum: a cautionary note about normative modeling of brain imaging data. Biol Psychiatry Cogn Neurosci Neuroimaging. 2025;10(6):666–72.
doi:10.1016/j.bpsc.2024.06.001