NeuroImage 245 (2021) 118715

Contents lists available at ScienceDirect

NeuroImage

journal homepage: www.elsevier.com/locate/neuroimage

Warped Bayesian linear regression for normative modelling of big data

Charlotte J. Fraza[a,b,*], Richard Dinga[a], Christian F. Beckmann[a,b,d], Andre F. Marquand[a,b,c]

[a] Donders Centre for Cognitive Neuroimaging, Donders Institute for Brain, Cognition and Behaviour, Radboud University, Kapittelweg 29, Nijmegen 6525 EN, the Netherlands
[b] Department of Cognitive Neuroscience, Radboud University Medical Centre, Nijmegen, the Netherlands
[c] Department of Neuroimaging, Centre for Neuroimaging Sciences, Institute of Psychiatry, King's College London, London, UK
[d] Centre for Functional MRI of the Brain (FMRIB), Nuffield Department of Clinical Neurosciences, Wellcome Centre for Integrative Neuroimaging, University of Oxford, Oxford, UK

ARTICLE INFO

Keywords:
Machine learning
UK Biobank
Big data
Bayesian linear regression
Normative modelling

ABSTRACT

Normative modelling is becoming more popular in neuroimaging due to its ability to make predictions of deviation from a normal trajectory at the level of individual participants. It allows the user to model the distribution of several neuroimaging modalities, giving an estimation for the mean and centiles of variation. With the increase in the availability of big data in neuroimaging, there is a need to scale normative modelling to big data sets. However, the scaling of normative models has come with several challenges.

So far, most normative modelling approaches used Gaussian process regression, and although suitable for smaller datasets (up to a few thousand participants) it does not scale well to the large cohorts currently available and being acquired. Furthermore, most neuroimaging modelling methods that are available assume the predictive distribution to be Gaussian in shape. However, deviations from Gaussianity can be frequently found, which may lead to incorrect inferences, particularly in the outer centiles of the distribution. In normative modelling, we use the centiles to give an estimation of the deviation of a particular participant from the 'normal' trend. Therefore, especially in normative modelling, the correct estimation of the outer centiles is of utmost importance, which is also where data are sparsest.

Here, we present a novel framework based on Bayesian linear regression with likelihood warping that allows us to address these problems, that is, to correctly model non-Gaussian predictive distributions and scale normative modelling elegantly to big data cohorts. In addition, this method provides likelihood-based statistics, which are useful for model selection.

To evaluate this framework, we use a range of neuroimaging-derived measures from the UK Biobank study, including image-derived phenotypes (IDPs) and whole-brain voxel-wise measures derived from diffusion tensor imaging. We show good computational scaling and improved accuracy of the warped BLR for certain IDPs and voxels if there was a deviation from normality of these parameters in their residuals.

The present results indicate the advantage of a warped BLR in terms of; computational scalability and the flexibility to incorporate non-linearity and non-Gaussianity of the data, giving a wider range of neuroimaging datasets that can be correctly modelled.

1. Introduction

Data from large-scale cohorts have become more widely available in neuroimaging (UK Biobank, ENIGMA, ABCD study, PNC, among others) (Casey et al., 2018; Satterthwaite et al., 2016; Sudlow et al., 2015; Thompson et al., 2014). We can use these data for modelling normal brain development, to estimate quantitive brain-behaviour mappings and capture deviations from such models to derive neurobiological markers of different psychiatric disorders. These neurobiological markers could move us closer towards individualized and precision medicine (Insel and Cuthbert, 2015). Until now, the neurobiological markers for psychiatric disorders have been mostly developed with studies that used classifiers trained in a case-control setting. Interestingly, yet counter-intuitively, an increase in sample size has shown to reduce the accuracy of classifying cases from controls for psychiatric disorders (Wolfers et al., 2015). One of the main reasons for this decrease in accuracy has been posed to be the heterogeneity in the participants both biologically and behaviorally, which can only fully be captured by a

* Corresponding author at: Donders Centre for Cognitive Neuroimaging, Donders Institute for Brain, Cognition and Behaviour, Radboud University, Kapittelweg 29, 6525 EN Nijmegen, the Netherlands.
E-mail address: charlotte.fraza@donders.ru.nl (C.J. Fraza).

https://doi.org/10.1016/j.neuroimage.2021.118715.
Received 14 May 2021; Received in revised form 4 November 2021; Accepted 5 November 2021
Available online 17 November 2021.

large data set (Wolfers et al., 2015). Normative modelling is an emerging method used to understand this heterogeneity in the population. Similar to growth charts in pediatric medicine, which describe the distribution of height or weight of children according to their age and sex, normative models can be used to model the distribution of neuroimaging derived phenotypes in a population, typically in terms of the mean and centiles of variation (Marquand et al., 2019), according to age, gender, or other demographic or clinical variables (Marquand et al., 2016a). The deviations from this normative range can be quantified statistically, for example as Z-scores and subsequently can be linked to several psychiatric disorders (Kaufmann et al., 2019; Lv et al., 2020; Marquand et al., 2019; 2016b; Wolfers et al., 2018; Zabihi et al., 2019).

There remain a variety of challenges in performing normative modelling on big neuroimaging data. First of all, most normative models assume the modelled distribution is Gaussian. However, distributions diverging from Gaussianity are frequently found in specific neuroimaging modalities, for example in diffusion MRI. These non-Gaussian signals cannot be accounted for using a standard normative model based on Gaussian process regression. We argue that modelling non-Gaussianity is important in general and is frequently overlooked by the neuroimaging community in that most regression methods used in practice –often implicitly– assume Gaussian residuals. Furthermore, normative models have been mainly developed using Gaussian process regression (Kia and Marquand, 2018). Gaussian process regression is flexible and accurate, but a drawback is its computational complexity, which is governed by the need to compute the exact inverse of the covariance matrix. This inversion scales poorly with an increase in data points (Rasmussen and Williams, 2005). Therefore, using these models on large datasets requires extensive computational power and is often not feasible (typically beyond a few thousand subjects). Thus, there is a need to develop methods that can flexibly handle the computational demand and non-Gaussianity of big data sets.

In this paper, we propose a framework based on Bayesian linear regression (BLR) to address these challenges. We introduce an extension of the BLR method for accurately modelling non-Gaussian distributions using a likelihood warping technique, giving a warped BLR model. The new framework has several benefits over previously used methods: (i) A BLR model can use a linear combination of non-linear basis functions (such as B-splines) which can be considered to provide a low-rank approximation of the Gaussian process regression models (Huertas et al., 2017). However, the BLR model has considerably better computational scaling, since the complexity of the model is fixed according to a set of basis functions. Therefore, the model can be scaled much more easily to large datasets. Also, similar to Gaussian process regression models, a set of model coefficients can be estimated that can easily be shared without the need to share the data (e.g. to compute a cross-covariance matrix for new data points), thus making it easier to make predictions on new datasets. (ii) The non-Gaussianity of the residuals can be modelled by the flexible warping of the Gaussian function, which gives more options to model different types of neuroimaging data that cannot be accurately modelled using a standard BLR. (iii) Efficient model selection criteria are naturally defined for the warped BLR through the marginal likelihood and can be calculated in closed form. The marginal likelihood gives a balance between model complexity and model fit. This can aid in choosing the optimal model for normative range estimates relative to a specific imaging modality.

We demonstrate this model by testing it on different types of neuroimaging data derived from the UK Biobank dataset. We consider UK biobank to be the best choice for the validation of this method since it is by far the largest dataset available from a single population, with strong protocols to reduce the number of different noise sources. The UK Biobank dataset has several magnetic resonance imaging (MRI) imaging modalities, including structural and functional data. With over 40,000 participants' MRI data from subjects with the age range 40 to 80 years old, this provides a rich set of different neuroimaging data and defines a benchmark for future population-based studies. In this work, we will present the warping function and recommend how to use it for several data modalities. First, we give an illustrative example using image-derived phenotypes (IDPs), which are convenient and widely used summary measures of brain function and structure (Alfaro-Almagro et al., 2018). Specifically, we will show a detailed example of estimating a normative model for white matter hyperintensities (WMHs). WMHs have been shown before to demonstrate quite non-Gaussian behaviour (Habes et al., 2020), and are therefore a good example where the warped BLR could be preferred over the BLR. Second, we show the scalability of this method by performing a whole-brain analysis using diffusion tensor imaging (DTI) data. We use DTI data, as there are large associations with age and we expect certain non-linear and non-Gaussian trends in the data (Cox et al., 2016).

Finally, we want to evaluate the link between brain imaging abnormality scores and behaviour. We always want to correlate brain data back to behaviour, as one could fit a perfect model to any type of neuroimaging data, but then if it does not say anything about behaviour it would not be beneficial for understanding different psychological disorders. With this correlation between Z-scores and cognitive scores, we want to give a first indication of how deviations from normal development for a single individual is related to a deviation in their behavioural, in this case, cognitive scores. Therefore, deviations from normal brain functioning are associated with cognitive functioning. The deviations are captured by Z-scores, which are shown to correlate with measures of intelligence in the UK Biobank dataset, such as; numerical memory, reaction time and visual memory.

This paper is partially methodological and partially applied in that we aim to introduce and evaluate a method specifically for neuroimaging. Our main contributions are: (i) a new comprehensive framework for big data normative modelling; (ii) the introduction of the novel methodological approach for modelling non-Gaussian response variables; (iii) an extensive evaluation of this framework on the UK Biobank cohort. Ultimately, we hope this paper will give deeper insight into the application of normative models on different types of neuroimaging modalities.

2. Materials and methods

2.1. Sample

All the data used came from the UK Biobank imaging dataset (Sudlow et al., 2015). Full details on the design of the study and the preprocessing steps can be found in papers (Alfaro-Almagro et al., 2018; Miller et al., 2016). Briefly, the data used contains 20,083 participants of the 2017 release and additional longitudinal data of around 5000 subjects of the 2020 release. The participants were between 40 and 80 years of age, with around 47% males, for the exact distribution of the covariates and number of participants currently in our dataset, see supplementary C.21.

In this study, two types of analyses were performed using different datasets. For the first analysis, a dataset containing IDPs was used. For consistency with existing work, the IDPs were processed using FUNPACK (McCarthy, 2020), which is an automatic normalisation, parsing and cleaning kit, developed at the Wellcome Centre for Integrative Neuroimaging. The IDPs include three main imaging modalities: structural, functional and diffusion MR data. Among these IDPs, there are very gross measures, such as the total amount of brain volume, to more detailed measurements, such as the connectivity between two brain regions. In total 819 neuroimaging IDPs were used for subsequent anal-

ysis, see E.1 for the list of IDPs used. Furthermore, we also tested our model on another set of IDPs; 150 FreeSurfer measures, which were preprocessed with FreeSurfer v6.1.0, using a $T2$-weighted image where available, see E.1 for the list of the FreeSurfer measures used.

For the second analysis, a whole-brain model was built, using voxel-wise fractional anisotropy (FA) and mean diffusivity (MD) measures. The data were processed using the UKB pipelines; including the DTI fitting tool DTIFIT and a tract-based spatial statistics (TBSS) style analysis, which gave us the skeletonised WM template. In total, 15,495 participants with dMRI-scans passed the quality control applied by the UK Biobank (Alfaro-Almagro et al., 2018). Afterwards, we added two extra exclusion criteria. First, participants were removed if their Z-score of the discrepancy between the dMRI image and the structural T1 image was higher than three, based on data-field 25,731 in the UK Biobank. Second, participants were removed if their Z-score of the number of outlier slices was higher than three, which is a reflection of the movement of the participant during the scan, based on data-filed 25746-2.0 in the UK Biobank. For the covariates we used age, gender and dummy coded site variables.

2.2. Cognitive data

We used the cognitive phenotypes that were extracted from the UK Biobank using FUNPACK (McCarthy, 2020) to evaluate the cognitive associations with the deviations from the normative model. These measures are derived from the 13 cognitive tests present in the UK Biobank, see the UKB showcase. The tests were administered using a touch-screen questionnaire and included numerical memory, reaction time, fluid intelligence, visual memory, prospective memory, executive function, declarative memory and non-verbal reasoning (Fawns-Ritchie and Deary, 2020). For full details on the different cognitive tests applied in UK Biobank see Lyall et al. (2016). An overview of all the measures used in this study is presented in the supplementary H.6. For the full brain model, we will reduce the dimensions of the cognitive phenotype in terms of the first eigenvector using principal component analysis (PCA), which has been shown to be correlated to the 'general' factor of cognitive ability or the 'g-factor' (Nave et al., 2019).

2.3. Normative model formulation

We use a flexible normative modelling framework to model different types of neuroimaging data. We have N subjects with brain data $\{\mathbf{y}_n\}_{n=1}^{N}$, each of dimension D (e.g. the number of voxels or IDPs) and acquired from one of S different scanning sites. We use $\mathbf{Y}$ to denote an $N \times D$ matrix containing these variables, where y_{nd} denotes the nth subject and dth neuroimaging variable. In the case of the IDPs, the dth neuroimaging variable is one IDP. In the case of the whole-brain model, the dth neuroimaging variable is one voxel. Since the neuroimaging variables are estimated separately here, we simplify the notation by using $\mathbf{y}$ to denote the vector of observations from a single variable (one IDP or one voxel for all the subjects) and y_n for a single observation. In general, we want to predict the distribution of the value for dependent variable ($\mathbf{y}$), from a set of covariates $\{\mathbf{x}_n\}_{n=1}^{N}$ (e.g. age, gender or site), the independent variables. In this paper, we adopt a straightforward approach to model nonlinear relationships, by applying a basis expansion to the independent variables. A common approach is to use polynomial functions, but these can be a poor choice, as they can induce global curvature (Fjell et al., 2010). Here we apply a common B-spline basis expansion (specifically, cubic splines with three evenly spaced knot points), although other approaches are also possible. We denote this expansion by $\phi(\mathbf{x})$, with $\mathbf{\Phi}$ an $N \times K$ matrix containing the basis expansion for all subjects. In the applied model, y is assumed to be the result of a linear combination of the B-spline basis function transformation plus a noise term:

$$y = \mathbf{w}^T \phi(\mathbf{x}) + \epsilon_s \tag{1}$$

With $\mathbf{w}$ the estimated vector of weights and $\epsilon_s = \mathcal{N}(0, \beta_s^{-1})$ a Gaussian noise distribution for site s, with mean zero and a noise precision term β_s (i.e. the inverse variance). All the noise precision terms from the different sites will be combined in a vector $\boldsymbol{\beta}$ and the site precision matrix $\mathbf{\Lambda}_{\boldsymbol{\beta}}$, which has $\boldsymbol{\beta}$ along the leading diagonal and is the inverse of the site covariance matrix $\mathbf{\Lambda}_{\boldsymbol{\beta}} = \mathbf{\Sigma}_{\boldsymbol{\beta}}{}^{-1}$. Note that we allow the noise precision to vary across sites in order to accommodate inter-site variation along with site-specific intercepts (i.e. dummy coded site regressors in the design matrix). We added the dummy coded site regressors to the resulting basis expansion matrix of the B-spline. This makes it a random intercept model, where we seek a single slope to estimate a single population-level developmental trajectory and there is, in this case, no reason to believe the different sites should differ in their developmental or ageing trajectory. We have shown previously that this approach provides an effective way to accommodate site effects in normative modelling (Bayer et al., 2021; Kia et al., 2020).

Following similar derivations as given by Huertas et al. (2017), we consider a BLR model, placing a Gaussian prior over our model parameters $p(\mathbf{w}|\boldsymbol{\alpha}) = \mathcal{N}(\mathbf{w}|0, \mathbf{\Lambda}_{\boldsymbol{\alpha}}{}^{-1})$, with $\boldsymbol{\alpha}$ the hyper-parameters that the weights depend on. The Gaussian prior is assumed to have a mean zero and a precision matrix $\mathbf{\Lambda}_{\boldsymbol{\alpha}}$, with the precision matrix the inverse of the covariance matrix $\mathbf{\Sigma}_{\boldsymbol{\alpha}} = \mathbf{\Lambda}_{\boldsymbol{\alpha}}{}^{-1}$. As shown in Huertas et al. (2017), $\mathbf{\Lambda}_{\boldsymbol{\alpha}}$ can be quite general, but here we use a simple isotropic precision matrix $\mathbf{\Lambda}_{\boldsymbol{\alpha}} = \alpha\mathbf{I}$. The Gaussian prior choice allows us to compute the posterior distribution of $\mathbf{w}$ in a closed form:

$$p(\mathbf{w}|\mathbf{y}, \mathbf{\Phi}, \boldsymbol{\alpha}, \boldsymbol{\beta}) = \frac{\text{likelihood} \times \text{prior}}{\text{marginal likelihood}} = \frac{\prod_n p(y_n|\mathbf{\Phi}, \boldsymbol{\beta}, \mathbf{w}) p(\mathbf{w}|\boldsymbol{\alpha})}{p(\mathbf{y}|\mathbf{\Phi}, \boldsymbol{\alpha}, \boldsymbol{\beta})} \tag{2}$$

The posterior for each subject can then be found using the standard derivations of the posterior (Bishop, 2006):

$$\begin{aligned} p(\mathbf{w}|\mathbf{y}, \mathbf{\Phi}, \boldsymbol{\alpha}, \boldsymbol{\beta}) &= \mathcal{N}(\mathbf{w}|\bar{\mathbf{w}}, \mathbf{A}^{-1}) \\ \mathbf{A} &= \mathbf{\Phi}^T \mathbf{\Lambda}_{\boldsymbol{\beta}} \mathbf{\Phi} + \mathbf{\Lambda}_{\boldsymbol{\alpha}} \\ \bar{\mathbf{w}} &= \mathbf{A}^{-1} \mathbf{\Phi}^T \mathbf{\Lambda}_{\boldsymbol{\beta}} \mathbf{y} \end{aligned} \tag{3}$$

We use a Type II maximum likelihood approach (i.e. empirical Bayes), optimizing the denominator of the posterior to find the optimal hyper-parameters $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$. This gives an automatic trade-off between model fit and model complexity. The marginal likelihood is maximized by minimizing the negative log likelihood (NLL):

$$\begin{aligned} \text{NLL} &= -log(p(\mathbf{y}|\boldsymbol{\alpha}, \boldsymbol{\beta})) \\ &= -log(\int p(\mathbf{y}|\mathbf{w}, \boldsymbol{\beta}) p(\mathbf{w}|\boldsymbol{\alpha}) d\mathbf{w}) \\ = -(\frac{N}{2} log|\mathbf{\Lambda}_{\boldsymbol{\beta}}| - \frac{ND}{2} log 2\pi - \frac{N}{2} log|\mathbf{\Lambda}_{\boldsymbol{\alpha}}| - \frac{N}{2} log|\mathbf{A}| \\ -\frac{1}{2} \sum_{n=1}^{N} (\mathbf{y_n} - \mathbf{\Phi}\bar{\mathbf{w}}_\mathbf{n})^T \mathbf{\Lambda}_{\boldsymbol{\beta}} (\mathbf{y_n} - \mathbf{\Phi}\bar{\mathbf{w}}_\mathbf{n}) - \bar{\mathbf{w}}_\mathbf{n}{}^T \mathbf{\Lambda}_{\boldsymbol{\alpha}} \bar{\mathbf{w}}_\mathbf{n}) \end{aligned} \tag{4}$$

The optimal hyper-parameters $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are often estimated using a conjugate gradient optimisation of the NLL, where the derivatives can be computed directly. However, here we used Powell's method to fit the hyper-parameters (Powell, 1964). Powell's method is a derivative-free method, which in this case is faster, because computing the derivatives of the marginal likelihood with respect to the hyper-parameters is computationally very expensive. Finally, the predictive distribution is given by:

$$\hat{y} = \mathcal{N}(\bar{\mathbf{w}}^T \phi(\mathbf{x}), \phi(\mathbf{x})^T \mathbf{A}^{-1} \phi(\mathbf{x}) + \beta_s^{-1}) \tag{5}$$

2.3.1. Likelihood warping

In order to model non-Gaussian error distributions, we employed a warped likelihood (Snelson et al., 2004). This involves applying a non-linear monotonic warping function φ_i to the input data during the model fit, with the index i indicating a different warping function (e.g. SinhArcsinh, Box-Cox etc.). This is similar to the classical statistical technique of variable transformation, but has the advantage that the parameters of

the transformation are optimised during model fitting, to provide the optimal mapping that ensures that model residuals have a Gaussian form. The warped functions are chosen such that they have a closed-form inverse and are differentiable, which has several benefits: first, non-Gaussian data can be mapped (i.e. warped) exactly to better match Gaussian modelling assumptions or the predictions can be warped back to the original non-Gaussian space; second, it allows inference, prediction and computation of error measures all in closed form; finally, we can construct compositions of functions from the invertible monotonic warping functions that can greatly improve the expressivity of the model in transforming non-Gaussian distributed data $\mathbf{y}$ to a Gaussian form, $\mathbf{z}$, where inference is straightforward (Rios and Tobar, 2019). This is done by applying a compositional warping function φ to the observations $\mathbf{y}$:

$$\varphi(.) = \varphi_i(\varphi_{i-1}(\dots(\varphi_2(\varphi_1(.)))\dots))$$
$$\mathbf{z} = \varphi(\mathbf{y};\boldsymbol{\gamma}) \tag{6}$$

With $\boldsymbol{\gamma}$ denoting the collection of the hyper-parameters of different warping functions. The warping transformation allows us to compute error measures in the warped space and to describe the deviations of subjects under a Gaussian error distribution in the form of pseudo Z statistics, even if the original data distribution is non-Gaussian.

The optimal hyper-parameters ($\boldsymbol{\alpha}, \boldsymbol{\beta}$ and $\boldsymbol{\gamma}$) are calculated by minimizing the warped NLL. We do not place a prior over γ, but estimate it from the data, making it an empirical Bayesian method. When $|\nabla\varphi(\mathbf{y})| > 0$, the warped NLL can be found by accounting for the change of variables in the probability density function (Rios and Tobar, 2019):

$$p_{\mathbf{y}}(\mathbf{y}) = p_{\mathbf{z}}(\varphi(\mathbf{y}))|\nabla\varphi(\mathbf{y})|$$

With $\nabla\varphi(.)$ the Jacobian of the transformation, which is diagonal and therefore we can simplify as a product of the individual terms:

$$p_{\mathbf{y}}(\mathbf{y}) = p_{\mathbf{z}}(\varphi(\mathbf{y}))\prod_{n=1}^{N}\frac{d\varphi(y_n)}{dy}$$

If we take the negative log of this equation the warped NLL will remain the same as Eq. (4), except for replacing the $\mathbf{y}$ by the transformed $\varphi(\mathbf{y})$ and the inclusion of the Jacobian term that takes the change of volume induced by the warping into account, thereby ensuring a valid probability measure (for details see Rios and Tobar, 2019):

$$\text{Warped NLL} = -log(p(\mathbf{y}|\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\gamma}))$$
$$= \text{NLL} - \sum_{n=1}^{N} log\frac{d\varphi(y_n)}{dy} \tag{7}$$

In order to further motivate the use of the BLR model, we provide a theoretical analysis of the computational complexity of the BLR model relative to Gaussian process regression in the appendix.

2.3.2. Warped composition function

Different elementary functions can be used to create the warped composition function φ. For this paper, we tested affine, Box-Cox and SinhArcsinh transformations and compositions of these transformations. We evaluate the BoxCox model because of its simplicity (i.e. having a single parameter) and because this transformation is quite common in the neuroimaging field as a preprocessing step. The SinArchsinh was chosen because it can model both skewness and kurtosis, which is equivalent to the SHASH distribution that shows good performance modelling non-Gaussianity in other fields, most notably in the estimation of growth charts for the World Health Organisation (Borghi et al., 2006), although we do not claim that this is the optimal choice, and we expect that for the types of relations evident in neuroimaging, it is possible that other basis expansions would also work relatively well (e.g. fractional polynomials).

$$\varphi_{\text{Affine}}(\mathbf{y};\boldsymbol{\gamma}) = a + b\mathbf{y}$$
$$\varphi_{\text{Box-Cox}}(\mathbf{y};\boldsymbol{\gamma}) = \frac{sgn(\mathbf{y})|\mathbf{y}|^{\lambda} - 1}{\lambda}$$
$$\varphi_{\text{SinhArcsinh}}(\mathbf{y};\boldsymbol{\gamma}) = sinh(b * arcsinh(\mathbf{y}) - a) \tag{8}$$

With $\boldsymbol{\gamma}$ the respective parameters of the different warping functions. For the SinhArcsinh warping we also applied a reparametrization (Jones and Pewsey, 2009), as this empirically gave more stable results:

$$\varphi_{\text{SinhArcsinh}}(\mathbf{y};\boldsymbol{\gamma}) = sinh(b * arcsinh(\mathbf{y}) + \epsilon * b)$$
$$a = -\epsilon * b$$

2.4. Model selection

We evaluate the models using different model selection criteria. First, we calculate the explained variance (EV) of the model per feature.

$$\text{EV} = 1 - \frac{Var(y - \hat{y})}{Var(y)} \tag{9}$$

Here y is the true value and $\hat{y}$ is the predicted values per feature. It is expected that the gain in fit for the warped BLR will be highly dependent on the flexibility of the model. Therefore, the Bayesian Information Criterion (BIC) is also considered in terms of the number of parameters induced by the parameterisations of non-linear warping functions used:

$$\text{BIC} = k * log(N) + 2 * NLL \tag{10}$$

The BIC penalises for model complexity. Here N denotes the number of participants in the training set, NLL the negative log-likelihood. k is the number of free parameters. Note that we use the marginalized form of the NLL, which already takes into account the number of estimated coefficients. Therefore, the BIC only needs to be corrected for the added complexity of the degrees of freedom of the model (i.e. the parameters that are not integrated out). For the standard BLR this is two, one for the precision over the weights and one for the precision over the noise (α and β respectively). For the warped SinhArcsinh BLR two extra degrees of freedom are added for the shape parameters (a and b). The BIC gives a good trade-off between the extra flexibility found in the warped BLR model and the better fit of the model. We also consider the mean standardized log-likelihood (MSLL) as a third model criterion. Per feature, this is calculated as:

$$\text{MSLL}_{\text{d}} = \frac{1}{N}\sum_{n=1}^{N}(\frac{1}{2}log(2\pi\sigma_*^2) + \frac{(y_{nd} - \hat{y}_{nd})^2}{2\sigma_*^2} - \frac{1}{2}log(2\pi\sigma_{dt}^2) - \frac{(y_{nd} - y_{dt})^2}{2\sigma_{dt}^2}) \tag{11}$$

Where N the number of subjects, σ_*^2 the sum of the prediction variance and the noise variance, $\hat{y}_{nd}$ the predicted mean and y_{nd} the true response, σ_{dt} and y_{dt} are the variance and mean for the training data of feature d. We compute the MSLL in the warped space for the warped BLR, since the targets may have a skewed distribution. The MSLL takes into account the mean error and the estimated prediction variance. Lastly, we consider the skewness and kurtosis for these models in the form of a QQ-plot of the residuals, as the benefit of the warped model is not necessarily a higher EV or lower MSLL, but mostly a better fit in the outer centiles, which is better visualized by looking at the residuals.

2.5. Deviance scores and correlation to cognitive phenotypes

We want to find a statistical estimate of how much each participant deviates from the normal range. This is done by computing a Z-score for each subject n, also denoting explicitly the dependence on each voxel or IDP d:

$$z_{nd} = \frac{y_{nd} - \hat{y}_{nd}}{\sqrt{\sigma_d^2 + (\sigma_*^2)_d}} \tag{12}$$

With, $\hat{y}_{nd}$ the predicted mean and y_{nd} the true response. Normalized by $\sigma_d^2 = (\beta_s^{-1})_d$ the estimated noise variance (i.e. reflecting variation in the data) and $(\sigma_*^2)_d = \phi(\mathbf{x})^T\mathbf{A}_d^{-1}\phi(\mathbf{x})$ the variance attributable to modelling uncertainty for the d-th voxel. For the warped statistic, we compute the Z-scores in the warped (i.e. Gaussian) space. The true response

variables are warped to the Gaussianised space to ensure the underlying assumption of normality is satisfied by the construction of the warping functions.

Afterwards, to ensure our model can also be applied for behavioural and clinical estimations, we look at the correlations between the Z-scores from the IDPs and the whole-brain analysis, and the cognitive scores of the UK Biobank. For the IDPs, we directly correlate the deviations and the cognitive phenotypes through a Spearman correlation. For the whole-brain analysis, we first make a summary statistic of the Z-scores by calculating an extreme value distribution. We model the extreme value distribution by looking at the mean of the top 1% of the deviations across the whole-brain (Zabihi et al., 2019). The extreme value statistics give the largest deviations per subject from the normal pattern, which have shown to be strongly correlated to behaviour (Marquand et al., 2016; Zabihi et al., 2019). The ordering of the extreme values will remain also in the warped space, as we imposed the warping to be positive monotonic (see above). We summarise the cognitive phenotype in terms of the first eigenvector using PCA. Lastly, we compute the Spearman coefficient between the first principal component and the summary deviation score.

2.6. Simulation study

In order to understand the performance of the warped BLR method under controlled conditions, we perform several simulations and fit the BLR model and under different conditions (e.g. skewed data, Platykurtic or Leptokurtic noise distributions). These analyses are described and presented in the appendix.

3. Results

3.1. Performance of the warped Bayesian linear regression model for IDPs

All the statistical analyses were performed in Python version 3.8, using the PCNtoolkit. The BLR algorithm from the PCNtoolkit was chosen for all experiments. We considered age, binary gender and binary site ID within the covariance matrix. We used a standard BLR or we transformed the age covariate with a B-spline of order three with five evenly spaced knots. We applied the B-spline transformation to both the covariance matrix of the BLR (B-spline BLR) and the SinArcsinh warped BLR (B-spline warped BLR). The Powell method was selected for the optimizer. We randomly split the dataset into 50% training and 50% test and reported all the error metrics on the test set. In the PCNtoolbox, several warpings can be chosen depending on the imaging modality one wants to model. We tested several warping functions (affine, Box-Cox and SinhArcsinh) and compositions of these warping functions. Preliminary testing showed that the SinhArcsinh warping gave the best fit compared to the alternatives evaluated. In the appendix a few of the results from the simulation studies can be found. Therefore, in this paper, only the results of the SinhArcsinh warping are presented.

In Fig. 1, Bland-Altman plots are shown comparing the standard BLR and the B-spline BLR. The figure presents different model selection criteria: MSLL and BIC (EV can be seen in supplement figure D.22). The plots demonstrate that for most IDPs a non-linear B-spline BLR model performs better than a standard BLR. Indicating that non-linearity is a key component that should be accounted for when one wants to accurately estimate the outer centiles using normative modelling with neuroimaging data.

In Fig. 2, Bland-Altman plots are shown that compare the B-spline BLR and the B-spline warped BLR models for all IDPs, using the MSLL and BIC (EV can be seen in supplement figure D.22). We also plotted the difference in absolute values of the skewness and kurtosis. In Fig. 3, the same plots are shown for the FreeSurfer measures. We included them separately, as they were preprocessed separately (i.e. we did not use the IDPs provided by UK Biobank and instead ran the Freesurfer reconstructions manually). The plots show that for specific IDPs the warped BLR performs better than the non-warped BLR. When we examined these IDPs more closely, it was noted that they demonstrated distinct non-Gaussian behaviour. An example of such behaviour is given down below with the WMHs (white matter hyperintensities). In the supplementary table F.3, we provide a summary of some of the results for different IDPs that can help inform which neuroimaging modalities are best modelled with the warped BLR. For an indication of the effect sizes of the model selection criteria for the different model settings, see supplementary tables G.4 and G.5. Note also that the MSLL and EV do not clearly reflect differences in the shape of the predictive distribution. For example, for the IDPs, there is no average difference between the warped and non-warped model (Fig. 2 panel A and supp. fig. D.22 panel B), yet the warped model consistently yields a predictive distribution –and resultant Z-score distribution– that is less (or equivalently) skewed and kurtotic (Fig. 2 panels C and D).

In Figs. 4 and 5, we show the results of an illustrative analysis predicting WMH load across ageing to demonstrate how the performance of the B-spline warped BLR model compares to a B-spline BLR. The figures show the BLR and warped BLR results for WMHs at one-time point and the longitudinal data of two-time points. The results demonstrate that (i) the non-linearity of the data is sufficiently captured with a B-spline transformed BLR (ii) the WMHs show a distinctly non-Gaussian variance pattern, which is better predicted by the warped BLR. Thus, indicating that if the data has a non-Gaussian distribution for the residuals a B-spline warped BLR is preferred over a B-spline BLR.

3.1.1. Correlation deviance scores WMHs and cognitive phenotypes

We also wanted to correlate the warped BLR model output of the WMHs to behavioural variables to ensure that the model can be used for behavioural predictions. We loaded all cognitive phenotypes available in UK Biobank according to the FUNPACK categorization, including: reaction time, numeric memory, prospective memory etc. (for a full list of the cognitive phenotypes used, see the supplementary table H.6). We calculated the deviance Z-scores according to formula 12. Afterwards, we calculated the Spearman correlation between the cognitive phenotypes and the Z-scores. Numeric memory (ID: 4259, 'Digits entered correctly') was modestly but significantly correlated with the warped Z-scores: $\rho = -0.0331$, $p = 0.0262$. In other words, if a participant's WMH deviation from common development increases the number of correctly remembered digits drops.

Lastly, to illustrate the value of normative models in a longitudinal context, we tested for an association between change in WMHs and change in cognitive phenotypes of the longitudinal data to see if WMH load is correlated to cognitive decline. We performed a statistical Wilcoxon rank-sum test on the participants' cognitive phenotypes contrasting subjects that have a difference in the Z-scores > 0.5, which corresponds to a difference in half a standard deviation, versus the participants that do not. Intuitively, this contrasts individuals who are following an expected trajectory of ageing with those who deviate from such a trajectory. Highly significant associations were found with the reaction time (ID: 404, 'Duration to first press of snap-button in each round') $W = 5.5641$, $p < .001$ and with the Trail Making Test (ID: 6771, 'Errors before selecting correct item in alphanumeric path (trail #2)') $W = 8.3105$, $p < .001$. The results show an association between the change in cognition and the change in WMH deviance scores.

3.2. Scalability to a whole-brain voxelwise based analysis

For the follow-up analysis, we evaluated the warped BLR approach on a whole-brain level for two dMRI imaging modalities (FA and MD). The results of these two modalities were very similar and therefore we will only present the results for FA here. We separated the entire dataset into 80% training data and 20% testing data. First, we computed the time complexity per model fit (e.g. for one voxel) with a varying number of subjects using the B-spline BLR model setting and compared it to the Gaussian process regression setting, keeping the design matrix and

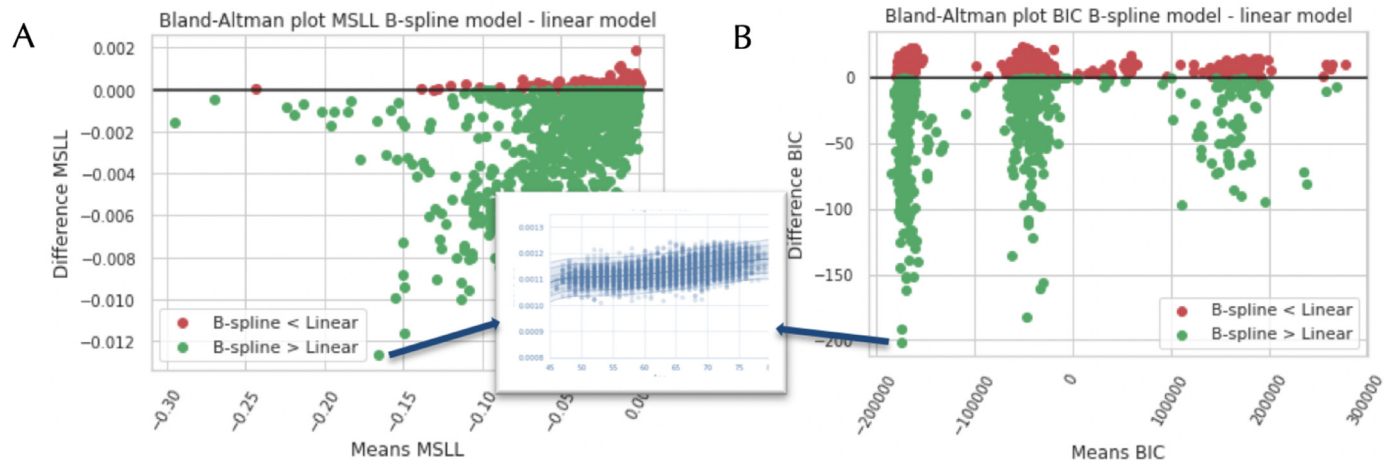

Fig. 1. Bland-Altman plots comparing the standard and B-spline Bayesian Linear Regression (BLR) models, using Image-Derived Phenotypes (IDPs). Each dot indicates one IDP. The models are compared according to the following model selection criteria: the Mean Standardized Log Loss (MSLL) (A) and the Bayesian Information Criteria (BIC) (B). The green colour indicates a better fit for the non-linear B-spline model compared to the linear model. We also plotted a zoomed-in view of the model fit for one of the IDPs.

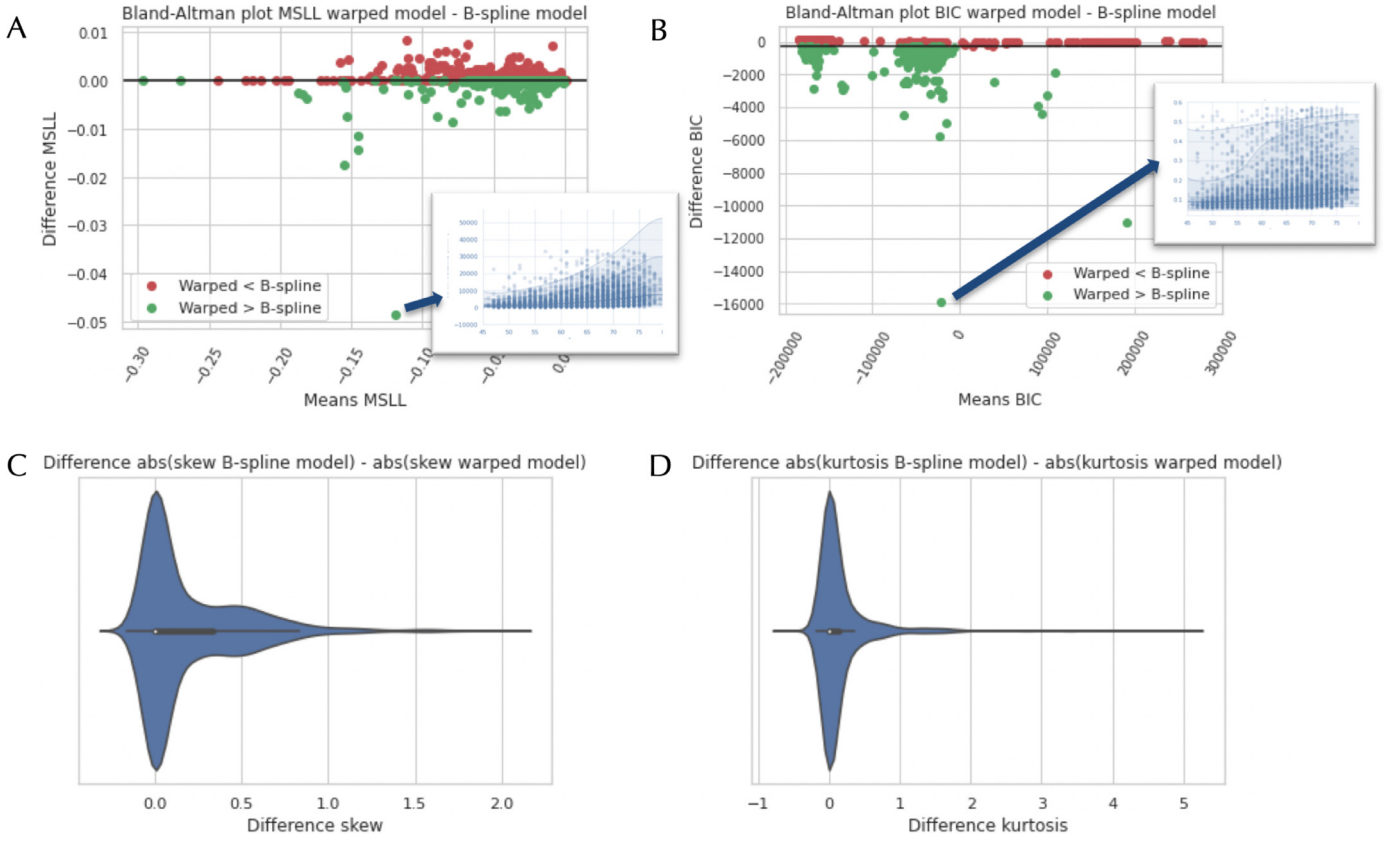

Fig. 2. Bland-Altman plots comparing the B-spline and B-spline warped Bayesian Linear Regression (BLR) models, using Image-Derived Phenotypes (IDPs). The models are compared according to the following model selection criteria: the Mean Standardized Log Loss (MSLL) (A) and the Bayesian Information Criteria (BIC) (B). The green colour indicates a better fit for the warped BLR model compared to the non-warped BLR model. We also plotted a zoomed-in view of the model fit for two of the IDPs. On images C and D, we show the difference in absolute values of the skewness and kurtosis between the B-spline and B-spline warped model. A more positive value indicates that the B-spline model had a higher skewness or kurtosis than the B-spline warped model.

the optimizer the same for both settings (Fig. 6). We see that the time to compute the model for a single voxel is lower for the BLR model compared to the Gaussian process regression model and this time difference increases with the addition of more subjects. This demonstrates the clear computational advantage of the BLR setting for the whole-brain analysis.

Afterwards, we tested different model settings for the imaging modalities including a standard BLR, B-spline BLR and a B-spline SinhArcsinh warped BLR. Fig. 7 shows the comparative results in a Bland-Altman plot for the FA dataset (which were similar for the MD dataset). The figure presents the EV, MSLL and the BIC for the B-spline BLR and the B-spline warped BLR. These results are consistent with the IDPs in

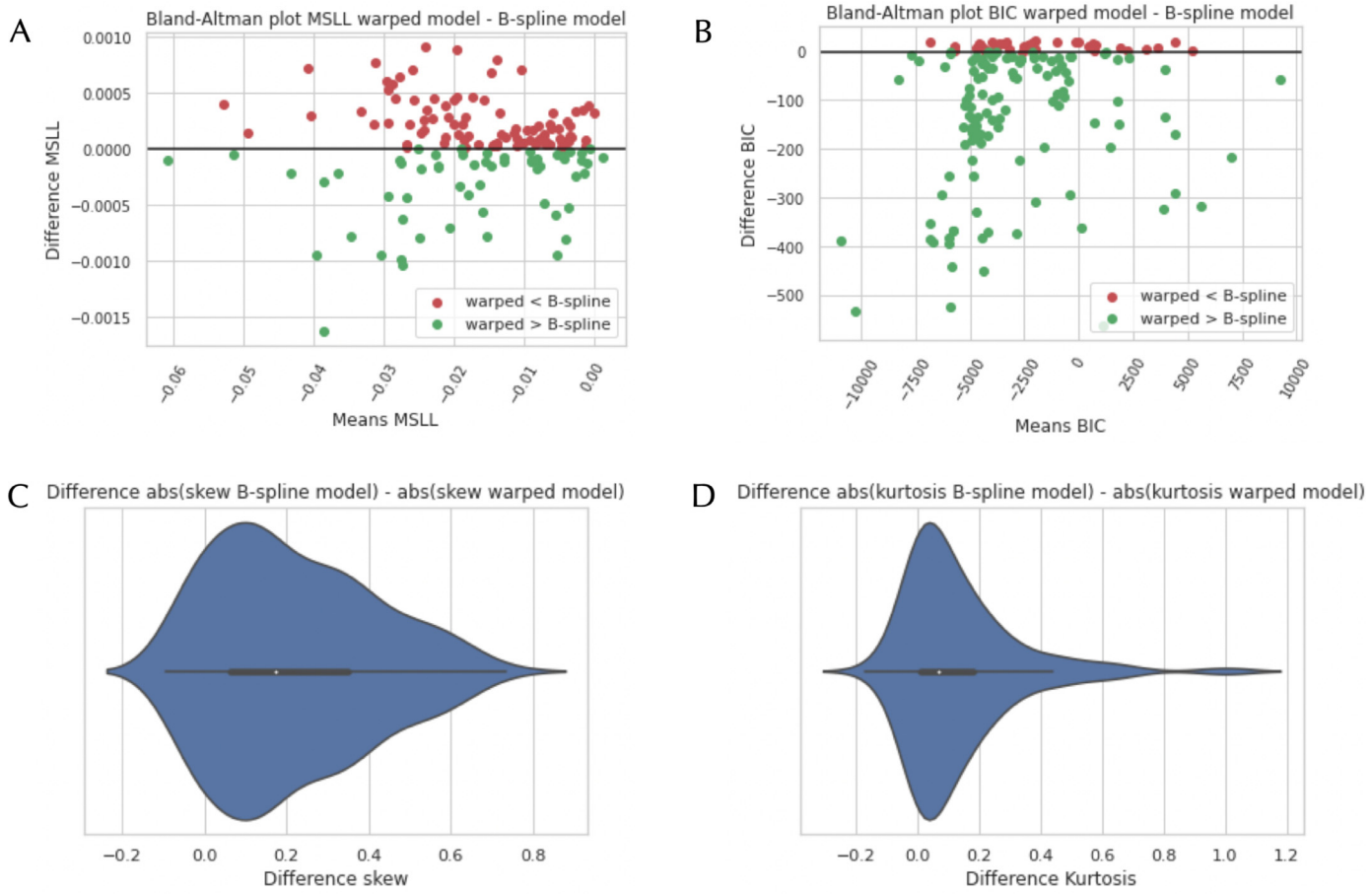

Fig. 3. Bland-Altman plots comparing the B-spline and B-spline warped Bayesian Linear Regression (BLR) models, using the FreeSurfer measurements. The models are compared according to the following model selection criteria: the Mean Standardized Log Loss (MSLL) (A) and the Bayesian Information Criteria (BIC) (B). On images C and D, we show the difference in absolute values of the skewness and kurtosis between the non-warped and warped models. A more positive number means a better fit for the B-spline warped model compared to the B-spline model.

that according to the EV and MSLL, the models perform quite similarly for most voxels. Although, we would argue that these measures are not necessarily sensitive for the added benefit of the warping of the likelihood, which will mostly affect the predictions in the outer centiles. For the BIC the results demonstrate that the warped BLR is preferred for certain voxels. The voxels where a warped model is favoured generally showed more non-Gaussian behaviour.

Finally, We used a paired-sample *t*-test, pairing the whole-brain results (EV, MSLL and BIC) of the different models to estimate the difference between performance measures of the warped and non-warped BLR. For MD the following effect sizes were found: $EV : d = 0.33$, $MSLL : d = 0.003$ and $BIC : d = -0.79$. For FA the following effect sizes were found: $EV : d = 0.028$, $MSLL : d = 0.017$ and $BIC : d = 0.55$. We can see that the difference between the methods is small. Indicating that the BLR and the warped BLR model are quite similar in their model fit for MD and FA.

3.2.1. Correlation deviance scores DTI and cognitive phenotypes

Finally, we correlated the Z-scores of the whole-brain warped BLR model for the MD dataset to the cognitive phenotypes. First, we scaled the cognitive data and performed a PCA. We selected the first component, which explained 29% of the variance in the data. Afterwards, we made a summary score of the Z-scores for each participant by looking at the largest deviations, which in the limit should follow an extreme value distribution (Fisher and Tippett, 1928). We fitted a generalized extreme value distribution to the top 1% of the absolute Z-scores of each subject. Subsequently, we computed a Spearman correlation between the extreme values and the first eigenvariate of the cognitive phenotypes, which gave $\rho = 0.158$, $p < .001$. The results demonstrate a clear correlation between the warped deviations from normal development and the cognitive phenotypes. This relationship will be explored further in future studies.

4. Discussion

In this paper, we presented a next-generation framework to scale normative models for large population-sized datasets based on warped Bayesian Linear Regression (BLR). Normative models can capture the heterogeneity in the population and model individual deviations from normal brain development. We demonstrated that the shift in normative modelling to a B-spline BLR with a likelihood warping gives several benefits. In this study we showed that: (i) The non-linearity of the model, incorporated by the B-spline, improves the fit and out of sample predictions for most variables. (ii) Non-Gaussianity of the data can be naturally included due to the incorporation of the likelihood warping in the algorithm, which allows for a wider range of datasets to be accurately modelled. (iii) Model selection criteria based on the marginal likelihood, such as the BIC, can be calculated in closed form and therefore a trade-off between model fit and model complexity can be chosen optimally from the training data, without cross-validation. (iv) Compared to Gaussian process regression, it is computationally much less demanding and is therefore scalable to big datasets. Furthermore, we demonstrated the use of the normative model with the warped BLR on different datasets from the UK Biobank, including image-derived pheno-

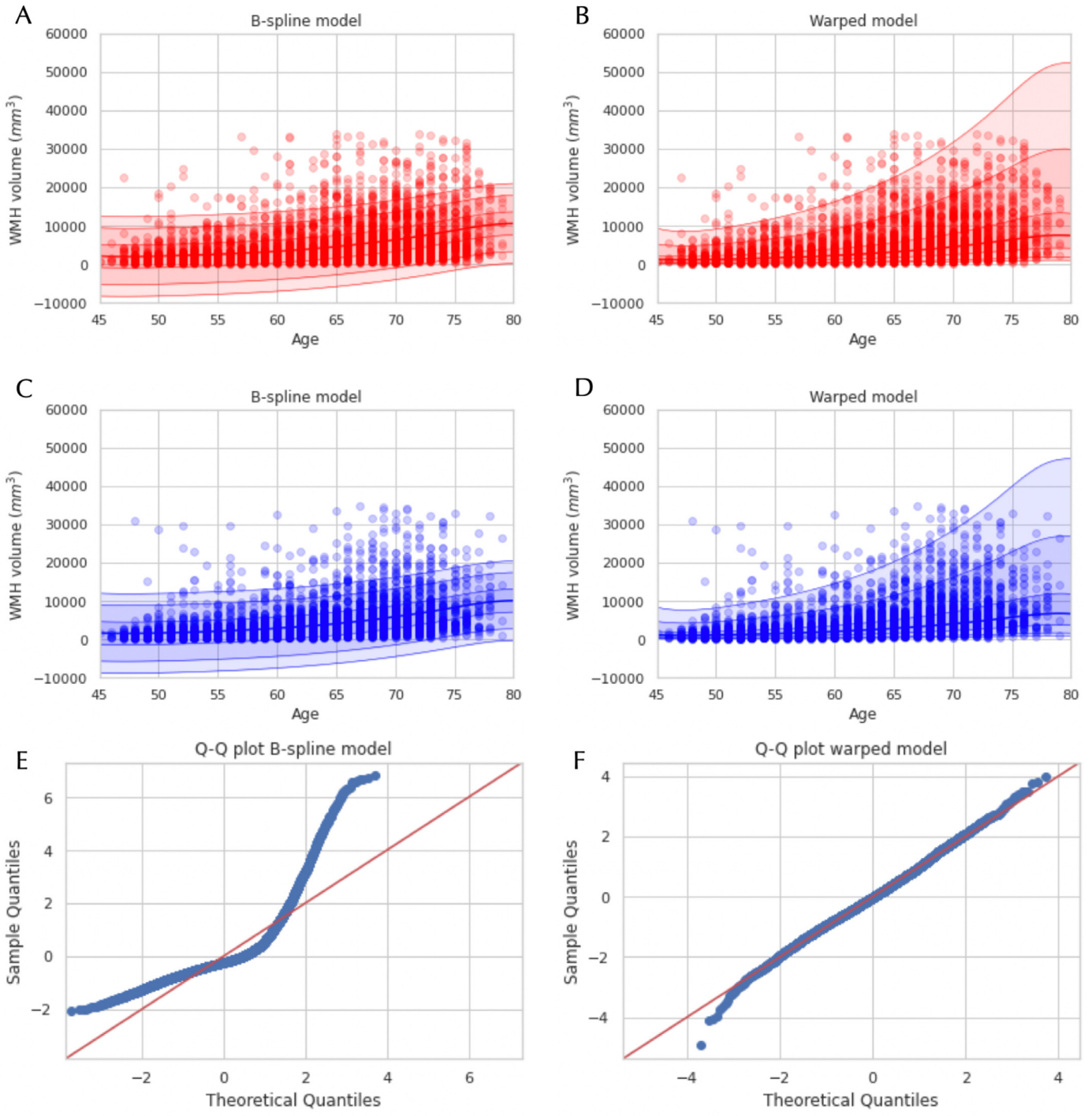

Fig. 4. White matter hyperintensities (WMHs) modelled as a function of age using a Bayesian Linear Regression (BLR) model. Images A and C demonstrate the model fit using a regular Gaussian B-spline BLR, for the female and male cohorts respectively, both visualizing the mean prediction and the centiles of variation for the WMHs. Images B and D show comparable fits for a B-spline SinhArcsinh warped BLR, for the female and male cohorts respectively. In images E and F quantile-quantile (QQ) plots of the two models are shown, demonstrating a better fit for the data using a warped BLR model.

types (IDPs); focusing on white matter hyperintensities (WMHs) as an example of non-Gaussianity and a diffusion tensor imaging (DTI) modality for a whole-brain model and show that the deviations scores from normal brain development can be meaningfully related to behaviour.

A major benefit of our method is the possibility of modelling non-Gaussian distributions by the use of the likelihood warping technique. This is important in general, as the aim of normative modelling is to accurately model the centiles of variation in addition to modelling the mean and is especially important for normative modelling of variables that are not approximately Gaussian distributed. For example, we showed that the WMHs show non-Gaussian behaviour that is well suited to uncover the benefits of the warped model over the standard model. We demonstrated the improved fit of the WMHs by including a B-spline transformation and a SinhArcsinh likelihood warping in the normative model, which was also exemplified for the longitudinal data. The same improvement in fit for other data modalities that showed more non-Gaussianity in their residuals was also demonstrated by comparing the B-spline warped BLR to the B-spline BLR for all the IDPs. Furthermore,

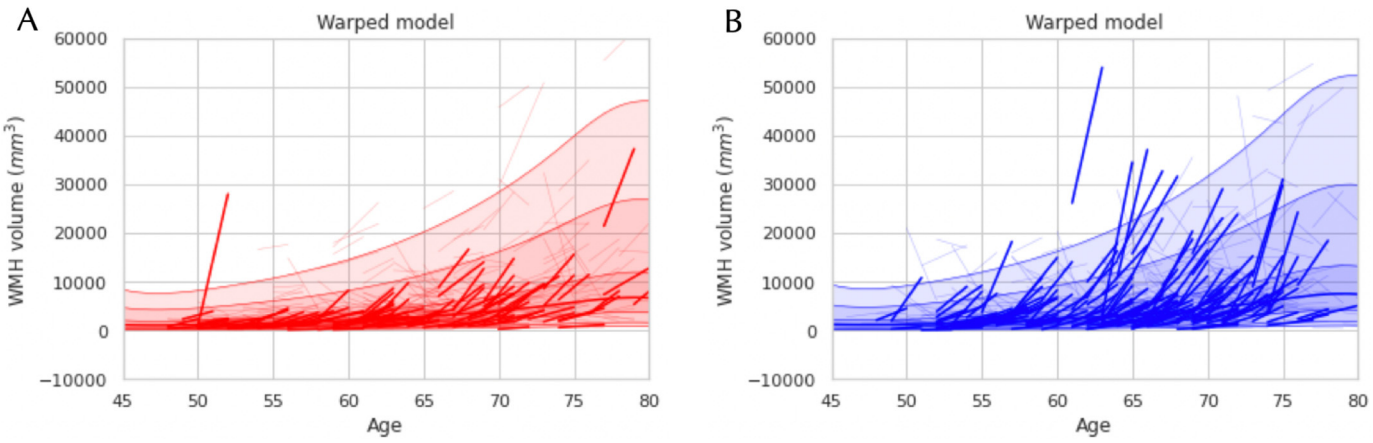

Fig. 5. Here the longitudinal follow-up data of the WMHs is plotted for female (A) and male (B) participants, using a B-spline SinhArcsinh warped BLR model.

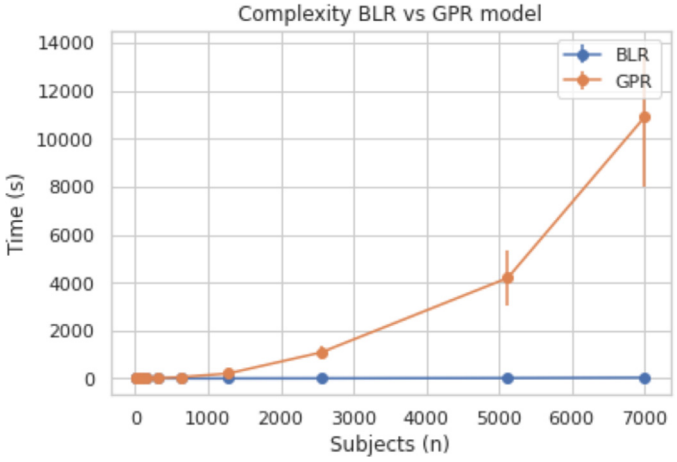

Fig. 6. Computational complexity comparison between the Bayesian Linear Regression (BLR) model setting and the Gaussian Process Regression (GPR) model setting, giving the mean and the standard error (SE) over ten runs.

it was shown on a whole-brain model of a DTI modality that for several voxels the warped BLR gives a better model performance than a B-spline BLR.

Aside from likelihood warping, there are other techniques that one can consider if non-Gaussianity is present in their data. One example is a pre-transformation of the dependent variable, such as a Box-Cox or log-transform. However, it is quite a challenge to find the correct transformation beforehand, which also remains consistent across multiple datasets. For a more in-depth discussion of data transformations and their advantages and disadvantages read (Keene, 1995). Thus, removing the added step of deciding on an appropriate transformation for each dataset and integrating this in the model through a likelihood warping is an added benefit of the warped BLR model, especially when multiple datasets are considered.

Furthermore, in our model, we captured non-linearity using a B-spline transformation. We choose the cubic B-spline over a polynomial implementation, as polynomial transformations can induce global curvature in space (Fjell et al., 2010). The world health organization has provided some comparisons of different basis transformations within normative modelling and they also recommend the cubic B-spline (Borghi et al., 2006). However, there are many other model options to capture non-linearity that could possibly do a better job for neuroimaging data. For example, mixture models, fractional polynomials or other types of splines functions. In future papers, we would like to explore in more detail how these choices affect the final model presented.

In our model, we accounted for the multi-modality that can be caused by the different sites in a similar manner as the Bayesian mixed effect models such as in Kia et al. (2020) and Bayer et al. (2021). The variance from different sites is captured by including site ID as a parameter in the covariance matrix (i.e. treating site as a fixed effect). This ensures that the variance from different sites does not influence the final Z-scores. We show that this approach does a good job of accounting for the multi-modality due to site effects in the data. The warp we use in most experiments is principally suited to unimodal data, it is the case that certain members of the SinArcsinh family of distributions can assume a multi-modal form for some parameter settings (Jones and Pewsey, 2009). However, we would recommend if there is still multi-modality present after the removal of the site effects to employ a combination of the warping function with models specifically made to handle multi-modality, such as mixture models or mixed-effects models in line with the recommendations in Jones and Pewsey (2009).

We emphasize that the addition of non-linear effects and non-Gaussianity makes the model more flexible which increase the need for model selection in order to avoid possible overfitting. We presented several model selection criteria that can be used to choose the optimal model settings for different neuroimaging modalities. It should be recognized that for some IDPs and voxels the B-spline BLR gives a better fit, showing that a more flexible model is not always needed. Therefore, we recommend carefully examining the type of data one wants to model and based on the data trends found for the residuals (Gaussian or non-Gaussian) to decide if a more flexible model is preferred. This can easily be checked by looking at the skewness and kurtosis of the distribution or making a QQ-plot. Additionally, different model selection criteria can sometimes contradict each other, as they are sensitive to different parts of the data. As we showed above, classical metrics such as EV and MSLL are not very sensitive to the shape of the predictive distribution. The consequence is that per task, we have to decide if we want a better EV, most sensitive to the mean fit and dependent on the flexibility of the model, or a better MSLL/BIC, which is more sensitive to the variance and penalizes the flexibility of the model. The variability in model selection criteria demonstrates that for different imaging modalities, different normative modelling settings are preferred and the added flexibility is confirmed to only give an advantage for response variables that show non-Gaussianity in their residuals.

Our proposed method makes it possible to apply normative modelling to considerably larger samples than was feasible before (Marquand et al., 2019; 2016a). The results from the computational experiments on the whole-brain model showed that the BLR method is scalable to population-sized data sets and fine-grained voxel-level data. In comparison, most normative models used Gaussian process regression, which due to its high computational complexity could only be used in

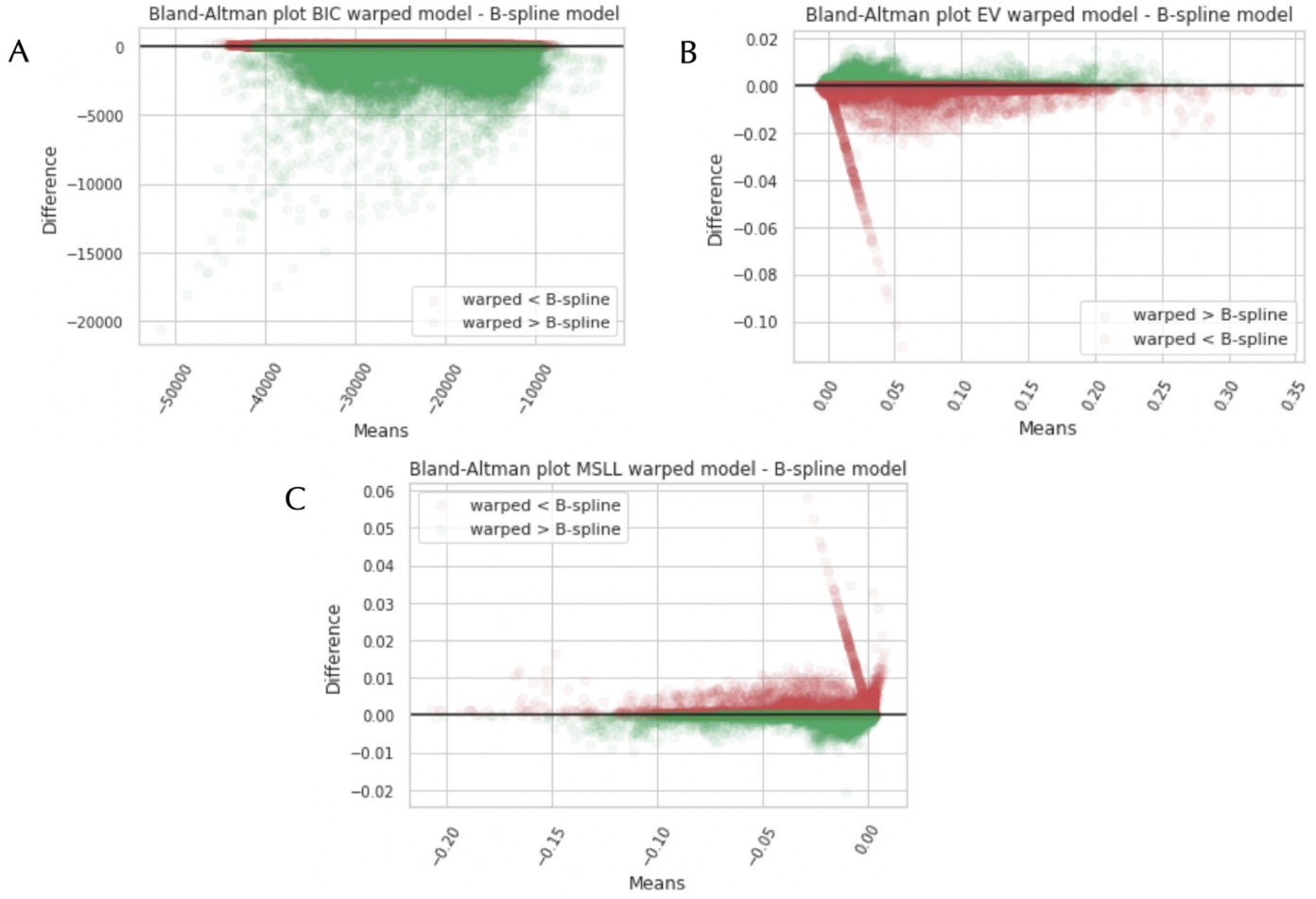

Fig. 7. Bland-Altman plots comparing the B-spline warped Bayesian Linear Regression (BLR) model to the B-spline BLR model, using Fractional Anisotropy (FA) data. The comparison is done according to the following model selection criteria: The Bayesian Information Criteria (BIC) (A), the Explained Variance (EV) (B), and the Mean Standardized Log Loss (MSLL) (C). The green colour indicates a better fit for the warped BLR.

studies with a relatively low sample size. This improvement is mainly because the approximation of the covariance matrix by a set of basis functions allowed us to account for non-linearity in a computationally less demanding way than the Gaussian process regression method, therefore making the B-spline BLR scalable for big datasets. Computationally scalable modelling of nonlinear effects is important since our experiments showed that a cubic B-spline transformation of the age covariate improved model fit compared to linear models for most neuroimaging modalities.

Finally, we confirmed that the deviations from the normative modelling framework can be meaningfully related to behaviour. We established a significant correlation between the warped deviance scores from the IDPs and several dimensions of the intelligence phenotype. These tests give an indication of the possible relationships between the deviations and behaviour. For the whole-brain model, the relationship with behaviour was shown with a significant correlation between an approximation to the g-factor in the form of the first principal component of the cognitive phenotypes and the warped deviance scores. This study confirms that the model could be extended to make predictive scores not only in the brain domain, but also for the behavioural phenotype for the individual, as has been demonstrated with other normative models (Marquand et al., 2016). In the future, the neurobiological markers of deviation from normal development can be extended to become markers of psychiatric disorders. This has already been done on a smaller scale, using normative modelling (Lv et al., 2020; Wolfers et al., 2019; 2018; Zabihi et al., 2020; 2019), but we would like to extend these studies to bigger data models, which include a wide variety of neuroimaging data modalities.

Some limitations of this method are worth mentioning. While we show an improved fit for most phenotypes, there are limits to the shape of distributions that a single warp is able to fit. For example, if the kurtosis is very high (e.g. due to extreme outliers), then it may be beneficial to compose multiple warps to improve the fit. However, in such cases, it is arguable whether modelling effort should be spent to capture extreme outliers given that they may be driven by artefacts rather than biological variation and it may be better to address these concerns using careful quality control procedures instead. See Rutherford et al. (2021) and Dinga et al. (2021) for further considerations and for examples showing how to train on big data and then transfer the hyper-parameters to new sites, similar to other methods (Dinga et al., 2021; Mostafa et al., 2021). Also, multi-modal data is difficult to handle in general with a single distribution or warp. Although the SinArcsinh warp we employed in this warp is principally suited for modelling unimodal distributions, there are certain members of the SinArcsinh family of distributions that can yield multi-modal distributions (Jones and Pewsey, 2009). However, we consider that it is better to model multimodality using purpose-built methods such as mixture models or random effect models, in line with other recommendations in the field (Jones and Pewsey, 2009). Indeed, we have shown that Bayesian mixed effect modelling approaches are a very efficient way to handle data that is multi-modal due to known causes (e.g. site effects) (Mostafa et al., 2021). Finally, there are also still open questions remaining that our model currently cannot address,

or are not fully explored; such as how to handle discrete data the possibility of applying warping compositions to handle more extreme cases of kurtosis and or skewness.

In conclusion, the current study suggests that non-linearity and non-Gaussianity are two parameters of big neuroimaging datasets that need to be captured to make accurate predictions for normal brain development. In this paper, we have done that through a warped BLR normative model. We have shown using several neuroimaging modalities the benefit of this model over more conservative models. Caution is essential when applying non-Gaussian models, as they can overfit and should mainly be used in the presence of non-normally distributed residuals. We recommend carefully assessing the distribution of residuals and the model selection parameters using the different model selection criteria mentioned in this paper that give a balance between model complexity and model fit.

Data and code availability statement

The UK Biobank data used in this paper are available through a procedure described at http://www.ukbiobank.ac.uk/using-the-resource/.

The methods developed here are available in the PCNtoolkit package at https://github.com/amarquand/PCNtoolkit.

Credit authorship contribution statement

Charlotte J. Fraza: Conceptualization, Methodology, Software, Writing – original draft. **Richard Dinga:** Validation, Writing – review & editing. **Christian F. Beckmann:** Supervision, Writing – review & editing. **Andre F. Marquand:** Supervision, Conceptualization, Methodology, Software, Writing – review & editing.

Acknowledgments

This research was supported by grants from the European Research Council (ERC, grant “MENTALPRECISION” 10100118), the Wellcome Trust under an Innovator award (“BRAINCHART”, 215698/Z/19/Z) and a Strategic Award (098369/Z/12/Z) and the Dutch Organisation for Scientific Research (VIDI grant 016.156.415). This research has been conducted using the UK Biobank resource under application number 23668.

Supplementary material

Supplementary material associated with this article can be found, in the online version, at 10.1016/j.neuroimage.2021.118715.

References

Alfaro-Almagro, F., Jenkinson, M., Bangerter, N.K., Andersson, J.L., Griffanti, L., Douaud, G., Sotiropoulos, S.N., Jbabdi, S., Hernandez-Fernandez, M., Vallee, E., et al., 2018. Image processing and quality control for the first 10,000 brain imaging datasets from UK Biobank. Neuroimage 166, 400–424. doi:10.1016/j.neuroimage.2017.10.034.

Bayer, J. M. M., Dinga, R., Mostafa Kia, S., Kottaram, A. R., Wolfers, T., Lv, J., Zalesky, A., Schmaal, L., Marquand, A., 2021. Accommodating site variation in neuroimaging data using normative and hierarchical Bayesian models. bioRxiv, 2021.02.09.430363. 10.1101/2021.02.09.430363

Bishop, C.M., 2006. Pattern Recognition and Machine Learning. Springer.

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., Bhandari, N., Norum, K. R., Bjoerneboe, G. E. A., Mohamed, A. J., Dewey, K. G., Belbase, K., Chumlea, C., Cole, T., Shrimpton, R., Albernaz, E., Tomasi, E., de Cássia Fossati da Silveira, R., Nader, G., Sagoe-Moses, I., Gomez, V., Sagoe-Moses, C., Taneja, S., Rongsen, T., Chetia, J., Sharma, P., Bahl, R., Baerug, A., Tufte, E., Alasfoor, D., Prakash, N. S., Mabry, R. M., Al Rajab, H. J., Helmi, S. A., Nommsen-Rivers, L. A., Cohen, R. J., Heinig, M. J., 2006. Construction of the World Health Organization child growth standards: Selection of methods for attained growth curves. https://pubmed.ncbi.nlm.nih.gov/16143968/. 10.1002/sim.2227

Casey, B., Cannonier, T., Conley, M.I., Cohen, A.O., Barch, D.M., Heitzeg, M.M., Soules, M.E., Teslovich, T., Dellarco, D.V., Garavan, H., et al., 2018. The adolescent brain cognitive development (ABCD) study: imaging acquisition across 21 sites. Dev. Cogn. Neurosci. 32, 43–54. doi:10.1016/j.dcn.2018.03.001.

Cox, S.R., Ritchie, S.J., Tucker-Drob, E.M., Liewald, D.C., Hagenaars, S.P., Davies, G., Wardlaw, J.M., Gale, C.R., Bastin, M.E., Deary, I.J., 2016. Ageing and brain white matter structure in 3,513 UK Biobank participants. Nat. Commun. 7, 1–13. doi:10.1038/ncomms13629.

Dinga, R., Fraza, C. J., Bayer, J. M. M., Kia, S. M., Beckmann, C. F., Marquand, A. F., 2021. Normative modeling of neuroimaging data using generalized additive models of location scale and shape. 10.1101/2021.06.14.448106

Fawns-Ritchie, C., Deary, I.J., 2020. Reliability and validity of the UK Biobank cognitive tests. PLoS ONE 15. doi:10.1371/journal.pone.0231627.

Fisher, R.A., Tippett, L.H.C., 1928. Limiting forms of the frequency distribution of the largest or smallest member of a sample. In: Mathematical Proceedings of the Cambridge Philosophical Society, Vol. 24. Cambridge University Press, pp. 180–190.

Fjell, A.M., Walhovd, K.B., Westlye, L.T., Østby, Y., Tamnes, C.K., Jernigan, T.L., Gamst, A., Dale, A.M., 2010. When does brain aging accelerate? Dangers of quadratic fits in cross-sectional studies. Neuroimage 50, 1376–1383. doi:10.1016/j.neuroimage.2010.01.061.

Habes, M., Pomponio, R., Shou, H., Doshi, J., Mamourian, E., Erus, G., Nasrallah, I., Launer, L.J., Rashid, T., Bilgel, M., et al., 2020. The brain chart of aging: machine-learning analytics reveals links between brain aging, white matter disease, amyloid burden, and cognition in the istaging consortium of 10,216 harmonized mr scans. Alzheimer’s Dement. doi:10.1002/alz.12178.

Huertas, I., Oldehinkel, M., van Oort, E.S., Garcia-Solis, D., Mir, P., Beckmann, C.F., Marquand, A.F., 2017. A Bayesian spatial model for neuroimaging data based on biologically informed basis functions. Neuroimage 161, 134–148. doi:10.1016/j.neuroimage.2017.08.009.

Insel, T.R., Cuthbert, B.N., 2015. Brain disorders? Precisely: precision medicine comes to psychiatry. Science 348, 499–500. doi:10.1126/science.aab2358.

Jones, M.C., Pewsey, A., 2009. Sinh-arcsinh distributions. Biometrika 96, 761–780. doi:10.1093/biomet/asp053.

Kaufmann, T., van der Meer, D., Doan, N.T., Schwarz, E., Lund, M.J., Agartz, I., Alnæs, D., Barch, D.M., Baur-Streubel, R., Bertolino, A., Bettella, F., Beyer, M.K., Bøen, E., Borgwardt, S., Brandt, C.L., Buitelaar, J., Celius, E.G., Cervenka, S., Conzelmann, A., Córdova-Palomera, A., Dale, A.M., de Quervain, D.J., Carlo, P.D., Djurovic, S., Dørum, E.S., Eisenacher, S., Elvsåshagen, T., Espeseth, T., Fatouros-Bergman, H., Flyckt, L., Franke, B., Frei, O., Haatveit, B., Håberg, A.K., Harbo, H.F., Hartman, C.A., Heslenfeld, D., Hoekstra, P.J., Høgestøl, E.A., Jernigan, T.L., Jonassen, R., Jönsson, E.G., Farde, L., Flyckt, L., Engberg, G., Erhardt, S., Fatouros-Bergman, H., Cervenka, S., Schwieler, L., Piehl, F., Agartz, I., Collste, K., Victorsson, P., Malmqvist, A., Hedberg, M., Orhan, F., Kirsch, P., Kłoszewska, I., Kolskår, K.K., Landrø, N.I., Hellard, S.L., Lesch, K.P., Lovestone, S., Lundervold, A., Lundervold, A.J., Maglanoc, L.A., Malt, U.F., Mecocci, P., Melle, I., Meyer-Lindenberg, A., Moberget, T., Norbom, L.B., Nordvik, J.E., Nyberg, L., Oosterlaan, J., Papalino, M., Papassotiropoulos, A., Pauli, P., Pergola, G., Persson, K., Richard, G., Rokicki, J., Sanders, A.M., Selbæk, G., Shadrin, A.A., Smeland, O.B., Soininen, H., Sowa, P., Steen, V.M., Tsolaki, M., Ulrichsen, K.M., Vellas, B., Wang, L., Westman, E., Ziegler, G.C., Zink, M., Andreassen, O.A., Westlye, L.T., 2019. Common brain disorders are associated with heritable patterns of apparent aging of the brain. Nat. Neurosci. 22, 1617–1623. doi:10.1038/s41593-019-0471-7.

Keene, O.N., 1995. The log transformation is special. Stat. Med. 14, 811–819. doi:10.1002/sim.4780140810. URL: https://pubmed.ncbi.nlm.nih.gov/7644861/

Kia, S. M., Huijsdens, H., Dinga, R., Wolfers, T., Mennes, M., Andreassen, O. A., Westlye, L. T., Beckmann, C. F., Marquand, A. F., 2020. Hierarchical Bayesian regression for multi-site normative modeling of neuroimaging data. arXiv preprint arXiv:2005.12055.

Kia, S.M., Marquand, A., 2018. Normative modeling of neuroimaging data using scalable multi-task Gaussian processes. In: Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 11072 LNCS, pp. 127–135. arXiv:1806.01047

Lv, J., Biase, M. D., Cash, R. F., Cocchi, L., Cropley, V., Klauser, P., Tian, Y., Bayer, J., Schmaal, L., Cetin-Karayumak, S., et al., 2020. Individual deviations from normative models of brain structure in a large cross-sectional schizophrenia cohort. bioRxiv. 10.1038/s41380-020-00882-5

Lyall, D.M., Cullen, B., Allerhand, M., Smith, D.J., Mackay, D., Evans, J., Anderson, J., Fawns-Ritchie, C., McIntosh, A.M., Deary, I.J., Pell, J.P., 2016. Cognitive test scores in UK biobank: data reduction in 480,416 participants and longitudinal stability in 20,346 participants. PLoS ONE 11, e0154222. doi:10.1371/journal.pone.0154222.

Marquand, A. F., Kia, S. M., Zabihi, M., Wolfers, T., Buitelaar, J. K., Beckmann, C. F., 2019. Conceptualizing mental disorders as deviations from normative functioning. 10.1038/s41380-019-0441-1

Marquand, A.F., Rezek, I., Buitelaar, J., Beckmann, C.F., 2016. Understanding heterogeneity in clinical cohorts using normative models: beyond case-control studies. Biol. Psychiatry 80, 552–561. doi:10.1016/j.biopsych.2015.12.023.

Marquand, A.F., Rezek, I., Buitelaar, J., Beckmann, C.F., 2016a. Understanding heterogeneity in clinical cohorts using normative models: beyond case-control studies. Biol. Psychiatry 80, 552–561. doi:10.1016/j.biopsych.2015.12.023.

Marquand, A. F., Wolfers, T., Mennes, M., Buitelaar, J., Beckmann, C. F., 2016b. Beyond lumping and splitting: a review of computational approaches for stratifying psychiatric disorders. 10.1016/j.bpsc.2016.04.002

McCarthy, P., 2020. funpack. 10.5281/zenodo.3761702

Miller, K.L., Alfaro-Almagro, F., Bangerter, N.K., Thomas, D.L., Yacoub, E., Xu, J., Bartsch, A.J., Jbabdi, S., Sotiropoulos, S.N., Andersson, J.L., Griffanti, L., Douaud, G., Okell, T.W., Weale, P., Dragonu, I., Garratt, S., Hudson, S., Collins, R., Jenkinson, M., Matthews, P.M., Smith, S.M., 2016. Multimodal population brain imaging in the UK Biobank prospective epidemiological study. Nat. Neurosci. 19, 1523–1536. doi:10.1038/nn.4393.

Mostafa, S., Huijsdens, H., Rutherford, S., Dinga, R., Wolfers, T., Mennes, M., Ole, A., 2021. Multi-site normative modeling using hierarchical Bayesian regression.

Nave, G., Jung, W.H., Karlsson Linnér, R., Kable, J.W., Koellinger, P.D., 2019. Are bigger brains smarter? Evidence from a large-scale preregistered study. Psychol. Sci. 30, 43–54. doi:10.1177/0956797618808470.

Powell, M.J., 1964. An efficient method for finding the minimum of a function of several variables without calculating derivatives. Comput. J. 7, 155–162.

Rasmussen, C. E., Williams, C. K., 2005. Approximation methods for large datasets.

Rios, G., Tobar, F., 2019. Compositionally-warped gaussian processes. Neural Netw. 118, 235–246. doi:10.1016/j.neunet.2019.06.012.

Rutherford, S., Fraza, C., Dinga, R., Kia, S. M., Wolfers, T., Zabihi, M., Berthet, P., Worker, A., Verdi, S., Andrews, D., et al., 2021. Charting brain growth and aging at high spatial precision. bioRxiv.

Satterthwaite, T.D., Connolly, J.J., Ruparel, K., Calkins, M.E., Jackson, C., Elliott, M.A., Roalf, D.R., Hopson, R., Prabhakaran, K., Behr, M., et al., 2016. The philadelphia neurodevelopmental cohort: a publicly available resource for the study of normal and abnormal brain development in youth. Neuroimage 124, 1115–1119. doi:10.1016/j.neuroimage.2015.03.056.

Snelson, E., Ghahramani, Z., Rasmussen, C.E., 2004. Warped Gaussian processes. In: Proceedings of the Advances in Neural Information Processing Systems, pp. 337–344.

Sudlow, C., Gallacher, J., Allen, N., Beral, V., Burton, P., Danesh, J., Downey, P., Elliott, P., Green, J., Landray, M., et al., 2015. Uk Biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. PloS Med 12, e1001779. doi:10.1371/journal.pmed.1001779.

Thompson, P.M., Stein, J.L., Medland, S.E., Hibar, D.P., Vasquez, A.A., Renteria, M.E., Toro, R., Jahanshad, N., Schumann, G., Franke, B., et al., 2014. The enigma consortium: large-scale collaborative analyses of neuroimaging and genetic data. Brain Imaging Behav. 8, 153–182. doi:10.1007/s11682-013-9269-5.

Wolfers, T., Beckmann, C.F., Hoogman, M., Buitelaar, J.K., Franke, B., Marquand, A.F., 2019. Individual differences V. The average patient: mapping the heterogeneity in ADHD using normative models. Psychol. Med. 50, 314–323. doi:10.1017/S0033291719000084. URL: https://pubmed.ncbi.nlm.nih.gov/30782224/

Wolfers, T., Buitelaar, J. K., Beckmann, C. F., Franke, B., Marquand, A. F., 2015. From estimating activation locality to predicting disorder: a review of pattern recognition for neuroimaging-based psychiatric diagnostics. 10.1016/j.neubiorev.2015.08.001

Wolfers, T., Doan, N.T., Kaufmann, T., Alnæs, D., Moberget, T., Agartz, I., Buitelaar, J.K., Ueland, T., Melle, I., Franke, B., Andreassen, O.A., Beckmann, C.F., Westlye, L.T., Marquand, A.F., 2018. Mapping the heterogeneous phenotype of schizophrenia and bipolar disorder using normative models. JAMA Psychiatry 75, 1146–1155. doi:10.1001/jamapsychiatry.2018.2467.

Zabihi, M., Floris, D.L., Kia, S.M., Wolfers, T., Tillmann, J., Arenas, A.L., Moessnang, C., Banaschewski, T., Holt, R., Baron-Cohen, S., Loth, E., Charman, T., Bourgeron, T., Murphy, D., Ecker, C., Buitelaar, J.K., Beckmann, C.F., Marquand, A., 2020. Fractionating autism based on neuroanatomical normative modeling. Transl. Psychiatry 10, 1–10. doi:10.1038/s41398-020-01057-0.

Zabihi, M., Oldehinkel, M., Wolfers, T., Frouin, V., Goyard, D., Loth, E., Charman, T., Tillmann, J., Banaschewski, T., Dumas, G., Holt, R., Baron-Cohen, S., Durston, S., Bölte, S., Murphy, D., Ecker, C., Buitelaar, J.K., Beckmann, C.F., Marquand, A.F., 2019. Dissecting the heterogeneous cortical anatomy of autism spectrum disorder using normative models. Biol. Psychiatry Cogn. Neurosci. Neuroimaging 4, 567–578. doi:10.1016/j.bpsc.2018.11.013.