Fairly evaluating the performance of normative models

Authors' reply

We appreciate the time and effort of Andre Marquand and colleagues dedicated to offering their perspective on our work.[1] In response, we conducted additional analyses using the same training and testing samples and procedures as in our manuscript. First, we compared algorithm performance based on the shape of the data distribution using the mean standardised log-likelihood (MSLL),[2] defined as:

$$MSLL=\frac{1}{N}\sum_{n=1}^{N}\left(\frac{1}{2}\log(2\pi\sigma_{sum}^2)+\frac{(y_{nd}-\hat{y}_{td})^2}{2\delta_{sum}^2}-\frac{1}{2}\log(2\pi\sigma_{td}^2)-\frac{(y_{nd}-\hat{y}_{td})^2}{2\sigma_{td}^2}\right)$$

where N is the number of individuals in new data, δ_{sum}^2 is the sum of the prediction variance of new data and the noise variance, $\hat{y}_{nd}$ and y_{nd} are the predicted mean and true value of new data, and σ_{td}^2 and y_{td} are the variance and mean for the training data. The results of this evaluation support the choice of the Multivariate Fractional Polynomial Regression as the preferred algorithm (appendix).

Second, we recomputed the deviation Z-scores using the recommended formula:[3,4]

$$z=\frac{y_{nd}-\hat{y}_{nd}}{\sqrt{\delta_{sum}^2}}=\frac{y_{nd}-\hat{y}_{nd}}{\sqrt{\delta_p^2+\delta_n^2}}$$

where N is the number of individuals in new data, δ_{sum}^2 is the sum of the prediction variance of new data and the noise variance δ_n^2, and $\hat{y}_{nd}$ and y_{nd} are the predicted mean and true value of new data. We compared these estimates to those derived by

$$z=\frac{y_{nd}-\hat{y}_{nd}}{RMSE}$$

in our original study, where *RMSE* is the root mean square error of the pretrained model. For this analysis we used independent samples (N=352) from the Southwest Longitudinal Imaging Multimodal Study,[5] the Queensland Twin Adolescent Brain Study,[6] and the Neurocognitive aging data.[7] The two formulas yielded nearly identical results (appendix). Finally, we provide the analysis code and access information for all the datasets used in the supplement of the Review.[1]

We declare no competing interests.

*Ruiyang Ge, Yuetong Yu, Denghuang Zhan, *Sophia Frangou*
sophia.frangou@gmail.com

Djavad Mowafaghian Centre for Brain Health, Department of Psychiatry, University of British Columbia, Vancouver BC V6T1Z3, Canada (RG, YY, SF); School of Population and Public Health, University of British Columbia, Vancouver BC, Canada (DZ); Department of Psychiatry, Icahn School of Medicine at Mount Sinai, New York NY, USA (SF)

1 Ge R, Yu Y, Qi YX, et al. Normative modelling of brain morphometry across the lifespan with CentileBrain: algorithm benchmarking and model optimisation. *Lancet Digit Health* 2024; **6:** e211–21.
2 Fraza CJ, Dinga R, Beckmann CF, Marquand AF. Warped Bayesian linear regression for normative modelling of big data. *Neuroimage* 2021; **245:** 118715.
3 Marquand AF, Rezek I, Buitelaar J, Beckmann CF. Understanding heterogeneity in clinical cohorts using normative models: beyond case-control studies. *Biol Psychiatry* 2016; **80:** 552–61.
4 Marquand AF, Kia SM, Zabihi M, Wolfers T, Buitelaar JK, Beckmann CF. Conceptualizing mental disorders as deviations from normative functioning. *Mol Psychiatry* 2019; **24:** 1415–24.
5 Southwest University. Southwest University longitudinal imaging multimodal (SLIM) brain data repository: a long-term test-retest sample of young healthy adults in southwest China. https://fcon_1000.projects.nitrc.org/indi/retro/southwestuni_qiu_index.html (accessed May 6, 2024).
6 OpenNEURO. Queensland twin adolescent brain (QTAB). https://openneuro.org/datasets/ds004146/versions/1.0.4 (accessed May 6, 2024).
7 OpenNEURO. Neurocognitive aging data release with behavioral, structural, and multi-echo functional MRI measures. https://openneuro.org/datasets/ds003592/versions/1.0.13 (accessed May 6, 2024).

For the **analysis code** see https://github.com/CentileBrain/CB-code

See **Online** for appendix