Normative Modeling of Brain Morphometry Across the Lifespan using CentileBrain: Algorithm Benchmarking and Model Optimization

Ruiyang Ge[1], Yuetong Yu[1], Yi Xuan Qi[1], Yunan Vera Fan[1], Shiyu Chen[1], Chuntong Gao[1], Shalaila S Haas[2], Amirhossein Modabbernia[2], Faye New[2], Ingrid Agartz[3,4,5], Philip Asherson[6], Rosa Ayesa-Arriola[7], Nerisa Banaj[8], Tobias Banaschewski[9], Sarah Baumeister[9], Alessandro Bertolino[10], Dorret I Boomsma[11], Stefan Borgwardt[12], Josiane Bourque[13], Daniel Brandeis[9,14], Alan Breier[15], Henry Brodaty[16], Rachel M Brouwer[17,18], Randy Buckner[19,20], Jan K Buitelaar[21], Dara M Cannon[22], Xavier Caseras[23], Simon Cervenka[5,24], Patricia J Conrod[25], Benedicto Crespo-Facorro[26,27], Fabrice Crivello[28], Eveline A Crone[29,30], Liewe de Haan[31], Greig I de Zubicaray[32], Annabella Di Giorgio[33], Susanne Erk[34], Simon E Fisher[35,36], Barbara Franke[37,38], Thomas Frodl[39], David C Glahn[40], Dominik Grotegerd[41], Oliver Gruber[42], Patricia Gruner[43], Raquel E Gur[13], Ruben C Gur[13], Ben J Harrison[44], Sean N Hatton[45], Ian Hickie[46], Fleur M Howells[47], Hilleke E Hulshoff Pol[17,48], Chaim Huyser[49], Terry L Jernigan[50], Jiyang Jiang[16], John A Joska[51], René S Kahn[2], Andrew J Kalnin[52], Nicole A Kochan[16], Sanne Koops[53], Jonna Kuntsi[6], Jim Lagopoulos[54], Luisa Lazaro[55], Irina S Lebedeva[56], Christine Lochner[57], Nicholas G Martin[58], Bernard Mazoyer[28], Brenna C McDonald[59], Colm McDonald[60], Katie L McMahon[61], Tomohiro Nakao[62], Lars Nyberg[63], Fabrizio Piras[8], Maria J Portella[27,64], Jiang Qiu[65,66,67], Joshua L Roffman[68], Perminder S Sachdev[16], Theodore D Satterthwaite[13], Andrew J Saykin[59], Gunter Schumann[69], Carl M Sellgren[5,70], Kang Sim[71], Jordan W Smoller[72], Jair Soares[73], Iris E Sommer[74], Gianfranco Spalletta[8], Dan J Stein[75], Christian K Tamnes[3,4,76], Sophia I Thomopolous[77], Alexander S Tomyshev[56], Diana Tordesillas-Gutiérrez[78], Julian N Trollor[16,79], Dennis van 't Ent[11], Odile A van den Heuvel[80,81], Theo GM van Erp[82], Neeltje EM van Haren[83], Daniela Vecchio[8], Dick J Veltman[80], Henrik Walter[34], Yang Wang[84], Bernd Weber[85], Dongtao Wei[65,66], Wei Wen[16], Lars T Westlye[86], Lara M Wierenga[30], Steven CR Williams[87], Margaret J Wright[88], Sarah Medland[88],Mon-Ju Wu[89], Neda Jahanshad[77], Paul M Thompson[77], Sophia Frangou[1,2]

1. Djavad Mowafagian Centre for Brain Health, University of British Columbia, Vancouver, BC, Canada
2. Department of Psychiatry, Icahn School of Medicine at Mount Sinai, New York, NY, USA
3. Norwegian Centre for Mental Disorders Research (NORMENT), Institute of Clinical Medicine, University of Oslo, Oslo, Norway
4. Department of Psychiatric Research, Diakonhjemmet Hospital, Oslo, Norway
5. Centre for Psychiatry Research, Department of Clinical Neuroscience, Karolinska Institutet & Stockholm Health Care Services, Region Stockholm, Stockholm, Sweden
6. Institute of Psychiatry, Psychology and Neuroscience, Social, Genetic and Developmental Psychiatry Center, King's College London, London, UK
7. Department of Psychiatry, Marqués de Valdecilla University Hospital, Valdecilla Biomedical Research Institute (IDIVAL), Santander, Spain
8. Laboratory of Neuropsychiatry, IRCCS Santa Lucia Foundation, Rome, Italy
9. Department of Child and Adolescent Psychiatry and Psychotherapy, Central Institute of Mental Health, Mannheim, Germany
10. Department of Basic Medical Science, Neuroscience and Sense Organs, University of Bari Aldo Moro, Bari, Italy
11. Department of Biological Psychology, Vrije Universiteit Amsterdam, Amsterdam, The Netherlands
12. Translational Psychiatry Unit, Department of Psychiatry and Psychotherapy, University of Lübeck, Lübeck, Germany
13. Department of Psychiatry, University of Pennsylvania, Philadelphia, Pennsylvania, USA

14. Department of Child and Adolescent Psychiatry, University of Zürich, Zurich, Switzerland
15. Department of Psychiatry, Indiana University School of Medicine, Indianapolis, Indiana, USA
16. Centre for Healthy Brain Ageing (CHeBA), School of Psychiatry, University of New South Wales, Sydney, Australia
17. Department of Psychiatry, UMC Brain Center, University Medical Center Utrecht, Utrecht University, Utrecht, The Netherlands
18. Complex Trait Genetics, Center for Neurogenomics and Cognitive Research, Vrije Universiteit Amsterdam, Amsterdam, The Netherlands
19. Department of Psychology, Center for Brain Science, Harvard University, Cambridge, Massachusetts, USA
20. Department of Psychiatry, Massachusetts General Hospital, Boston, Massachusetts, USA
21. Department of Cognitive Neuroscience, Donders Institute for Brain, Cognition and Behaviour, Radboud University Medical Center, Nijmegen, The Netherlands
22. Clinical Neuroimaging Laboratory, National University of Ireland Galway, Galway, Ireland
23. MRC Centre for Neuropsychiatric Genetics and Genomics, Division of Psychological Medicine and Clinical Neurosciences, Cardiff University, Cardiff, UK
24. Department of Medical Sciences, Psychiatry, Uppsala University, Uppsala, Sweden
25. Department of Psychiatry and Addiction, Université de Montréal, CHU Ste Justine, Montréal, Canada
26. University Hospital Virgen del Rocio, Seville, Spain; Department of Psychiatry, University of Seville, Institute of Biomedicine of Seville (IBIS), Seville, Spain
27. Mental Health Research Networking Center (CIBERSAM), Madrid, Spain
28. Institut des Maladies Neurodégénératives, Université de Bordeaux, Bordeaux, France
29. Department of Psychology, Education and Child Studies, Erasmus University Rotterdam, Rotterdam, The Netherlands
30. Institute of Psychology, Leiden University, Leiden, The Netherlands; Leiden Institute for Brain and Cognition, Leiden University, Leiden, The Netherlands
31. Department of Psychiatry, Amsterdam UMC, Amsterdam, The Netherlands
32. School of Psychology & Counselling, Faculty of Health, Queensland University of Technology, Brisbane, Australia
33. Laboratory of Biological Psychiatry, Fondazione IRCCS Casa Sollievo della Sofferenza, San Giovanni Rotondo, Italy
34. Division of Mind and Brain Research, Department of Psychiatry and Psychotherapy, Charité-Universitätsmedizin Berlin, Berlin, Germany
35. Language and Genetics Department, Max Planck Institute for Psycholinguistics, Nijmegen, The Netherlands
36. Donders Institute for Brain, Cognition and Behaviour, Radboud University Medical Center, Nijmegen, The Netherlands
37. Department of Human Genetics, Radboud University Medical Center, Nijmegen, The Netherlands
38. Department of Psychiatry, Donders Institute for Brain, Cognition and Behaviour, Radboud University Medical Center, Nijmegen, The Netherlands
39. University Clinics and Clinics for Psychiatry, Psychotherapy and Psychosomatic Medicine, RWTH Aachen University, Aachen, Germany

40. Department of Psychiatry, Tommy Fuss Center for Neuropsychiatric Disease Research Boston Children's Hospital, Harvard Medical School, Boston, Massachusetts, USA
41. Department of Psychiatry and Psychotherapy, University of Muenster, Muenster, Germany
42. Section for Experimental Psychopathology and Neuroimaging, Department of General Psychiatry, Heidelberg University, Heidelberg, Germany
43. Department of Psychiatry, Yale University, New Haven, Connecticut, USA
44. Melbourne Neuropsychiatry Centre, Department of Psychiatry, The University of Melbourne & Melbourne Health, Melbourne, Australia
45. Center for Multimodal Imaging and Genetics, University of California San Diego, La Jolla, California, USA
46. Brain and Mind Centre, University of Sydney, Sydney, Australia
47. Neuroscience Institute, University of Cape Town, Cape Town, South Africa
48. Department of Psychology, Utrecht University, Utrecht, The Netherlands
49. Department of Child and Adolescent Psychiatry, Academic Medical Centre/De Bascule, Amsterdam, The Netherlands
50. Center for Human Development, Departments of Cognitive Science, Psychiatry, and Radiology, University of California, San Diego, USA
51. Department of Neuropsychiatry, University of Cape Town, Cape Town, South Africa
52. Department of Radiology, The Ohio State University College of Medicine, Columbus, Ohio, USA
53. Department of Psychiatry and Brain Center Rudolf Magnus, University Medical Center Utrecht, Utrecht, The Netherlands
54. Sunshine Coast Mind and Neuroscience - Thompson Institute, University of the Sunshine Coast, Queensland, Australia
55. Department of Child and Adolescent Psychiatry and Psychology, Hospital Clínic Barcelona, Barcelona, Spain
56. Mental Health Research Center, Moscow, Russia
57. SA MRC Unit on Risk and Resilience in Mental Disorders, Department of Psychiatry, Stellenbosch University, Stellenbosch, South Africa
58. Queensland Institute of Medical Research, Berghofer Medical Research Institute, Brisbane, Australia
59. Department of Radiology and Imaging Sciences, Indiana University School of Medicine, Indianapolis, Indiana, USA
60. Centre for Neuroimaging & Cognitive Genomics (NICOG), NCBES Galway Neuroscience Centre, National University of Ireland Galway, Galway, Ireland
61. School of Clinical Sciences, Centre for Biomedical Technologies, Queensland University of Technology, Brisbane, Australia
62. Department of Neuropsychiatry, Kyushu University, Fukuoka, Japan
63. Department of Radiation Sciences, Umeå Center for Functional Brain Imaging, Umeå University, Umeå, Sweden; Department of Integrative Medical Biology, Umeå University, Umeå, Sweden
64. Department of Psychiatry, Hospital de la Santa Creu iSant Pau, Institutd' Investigació Biomèdica SantPau, Universitat Autònomade Barcelona (UAB), Barcelona, Spain
65. Key Laboratory of Cognition and Personality, Southwest University, Ministry of Education, Chongqing, PR China
66. Faculty of Psychology, Southwest University, Chongqing, PR China
67. Southwest University Branch, Collaborative Innovation Center of Assessment Toward Basic Education Quality, Beijing Normal University, Beijing, PR China

68. Department of Psychiatry, Massachusetts General Hospital, Harvard Medical School, Boston, Massachusetts, USA
69. Centre for Population Neuroscience and Precision Medicine (PONS), Institute of Psychiatry, Psychology, and Neuroscience, Social, Genetic & Developmental Psychiatry Centre, King's College London, London, UK; Institute for Science and Technology of Brain-inspired Intelligence, Fudan University, Shanghai, PR China; Centre for Population Neuroscience and Stratified Medicine (PONS), Charite Mental Health, Department of Psychiatry and Psychotherapy, CCM, Charite Universitätsmedizin Berlin, Berlin, Germany
70. Department of Physiology and Pharmacology, Karolinska Institute, Stockholm, Sweden
71. Institute of Mental Health, Singapore
72. Center for Genomic Medicine, Massachusetts General Hospital, Boston, Massachusetts, USA
73. University of Texas Health Harris County Psychiatric Center, Houston, Texas, USA
74. Department of Biomedical Sciences of Cells and Systems, Rijksuniversiteit Groningen, University Medical Center Groningen, Groningen, The Netherlands
75. SA MRC Unit on Risk and Resilience in Mental Disorders, Department of Psychiatry and Neuroscience Institute, University of Cape Town, Cape Town, South Africa
76. PROMENTA Research Center, Department of Psychology, University of Oslo, Oslo, Norway
77. Genetics Center, Stevens Institute for Neuroimaging and Informatics, Keck USC School of Medicine, Marina del Rey, California, USA
78. Department of Radiology, Marqués de Valdecilla University Hospital, Valdecilla Biomedical Research Institute (IDIVAL), Santander, Spain; Advanced Computing and e-Science, Instituto de Física de Cantabria (UC-CSIC), Santander, Spain
79. Department of Developmental Disability Neuropsychiatry, School of Psychiatry, University of New South Wales, Sydney, Australia
80. Department of Psychiatry, Amsterdam UMC, Vrije Universiteit Amsterdam, Amsterdam Neuroscience, Amsterdam, The Netherlands
81. Department of Anatomy & Neurosciences, Amsterdam UMC, Vrije Universiteit Amsterdam, Amsterdam Neuroscience, Amsterdam, The Netherlands
82. Department of Psychiatry and Human Behavior, University of California, Irvine, California, USA
83. Department of Child and Adolescent Psychiatry/Psychology, Erasmus University Medical Centre, Rotterdam, The Netherlands
84. Department of Radiology, Medical College of Wisconsin, Milwaukee, Wisconsin, USA
85. Institute for Experimental Epileptology and Cognition Research, University of Bonn Germany, Bonn, Germany; University Hospital Bonn, Bonn, Germany
86. Department of Psychology, University of Oslo, Oslo, Norway
87. Department of Neuroimaging, Institute of Psychiatry, Psychology and Neuroscience, King's College London, London, UK
88. Queensland Brain Institute, University of Queensland, Brisbane, Queensland, Australia
89. Department of Psychiatry and Behavioral Science, University of Texas Health Science Center, Houston, Texas, USA

For correspondence: Prof. Sophia Frangou (sophia.frangou@gmail.com)

ABSTRACT

Normative modeling is a statistical approach to quantify the degree to which a particular individual-level measure deviates from the pattern observed in a normative reference population. When applied to human brain morphometric measures it has the potential to inform about the significance of normative deviations for health and disease. Normative models can be implemented using a variety of algorithms that have not been systematically appraised. To address this gap, eight algorithms were compared in terms of performance and computational efficiency using brain regional morphometric data from 37,407 healthy individuals (53.33% female; aged 3 to 90 years) collated from 86 international MRI datasets. Performance was assessed with the mean absolute error (MAE) and computational efficiency was inferred from central processing unit (CPU) time. The algorithms evaluated were Ordinary Least Squares Regression (OLSR), Bayesian Linear Regression (BLR), Generalized Additive Models for Location, Scale, and Shape (GAMLSS), Parametric Lambda, Mu, Sigma (LMS), Multivariable Fractional Polynomial Regression (MFPR), Warped Bayesian Linear Regression (WBLG), and Hierarchical Bayesian Regression (HBR). Model optimization involved testing 9 covariate combinations pertaining to acquisition features, parcellation software versions, and global neuroimaging measures (i.e., total intracranial volume, mean cortical thickness, and mean cortical surface area). Statistical comparisons across models at $P_{FDR}<0.05$ indicated that MFPR-derived sex and region-specific models with nonlinear polynomials for age and linear effect of global measures had superior predictive accuracy; for models of regional subcortical volume the MAE range was 70-80 mm^3 and the corresponding range for regional cortical thickness and regional cortical surface area was 0.09-0.26 mm; and 24-660 mm^2 respectively. The MFPR-derived models were also computationally more efficient with a CPU time below one second compared to a range of 2 seconds to 60 minutes for the other algorithms. MFPR-derived models were robust across distinct age groups when sample sizes exceeded 3,000. These results support CentileBrain (https://centilebrain.org/) an empirically benchmarked framework for normative modeling that is freely available to the scientific community.

1. INTRODUCTION

Normative modeling is a class of statistical methods to quantify the degree to which a particular individual-level measure deviates from the pattern observed in a normative reference population. Normative modeling of neuroimaging phenotypes is an emerging area of research - most applications focus on brain morphometry given the wide availability and validity of structural magnetic resonance imaging (Potvin et al., 2017; Potvin et al., 2016; Bethlehem et al., 2022; Dima et al., 2022; Frangou et al., 2022; Potvin et al., 2022) with recent extensions into diffusion MRI (Villalon-Reina, 2022). In parallel, an increasing number of studies are applying normative modeling to the study of neuropsychiatric disorders (Wolfers et al., 2018; Zabihi et al., 2019; Bethlehem et al., 2020; Wolfers et al., 2020; Zabihi et al., 2020; Lv et al., 2021; Bayer et al., 2022; Haas et al., 2022).

Current normative modeling studies employ a range of linear, nonlinear, and Bayesian algorithms that reflect researchers' preferences (Bethlehem et al., 2020; Frangou et al., 2022; Kia et al., 2022). At present, there is no systematic comparative evaluation of the performance of these algorithms and no empirical determination of the minimum sample size necessary for reliable normative estimates of brain morphometric measures. With few exceptions (Bethlehem et al., 2022; Dima et al., 2022; Frangou et al., 2022; Lawrence et al., 2021; Rutherford et al., 2022a), the size of the sample used for the reference population is small to modest (range: 145-870) (Wolfers et al., 2018; Marquand et al., 2019; Zabihi et al., 2019; Bethlehem et al., 2020; Wolfers et al., 2020; Zabihi et al., 2020; Lv et al., 2021; Bayer et al., 2022).

The aim of the current study is to provide a systematic comparison of the statistical methods for normative modeling of brain morphometric data as a guide for evaluating prior literature and informing future study designs. Regional measures of subcortical volume, cortical thickness, and cortical surface area were pooled into a multisite sample of 37,407 healthy individuals (53% female, age range: 3-90 years) which was then split into a training and a test subset. Eight algorithms, representing the range of methods currently used in normative modeling studies, were evaluated in terms of their accuracy and computational efficiency, in sex-specific models for each brain morphometric feature. For each regional morphometric measure, model optimization involved 9 combinations of covariates pertaining to acquisition and parcellation methods, and linear and non-linear effects of age and global neuroimaging measures. All models were tested for robustness to sample composition (size and age range). The results form the basis of CentileBrain (https://centilebrain.org/) an empirically benchmarked framework for normative modeling that is made available to a scientific community through a dedicated web platform.

2. MATERIALS AND METHODS

2.1 Samples

We collated de-identified data from 86 datasets from Europe, Australia, USA, South Africa, and East Asia (*Supplementary Figure S1*; *Supplementary Table S1*). Data use for the current analyses was governed by the policies of the ENIGMA Lifespan Working Group (https://enigma.ini.usc.edu/ongoing/enigma-lifespan/), as well as by the policies of individual studies and national repositories. Based on the information provided in each dataset, data were further selected to include high-quality neuroimaging measures (Supplement S2) from participants that were free of psychiatric, medical, and neurological morbidity and cognitive impairment at the time of scanning. Only scans acquired at baseline were included from datasets with multiple scanning assessments.

2.2 Brain Morphometry

Acquisition protocols and scanner vendors varied across datasets but morphometric feature extraction from whole-brain T_1-weighted images was implemented using the standard pipelines in the FreeSurfer image analysis suite (http://surfer.nmr.mgh.harvard.edu/) (Supplementary Table S1) to yield global measures of total intracranial volume (ICV), mean cortical thickness and mean surface area and regional measures of cortical thickness (N=68) and cortical surface area (N=68) based on the Desikan-Killiany atlas (Desikan et al., 2006), and 14 subcortical volumetric measures based on the *aseg* outputs in FreeSurfer (Fischl et al., 2002). Sex-specific normative models were developed separately for each of the 150 regional morphometry measures to accommodate sex differences in brain morphometry (Ge et al., 2021; Ruigrok et al., 2014; Wierenga et al., 2022).

2.3 Optimization of normative models

The procedures used to generate optimized sex-specific models for each brain morphometric measure are illustrated in **Figure 1** and consisted of the following steps: **(I)** ***Data preparation***: Sex-specific subsamples of the study sample were randomly split into a training subset (80%) and a test subset (20%) stratified by scanning site. In each subset, data were mean-centered after extreme values in each site-dataset were identified and removed using the interquartile range (IQR) method;

II) ***Algorithm Selection***: The data for each morphometric measure were analyzed with the following algorithms:

(1) Ordinary Least Squares Regression (**OLSR**) (implemented using the "*lm*" function in R): this is a linear regression model that aims to minimize the sum of squared differences between the observed and predicted values;

(2) Bayesian Linear Regression (**BLR**) (implemented using the "*stan*" package in R): this is a linear model in which the outcome variable and the model parameters are assumed to be drawn from a probability distribution (Gelman et al., 2008);

(3) Generalized Additive Models for Location, Scale, and Shape (**GAMLSS**) (Rigby & Stasinopoulos, 2005): this framework can model heteroskedasticity, non-linear effects of variables, and hierarchical structure of the data; this algorithm was implemented using the "*caret*" package in R following prior recommendations (Dinga et al., 2021);
(4) Parametric Lambda (λ), Mu (μ), Sigma (σ) (**LMS**) method: this subclass of GAMLSS assumes that the outcome variable follows the Box-Cox Cole and Green distribution (Cole and Green, 1992); this algorithm was implemented using the "*gamlss*" package in R (Dong et al., 2020);
(5) Gaussian Process Regression (**GPR**) (Williams & Barber, 1998): this is a nonparametric regression model that follows Bayesian principles, and was implemented using the "*kernlab*" package in R and the "*sigest*" function for estimating the hyperparameter sigma;
(6) Multivariable Fractional Polynomial Regression (**MFPR**) (Royston & Altman, 1994): this algorithm enables the determination of the functional form of a predictor variable by testing a broad family of shapes and multiple turning points while simultaneously providing a good fit at the extremes of the covariates; this algorithm was implemented using the "*mfp"* package in R and the closed test procedure (known as RA2) to select the most appropriate fractional polynomial (Ambler et al., 2001);
(7) Warped Bayesian Linear Regression (**WBLR**) (Fraza et al., 2021): this framework is based on Bayesian linear regression with likelihood warping and was implemented using the "PCNtoolkit" in Python with default settings (Bayer et al., 2022);
(8) Hierarchical Bayesian Regression (**HBR**) (Kia et al., 2021): this approach also uses Bayesian principles and is considered particularly useful when variance from multiple hierarchical levels, including the scanning protocol or site effects; this algorithm was implemented using the "PCNtoolkit" in Python with default settings (Bayer et al., 2022).

For the HBR, WBLR, GPR, and GAMLSS, site effects were modeled as random effects as recommended by prior literature (Dinga et al., 2021; Bayer et al., 2022; Rutherford et al., 2022b). For all other algorithms, site-harmonization used ComBat-GAM (Pomponio et al., 2020) as this approach has the potential to generate models without having to recalculate model parameters for each unseen data.

The processes outlined above generated 8 models for each regional morphometric measure per sex. These models were trained using 5-fold cross-validation (5F-CV) in the corresponding sex-specific training subset with age being the only explanatory variable and then model parameters were tested in the corresponding sex-specific test subset. The cross-validated Mean Absolute Error (MAE), which is the average of the absolute differences (i.e., errors) between the predicted and the actual value of the outcome variable, was the main measure of model performance supplemented by the Root Mean Square Error (RMSE), which is the standard deviation of the prediction errors. The computational efficiency of each model was assessed using the Central

Processing Unit (CPU) time of the supercomputing infrastructure of the Icahn School of Medicine at Mount Sinai (https://labs.icahn.mssm.edu/minervalab/resources/hardware-technical-specs/).

Across regional morphometric measures (and separately in males and females), the MAEs and RMSE of the models generated by each algorithm were concatenated as a single vector to enable pairwise comparisons between algorithms. Results were considered significant across comparisons at $P_{FDR}<0.05$ using false discovery rate (FDR) correction for multiple testing.

(III) ***Model Optimization***: After selecting the best-performing algorithm, model optimization involved the evaluation of improvements in the MAE (and RMSE) by adding the following explanatory variables: (1) global neuroimaging measures, i.e., ICV, mean cortical thickness or mean cortical surface area, as appropriate - both linear and non-linear contributions from these variables were considered; (2) scanner vendor type; (3) FreeSurfer version; and (4) combinations of these variables. Each of these models was trained using 5F-CV in the corresponding sex-specific training subset and then tested in the corresponding sex-specific test subset. Variables that significantly improved performance were retained.

Upon completion of steps (i)-(iii), *optimized* sex-specific and region-specific models were defined based on the best-performing algorithm and covariate combination.

2.4 Sensitivity Analyses

Sample Size: The study sample was partitioned into 75 sex-specific random subsets comprising between 200 to 15,000 participants in increments of 200. The robustness of the optimized sex- and region-specific models to sample size in terms of MAE and RMSE was assessed in each partition using 5F-CV.

Distinct age bins: Model accuracy may be influenced by the sample's age range and by distinct challenges encountered in scanning different age groups such as higher levels of motion in pediatric populations (De Bie et al., 2010). Accordingly, the study sample was divided into 9 sex-specific age-bins (i.e., age≤10 years; 10<age≤20 years; 20<age≤30 years; 30<age≤40 years; 40<age≤50 years; 50<age≤60 years; 60<age≤70 years; 70<age≤80 years; 80<age≤90 years). The MAE and RMSE of each optimized sex and region-specific model were estimated in each age bin using 5F-CV. Subsequently, Pearson's correlation coefficients were computed between the MAE and RMSE values of the models within each sex-specific age bin with those derived from the sex-specific subset of the entire sample.

Model Comparison: The optimized sex and region-specific models of the best-performing algorithm were compared to those based on the other algorithms through a series of pairwise tests between the concatenated MAE and RMSE of the models across all regional measures.

3. RESULTS

3.1 Identification of the best-performing algorithm: The MAE, RMSE and CPU time for each sex and region-specific model are shown in *Supplemental Tables S2* to *S7 and Supplementary Figures S3 and S4*. As the same general pattern was noted across all models in both sexes, the MAE, RMSE, and CPU of the models for the left thalamic volume and left medial orbitofrontal cortical thickness and surface area are shown as exemplars in Figure 2 and Supplementary Figure S2. Across all sex- and region-specific models, the LMS, GPR, WBLR, and MFPR had comparable values for MAE and RMSE that were statistically better at $P_{FDR}<0.05$ than those for BLR, OLSR, and HBR. Across all models, the OLSR and MFPR had the shortest CPU times (less than one second) while GPR had the longest (25-60 minutes) (Figure 2). Accordingly, the MFPR emerged as the preferred algorithm given its combined advantages in accuracy and computational efficiency.

Selection of explanatory variables for model optimization: Based on the preceding results, we assessed improvements in the MAE and RMSE of sex and region-specific models generated by the MFPR after adding scanner vendor, FreeSurfer version, and global neuroimaging measures (i.e., ICV, mean cortical thickness, or mean cortical surface area, as indicated) and their linear and non-linear combinations as covariates. The results are shown in Supplemental Tables S8 to S13 and Supplementary Figures S6 and S7. As the same pattern was observed for all sex- and region-specific models, we continue to use the left thalamic volume and left medial orbitofrontal cortex thickness and area as exemplars in Figure 3 and Supplementary Figure S5. The impact of the scanner vendor and FreeSurfer version on model performance was minimal while the opposite was the case for the global neuroimaging measures. Accordingly, we define as “optimal models” the sex and region-specific models that were based on the MFPR algorithm with nonlinear fractional polynomials of age and linear effects of the appropriate global neuroimaging measure (i.e., ICV for models of regional subcortical volumes, mean cortical thickness for models of regional cortical thickness and mean cortical surface area for models of regional cortical surface area).

3.2 Sensitivity Analyses

***Sample Size*:** The MAE and RMSE values of the optimal sex and region-specific models plateaued at a sample size of approximately 3,000 participants as shown in Figure 4 and Supplementary Figure S8.

Distinct Age Bins: The MAE and RMSE values of the optimal sex and region-specific models in each of the 9 age bins are presented in Figure 5, Supplementary Figure S9, and in Supplementary Tables S14 to S19. Across all age bins, the correlation coefficient between the MAE or RMSE values of the sex and region-specific models obtained from the full study sample and MAE or RMSE values of the corresponding models estimated in each age bin were all greater than 0.98, suggesting the robustness of the model accuracy across all age groups.

Model Comparison

To test the advantage of the optimal sex- and region-specific models further, we computed models with the same covariates using the HBR, GPR, GAMLSS, and WBLR algorithms. We focused on these algorithms as they have been applied to other large normative datasets (Dinga et al., 2021; Bethlehem et al., 2022; Rutherford et al., 2022a; Rutherford et al., 2022b). The MAE, RMSE, and CPU time for each of the sex and region-specific model derived from these five algorithms are shown in Supplemental Tables S20 to S25 and Supplementary Figures S11 and S12. Pairwise statistical comparison of the models from each algorithm indicated superior accuracy at $P_{FDR}<0.05$ for the optimal MFPR-derived models. We illustrate these findings in Figure 6 and *Supplementary Figure 10* using the left thalamic volume and left medial orbitofrontal cortical thickness and surface area as exemplars. In addition to having better accuracy, the MFPR-derived models remained the most computationally efficient with CPU times of less than a second.

4. Discussion

The present study undertook comparative evaluation of eight algorithms commonly used for normative modeling using morphometric data (i.e., 14 regional subcortical volumes, 68 measures of cortical thickness, and cortical surface area) from a multisite sample of 37,407 healthy individuals. Models based on the MFPR algorithm with nonlinear fractional polynomials of age and linear global neuroimaging measures emerged as optimal based on their performance and computational efficiency, the latter being an important consideration when analyzing large datasets. These models were robust to variations in sample composition with respect to age and their performance plateaued at sample sizes of ≈ 3,000. Establishing a minimum sample size requirement for normative models of brain morphometry provides a useful point of reference when evaluating prior literature or designing new studies.

The findings presented validate our choice to use MFPR in our previous normative studies on brain morphometry across the lifespan (Dima et al., 2022; Frangou et al., 2022) and on white matter microstructural measures based on diffusion-weighted magnetic resonance imaging (Nabulsi et al., 2020; Lawrence et al., 2022). Further, after testing the impact of 9 combinations of explanatory variables on model performance we found that global morphometric measures (i.e., ICV, mean cortical thickness, mean cortical surface area) had the most significant effect. This observation is aligned with prior literature on the contribution of ICV in explaining the variance of regional subcortical volumes and cortical surface area measures (Potvin et al., 2016; Potvin et al., 2017). The present study extended these findings by demonstrating that mean cortical thickness and mean surface area outperformed ICV as explanatory variables in normative models of regional cortical thickness and cortical surface area respectively. Accordingly, the optimal normative models for brain morphometry consisted of an MFPR algorithm and a combination of explanatory variables that comprised nonlinear fractional polynomials of age and linear global measures of ICV, mean cortical thickness, and mean cortical surface area for models of regional subcortical volume, regional cortical thickness, and regional cortical surface area respectively. Sensitivity analyses across different age bins supported the applicability of the models developed in the whole study sample, which spanned an age range of 3-90 years, to groups with a more restricted age range and at different points in their life trajectory.

Site effects are a major challenge when aggregating multisite data as they may confound or bias results. The most common methods for minimizing site effects involve either site harmonization using ComBat-GAM (Pomponio et al., 2020) prior to normative modeling or the inclusion of site as an explanatory variable in the normative models. A recent publication that used a smaller sample (569 healthy participants) and a narrower age range (6 to 40 years) suggested the HBR with site as an explanatory variable may be superior to ComBat-based site harmonization for the normative modeling of brain morphometry (Bayer et al., 2022). We found no support for this assertion, as in this study MFPR-derived models using Combat-GAM for site harmonization outperformed HBR-

derived models with site as a random effect. An additional advantage of using ComBat-GAM is that it removes the requirement for calibration and model parameter adaptation every time the model is applied to data from a new site. In the HBR models, by contrast, pre-trained parameters can be used for new data if they originate from one of the sites in the training dataset (Bayer et al., 2022) or under the assumption that the variation accounted for by an unseen site should align with that of the sites in the training dataset (Kia et al., 2022).

Prior studies have shown that sex accounts for a significant amount of variance in brain morphology (Potvin et al., 2016; Potvin et al., 2017), both cross-sectionally (Ge et al., 2021; Kaczkurkin et al., 2019) and with regard to age-related changes (Ching et al., 2022; Forde et al., 2020; Wierenga et al., 2022). In view of that, we developed sex-specific models for each brain morphometric measure thus extending prior normative studies that considered males and females together (Bayer et al., 2022; Rutherford et al., 2022a). Additionally, we provide normative models for regional cortical surface area measures; changes in the cortical surface area have important functional implications for cognition during development (Cafiero et al., 2019; Wierenga et al., 2014; Vijayakumar et al., 2016) and aging (Lemaitre et al., 2012) but have been overlooked in prior normative studies (Bethlehem et al., 2022; Rutherford et al., 2022a). We note that the current normative model is compiled cross-sectionally, from people of different ages who experienced different exposures to factors that can influence brain health. In later life, there is also attrition of individuals from the control samples through mortality and morbidity. As such the trajectory of brain aging for specific individuals may differ from the time-dependent charts implied by these centile curves; future tests with longitudinal data from people of different ages will be of interest to confirm this.

The present study did not include an exhaustive list of potential explanatory variables. It could be argued that the inclusion of other variables, such as childhood adversity, premature birth, or socioeconomic status that are known to influence brain morphometry (Brito and Noble, 2014; Frodl et al., 2017), could have further improved model performance. Exploring this possibility further could be best achieved within the context of single large-scale studies where such variables would be consistently recorded in all participants. Future work will also attempt to model the effect of environment and early life adversity on brain measures. On the other hand, the inclusion of multiple explanatory variables in the normative model itself could restrict the applicability of normative models only to those datasets where all such features were assessed.

In conclusion, this study presents a detailed evaluation of the comparative performance of the eight algorithms available for normative modeling and of the influence on the predictive accuracy of key parameters pertaining to site effects, covariates, sample size, and sample composition with respect to age.

Data Availability

Due to the multisite nature of the data, access to individual participant data from each dataset would need to be made available through the principal investigators of the original studies or through access requests addressed to the appropriate data repositories. Details are provided in Supplemental Table S1.

Code availability

A dedicated web portal (https://centilebrain.org) provides the optimal model parameters, as pre-trained models, to be applied to any user-specified dataset in the context of open science (Supplement S6).

Acknowledgments

Analyses were conducted using high-performance computing infrastructure of the Icahn School of Medicine at Mount Sinai and the University of British Columbia.

5. REFERENCES

Ambler, G., et al. (2001). Fractional polynomial model selection procedures: investigation of Type I error rate. *Journal of Statistical Computation Simulation*, 69(1), 89-108.

Bayer, J.M.M., et al. (2022). Marquand A. Accommodating site variation in neuroimaging data using normative and hierarchical Bayesian models. *NeuroImage*, 264, 119699.

Bethlehem, R. A., et al. (2020). A normative modelling approach reveals age-atypical cortical thickness in a subgroup of males with autism spectrum disorder. *Communications Biology,* 3(1), 1-10.

Bethlehem, R. A., et al. (2022). Brain charts for the human lifespan. *Nature*, 604*,* 525-533.

Brito, N. H., (2014). Socioeconomic status and structural brain development. *Frontiers in Neuroscience*, 8, 276.

Cafiero, R., et al. (2019). The concurrence of cortical surface area expansion and white matter myelination in human brain development. *Cerebral Cortex*, 29(2), 827-837.

Ching, C., et al. (2022). Sex-dependent age trajectories of subcortical brain volume: A UK Biobank study (N=39,544). *Alzheimer's & Dementia*, 17, e055854.

De Bie HMA, Boersma M, Wattjes MP, Adriaanse S, Vermeulen RJ, Oostrom KJ, et al. Preparing children with a mock scanner training protocol results in high quality structural and functional MRI scans. Eur J Pediatr. (2010) 169:1079–85.

Desikan, R. S., et al. (2006). An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest. *NeuroImage, 31*(3), 968-980.

Dima, D., et al. (2022). Subcortical volumes across the lifespan: Data from 18,605 healthy individuals aged 3–90 years. *Human Brain Mapping, 43*(1), 452-469.

Dinga, R., et al. (2021). Normative modeling of neuroimaging data using generalized additive models of location scale and shape. *bioRxiv*.

Dong, H.-M., et al. (2020). Charting brain growth in tandem with brain templates at school age. 65(22), 1924-1934.

Fischl, B., et al. (2002). Whole brain segmentation: automated labeling of neuroanatomical structures in the human brain. *Neuron, 33*(3), 341-355.

Forde, N. J., et al. (2020). Sex differences in variability of brain structure across the lifespan. *Cerebral Cortex*, 30(10), 5420-5430.

Frangou, S., et al. (2022). Cortical thickness across the lifespan: Data from 17,075 healthy individuals aged 3–90 years. *Human Brain Mapping, 43*(1), 431-451.

Fraza, C., et al. (2021). Warped Bayesian linear regression for normative modelling of big data. *NeuroImage, 245*, 118715.

Frodl, T., Childhood adversity impacts on brain subcortical structures relevant to depression, *Journal of Psychiatric Research*, 86, 58-65.

Ge, R., et al. (2021). Sex effects on cortical morphological networks in healthy young adults. *NeuroImage, 233*, 117945.

Haas et al. Normative modeling of brain morpjometry in individuals at clinical high risk for psychosis; submission to BioRx pending

Kaczkurkin, A. N., et al. (2019). Sex differences in the developing brain: insights from multimodal neuroimaging. *Neuropsychopharmacology*, 44(1), 71-85.

Kia, S., Federated multi-site normative modeling using hierarchical Bayesian regression. *bioRxiv*. 2021.

Lawrence, K., et al. (2022). Age and sex effects on advanced white matter microstructure measures in 15,628 older adults: A UK biobank study. *Brain Imaging and Behavior, 15*, 2813-2823.

Lemaitre, H., et al. (2012). Normal age-related brain morphometric changes: nonuniformity across cortical thickness, surface area and gray matter volume? *Neurobiology of Aging*, 33(3), 617. e1-617. e9.

Lv, J., et al. (2021). Individual deviations from normative models of brain structure in a large cross-sectional schizophrenia cohort. *Molecular Psychiatry, 26*(7), 3512-3523.

Marquand, A. F., et al. (2019). Conceptualizing mental disorders as deviations from normative functioning. *Molecular Psychiatry, 24*(10), 1415-1424.

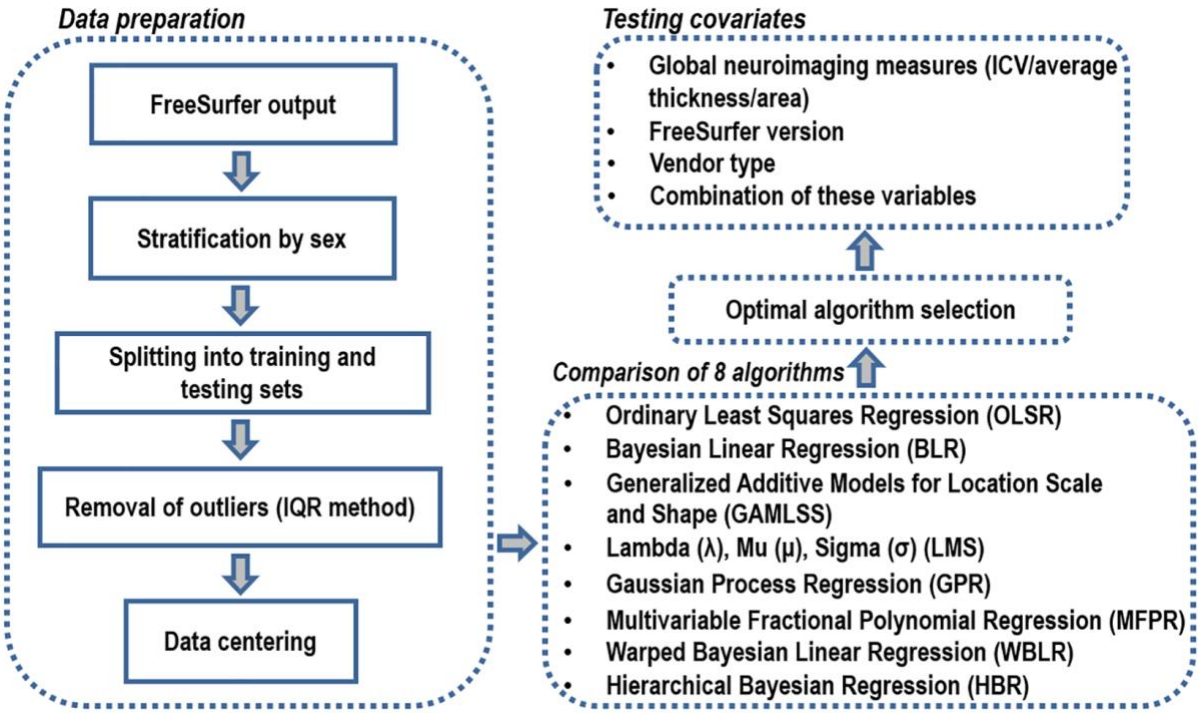
Nabulsi, L., et al. (2020). Exogenous sex hormone effects on brain microstructure in women: a diffusion MRI study in the UK Biobank. *bioRxiv*.

Pomponio, R., et al. (2020). Harmonization of large MRI datasets for the analysis of brain imaging

patterns throughout the lifespan. *NeuroImage, 208*, 116450.
Potvin, O., et al. (2016). Normative data for subcortical regional volumes over the lifetime of the adult human brain. *NeuroImage, 137*, 9-20.
Potvin, O., et al. (2017). Normative morphometric data for cerebral cortical areas over the lifetime of the adult human brain. *NeuroImage, 156*, 315-339.
Potvin, O., et al. (2022). NOMIS: Quantifying morphometric deviation from normality over the lifetime in the adult human brain. *bioRxiv*.
Rigby, R. A., & Stasinopoulos, D. M. (2005). Generalized additive models for location, scale and shape. Journal of the Royal Statistical Society: Series C, 54(3), 507-554.
Royston, P., & Altman, D. G. (1994). Regression using fractional polynomials of continuous covariates: parsimonious parametric modelling. *Journal of the Royal Statistical Society: Series C, 43*(3), 429-453.
Ruigrok, A. N., et al. (2014). A meta-analysis of sex differences in human brain structure. *Neuroscience Biobehavioral Reviews, 39*, 34-50.
Rutherford, S., et al. (2022a). Charting brain growth and aging at high spatial precision. *Elife, 11*, e72904.
Rutherford, S., et al. (2022b). The Normative modeling framework for computational psychiatry. *Nature Protocols, 17*, 1711–1734.
Vijayakumar, N., et al. (2016). Brain development during adolescence: A mixed-longitudinal investigation of cortical thickness, surface area, and volume. *Human Brain Mapping*, 37(6), 2027-2038.
Villalón-Reina, J., et al. (2022). Multi-site normative modeling of diffusion tensor imaging metrics using hierarchical Bayesian regression. *MICCAI 2022: Medical Image Computing and Computer Assisted Intervention*, 207-217.
Wierenga, L. M., et al. (2014). Unique developmental trajectories of cortical thickness and surface area. *NeuroImage*, 87, 120-126.
Wierenga, L. M., et al. (2022). Greater male than female variability in regional brain structure across the lifespan. *Human Brain Mapping, 43*(1), 470-499.
Williams, C. K., & Barber, D. (1998). Bayesian classification with Gaussian processes. *IEEE Transactions on Pattern Analysis Machine Intelligence, 20*(12), 1342-1351.
Wolfers, T., et al. (2020). Individual differences v. the average patient: mapping the heterogeneity in ADHD using normative models. *Psychological Medicine, 50*(2), 314-323.
Wolfers, T., et al. (2018). Mapping the heterogeneous phenotype of schizophrenia and bipolar disorder using normative models. *JAMA Psychiatry, 75*(11), 1146-1155.
Zabihi, M., et al. (2020). Fractionating autism based on neuroanatomical normative modeling. *Translational psychiatry, 10*(1), 1-10.
Zabihi, M., et al. (2019). Dissecting the heterogeneous cortical anatomy of autism spectrum disorder using normative models. *Biological Psychiatry: Cognitive Neuroscience, 4*(6), 567-578.

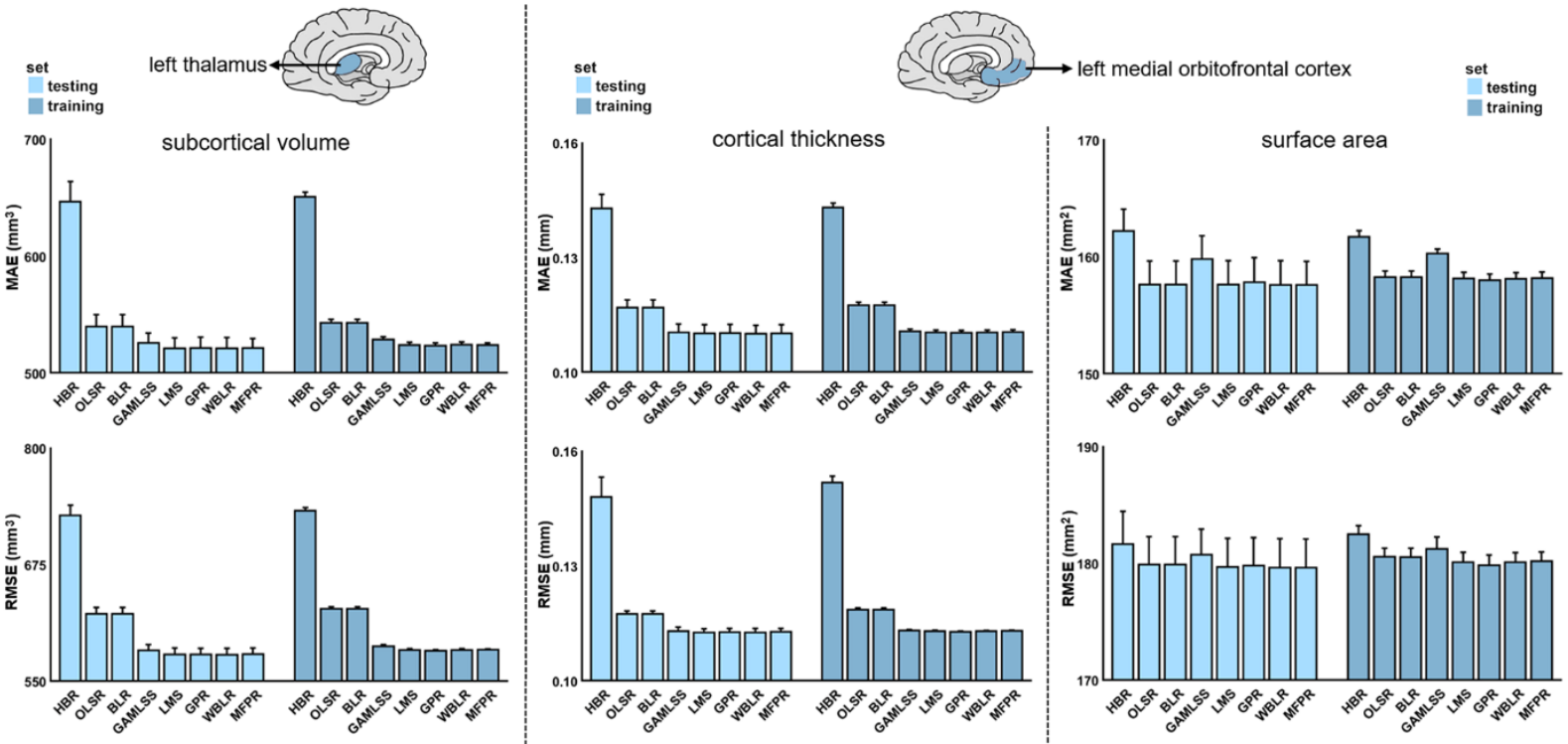
6. FIGURE CAPTIONS

Figure 1. Flowchart of normative model optimization



(1) The study sample was stratified by sex and split into training (80%) and testing (20%) datasets, followed by outlier removal, and mean-centering. (2) Normative models were generated using 8 different algorithms and compared in terms of accuracy and computational efficiency; (3) Explanatory variables were added to the optimal algorithm to identify the appropriate combination for optimal model performance.
ICV=intracranial volume; IQR=interquartile range

Figure 2. Illustrative Examples of Comparative Algorithm Performance



algorithm	HBR	OLSR	BLR	GAMLSS	LMS	GPR	WBLR	MFPR
CPU time (sec.)	345.87 ±403.58	0.01 ±0.001	6.65 ±0.43	4.65 ±0.29	4.66 ±0.17	2920.90 ±26.15	1.00 ±0.03	0.42 ±0.02

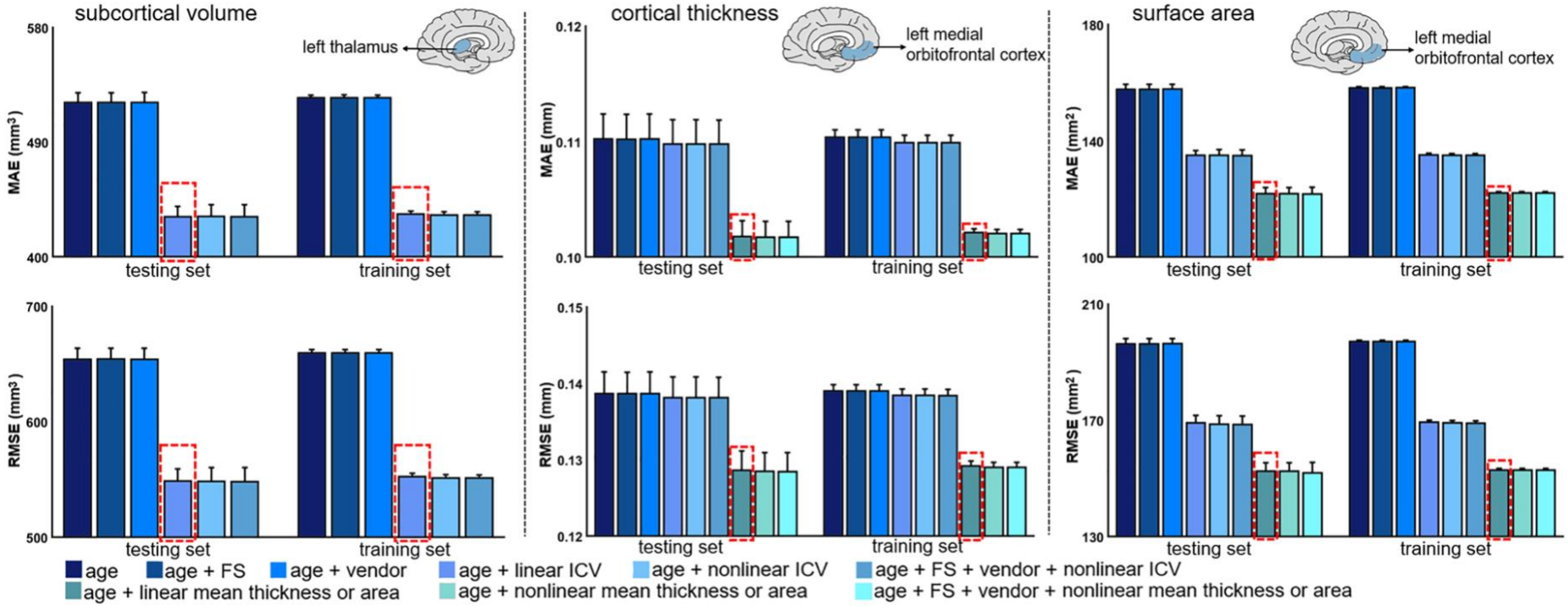
algorithm	HBR	OLSR	BLR	GAMLSS	LMS	GPR	WBLR	MFPR
CPU time (sec.)	52.68 ±8.74	0.01 ±0.001	4.99 ±0.13	2.12 ±0.19	3.26 ±0.07	2858.70 ±23.09	0.88 ±0.10	0.52 ±0.06

algorithm	HBR	OLSR	BLR	GAMLSS	LMS	GPR	WBLR	MFPR
CPU time (sec.)	278.26 ±39.59	0.01 ±0.001	5.34 ±0.42	2.80 ±0.06	2.87 ±0.06	2543.72 ±11.82	2.23 ±0.07	0.41 ±0.01

Algorithm performance for each regional morphometric measure was assessed separately in females (for results in males, see Supplementary Figure S2) using the mean absolute error (MAE), the root mean square error (RMSE), and the central processing unit (CPU) time. Details of all the models are presented in Supplementary Tables S2-S7 and Supplementary Figures S3 and S4. The pattern identified was the same across all region-specific models and in both sexes. The MAE, RSME, and CPU times of the models for left thalamic volume (left panel), the left medial orbitofrontal cortical thickness (middle panel) and surface area (right panel) as exemplars.

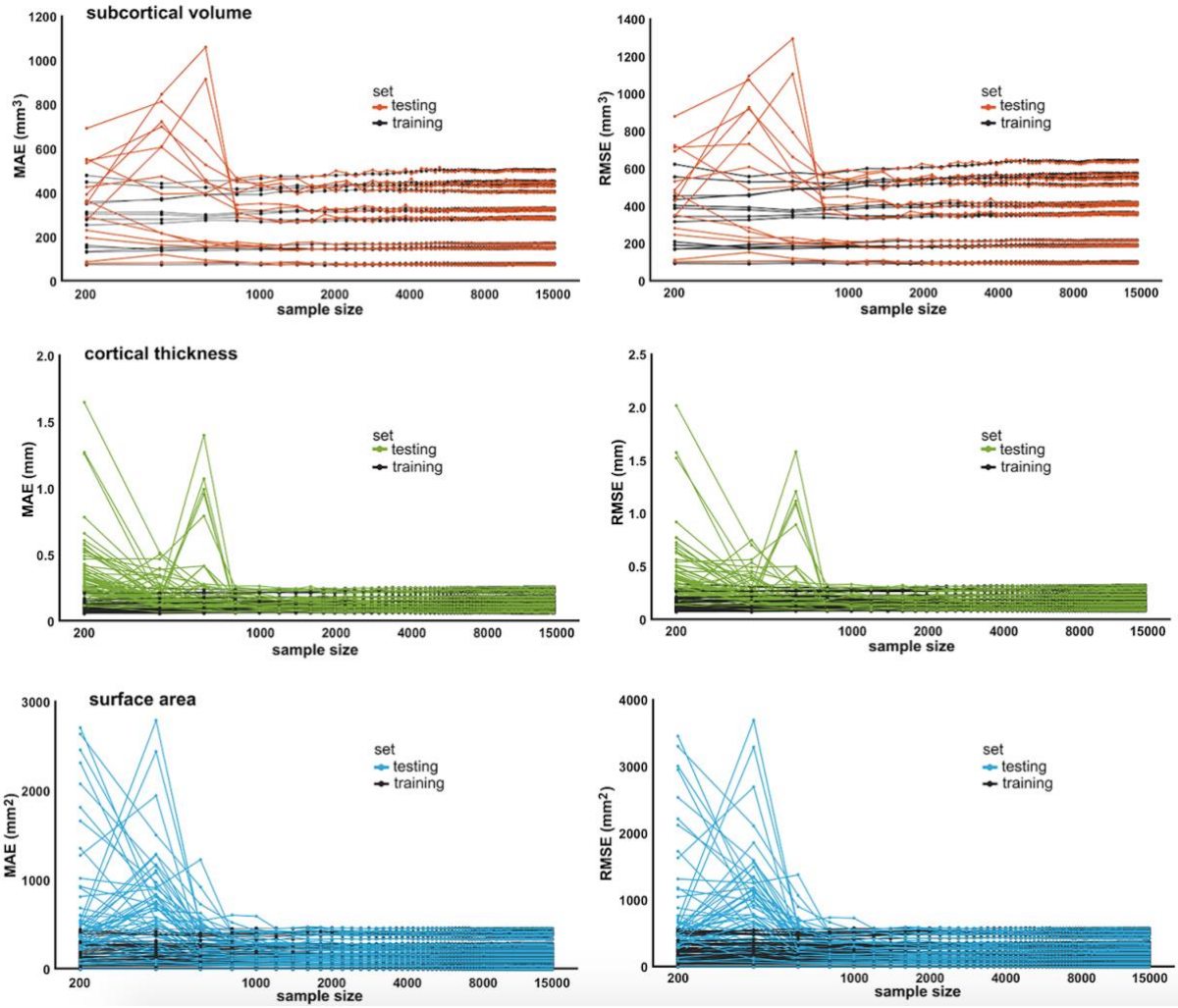
HBR=Hierarchical Bayesian Regression; OLSR=Ordinary Least Squares Regression; BLR=Bayesian Linear Regression; GAMLSS=Generalized Additive Models for Location, Scale and Shape; LMS=Lambda (λ), Mu (μ), Sigma (σ) Quantile Regression; GPR=Gaussian Process Regression; WBLR=Warped Bayesian Linear Regression; MFPR=Multivariable Fractional Polynomial Regression.

Figure 3. Illustrative examples of the performance of MFPR-derived models as a function of explanatory variables



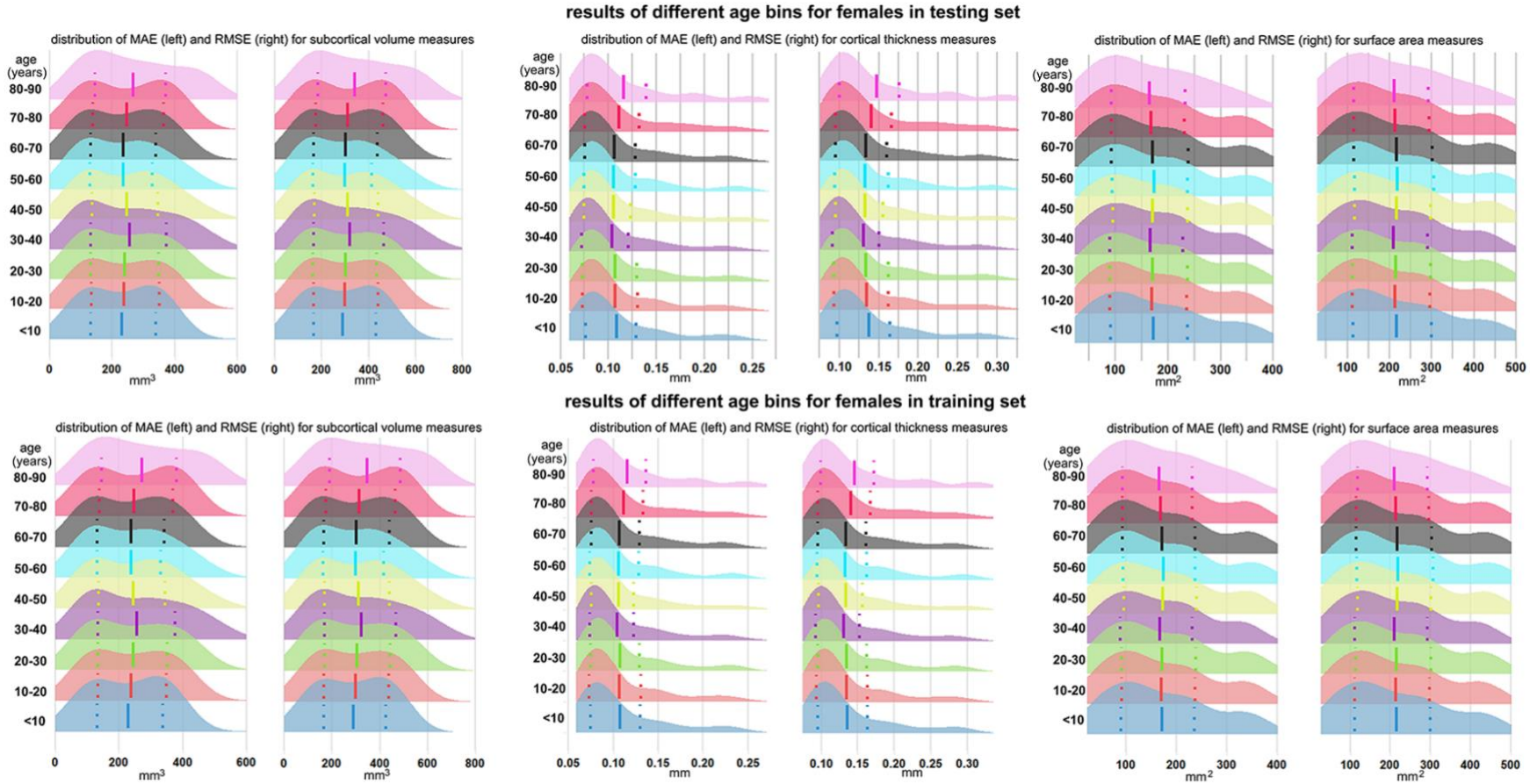
For each regional morphometric measure, sex-specific models derived from Multivariable Fractional Polynomial Regression (MFPR) were trained and tested using 9 different covariate combinations that included linear and non-linear effects of age FreeSurfer version (FS), scanner vendor, intracranial volume (ICV) and global estimates of mean cortical thickness or area. The Mean Absolute Error (MAE) and Root Mean Square Error (RMSE) of all the models in males and females are shown in Supplemental Tables S8 to S13 and Supplementary Figures S6 and S7. In both sexes, the pattern identified was identical for all region-specific models. The MAE and RSME of the models for left thalamic volume (left panel), the left medial orbitofrontal cortical thickness (*middle panel*), and surface area (*right panel*) as exemplars here for females and in Supplementary Figure S5 for males. The optimal variable combination is marked with a dashed frame.

Figure 4. Performance of region-specific MFPR-derived models as a function of sample size



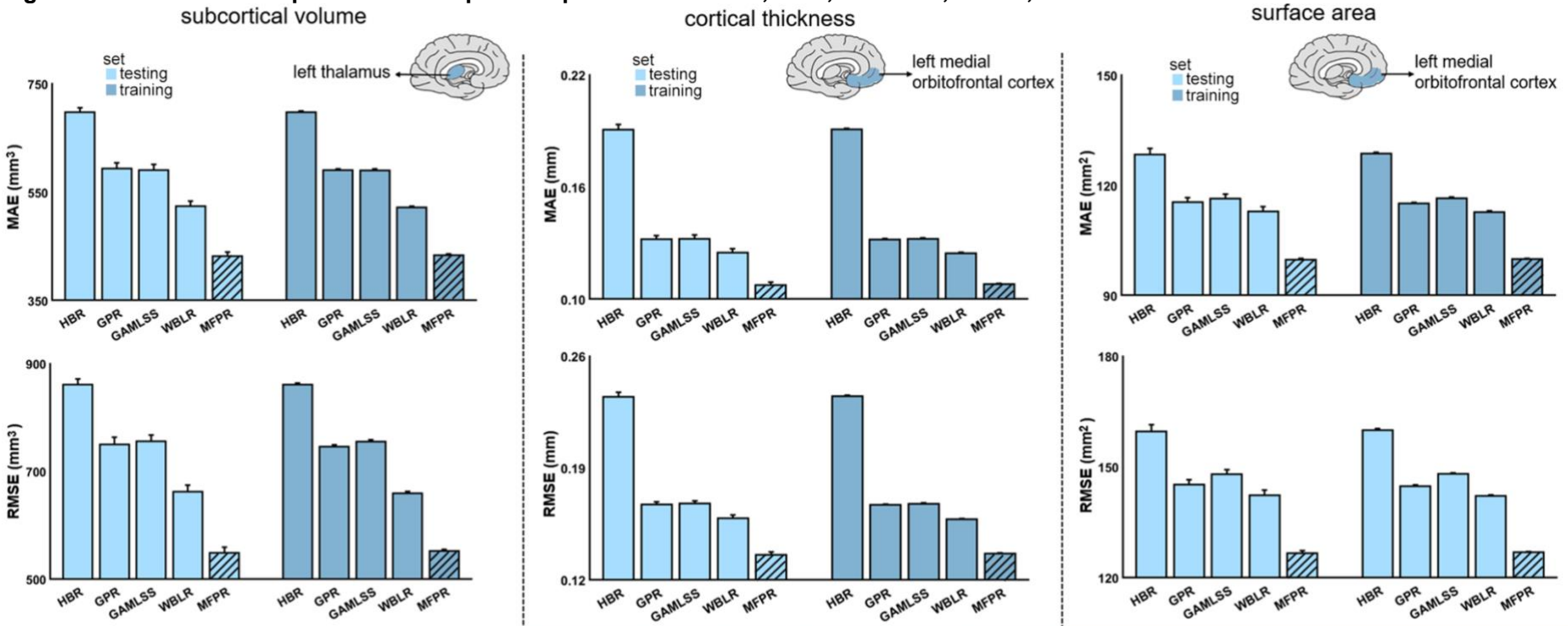
Models for each regional morphometric measure were estimated in random sex-specific subsets of 200 to 15,000 participants, at increments of 200, generated from the study sample. Each line represents the values of the mean absolute error (MAE) or root mean square error (RMSE) derived from the optimized Multivariable Fractional Polynomial Regression (MFPR) models of each regional morphometric measure as a function of sample size. The pattern identified was identical in box sexes. The data from females are shown here and in Supplementary Figure S8 for males.

Figure 5 Performance of region-specific models in distinct age groups



Sex and region-specific models for all morphometric measures for different age groups were estimated using by partitioning the sex-specific training and testing subsets of the study sample into 9 age bins (9 age-bins (i.e., age≤10 years; 10<age≤20 years; 20<age≤30 years; 30<age≤40 years; 40<age≤50 years; 50<age≤60 years; 60<age≤70 years; 70<age≤80 years; 80<age≤90 years). Details are provided in Supplementary Tables S14 to S19. The pattern was identical in both sexes. The figure presents the distribution of the Mean Absolute Error (MAE) and the Root Mean Square Error (RMSE) across all region-specific models in females in the training (upper panel) and test subset (lower panel). The results in males are presented in Supplementary Figure S9.

Figure 6. Illustrative examples of the comparative performance HBR, GPR, GAMLSS, WBLR, and MFPR-derived models.



algorithm	HBR	GPR	GAMLSS	WBLR	MFPR
CPU time(sec.)	273.69 ± 68.49	1686.60 ± 46.54	5.47 ± 0.22	1.58 ± 0.41	0.42 ± 0.02

algorithm	HBR	GPR	GAMLSS	WBLR	MFPR
CPU time (sec.)	2090.99 ± 436.84	11423.1 ± 1986.09	3.14 ± 0.28	1.32 ± 0.27	0.52 ± 0.06

algorithm	HBR	GPR	GAMLSS	WBLR	MFPR
CPU time(sec.)	3434.79 ± 644.21	1226.76 ± 43.24	4.04 ± 0.17	1.12 ± 0.03	0.41 ± 0.01

Region-specific models with the optimized covariate combination were estimated in males and females separately using Hierarchical Bayesian Regression (HBR), Gaussian Process Regression (GPR); Generalized Additive Models for Location, Scale, and Shape (GAMLSS), Warped Bayesian Linear Regression (WBLR), and Multivariable Fractional Polynomial Regression (MFPR). Model performance was assessed in terms of mean absolute error (MAE), root mean square error (RMSE), and central processing unit (CPU). In both sexes and across all regional morphometric measures, MFPR-derived models outperformed all others at $P_{FDR}<0.05$ (Supplemental Tables S20 to S25 and Supplementary Figures S11 and S12). The MAE, RSME, and CPU time of the models for left thalamic volume (left panel), the left medial orbitofrontal cortical thickness (middle panel), and surface area (right panel) as exemplars here for females and in Supplementary Figure S10 for males.