Research Square

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

Cross-release transfer of Riemannian EEG pipelines: nearly lossless, but no robust gain from domain adjustment

Xiangzhu Li
Shenzhen Hospital of Integrated Traditional Chinese and Western Medicine
Huang Tan
University of Electronic Science and Technology of China
Xiaoyu Yang
Shenzhen Hospital of Integrated Traditional Chinese and Western Medicine
Yuhao Liu
Shenzhen Hospital of Integrated Traditional Chinese and Western Medicine
Huaxiao Li
Shenzhen Hospital of Integrated Traditional Chinese and Western Medicine
Li Li
Shenzhen Hospital of Integrated Traditional Chinese and Western Medicine
Haojian Mai
Shenzhen Hospital of Integrated Traditional Chinese and Western Medicine
Qihui Zhang
Shenzhen Hospital of Integrated Traditional Chinese and Western Medicine
Gaozhan Xie
Nanjing University of Posts and Telecommunications
Li Zhang
zhanglith@163.com

Shenzhen Hospital of Integrated Traditional Chinese and Western Medicine

Article

Keywords: EEG, brain age, domain generalization, Riemannian geometry, harmonization, predictive reproducibility

Posted Date: November 13th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11504555/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Cross-release transfer of Riemannian EEG pipelines: nearly lossless, but no robust gain from domain adjustment

Xiangzhu Li[1,2], Huang Tan[3], Xiaoyu Yang[1,2], Yuhao Liu[1,2], Huaxiao Li[1,2], Li Li[1,2], Haojian Mai[1,2], Qihui Zhang[1,2], Gaozhan Xie[4,*], Li Zhang[1,2,*]

[1] Shenzhen Hospital of Integrated Traditional Chinese and Western Medicine, Shenzhen, Guangdong, China.

[2] Shenzhen Clinical College of Integrated Chinese and Western Medicine, Guangzhou University of Chinese Medicine, Shenzhen, Guangdong, China.

[3] University of Electronic Science and Technology of China, Chengdu, Sichuan, China.

[4] State Key Laboratory of Flexible Electronics (LoFE) & Institute of Advanced Materials (IAM), Nanjing University of Posts and Telecommunications, Wenyuan Road, Nanjing, China.

*Corresponding authors: Gaozhan Xie (iamgzxie@njupt.edu.cn); Li Zhang(zhanglith@163.com).

Keywords: EEG; brain age; domain generalization; Riemannian geometry; harmonization; predictive reproducibility

Abstract

Whether resting-state electroencephalography (EEG) pipelines developed within one dataset transfer across independently collected data batches is rarely tested, although developmental cohorts are assembled from multiple releases. We evaluated a filter-band tangent-space pipeline on n = 2,525 participants (ages 5–21) from ten releases of the Healthy Brain Network (HBN)–EEG dataset, using leave-one-release-out (LORO) generalization with ridge regression on log-

Euclidean and Riemannian tangent-space features (225 and 11,545) under five domain-handling conditions. Brain-age prediction generalized with almost no loss: LORO R^2 was 0.596 (log-Euclidean) and 0.397 (Riemannian) versus pooled upper bounds of 0.602 and 0.400, with positive R^2 in all 10/10 held-out releases. Domain adjustment gave no robust gain: paired δR^2 depended on representation, with rotation-based alignment gaining +0.111 under Riemannian features but losing −0.062 under log-Euclidean features; re-centering left Pearson r exactly unchanged ($\delta r \equiv 0$) while producing small nonzero δR^2. Four phenotype tasks showed no detectable signal (all $|r| \leq 0.194$ versus an 80%-power minimum detectable effect of 0.248), and 9/10 representation × condition combinations were equivalent within ±0.02. Cross-release generalization is thus nearly lossless, whereas harmonization is not a reliable improvement; the phenotype nulls are reported as undetectable rather than absent.

1 Introduction

Cross-cohort reproducibility has become an explicit concern in developmental neuroimaging: brain-age and normative-modeling pipelines are routinely developed and evaluated on a single dataset, and their reported accuracy may not survive a change of acquisition batch [1,2]. Within-dataset accuracies for EEG brain-age prediction can be as high as $R^2 = 0.60–0.74$ across four international cohorts [3], yet external validation across datasets suggests a markedly different picture, with performance reported to degrade sharply when entire datasets are held out [4]. Independent benchmark work on SPD-matrix learning likewise reports that between-dataset differences in method ranking shrink when an entire dataset is held out [5]. Metrics matter as well: a systematic evaluation of brain-age performance measures shows that widely used summaries can be misleading when the evaluated quantity is a difference rather than a raw prediction [2].

Harmonization is the standard remedy assumed to transfer such models across batches, and Riemannian geometry has become the dominant coordinate system for this purpose, with recent work applying covariance-based alignment techniques specifically to boost predictive regression on M/EEG [6] and even optimizing site intercepts jointly within the Riemannian representation [7]. Two cautions apply. First, comparative evidence does not support a general claim that Riemannian pipelines outperform convolutional ones: systematic benchmarking reports no consistent advantage for the Riemannian approach [8], and the relationship between the Riemannian geometry and Euclidean summaries of SPD matrices is better understood as a choice of representation than as an inherent advantage [9]. Second, the claim that harmonization reliably improves prediction is itself under pressure, since over-correction can remove genuine signal.

Harmonization therefore needs to be evaluated as an empirical question on each target dataset, with the direction of its effects reported rather than assumed.

The Healthy Brain Network (HBN) EEG dataset offers an unusual opportunity to address these issues. Its resting-state recordings were released in ten sequential releases, and the released participant metadata contain an explicit release number but no collection-site field, which makes the release the only observable batch variable in the published data. Each release therefore constitutes an independently collected batch whose membership is not confounded with any other recorded site characteristic. To our knowledge, this batch structure has not been used to evaluate EEG harmonization, and the question of whether such pipelines transfer across releases remains open.

We evaluated a filter-band tangent-space pipeline on n = 2,525 participants and asked four questions. (RQ1) How much brain-age prediction accuracy is lost when an entire release is held out relative to a pooled upper bound, and does any individual release collapse? (RQ2) Do domain-handling steps—re-centering, rotation-based alignment, moment correction, and transductive centering—improve leave-one-release-out (LORO) prediction over no adjustment, and is any gain consistent across representations? (RQ3) What accounts for the measurable differences between releases? (RQ4) Do the four dimensional phenotype targets show any cross-release predictive signal, and can alignment steps be shown to leave those associations unchanged within a prespecified equivalence margin?

Our contributions are threefold. First, we provide a factorial evaluation of two tangent-space representations against five domain-handling conditions under a within-dataset batch structure, with all scores reproducible from released products. Second, we derive and then verify two algebraic identities that render two widely used domain steps inert under this protocol, which

explains why two widely used domain steps cannot produce genuine gains under this protocol. Third, we report the phenotype tasks with an explicit detectability bound and a paired equivalence test, positioning the null result as a statement about detectability rather than about the absence of association.

2 Results

2.1 Cross-release generalization of brain-age prediction is nearly lossless

Under LORO(**Fig. 1d**), brain-age prediction retained almost all of the accuracy achievable when release membership was ignored (**Fig. 2a**). Pooled ten-fold cross-validation gave mean R^2 = 0.602 (log-Euclidean, SD 0.067 across folds) and R^2 = 0.400 (Riemannian, SD 0.045), with mean Pearson r = 0.781 and r = 0.670 respectively. LORO with no adjustment gave R^2 = 0.596 (SD 0.095) and R^2 = 0.397 (SD 0.060), with r = 0.780 and r = 0.683. The LORO-minus-pooled R^2 difference was −0.006 for log-Euclidean and −0.003 for Riemannian.

Release by release, R^2 was positive in all 10 of 10 held-out releases for both representations (**Fig. 2b**), spanning 0.407–0.694 (log-Euclidean) and 0.323–0.510 (Riemannian), with no single-release collapse. Within this dataset and pipeline, therefore, a brain-age model trained on age transfers to unseen releases with no substantive loss relative to the release-agnostic upper bound. One qualification is required: the log-Euclidean representation was more accurate than the Riemannian one for this task by about 0.20 in R^2, which is a within-sample comparison and is not generalized here into a claim about the two geometries.

2.2 Domain adjustment yields no robust gain, and its sign depends on the representation

Paired δR^2 relative to no adjustment for brain age under LORO (**Fig. 3a**) showed no domain-handling condition that improved prediction robustly across both representations. Under

Riemannian features, the rotation-based alignment gave a mean δR^2 of +0.111 (bootstrap 95% CI [+0.063, +0.161]), positive in 9 of 10 releases (**Fig. 3b**), with gains concentrated in R3 (+0.222) and R8 (+0.252). Re-centering was slightly negative (−0.007, [−0.014, −0.001]), moment correction was indistinguishable from no adjustment (+0.002, [−0.015, +0.017]), and transductive centering was algebraically identical to no adjustment ($|\delta R^2| < 10^{-14}$). Under log-Euclidean features the same rotation-based alignment was significantly negative (−0.062, [−0.107, −0.021]), as was moment correction (−0.097, [−0.136, −0.062]), with per-release disagreement (positive in only 2 of 10 releases, −0.220 in R1); re-centering was again slightly negative (−0.010, [−0.020, −0.0001]).

The same alignment step therefore helped under one representation and hurt under the other, which identifies the effect as an interaction between representation and procedure rather than evidence that alignment works. Moreover, even the single significantly positive combination (Riemannian with rotation-based alignment, mean $R^2 \approx 0.51$) remained below the unadjusted log-Euclidean condition ($R^2 = 0.596$). On this dataset the simplest unadjusted condition gave the best or near-best cross-release brain-age performance, and introducing domain handling brought no robust improvement.

Consistent with **equation (1)**, re-centering produced a δr of exactly zero at machine precision while its δR^2 moved slightly; this residual movement reflects a constant prediction offset acting on evaluation residuals, not a benefit of domain transport.

2.3 Release-to-release drift is structural rather than demographic

Quantifiable differences between releases arose mainly from measurement arrangement and phenotype availability rather than from age composition (**Fig. 4–7**): age drift was mild ($|d| \leq 0.15$, Kruskal–Wallis $p = 0.036$), whereas the depth of phenotype assessment and task

availability depended clearly on release. This structure is compatible with the near-lossless transfer above, because the core population variable is stable across releases while feature geometry changes only modestly; it also indicates that the phenotype tasks are directly exposed to between-release differences in assessment.

2.4 Phenotype tasks show no cross-domain signal, and equivalence is limited by detectability

The four non-BAG phenotype tasks showed no usable predictive signal under LORO with no adjustment (**Fig. 2c**): mean Pearson r was near zero throughout, reaching 0.037 (p-factor), 0.052 (attention), 0.194 (internalizing) and 0.103 (externalizing) under log-Euclidean features and 0.051, 0.028, 0.157 and 0.172 under Riemannian features, with all $|r| \leq 0.194$ and corresponding mean R^2 between −0.009 and +0.027. Pooled cross-validation showed the same absence of signal (R^2 from −0.009 to +0.026, $|r| \leq 0.18$), indicating that the failure is not driven mainly by domain shift but by the limited linear predictability of these phenotypes from the present EEG features. We therefore do not claim to have detected any linear predictability of the dimensional phenotypes from the present EEG features.

To test whether alignment materially changed these near-zero associations, we applied paired Fisher-z TOST on 40 within-task, within-release pairs per cell, with a margin of ±0.02 in r. Nine of ten representation × condition cells were equivalent (**Table S6; Fig. S1**). Six cells were structurally zero by the identities derived in Methods: no adjustment trivially, re-centering through equation (1), and transductive centering by exact equivalence. The three conditions that genuinely transported the distribution—log-Euclidean moment correction (90% CI [−0.009, +0.002]), log-Euclidean rotation-based alignment ([−0.013, +0.001]) and Riemannian moment correction ([−0.004, +0.006])—had increments inside the margin. The single exception was

Riemannian with rotation-based alignment (mean dz = −0.013, SD 0.097, CI [−0.038, +0.012]), whose across-release increment was the most dispersed of the ten cells, consistent with the high across-release variability of that alignment under Riemannian features in **Fig. 3**. Taken together, the MDE and the equivalence tests support reporting no detectable phenotype signal and no evidence that alignment materially altered that conclusion, except for Riemannian with rotation-based alignment; this is a statement about detectability, not a demonstration of any brain–behavior association.

2.5 Within-release small-sample baselines are unstable

Within-release cross-validation served only as a reference and carries no main conclusion, but its instability supports LORO as the primary analysis (**Fig. 8**). Across 50 folds the log-Euclidean brain-age R^2 had mean 0.017 and SD 0.957, with 18% of folds negative and four releases (releases 1, 2, 3 and 6) showing negative five-fold mean R^2; release 3 showed catastrophic extrapolation (worst fold $R^2 = -5.785$). The Riemannian representation was more stable (mean 0.135, SD 0.089, 6% negative, worst −0.089) but at a low overall level. The cause is small per-release training size with high feature dimensionality, which inflates fold-to-fold extrapolation variance. LORO trains on all remaining releases and is immune to this instability.

3 Discussion

We evaluated a Riemannian tangent-space pipeline across ten sequential data releases using leave-one-release-out generalization, and obtained three findings. First, cross-release transfer of brain-age prediction was nearly lossless: LORO accuracy matched the release-agnostic pooled upper bound to within 0.006 in R^2, and no held-out release collapsed. This is notable because published external validations of EEG brain-age report substantial degradation when datasets are

swapped [4]; the difference here is that releases share a cohort definition and acquisition protocol while differing in when and by whom recordings were made, so the shift is batch-like but not site-institution-like. Second, domain adjustment did not yield a robust gain: the sign of its effect depended on the tangent-space representation, and the best adjusted condition did not reach the unadjusted baseline. Third, two widely used domain steps proved algebraically inert in this design, which provides a concrete explanation for one class of apparently interpretable but spurious harmonization effects.

The most actionable methodological consequence concerns re-centering. Under leave-one-domain-out evaluation with an intercept-bearing linear model, a constant feature shift on the evaluation side leaves Pearson correlation exactly unchanged, so any reported effect on prediction must come through R^2 alone, and that channel is sensitive to a constant prediction offset acting on evaluation residuals, not to domain transport. Investigators using this design should report δr alongside δR^2 and should not read small re-centering δR^2 values as harmonization gains. Our data show this pattern directly, with δr at machine precision alongside δR^2 of order 0.007–0.010. Transductive centering, when applied as a single shift to both sides, is exactly redundant under the same model and should be removed from comparison designs.

The representation dependence of the rotation-based alignment is equally consequential. The same procedure gained +0.111 in δR^2 under Riemannian features and lost −0.062 under log-Euclidean features, and neither sign is the alignment's intrinsic property. We interpret this as an interaction between the procedure and the geometry in which the rotation is performed, consistent with recent evidence that the choice of SPD representation is itself a modeling decision rather than an inherent advantage [9], and with benchmark work in which method differences shrink when whole datasets are held out [5]. Because covariance-based alignment for

predictive regression is an active area [6] and geodesically optimized alignment is now available [7], our re-centering and rotation conditions should be read as simplified variants; the latter is the one condition that failed the equivalence test, and its per-release dispersion suggests that a jointly learned alignment might behave differently rather than identically.

For the phenotype tasks, we report no detectable association. The relevant point is that this is a statement about detectability at the available sample size, not about the absence of a brain–behavior relationship. With a single-release minimum detectable effect of $r = 0.248$ and all observed $|r| \leq 0.194$, the present design could not have detected moderate phenotype associations, and the pooled setting showed the same absence of signal, indicating the limitation is predictability from these features rather than cross-release shift. Framing such results in terms of out-of-sample generalization rather than in terms of significance testing follows established practice in neuroimaging methodology [14]. This direction is also consistent with work on bias correction in brain-age prediction, which shows that age-related bias can account for apparent associations [15], with evaluations of brain-age performance metrics that caution against over-interpreting derived quantities [2], and with methods work estimating brain age deltas from imaging [1]. Establishing a floor on observable association would require either far larger evaluation samples, or measurement reliability information that caps attainable correlation, and we regard this as the natural next step for a multi-release cohort.

Several limitations should be stated. We did not implement component-based artifact removal or channel-level repair, and retained bad channels to keep the channel dimension constant; this affects absolute accuracy and is not addressed by the relative comparisons we draw. A geodesic intercept-optimizing alignment was not implemented and is reported as missing rather than approximated. The projection basis and global tangent reference are label-

free geometric constants estimated once, not re-estimated within each fold, so a strictly fold-only variant remains untested. The within-release baseline is unstable by construction and carries no conclusion. Finally, our release-level batches differ in composition along measurement-related rather than demographic axes, so the near-lossless transfer observed here should not be extrapolated to between-site shifts or to other cohorts.

4 Methods

4.1 Data and cohort

Resting-state EEG from the HBN developmental cohort was obtained from OpenNeuro as ten sequential releases (ds005505–ds005512, ds005514, ds005515), denoted R1–R10 in order of release [10]. Recordings are EEGLAB files with 129 channels (E1–E128 plus Cz) sampled at 500 Hz with a Cz reference; the released metadata contain no collection-site field, so release membership is the only observable batch variable. Because the cohort spans ages 5–21 and was assembled incrementally over roughly a decade, differences in age composition between releases are expected a priori and are quantified in the Results. Large developmental cohorts of this kind have made individual-level anatomical and functional reference curves available [11]; we do not attempt normative modelling here and use the cohort only as a test of transfer across batches.

Subject-level quality control was applied independently in the preprocessing and feature stages, and a subject was retained only if both stages reported success, all 129 channels were present, the smallest eigenvalue of every covariance matrix satisfied a relative positive-definiteness criterion, at least 60 windows survived sampling, fewer than half of the windows were rejected as artifact, and the median across-channel median peak amplitude was at most 100 μV. Of 2,725 participants listed in the release manifests, 2,597 had a usable resting-state file; 72 of these failed preprocessing or quality control, leaving 2,525 participants for analysis (R1 135, R2 149, R3 178, R4 314, R5 319, R6 131, R7 358, R8 211, R9 259, R10 471), corresponding to 1,875,615 covariance matrices. The four dimensional phenotypes (p-factor, attention, internalizing, externalizing) were available for 2,490 participants each. **Fig. 4** summarizes data availability and task-flag composition across releases, **Fig. 5** the demographic drift, **Fig. 6** phenotype missingness, and **Fig. 7** the per-release age and p-factor distributions.

4.2 Preprocessing and covariance features

The full analysis chain, from cohort assembly to the LORO split and the five domain-handling conditions, is summarized in **Fig. 1**. Each recording was read, band-pass filtered from 1 to 40 Hz with zero phase, notch filtered at 60 Hz, resampled to 250 Hz, and divided into non-overlapping 2-s windows (**Fig. 1b**). Windows were rejected when the across-channel median peak amplitude exceeded 100 μV, and surviving windows were subsampled at equal intervals to at most 150 windows per participant. We chose the across-channel median rather than a per-channel threshold because the latter is dominated by a small number of bad channels: in one participant the median channel had a peak amplitude of 1,930 μV across almost the entire recording while a typical participant sat at 122 μV, so a per-channel rule still rejected 63% of that participant's windows. Note that a released "available" flag does not guarantee a clean amplitude trace, which is why pipeline-level window rejection is not redundant with the released flags. Bad channels were not removed, keeping the channel dimension constant across participants; channel-level repair and component-based artifact removal were not implemented.

Each window and frequency band (delta 2–4, theta 4–8, alpha 8–12, beta 12–25, gamma 25–40 Hz) yielded a 129 × 129 symmetric positive definite covariance matrix. Band-pass filtering was applied within single windows after one second of reflection padding to suppress edge transients in short segments. Positive definiteness was assessed with a relative criterion: the smallest eigenvalue divided by the mean diagonal had to exceed -10^{-4}, since float32 storage introduces rounding-scale negative eigenvalues of order 10^{-15}. Low-frequency bands are structurally rank-deficient in 2-s windows, so positive definiteness is guaranteed by a small trace-relative ridge rather than by the data themselves.

4.3 Geometric representation and tangent-space projection

For each subject, band, and metric (Riemannian, log-Euclidean, Euclidean) a Riemannian domain mean was computed over that subject's windows. Tangent-space coordinates were taken relative to a fixed global reference, the domain mean of all subjects, and the upper triangular part of the whitened matrix was used, giving D = 8,385 features per band [3,6]. Coordinates for all subjects were stacked and projected onto the leading eigenvectors of the pooled within-band scatter matrix, retained at a cumulative explained variance of 0.999 and shared across participants as a fixed feature extractor. The retained dimensionality was k = 2,309 for Riemannian and k = 45 for log-Euclidean features, giving 11,545 and 225 features respectively.

Because the tangent-space origin is a global reference computed across all releases, release-specific drift is retained in the features under the unadjusted condition rather than removed; this is the object of the LORO evaluation and not a defect. Euclidean features were excluded from the reported analyses: their standardized condition number is 1.3×10^7 (versus 4.1×10^2 for log-Euclidean), and ridge regression extrapolated catastrophically, with R^2 ranging from −1.2 to −28.6. All reported results therefore concern the Riemannian and log-Euclidean representations.

4.4 Regression

For each held-out release r, the training set T comprised the remaining nine releases, the target-domain auxiliary set A was an unlabelled half of the held-out release, and the evaluation set E was the remainder. The predictor was ridge regression with an intercept, on features standardized using training-set statistics only, with the penalty selected by generalized cross-validation over a logarithmic grid from 10^{-4} to 10^6. Standardization is a requirement rather than a convenience: the raw scale of tangent-space features differs by ten orders of magnitude across metrics, and without it the small-scale log-Euclidean features were compressed to a constant with

289 a ratio of predicted to observed standard deviation of 5×10^{-13}. Pearson correlation would still
290 have returned a number in that regime, because it is scale invariant, which is precisely the failure
291 mode that standardization prevents.

292 ### *4.5 Evaluation protocols*

293 The main analysis used LORO, holding out one release at a time and training on all others
294 (**Fig. 1d**); training-set size ranged from 2,024 (phenotype tasks) to 2,394 (BAG) and evaluation-
295 set size from 64 (phenotype tasks) to 236 (BAG) participants (median 118 for BAG). The same
296 split was shared across all domain-handling conditions so that paired differences are computed
297 within release. Two benchmark protocols accompanied it: pooled ten-fold subject-level cross-
298 validation ignoring release membership, which upper-bounds performance in the absence of
299 release shift, and within-release five-fold cross-validation, which serves as a small-sample
300 within-release reference. Five domain-handling conditions were evaluated (**Fig. 1d**): no
301 adjustment; re-centering on the unlabelled auxiliary mean; a tangent-space re-centering-plus-
302 orthogonal-Procrustes alignment(rotation-based Procrustes alignment, RPA) estimated from
303 unlabelled target-domain features; a simplified Riemannian-ComBat-style moment correction
304 inspired by established ComBat methodology [12]; and transductive centering. A geodesically
305 optimized intercept-learning alignment was not implemented and is therefore reported as such
306 rather than compared [7]. The scored dataset comprises 1,100 rows: 500 LORO, 100 pooled, and
307 500 within-release, all with successful status.

308 Brain age gap (BAG) was operationalized as regression on age itself, so no ground-truth
309 leakage occurs. The four phenotype tasks used the released phenotype scores as targets. Only the
310 primary task and its within-release benchmark are shown in the main figures; the remaining
311 figures are provided in the Supplementary Information.

4.6 Why two domain steps are algebraically inert

Under LORO the target-domain center cannot be estimated from labeled evaluation data, so re-centering uses only the unlabelled auxiliary mean μ_A. The evaluation features are then shifted by a single constant vector $c = -\mu_A$ that is identical for every evaluation subject, so predictions shift by a subject-independent constant $c_0 = \hat{w}^T c$ and, because Pearson correlation is invariant under addition of a constant,

$$r' = \mathrm{corr}(\hat{y} + c_0, y) = r \;\Rightarrow\; \delta r \equiv 0 \qquad (1)$$

The coefficient of determination has no such invariance: with $SSE = \Sigma_e (y_e - \hat{y}_e)^2$, the residual sum of squares changes as $SSE' = SSE - 2c_0 \Sigma_e (y_e - \hat{y}_e) + n_{te} \cdot c_0^2$, which is generally nonzero because residuals from a training-fitted intercept do not sum to zero. Re-centering therefore has $\delta r \equiv 0$ exactly while δR^2 is a small nonzero quantity, as the data confirm (**Fig. 3**). Transductive centering applies the *same* shift to training and evaluation features, which the ridge model with intercept absorbs entirely, so $\delta R^2 = \delta r = 0$ to machine precision.

4.7 Effect sizes, interval estimation, and equivalence

Effects of an alignment condition relative to no adjustment were computed as paired differences within metric, task, and held-out release, and their means across releases were accompanied by bootstrap 95% confidence intervals from 2,000 resamples. Correlations were transformed with the Fisher z-transform. To make the null phenotype results interpretable, we computed the minimum detectable effect (MDE) of a single-release correlation test at 80% power for a two-sided $\alpha = 0.05$, using the mean evaluation-set size across phenotype tasks and conditions ($\bar{n}_{te} = 124.8$):

$$r_{MDE} = \tanh((1.96 + 0.84) / \sqrt{(\bar{n}_{te} - 3)}) = \tanh(2.80 / \sqrt{121.8}) \approx 0.248 \qquad (2)$$

Whether alignment changed the already near-zero phenotype associations was assessed by two one-sided tests (TOST) on the paired Fisher-z increments $dz_j = z_{a,j} - z_{none,j}$ within the same task and held-out release, pooled over 40 pairs per metric × condition cell, with a 90% confidence interval for the mean increment and an equivalence margin of $\Delta = \operatorname{arctanh}(0.02)$ in z units. A cell was declared equivalent when the entire interval lay inside the margin.

4.8 Reproducibility

All results derive from a single released run (n = 2,525) with a fixed seed. Riemannian operations were computed with pyRiemann version 0.12 [13]; this version is locked because a later release changed the tangent-space rotation implementation, which would make earlier results non-comparable. The evaluation was re-run end to end on an independent Linux workstation from the harmonized intermediate products; all four output tables reproduced bit-identically, as did the reported TOST count and detectability bound. Code is provided as Supplementary Information.

4.9 Ethics statement

This study analysed only de-identified resting-state EEG data that were collected by the Child Mind Institute and publicly released under CC-BY-SA 4.0 as part of the Healthy Brain Network (HBN)–EEG dataset (OpenNeuro accessions ds005505–ds005512, ds005514 and ds005515). Because the present work involved no new recruitment, no new collection of data from human participants and no intervention, and because it used only already-de-identified public data, ethical review approval and informed consent were not required for this analysis.

4.10 Use of generative AI

During the preparation of this manuscript the authors used DeepSeek V4.1 and MiniMax M3.1 for literature retrieval and language polishing. The tools did not perform data analysis or generate scientific results. All scientific content, analyses, interpretations and conclusions are the authors' own and were reviewed and verified by the authors. The tools are not listed as authors and do not satisfy authorship criteria.

Data availability

The HBN–EEG dataset is publicly available on OpenNeuro (accessions ds005505–ds005512, ds005514 and ds005515). All figures and reported tables derive from a single frozen evaluation run (n = 2,525). The code and derived products needed to reproduce every number, figure and table in this manuscript are provided as Supplementary Information, together with the verification script used to confirm cross-platform reproduction.

References

[1] Smith, S. M., Vidaurre, D., Alfaro-Almagro, F., Nichols, T. E. & Miller, K. L. Estimation of brain age delta from brain imaging. *Neuroimage* **200**, 528–539 (2019). https://doi.org/10.1016/j.neuroimage.2019.06.017

[2] de Lange, A.-M. G. et al. Mind the gap: performance metric evaluation in brain-age prediction. *Hum. Brain Mapp.* **43**, 3113–3129 (2022). https://doi.org/10.1002/hbm.25837

[3] Engemann, D. A. et al. A reusable benchmark of brain-age prediction from M/EEG resting-state signals. *Neuroimage* **262**, 119521 (2022). https://doi.org/10.1016/j.neuroimage.2022.119521

[4] Tveitstøl, T. et al. Assessing the robustness of deep learning based brain age prediction models across multiple EEG datasets. *IEEE Trans. Biomed. Eng.* **73**, 2582–2601 (2026). https://doi.org/10.1109/TBME.2025.3639477

[5] Ju, C., Collas, A., Bouchard, F. & Thirion, B. Benchmarking external generalization of SPD matrix learning for resting-state fMRI connectome prediction. Preprint at https://arxiv.org/abs/2608.30418 (2026).

[6] Mellot, A., Collas, A., Rodrigues, P. L. C., Engemann, D. & Gramfort, A. Harmonizing and aligning M/EEG datasets with covariance-based techniques to enhance predictive regression modeling. *Imaging Neurosci.* **1** (2023). https://doi.org/10.1162/imag_a_00040

[7] Mellot, A., Collas, A., Chevallier, S., Gramfort, A. & Engemann, D. A. Geodesic optimization for predictive shift adaptation on EEG data. Preprint at https://arxiv.org/abs/2407.03878 (2024).

[8] Eder, M., Xu, J. & Grosse-Wentrup, M. Benchmarking brain–computer interface algorithms: Riemannian approaches vs convolutional neural networks. *J. Neural Eng.* **21**, 044002 (2024). https://doi.org/10.1088/1741-2552/ad6793

[9] de Surrel, T., Yger, F., Lotte, F. & Chevallier, S. A probabilistic view on Riemannian machine learning models for SPD matrices. Preprint at https://arxiv.org/abs/2505.02402 (2025).

[10] Child Mind Institute. Healthy Brain Network (HBN)–EEG: multi-release resting-state EEG. OpenNeuro, accessions ds005505–ds005515 (2026). https://openneuro.org/datasets/ds005505

[11] Bethlehem, R. A. I. et al. Brain charts for the human lifespan. *Nature* **604**, 525–533 (2022). https://doi.org/10.1038/s41586-022-04554-y

[12] Fortin, J.-P. et al. Harmonization of multi-site diffusion tensor imaging data. *Neuroimage* **161**, 149–170 (2017). https://doi.org/10.1016/j.neuroimage.2017.08.047

[13] pyRiemann developers. pyRiemann version 0.12. Software release (2025).

[14] Bzdok, D., Varoquaux, G. & Thirion, B. Neuroimaging research: from null-hypothesis falsification to out-of-sample generalization. *Educ. Psychol. Meas.* **77**, 868–880 (2017). https://doi.org/10.1177/0013164416667982

[15] Zhang, B., Zhang, S., Feng, J. & Zhang, S. Age-level bias correction in brain age prediction. *Neuroimage Clin.* **37**, 103319 (2023). https://doi.org/10.1016/j.nicl.2023.103319

Acknowledgements

The authors thank the Child Mind Institute for releasing the Healthy Brain Network (HBN)–EEG data under CC-BY-SA 4.0, and the participants and their families who contributed to the study.

Funding

This study was supported in part by the National Natural Science Foundation of China (22105104), the Natural Science Fund for Colleges and Universities in Jiangsu Province (24KJB430030), and the Natural Science Foundation of Nanjing University of Posts and Telecommunications (NY224151).

Author contributions

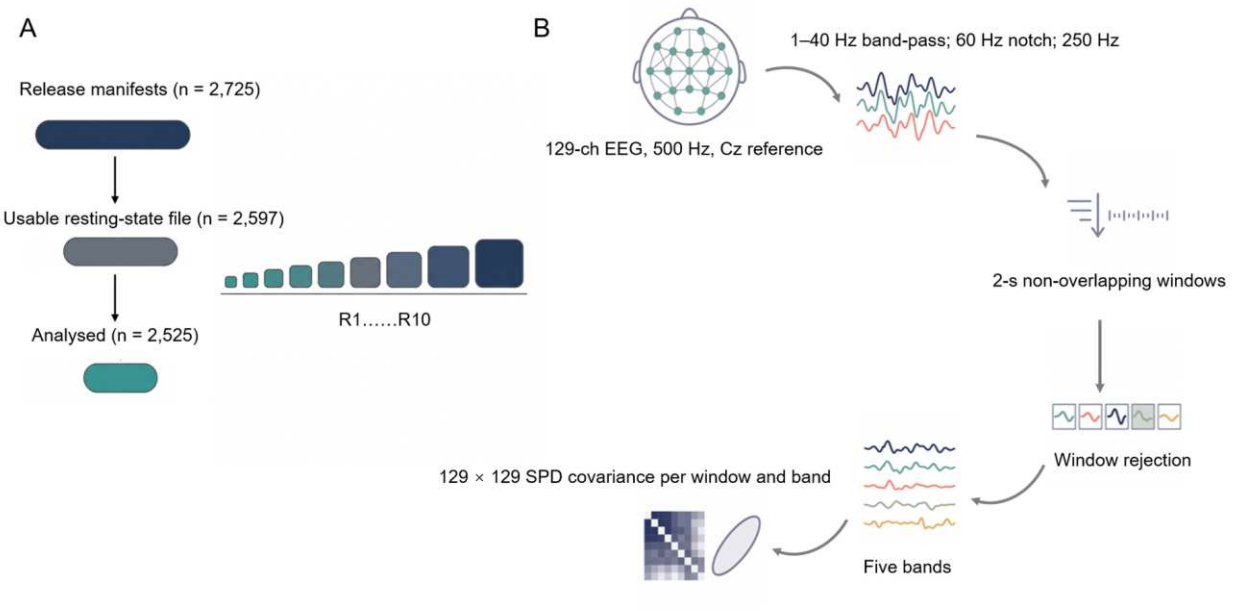
X.L. and L.Z. conceived the study. X.L. processed the ten EEG releases, ran the LORO evaluation, and wrote the first draft. H.T. contributed to the statistical design and checked the paired and equivalence analyses. X.Y., Y.L. and H.L. curated the releases and implemented preprocessing, quality control and covariance feature extraction. L.L. validated the regression

results. H.M. prepared the figures. Q.Z. reviewed the literature and edited the manuscript. G.X. and L.Z. supervised the project, checked the analysis and revised the manuscript. All authors read and approved the final manuscript.

Competing interests

The author(s) declare no competing interests.

Figure legends



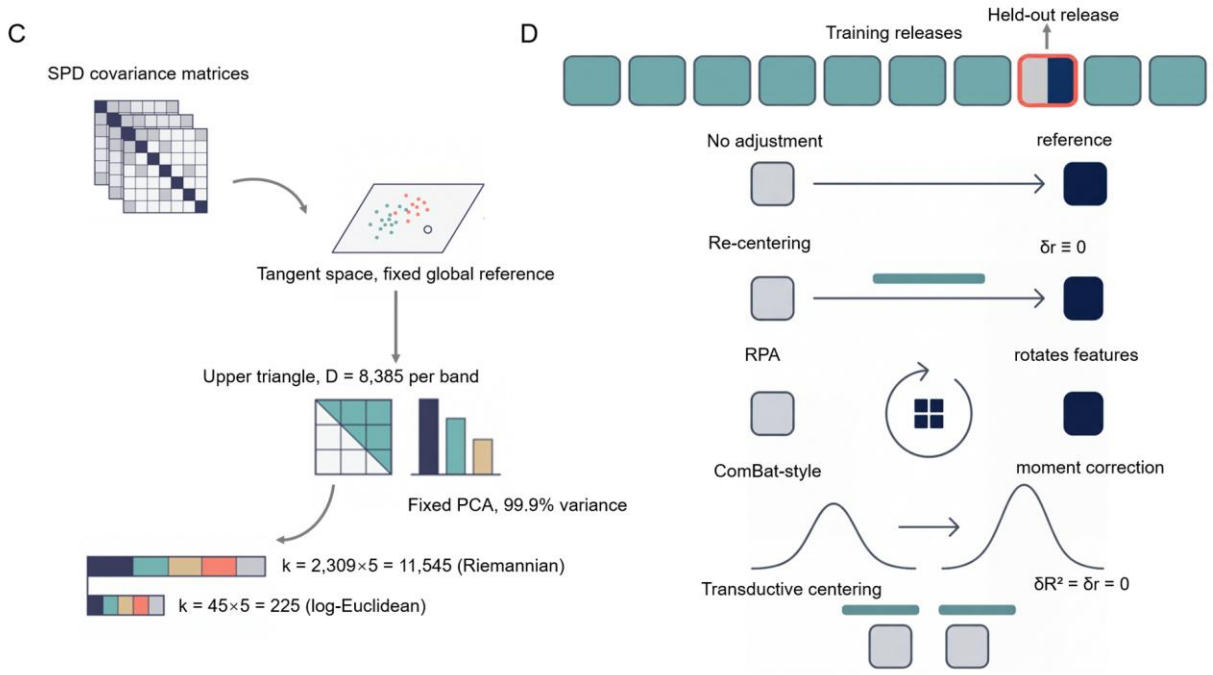


Figure 1. Analysis pipeline and leave-one-release-out evaluation design. (A) Cohort assembly from the ten sequential HBN-EEG releases (R1–R10). Of 2,725 participants listed in the release manifests, 2,597 had a usable resting-state file; 72 of these failed preprocessing or quality control, leaving 2,525 for analysis (R1 135, R2 149, R3 178, R4 314, R5 319, R6 131, R7 358, R8 211, R9 259, R10 471). Block size is proportional to the number of analysed participants. The released metadata contain no collection-site field, so release membership is the only observable batch variable. (B) Preprocessing and covariance features. Each recording was band-pass filtered (1–40 Hz), notch filtered (60 Hz), resampled to 250 Hz, and cut into non-overlapping 2-s windows. Windows whose across-channel median peak amplitude exceeded 100 μV were rejected. Each surviving window and frequency band (δ 2–4, θ 4–8, α 8–12, β 12–25, γ 25–40 Hz) produced a 129 × 129 symmetric positive-definite covariance matrix. Bad channels were retained so that the covariance dimension is the same for every participant. (C) Tangent-space representation. For each subject and band, covariance matrices were projected into the tangent

441 space at a single fixed reference, the domain mean of all subjects, and the upper triangle of the
442 whitened matrix was kept (D = 8,385 per band). A fixed PCA basis reduced each band to k =
443 2,309 (Riemannian) or k = 45 (log-Euclidean) components, and concatenating five bands gave
444 11,545 or 225 features. The same 99.9% cumulative-variance threshold was used for both
445 representations; the retained k differs because the two spectra decay at different rates, not
446 because of a tuning choice. Because the reference is estimated once across all releases, release-
447 specific drift remains in the features under no adjustment, which is what the LORO evaluation
448 measures. The reference and basis are label-free constants and were not re-estimated inside any
449 fold. (D) LORO split and domain-handling conditions. For each held-out release, the other nine
450 releases formed the training set, and the held-out release was divided into an unlabelled auxiliary
451 half and an evaluation half. The auxiliary half supplies target-domain statistics only; its labels
452 were never used. Five conditions were compared: no adjustment (reference); re-centering on the
453 auxiliary mean; re-centering plus an orthogonal Procrustes alignment (RPA) estimated from the
454 unlabelled auxiliary and evaluation features; a ComBat-style moment correction using auxiliary
455 means and variances; and transductive centering, which applies one shift to training and
456 evaluation features. All conditions share the same split, so differences are paired within release,
457 and evaluation labels never entered any fit. Re-centering translates every evaluation prediction
458 by a subject-independent constant, so Pearson r is exactly unchanged ($\delta r \equiv 0$) while δR^2 is small
459 and generally nonzero. RPA additionally applies an orthogonal map that mixes feature columns,
460 so the training-fitted weights are not invariant and its effect must be measured empirically; its
461 sign depends on the tangent-space representation. Transductive centering is absorbed exactly by
462 the unpenalized intercept, giving $\delta R^2 = \delta r = 0$ to machine precision.

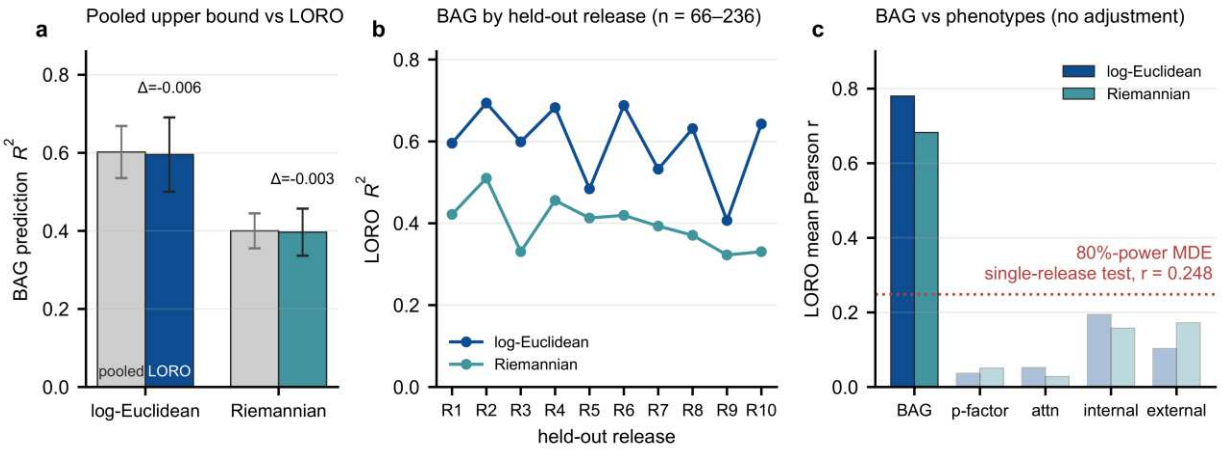


Figure 2. Cross-release generalization of brain-age prediction. (a) Pooled ten-fold upper bound and the LORO unadjusted condition, with bars showing means across folds or across held-out releases and error bars the standard deviation; Δ denotes the LORO-minus-pooled difference (−0.006 log-Euclidean, −0.003 Riemannian). (b) LORO R^2 for each held-out release, positive in 10 of 10 releases for both representations (evaluation n = 64–236). (c) Mean LORO Pearson r for the five targets; the four dimensional phenotypes satisfy $|r| \leq 0.20$ (maximum 0.194) and all lie below the 80%-power minimum detectable effect of a single-release correlation test (dashed red line, r = 0.248). Axis abbreviations: attn, attention; internal, internalizing; external, externalizing; p-factor, general psychopathology factor.

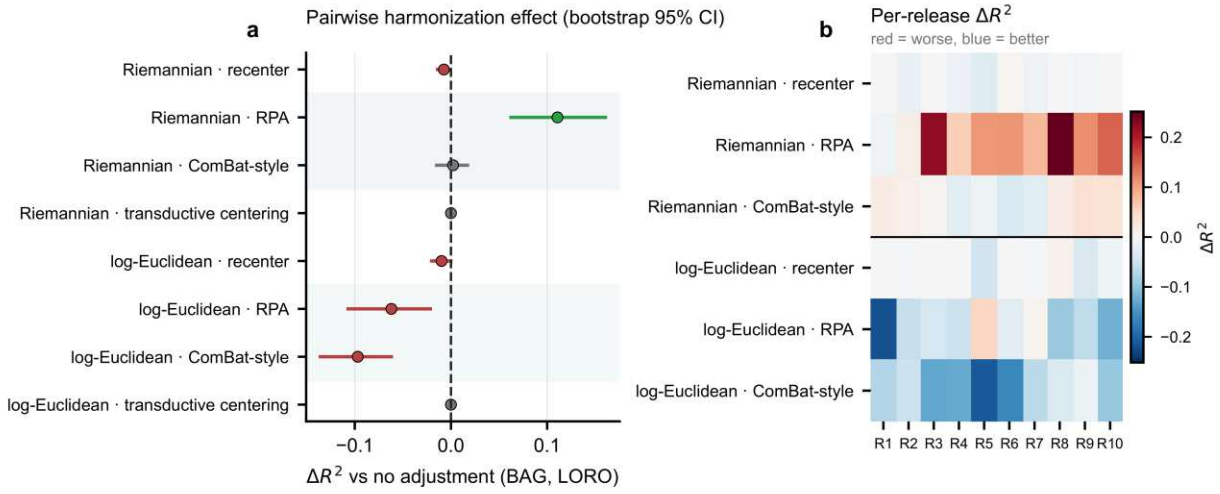


Figure 3. Paired effect of domain handling on brain-age prediction. (a) Mean δR^2 relative to no adjustment with bootstrap 95% confidence intervals, under Riemannian (left) and log-Euclidean (right) features; no condition improves robustly across both representations. (b) Per-release δR^2 for each condition; the positive Riemannian gain of the rotation-based alignment concentrates in two releases, and the transductive centering is zero to machine precision; re-centering leaves Pearson r unchanged while ΔR^2 is small and nonzero. RPA：rotation-based Procrustes alignment.

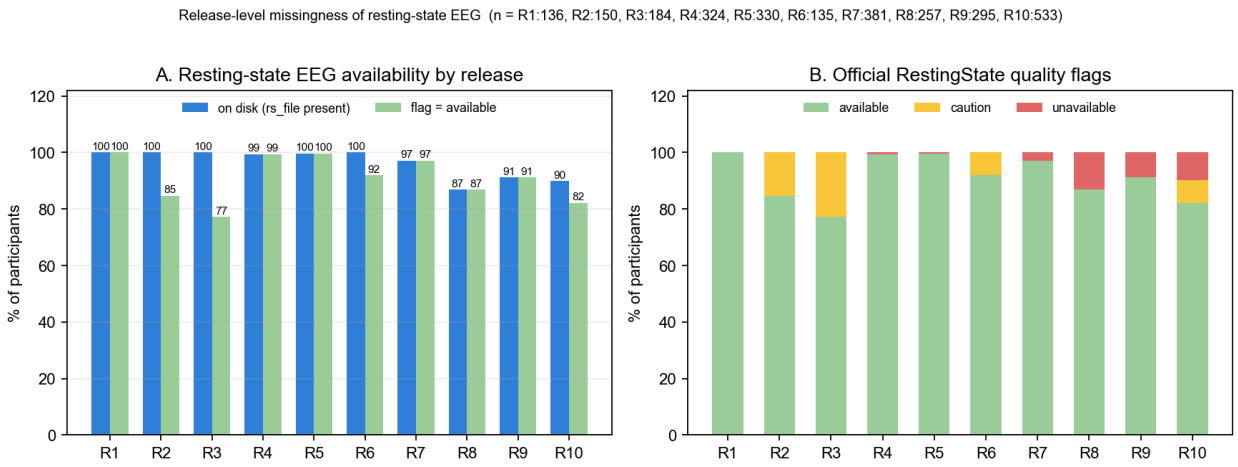
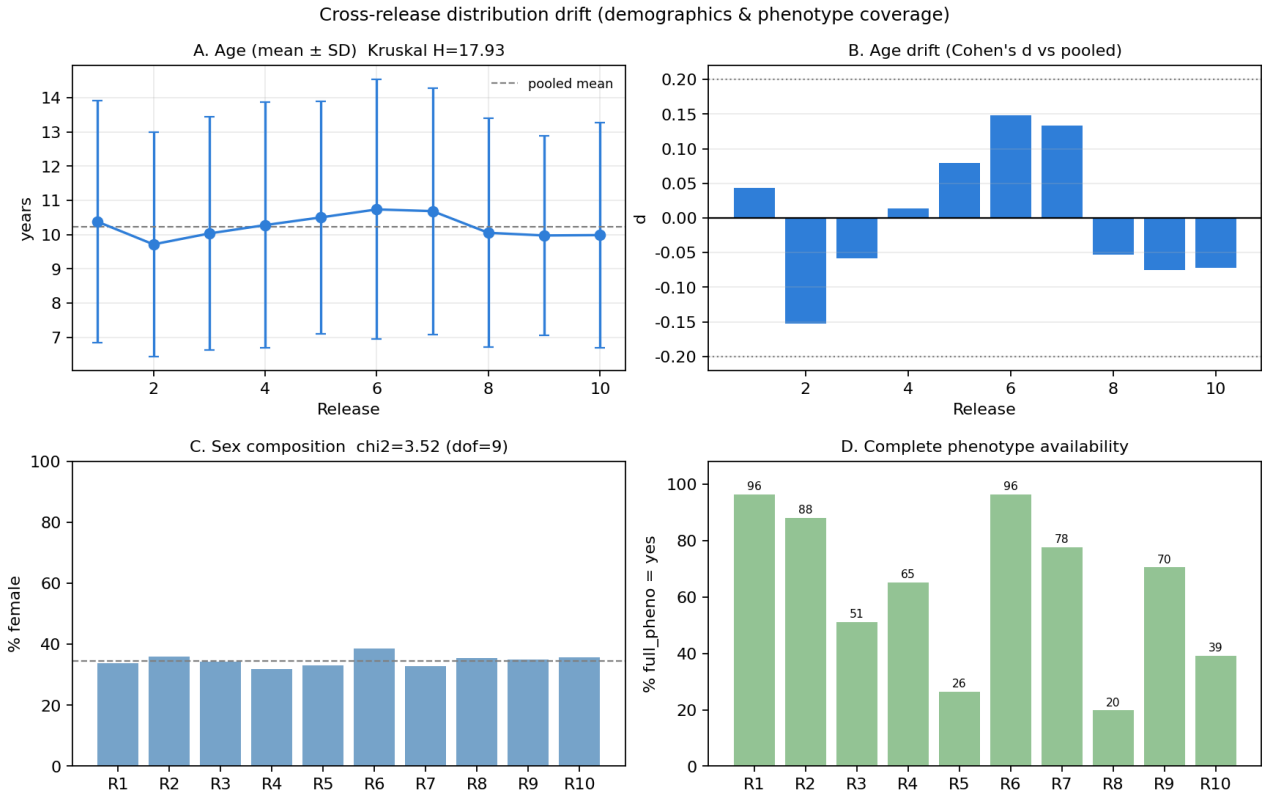


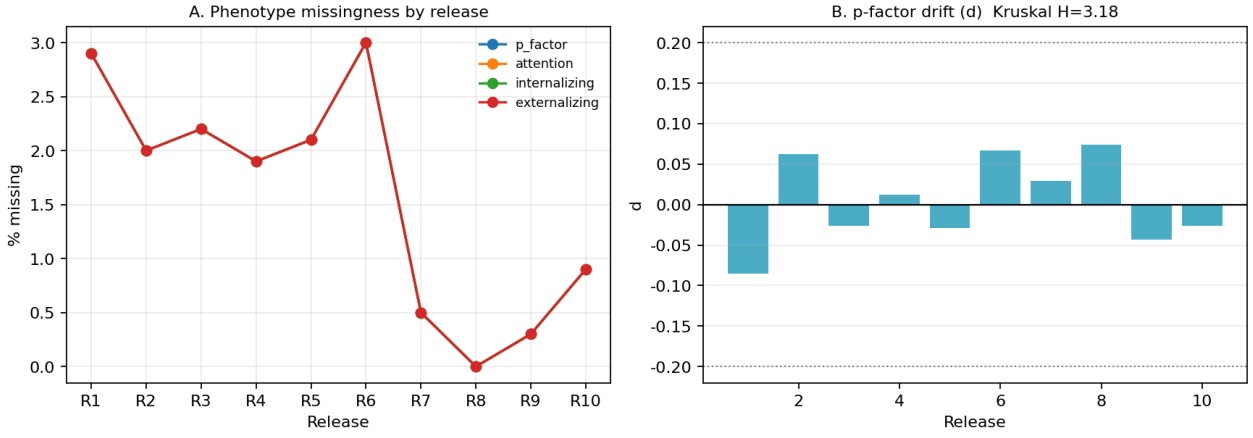
Figure 4. Data availability across the ten releases. Panels report the on-disk availability of resting-state EEG per release together with the composition of released quality flags. Counts are participants listed in the release manifests before quality control; all predictive analyses use the final analysed sample of n = 2,525.



485

486 **Figure 5.** Demographic drift across releases. Age, sex ratio, and availability of complete

487 phenotype panels, with the corresponding omnibus test statistics.



488

489 **Figure 6.** Phenotype missingness across releases. Missingness rates for each phenotype and

490 release-specific drift in the general psychopathology p-factor.

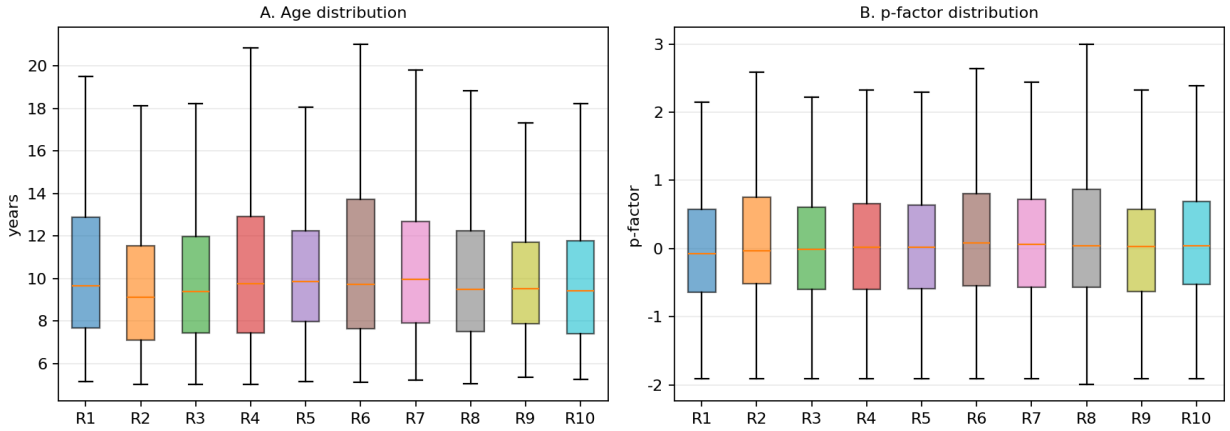


Figure 7. Per-release distributions of subject-level age and p-factor (box plots; outliers not drawn).

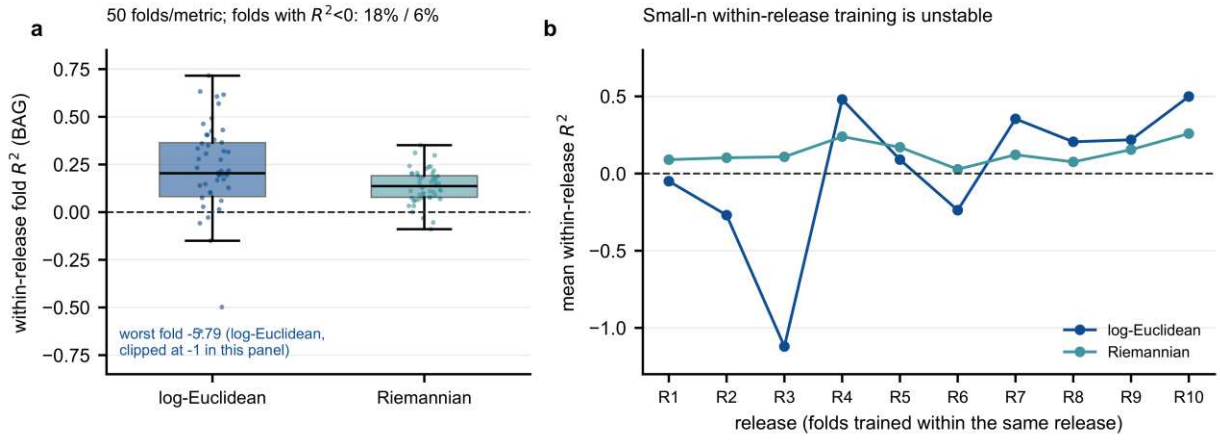


Figure 8. Stability of within-release five-fold cross-validation, shown as a reference baseline rather than a main result. Per-release and per-fold R^2 for brain-age prediction under both representations; the log-Euclidean representation shows large fold-to-fold variance and occasional extrapolation blow-up, whereas LORO (**Fig. 2)** trains on all remaining releases.

Supplementary Files

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

- SupportingInformation.docx