Fairly evaluating the performance of normative models

We write in response to the recent article in *The Lancet Digital Health* by Ruiyang Ge and colleagues.[1] We would first like to commend the authors on assembling a large multisite dataset, having harmonised protocols, and for their evaluation of many different algorithms for normative modelling in their experiments. However, we would like to express our concern about several aspects of the evaluation of the different algorithms presented in the paper and we believe the evidence presented in the manuscript does not support the conclusions derived.

First, the evaluation metrics used, namely the root mean squared error (RMSE), the mean absolute error (MAE) and the explained variance, are not sufficient to assess the fit of normative models.[2,3] These metrics only measure the accuracy of the estimated centre of the distribution, but ignore its shape, that is, whether individual centiles are well approximated. Adequate calibration of the overall distribution to the data is crucial for reliable inferences derived from normative models, which is well known both in the classical growth-charting literature[4] and in neuroimaging.[2,3,5] This requirement is particularly true in the outer centiles, which are often of primary clinical interest. We illustrate this schematically in the appendix, showing that even the model that was used to generate the data has worse performance according to MAE, RMSE and explained variance, compared with a severely miscalibrated model, due to the inadequacy of these measures.

Second, the computation of the Z statistics in the paper might not yield a quantity that corresponds to the quantiles of a standard normal distribution, which is essential for accurate statistical inferences. This issue arises because the Z-statistics are computed by dividing the residual errors by the RMSE (which can be seen as an estimate of the error variance) rather than fully accounting for the estimated error distribution as proposed elsewhere,[3-5] which can invalidate inference based on Z-scores if the errors have heteroskedastic or non-Gaussian distributions.[2,3,5]

In order to address these issues, we recommend that the authors: (1) comprehensively evaluate the relative performance using metrics that are sensitive to the shape (eg, in terms of skew and kurtosis)[2] of the distribution used to model the data;[3] (2) evaluate the fit of resulting Z-statistics to the centiles of a standard normal distribution including for image-derived phenotypes having non-Gaussian distributions; and (3) share the analysis code and preferably also preprocessed publicly available data, to allow other researchers independently validate the results of the study.

AM received fees for lecturing from Wiegerink BV Netherlands and payments for serving on the board of editors for eLife. All other authors declare no competing interests.

*Andre Marquand†, Saige Rutherford†, Richard Dinga†
andre.marquand@donders.ru.nl

†Contributed equally

Donders Institute for Brain, Cognition and Behaviour, Radboud University Medical Centre, Nijmegen 6525EN, Netherlands (AM, SR); Department of Cognitive Science and Artificial Intelligence, Tilburg University, Tilburg, Netherlands (RD)

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 Dinga R, et al. Normative modeling of neuroimaging data using generalized additive models of location scale and shape. *bioRxiv* 2021; published online June 14. https://doi.org/10.1101/2021.06.14.448106 (preprint).
4 Borghi E, de Onis M, Garza C, et al. Construction of the World Health Organization child growth standards: selection of methods for attained growth curves. *Stat Med* 2006; **25:** 247–65.
5 De Boer AAA, Bayer JMM, Kia SM, et al. Non-Gaussian normative modelling with hierarchical Bayesian regression. *Imaging Neurosci (Camb)* 2024; **2:** 1–36.

See **Online** for appendix