1 Supplementary Material

2 Contents

- 3 I. Demographic imbalances are common across sites in consortia and multisite samples
- 4 II. Results including ComBat without covariate preservation
- 5 III. The magnitude of sex's effect on variance depends on the type of brain feature measured
- 6 IV. Replication without Extreme Centiles
- 7 V. Z-score analyses
 - A. Main Analyses
 - B. Varying M:F ratios
 - C. Replication without Extreme Z-scores
- 11 VI. Comparisons of ComBatLS and ComBat-GAM for harmonizing consortium data
 - A. Main Analyses
 - B. Validating generalizability to other normative models
- 14 VII. References

16 Section I. Demographic imbalances are common across sites in consortia and multisite samples

Like all ComBat methods, ComBatLS is designed to retain biological information that may be confounded by site effects. While some multi-site studies ensure that demographics are relatively uniform across sites, substantial variability in subject-level factors like sex and age are common (SFigs 1 and 2). These imbalances become even more prevalent when compiling datasets to create large, representative consortium samples, as is often done for normative modeling (SFig 3). Thus, is it cruical that harmonization is able to disambiguate and preserve covariate effects from nuisance site effects, both in brain feature distributions' locations (all ComBat methods) and scales (ComBatLS).

24

8

9

10

12

13



25

26 Supplemental Figure 1) In many multisite studies, the sex-ratios of each site's sample varies

considerably. Bar plots indicate the percentage of each site's sample reported as Female or Male across six
 multisite studies included in the LBCC. Though the sex-ratios remain fairly stable in (A) the Adolescent Brain

29 Cognitive Development (ABCD) and (B) UK Biobank, (C-F) the remaining studies's sex-ratios vary

30 substantially across sites. Bars are labeled to show the count of males and females, respectively, in each site.



Supplemental Figure 2) Sample's ages are often inconsistent across studies' sites. Box and whisker
plots indicating the age (years post-birth) of each site's sample across six multisite studies in the LBCC. (A and
B) While some studies' sites have very similar age distributions, (C-F) the subjects' age distributions are often
vary greatly from one site to another.



38

Supplemental Figure 3) Consortium samples are particularly likely to show site-level variability in
 sample demographics. A) Sample sex-ratios by site and study. B) Age distribution of each site's sample

40 sample demographics. A) Sample sex-ratios by site and study. B) Age distribution of each site's sample,
41 grouped by study.

42

43 Section II. Results across 100 sampling replications including ComBat without Covariate Preservation

44 We included an application of ComBat in which no covariate effects are preserved to serve as a benchmark

- 45 harmonization method against which we could assess the effects of increasingly complex covariate
- 46 preservation: linear ComBat, ComBat-GAM, and ComBatLS. While tests with this method are controlled for
- 47 throughout the analyses as part of our FDR corrections, we chose not to include results in the main text to

48 facilitate easy comparison of our primary methods of interest. Results from all four methods are presented

49

below.

50

Pairwise comparisons of ComBat methods' absolute centile errors for 208 brain features across 100 replications		
Method Producing Smaller Absolute Errors	N Features	% Features
ComBat-GAM vs ComBat w/o Covariates		
ComBat w/o Covariates	190	0.91%
ComBat-GAM	20605	99.06%
ComBatLS vs ComBat w/o Covariates		
ComBat w/o Covariates	155	0.75%
ComBatLS	20638	99.22%
ComBatLS vs ComBat-GAM		
ComBat-GAM	1503	7.23%
ComBatLS	19115	91.90%
Linear ComBat vs ComBat w/o Covariates		
ComBat w/o Covariates	190	0.91%
Linear ComBat	20604	99.06%
Linear ComBat vs ComBat-GAM		
ComBat-GAM	8093	38.91%
Linear ComBat	9918	47.68%
Linear ComBat vs ComBatLS		
ComBatLS	19146	92.05%
Linear ComBat	1466	7.05%

51 Supplemental Table 1) Pairwise tests of absolute centile errors within each of 208 brain features replicated
 53 across 100 subject resamplings. All tests were conducted as pairwise, two-tailed t-tests of ranks with Welch's
 54 correction. FDR-corrected across 1248 tests (208 features x 6 ComBat method pairings) within each sampling

55 permutation.



57

58 Supplemental Figure 4. Absolute centile errors across brain features and ComBat methods. Violin plots

59 of absolute centile errors across 208 brain features.



61

62 Supplemental Figure 5. ComBatLS recapitulates true centile scores more accurately than other

63 ComBat methods. A) Absolute centile errors within each brain feature compared pairwise between four
 64 ComBat methods, replicated across 100 sampling permutations. Fill indicates the ComBat method that

65 produces significantly smaller absolute centile errors, FDR-corrected across 1248 tests (208 features x 6

66 ComBat method pairings) within each permutation. B) Proportion of sampling replications in which ComBatLS

67 produces significantly smaller absolute centile errors for a cortical thickness feature than an alternative

68 ComBat method. Significant differences were assessed using pairwise t-tests between ComBat methods,

69 FDR-corrected. Abbrv: CT, cortical thickness; SA, surface area.



Supplemental Figure 6. Subjects' mean absolute centile error across ComBat configurations and 100

sampling replications. Violin plots of absolute centile error for 208 features averaged within subject. Fill corresponds to sampling replication.



Supplemental Figure 7. Brain features with significant sex effects in scale. Features in gray are those for
 which sex does not significantly impact the second moment of a gamlss brain chart. Fill represents the
 difference in males' and females' predicted variance at the sample's mean age (64.94 years), standardized by

79 dividing by females' predicted variance. Positive effects indicating that males' variance is higher than females'.



81 Supplemental Figure 8. Significant differences in males' and females' median centile errors across

brain features and ComBat methods. Positive centile errors (green) indicate that males' centiles tend to be
 overestimated relative to females'. Abbrv: CT, cortical thickness; SA, surface area. Results did not change
 when correcting for multiple comparisons.



87 Supplemental Figure 9. Differences between males' and females' median centile errors in 208 brain

features across 100 sampling replications. Feature counts are plotted by brain tissue types. Fill represents
 sampling replication. Abbrv: CT, cortical thickness; SA, surface area.



92

93 Supplemental Figure 10. Density plots of median sex differences in centile errors induced by different

94 ComBat methods within phenotype categories across 100 replications. Data shows medians of the
 95 distributions of each replication plotted in Supplemental Figure 5. Abbrv: CT, cortical thickness; SA, surface

96 area; ***, p < 0.001; **, p < 0.01. Results did not change when correcting for multiple comparisons.



98 99 Supplemental Figure 11. ComBat methods may lead females to be slightly over- or under-represented among individuals with extreme phenotypes. Bias in the proportion of females with low (<20th percentile) or 100 high (>80th) mean centiles across 100 sampling replications. Positive values indicate a higher proportion of 101 102 females than "true" mean centiles calculated from unharmonized data (dashed line). Abbrv: ***, p < 0.001; **, p 103 < 0.01.



105

106 Supplemental Figure 12. Magnitude centile errors compared pairwise between ComBat methods

across varying levels of sex-imbalances in simulated sites. Fill indicates ComBat method with significantly
 lower absolute centile errors for a given feature, FDR-corrected across pairwise combinations, brain features,
 and 11 samplings. Abbrv: CT, cortical thickness; SA, surface area.



111 Supplemental Figure 13. Sex-biases in centile errors induced by various ComBat methods across

varying degrees of sex-imbalance. A) Illustrative example of differences in males' and females' distributions 112 of centile errors across global brain features when harmonized with ComBat-GAM and ComBatLS. B) Points 113 show differences in males' and females' median centile errors in each brain feature. Boxplots show median 114 115 male - median female centile errors across these features when centiles are derived from data harmonized by different ComBat methods. ComBat without covariate preservation induces strong biases wherein males' 116 centiles are underestimated relative to females', particularly as simulated sites become more imbalanced for 117 sex. Results for both panels are highly similar when applying FDR correction across 11 M:F ratios. Abbrv: ***, 118 119 p < 0.001; **, p < 0.01.

- 120
- 121

122 Section III. The magnitude of sex's effect on variance depends on the type of brain feature measured

- 123 Motivated by prior literature, we used UKB data to assess how sex's effects on scale vary by brain feature
- type. Sex's effects in variance were calculated from brain charts and standardized across features (see
- 125 Methods). We performed a nonparametric Kruskal-Wallis test to establish that the distributions of standardized
- sex effects varied across brain feature types: global tissue volumes (4 features), regional cortical thickness (68
- 127 features), regional surface area (68 features), and regional volume (68 features). We then conducted pairwise
- 128 Wilcox tests (FDR-corrected) which show that cortical thickness features's scales are significantly less
- impacted by sex than cortical surface area or cortical regional volume features. Sex-effects on cortical
- thickness features' scales are not significantly smaller than those of global volumes, though this analysis is
- 131 limited by small number of global volume features.



138 Section IV. Replication without Extreme Centiles

- 139 To determine whether a small number of subjects with very high or very low centile scores drove differences
- 140 across ComBat methods, we repeated our statistical comparisons after removing subjects with "extreme"
- 141 centiles in a given feature. We defined extreme centiles as those >95% or <5% when calculated from raw,
- 142 "unharmonized" data. First, we identified and removed extreme centiles across our 100 replications. We then
- compared the remaining centile errors across ComBat methods, using two-tailed t-tests of centile error ranks
- 144 with Welch's correction, controlling FDR for 1248 comparisons within each replication. We also repeated our
- assessments of sex-differences in centile errors and ComBat-induced sex biases in centile displacement after
- taking the median of sex-differences within feature categories. Second, we applied these procedures to our 11
- samples of synthetic sites with varying Male:Female ratios to assess whether extreme subjects drove
- differences in ComBat methods' performance across degrees of sex imbalance.



149

Supplemental Figure 15. Extreme phenotypes do not drive differences in absolute centile errors
 between ComBat methods. Comparison of pairwise tests of absolute centile errors between ComBat

methods when centiles with raw values above 95% or below 5% are excluded. Absolute centile errors were
 compared within each brain featureacross 100 sampling permutations. Fill indicates the ComBat method that

produces significantly smaller absolute centile errors, FDR-corrected across 1248 tests (208 features x 6

155 ComBat method pairings) within each permutation.



157 158 Supplemental Figure 16. Significant differences in males' and females' median centile errors across brain features and ComBat methods when extreme features are excluded. Positive centile errors (green) 159 160 indicate that males' centiles tend to be overestimated relative to females'. Results do not change when

161 correcting for multiple comparisons. Abbrv: CT, cortical thickness; SA, surface area.



Supplemental Figure 17. Density plots of median sex differences in centile errors induced by different
 ComBat methods across 100 replications when extreme phenotypes are excluded. Results are

166 consistent when applying FDR-correction across 4 phenotype classes and 4 harmonization methods. Abbrv:

167 CT, cortical thickness; SA, surface area; ***, p < 0.001; **, p < 0.01.

171 Section V. Z-score analyses

Z-scores for each feature derived from centile scores using R's *qnorm()* function. To prevent infinite z-scores,
centiles of 0 and 1 were estimated as 1e-25 and 0.9999999999999999994, respectively. As with centile scores,
analyses of z-scores were repeated without extreme scores, here defined as z-scores less than -2 or greater
than 2.

176

178

177 A) Main Results

Pairwise comparisons of ComBat methods' absolute Z-score errors for 208 brain features across 100 replications

Method Producing Smaller Absolute Errors	N Features	% Features
ComBat-GAM vs ComBat w/o Covariates		
ComBat w/o Covariates	191	0.92%
ComBat-GAM	20600	99.04%
ComBatLS vs ComBat w/o Covariates		
ComBat w/o Covariates	156	0.75%
ComBatLS	20634	99.20%
ComBatLS vs ComBat-GAM		
ComBat-GAM	1617	7.77%
ComBatLS	18992	91.31%
Linear ComBat vs ComBat w/o Covariates		
ComBat w/o Covariates	189	0.91%
Linear ComBat	20599	99.03%
Linear ComBat vs ComBat-GAM		
ComBat-GAM	7639	36.73%
Linear ComBat	10398	49.99%
Linear ComBat vs ComBatLS		
ComBatLS	19030	91.49%
Linear ComBat	1580	7.60%

179 **Supplemental Table 2**. Pairwise tests of absolute z-score errors within each of 208 brain features replicated

across 100 subject resamplings. All tests conducted as pairwise, two-tailed t-tests of ranks with Welch's

correction. FDR-corrected across 1248 tests (208 features x 6 ComBat method pairings) within each sampling
 permutation.



183

Supplemental Figure 18. Absolute z-score errors across brain features and ComBat methods. Violin plots of absolute z-score errors across 208 brain features. Abbrv: CT, cortical thickness; SA, surface area.



187

Supplemental Figure 19. Pairwise comparisons of absolute z-score errors across ComBat methods
 within each brain feature across 100 sampling replications. Fill indicates the ComBat method that
 produces significantly smaller absolute z-score errors, FDR-corrected across 1248 tests (208 features x 6
 ComBat method pairings) within each permutation.





194

Supplemental Figure 20. Density plots of median sex differences in z-score errors induced by different
 ComBat methods within phenotype categories across 100 replications. Results are consistent when
 applying FDR correction across 4 phenotype categories and 4 harmonization methods. Abbrv: CT, cortical
 thickness; SA, surface area; ***, p < 0.001; **, p < 0.01.



Supplemental Figure 21. Significant differences in males' and females' median z-score errors across
 brain features and ComBat methods. Positive centile errors (green) indicate that males' z-scores tend to be
 overestimated relative to females'. Results are consistent when applying FDR correction across regions.
 Abbrv: CT, cortical thickness; SA, surface area.



Supplemental Figure 22. Subjects' mean absolute z-score error across ComBat configurations and 100 sampling replications. Violin plots of absolute z-score error for 208 features averaged within subject. Fill corresponds to sampling replication.



Supplemental Figure 23. Over- or under-representation of females among individuals with extreme zscores across ComBat methods. Bias in the proportion of females with low (<-1.5) or high (>1.5) mean zscores across 100 sampling replications. Positive values indicate a higher proportion of females than "true" mean z-scores calculated from unharmonized data (dashed line). Results are consistent when applying FDR

215 correction. Abbrv: ***, p < 0.001; **, p < 0.01.



lower absolute z-score errors for a given feature, FDR-corrected across simulated sex ratios. Abbrv: CT,
 cortical thickness; SA, surface area.



224

% Females in Imbalanced Site

225 Supplemental Figure 25. Sex-biases in z-score errors induced by various ComBat methods across 226 varying degrees of sex-imbalance. Points show brain features with significant differences in the distributions 227 of males' and females' z-score errors (FDR corrected). Boxplots show median male - median female z-score 228 errors across these features when centiles are derived from data harmonized by different ComBat methods. 229 ComBat without covariate preservation induces strong biases wherein males' z-scores are underestimated 230 relative to females', particularly as simulated sites become more imbalance for sex. Results are highly similar 231 when applying FDR correction for 11 M:F ratios. Abbrv: ***, p < 0.001; **, p < 0.01.

232

233 **C) Without Extreme Z-scores**



234

Supplemental Figure 26. Extreme phenotypes do not drive differences in absolute z-score errors between ComBat methods. Comparison of pairwise tests of absolute z-score errors between ComBat methods when z-scores with raw values above 2 or below -2 are excluded. Absolute z-score errors were compared within each brain feature across 100 sampling permutations. Fill indicates the ComBat method that produces significantly smaller absolute errors, FDR-corrected across 1248 tests (208 features x 6 ComBat method pairings) within each permutation.



Supplemental Figure 27. Significant differences in males' and females' median z-score errors across
 brain features and ComBat methods when extreme features are excluded. Positive z-score errors (green)

indicate that males' z-score tend to be overestimated relative to females'. Results are consistent when applying
 FDR correction across features. Abbrv: CT, cortical thickness; SA, surface area.



Supplemental Figure 28. Density plots of median sex differences in z-score errors induced by different

ComBat methods across 100 replications when extreme phenotypes are excluded. Results are

consistent when applying FDR correction across 4 phenotype classes and 4 harmonization methods. Abbrv:

250 CT, cortical thickness; SA, surface area; ***, p < 0.001; **, p < 0.01.

251

252

Section VI. Comparisons of ComBatLS and ComBat-GAM for harmonizing consortium data A) Main Results



Supplemental Figure 29. Residual effects of study following harmonization with ComBatLS or ComBat-GAM relative to unharmonized data. Effect size for study in each brain feature's gamlss growth chart after

harmonization. Abbrv: CT, cortical thickness; SA, surface area.



Supplemental Figure 30. Absolute differences in ComBatLS and ComBat-GAM-derived centiles are

265 related to age's distribution across batches. A) The mean age of a study's subjects is significantly 266 associated with the magnitude difference in their centile scores, such that centiles for individuals from older 267 studies tend to be more impacted by the choice of harmonization method (Beta=0.0056 centiles, p < 0.001), B) Absolute differences in centile scores are also associated with the range of ages included in a batch's sample. 268 269 with subjects from studies with broader age ranges being more impacted by harmonization method (Beta=0.0063 centiles, p < 0.001). C) Finally, how greatly any given individual's centiles differ when using 270 271 ComBatLS or ComBat-GAM depends on the magnitude of difference between their age and that of their study sample's mean, with subjects who are less well-described by their study's mean age being impacted most 272 273 greatly (Beta=0.031, p < 0.001). Y-axes shows mean absolute difference in a subject's centile scores across brain features when derived from ComBatLS- or ComBat-GAM-harmonized data. Each point represents one 274 275 individual while fill represents primary study. All trend lines indicate marginal effects for study's sample size. Gray bands (not visible) represents 95% confidence interval of marginal association. 276

277

278 B) Validating generalizability to other normative models

279

We designed ComBatLS with normative modeling in mind, as these models and their resulting scores 1) 280 281 depend strongly on their input data accurately representing phenotypes' distributions across a population and 282 require vast enough datasets that some form of harmonization across batches is almost inevitable. 283 However, there are numerous statistical approaches to fitting normative models (see (Borghi et al., 2006; Ge et al., 2024). Thus, we sought to validate that the benefits we observed in ComBatLS on normative scores 284 generalized to other methods for model fitting that also require data be harmonzied. We therefore borrowed the 285 286 approach popularized by (Frangou et al., 2022). As with our primary analyses in the LBCC, we fit these models on ComBat-GAM and ComBatLS-harmonized data, then assessed 1) how well each method was able to 287 harmonize the dataset, as evidenced by low residual site effects in the derived models, and 2) whether 288 differences in ComBat-GAM and ComBatLS-derived normative scores are related to batch demographics. 289 290 which suggests ComBatLS may improve the accuracy of such scores.

291

292 Using the same LBCC data which was harmonized using ComBat-GAM and ComBatLS (see Section 2.3 of the main text), we fit normative models covering ages 3 to 90 years for each brain phenotype using the LMS 293 294 method (Cole & Green, 1992) as in (Dima et al., 2021; Frangou et al., 2022) and implemented at https://centilebrain.org/#/tutorial. Notably, this approach requires that models are fit separately in each sex. 295 296 with the only covariate being a smooth effect of age. As above, we also refit these models with additional 297 'study' term, which we used to estimate residual batch effects in each dataset. As with our main analyses, we also obtained centile scores across every phenotype, which we used to determine changes in whether subjects 298 299 were classified as "extreme" and the average differences in each subjects' ComBat-GAM and ComBatLS-300 derived centiles. Finally, we tested whether subjects' average absolute centile differences were related to their batch's mean age, age range, or the offset from the batch's mean age. 301

302

303 As in our main analyses, we found that ComBatLS and ComBat-GAM mitigated batch effects comparably across all features with small residual study effects in the resultant normative models (Cohen's F-squared: 304 ComBatLS median=0.013, IQR=0.012; ComBat-GAM median=0.012, IQR=0.012; Unharmonized median = 305 0.200. IQR = 0.33; SFig 31A). We again found differences in ComBatLS or ComBat-GAM-derived centiles 306 (mean absolute difference in centile scores = 0.443, range = 0.06 - 7.16 centiles) with 51.7% of subjects 307 having discrepant categorization of extremely high (< 5%) or low (> 95%) centiles in at least one feature 308 (mean=0.954 features per subject, range = 0 - 40 features). Finally, as in our exploratory analyses in the main 309 text, we found that subjects' ComBatLS- and ComBat-GAM-harmonized centile scores varied with the mean 310

age of the study sample (Beta=0.005, p<0.001; **SFig 31B**), the range of ages included in a sample

- (Beta=0.006, p<0.001; SFig 31 C), and how much an individuals' age deviated from that sample's mean age
- (Beta=0.032, p<0.0001; **SFig 31 D**) when controlling for the sample's size.



316 Supplemental Figure 31. ComBatLS's performance generalizes to alternate normative modeling

317 methods. A) Effect size for study in each brain feature's LMS growth chart after harmonization. B) Abbrv: CT,

- cortical thickness; SA, surface area. B-D) Absolute differences in ComBatLS and ComBat-GAM-derived
- centiles vary with age's distribution across batches when controlling for batch size, showing positive
- relationships with (B) studies' mean ages (Beta=0.005, p<0.001), (C) studies' age ranges (Beta=0.006,
- p<0.001), and (D) the offset between a subject and their study's mean age (Beta=0.032, p<0.0001). Each point represents one individual while fill represents primary study. All trend lines indicate marginal effects analyses
- represents one individual while fill represents primary study. All trend lines indicate marginal effects analyses controlling for study sample size. Gray bands (not visible) represent 95% confidence interval of marginal
- 324 association.
- 325
- 326

327 Section VII. References

328

Borghi, E., de Onis, M., Garza, C., Van den Broeck, J., Frongillo, E. A., Grummer-Strawn, L., Van Buuren, S.,

- Pan, H., Molinari, L., Martorell, R., Onyango, A. W., Martines, J. C., Pinol, A., Siyam, A., Victoria, C. G.,
- Bhan, M. K., Araújo, C. L., Lartey, A., Owusu, W. B., ... Heinig, M. J. (2006). Construction of the World
- Health Organization child growth standards: Selection of methods for attained growth curves. *Statistics*
- 333 *in Medicine*, 25(2), 247–265. https://doi.org/10.1002/SIM.2227
- Cole, T. J., & Green, P. J. (1992). Smoothing reference centile curves: The lms method and penalized likelihood. *Statistics in Medicine*, *11*(10), 1305–1319. https://doi.org/10.1002/sim.4780111005
- Dima, D., Modabbernia, A., Papachristou, E., Doucet, G. E., Agartz, I., Aghajani, M., Akudjedu, T. N., Albajes-
- Eizagirre, A., Alnæs, D., Alpert, K. I., Andersson, M., Andreasen, N. C., Andreassen, O. A., Asherson,
- P., Banaschewski, T., Bargallo, N., Baumeister, S., Baur-Streubel, R., Bertolino, A., ... Frangou, S.
- (2021). Subcortical volumes across the lifespan: Data from 18,605 healthy individuals aged 3–90 years.
 Human Brain Mapping, *43*(1), 452–469. https://doi.org/10.1002/hbm.25320
- Frangou, S., Modabbernia, A., Williams, S. C. R., Papachristou, E., Doucet, G. E., Agartz, I., Aghajani, M.,
- Akudjedu, T. N., Albajes-Eizagirre, A., Alnaes, D., Alpert, K. I., Andersson, M., Andreasen, N. C.,
- Andreassen, O. A., Asherson, P., Banaschewski, T., Bargallo, N., Baumeister, S., Baur-Streubel, R., ...
- Dima, D. (2022). Cortical thickness across the lifespan: Data from 17,075 healthy individuals aged 3-
- 345 90 years. *Human Brain Mapping*, *43*(1), 431–451. https://doi.org/10.1002/hbm.25364
- Ge, R., Yu, Y., Qi, Y. X., Fan, Y., Chen, S., Gao, C., Haas, S. S., New, F., Boomsma, D. I., Brodaty, H.,
- Brouwer, R. M., Buckner, R., Caseras, X., Crivello, F., Crone, E. A., Erk, S., Fisher, S. E., Franke, B.,
- Glahn, D. C., ... Yu, K. (2024). Normative modelling of brain morphometry across the lifespan with
- CentileBrain: Algorithm benchmarking and model optimisation. *The Lancet Digital Health*, 6(3), e211–
- 350 e221. https://doi.org/10.1016/S2589-7500(23)00250-9