bioRxiv preprint doi: https://doi.org/10.1101/2021.02.09.430363; this version posted February 11, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

ACCOMMODATING SITE VARIATION IN NEUROIMAGING DATA USING HIERARCHICAL AND BAYESIAN MODELS

A PREPRINT

Johanna M. M. Bayer
Orygen Centre for Youth Mental Health, Melbourne, Australia
The University of Melbourne, Melbourne, Australia
bayerj@student.unimelb.edu.au

Richard Dinga
Donders Institute, Radboud University, Nijmegen, the Netherlands
Radboud University Medical Centre, Nijmegen, the Netherlands

Seyed Mostafa Kia
Donders Institute, Radboud University, Nijmegen, the Netherlands
Radboud University Medical Centre, Nijmegen, the Netherlands

Akhil R. Kottaram
Orygen Centre for Youth Mental Health, Melbourne, Australia

Thomas Wolfers
Radboud University Medical Centre, Nijmegen, the Netherlands
Department of Psychology, University of Oslo, Norway

Jinglei Lv
School of Biomedical Engineering
Brain and Mind Center, University of Sydney, Sydney, Australia

Andrew Zalesky
Melbourne Neuropsychiatry Centre, The University of Melbourne
Melbourne Health, Melbourne, Australia
Department of Biomedical Engineering, The University of Melbourne, Australia

Lianne Schmaal *
Orygen Centre for Youth Mental Health,Melbourne, Australia
The University of Melbourne, Melbourne, Australia

Andre Marquand *
Donders Institute, Radboud University, Nijmegen, the Netherlands
Radboud University Medical Centre, Nijmegen, the Netherlands
Institute of Psychiatry, Kings College London, London, UK

February 9, 2021

*shared last author

bioRxiv preprint doi: https://doi.org/10.1101/2021.02.09.430363; this version posted February 11, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

ABSTRACT

The potential of normative modeling to make individualized predictions has led to structural neuroimaging results that go beyond the case-control approach. However, site effects, often confounded with variables of interest in a complex manner, induce a bias in estimates of normative models, which has impeded the application of normative models to large multi-site neuroimaging data sets. In this study, we suggest accommodating for these site effects by including them as random effects in a hierarchical Bayesian model. We compare the performance of a linear and a non-linear hierarchical Bayesian model in modeling the effect of age on cortical thickness. We used data of 570 healthy individuals from the ABIDE (autism brain imaging data exchange, `http://preprocessed-connectomes-project.org/abide/`) data set in our experiments. We compare the proposed method to several harmonization techniques commonly used to deal with additive and multiplicative site effects, including regressing out site and harmonizing for site with ComBat, both with and without explicitly preserving variance related to age and sex as biological variation of interest. In addition, we make predictions from raw data, in which site has not been accommodated for. The proposed hierarchical Bayesian method shows the best performance according to multiple metrics. Performance is particularly bad for the regression model and the ComBat model when age and sex are not explicitly modeled. In addition, the predictions of those models are noticeably poorly calibrated, suffering from a loss of more than 90 % of the original variance. From these results we conclude that harmonization techniques like regressing out site and ComBat do not sufficiently accommodate for *multi-site effects* in pooled neuroimaging data sets. Our results show that the complex interaction between site and variables of interest is likely to be underestimated by those tools. One consequence is that harmonization techniques removed too much variance, which is undesirable and may have unpredictable consequences for subsequent analysis. Our results also show that this can be mostly avoided by explicitly modeling site as part of a hierarchical Bayesian Model. We discuss the potential of z-scores derived from normative models to be used as site corrected variables and of our method as site correction tool.

***K*eywords** neuroimaging · normative modeling · site effects · Hierarchical Bayesian Modeling

1 Introduction

The most prominent paradigm in clinical neuroimaging research has for a long time been case-control approaches which compare averages of groups of individuals on brain imaging measures. Case-control inferences can be clinically meaningful under some circumstances when the group mean is a good representation of each individual in the group. However, this pre-condition has been challenged recently, demonstrating that the biological heterogeneity within clinical groups can be substantially large [Marquand et al., 2016]. For example, the structure and morphology of the brain have been found to vary between individuals in dynamic phases like adolescence [Foulkes and Blakemore, 2018] and within clinical groups, such as bipolar disorder and schizophrenia [Wolfers et al., 2018a] and attention deficit disorder [Wolfers et al., 2019]. In addition, inter-individual differences have shown to not necessarily be in line with results obtained via the group comparison approach [Wolfers et al., 2019]. Such heterogeneity has been considered a potential cause for the lack of differences between clinical groups and controls within the standard group comparison approach [Feczko et al., 2019] and the failure to replicate findings between studies [Fried, 2017]. As a consequence, there has been a shift in focus towards taking into account variation at the individual level [Marquand et al., 2019]. This is in line with a trend towards personalized medicine or "precision medicine" [Mirnezami et al., 2012], where characteristics of the individual are used to guide the treatment of mental disorders.

This shift has been accompanied by a trend towards approaches that go beyond comparing averages of distinctly labeled groups [Insel et al., 2010, Insel, 2014], for an overview of methods see [Marquand et al., 2016]. Among them, normative modeling has been successfully used to capture inter-individual variability and make predictions at the individual level. The strength of normative modeling lies within the ability to map variation along one dimension (e.g., brain volume) onto a second co-varying variable (e.g., age), redefining the variation in the first dimension as explained by this new covariate of interest. This concept allows to describe the normative variation, thus the range containing e.g., 95 % of all individuals, as a function of the covariate and considers each individual's score in relation to the variation in the reference group defined by the covariate score. The concept is similar to the use of growth charts in pediatric medicine, in which height and weight are expressed as a function of age. Hence, in this setting, an individual's height or weight is not considered by its absolute value, but expressed as a percentile score of deviation fluctuating with age, with the median line corresponding to the 50% percentile and defining the norm, or average height.

bioRxiv preprint doi: https://doi.org/10.1101/2021.02.09.430363; this version posted February 11, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

In neuroimaging, normative models have been applied to clinical and non-clinical problems using various covariates, statistical modeling approaches (for an overview see [Marquand et al., 2019]) and targeting a variety of response variables. In general, any variable can be used as a covariate in a normative model targeting neuroimaging measures, as long as the variation along the co-varying dimension is not zero. However, normative models with age and sex as covariates and brain volume as response variable are currently more frequently found in the literature, [Wolfers et al., 2020, Wolfers et al., 2018b, Zabihi et al., 2019, Kessler et al., 2016]. These implement the growth charting idea applied to high dimensional brain imaging data. For example, a normative model of a brain structure can be created based on the variation of individuals in population based cohorts. The estimated norm can be used to infer where individuals with clinical symptoms can be placed with respect to the reference defined by the normative model. This has been the recipe of many recently published studies using the normative modeling framework [Wolfers et al., 2020, Bethlehem et al., 2018, Wolfers et al., 2019, Lv et al., 2020]. Underlying this approach is the assumption that the individually derived patterns of deviation uncover associations to clinical/behavioral variables that would be obscured by averaging across groups of individuals. However, the amount of data necessary to create normative models poses a challenge to normative modeling in neuroimaging, as the cost and time factor associated with neuroimaging data impedes the collection of large neuroimaging samples in a harmonized way. One exceptional example where large scale data collection succeeded and included both harmonized scanners and scanning protocols, is the UK Biobank initiative, which, when launched in 2006, aimed to scan 100,000 individuals at four different scanning locations [`https://www.ukbiobank.ac.uk/explore-your-participation/contribute-further/imaging-study`][Miller et al., 2016]. Other neuroimaging initiatives have also taken on the challenge to collect neuroimaging data in large scale quantities and have relied on harmonized scanning protocols, but did not collect the data using harmonized scanners (i.e. ADNI, [Mueller et al., 2005], ABCD study [Volkow et al., 2018]). Nonetheless, the restricted age ranges (e.g., 40-69 years in UK Biobank [Miller et al., 2016]), or focus on a particular (clinical) cohort (e.g. Alzheimer's in ADNI, [Mueller et al., 2005]) limit their utility for estimating normative models mapping the normative association between, for example, age and brain structure or function.

An alternative way to obtain large neuroimaging data sets and assess data from a large number of subjects is by pooling or sharing data that has already been collected. One example is the Enhancing NeuroImaging and Genetics through Meta-Analysis (ENIGMA) consortium [Thompson et al., 2020]. ENIGMA succeeded in pooling neuroimaging and genetics data of thousands of individuals, including healthy individuals and individuals with psychiatric or neurological disorders. The strategy of data sharing initiatives like ENIGMA is to collect already collected data from different cohorts and different scanning sites and harmonize preprocessing and statistical analysis with standardized protocols. However, a major disadvantage is the presence of confounding "scanner effects" [Fortin et al., 2018] (e.g., differences in field strength, scanner manufacturer etc. [Han et al., 2006])). These confounding effects present as site correlated biases that cannot be explained by biological heterogeneity between samples. An example of those effects on derived measures of cortical thickness can be found in Fig. 1a. They result from a complex interaction between site and variables of interest, manifesting in biases on lower and higher order properties of the distribution of interest, such as differences in mean and standard deviations, skewness and spatial biases Fig. (1a, 1b), and cannot be explained by e.g., differences in age or sex Fig. (1c). As the origin of these effects might not only be related to the *scanner per se*, but extend to various factors related to a single acquisition site [Gronenschild et al., 2012], we will refer to them as *site effects* from here on.

As outlined in the previous paragraph, the effort to create large samples to capture between subject variability often induces site-driven variability. This issue of site-driven variability in shared neuroimaging data has been acknowledged and has led to the development of harmonization methods at a statistical level. A common approach to deal with site effects is through "harmonizing" by, e.g., confound regression. One example of this approach is a set of algorithms summarized under the name "ComBat" [Fortin et al., 2017]. The method had originally been developed by [Johnson et al., 2007], who used empirical Bayes to estimate "batch effects", referring to non-biological variation added due to the handling of petri-dishes in micro-array experiments on the results of gene expression data. Fortin and colleagues adapted the framework to apply to neuroimaging data [Fortin et al., 2017]. In ComBat, additive and multiplicative site effects on a particular target unit (e.g., a particular brain voxel for one participant) are estimated using empirical Bayes and by placing a prior distribution over estimates for these units. The etsimate of the scanner effect is then used to adjust the prediction. Newer versions also allow to preserve variance of interest in the model, for example for age, sex or diagnosis [Fortin et al., 2017, Fortin et al., 2018]. ComBat has been applied to several types of neuroimaging data, including diffusion tensor imaging data (DTI, [Fortin et al., 2017]) and structural magnetic resonance imaging data [Fortin et al., 2018].

However, the reliability of harmonization strategies is grounded on the condition that site effects are orthogonal to the effect of interest and uncorrelated with other covariates in the model [Chen et al., 2014]. In reality, however, data pooled from several sites is often confounded with co-linear effects. Many individual neuroimaging samples,

bioRxiv preprint doi: https://doi.org/10.1101/2021.02.09.430363; this version posted February 11, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

for example, are restricted to a specific age range, leading to age being correlated with site effects. In this scenario, removing an estimate of the scanner effect can lead to excluding (biological) variation that would be of interest.

With this paper we suggest an alternative approach to deal with site effects in neuroimaging, which is relatively generally applicable. However, here we focus in particular on normative modeling. We propose a hierarchical Bayesian approach in which we include site as a covariate into the model, avoiding the exclusion of meaningful variance correlated to site by predicting site effects as part of the model instead of removing them from the data.

This approach is similar to the approach by [Kia et al., 2020], who used hierarchical Bayesian regression (HBR) in a similar way for multi-site modeling in a pooled neuroimaging data set, which contained 7499 participants that were scanned with 33 different scanners. [Kia et al., 2020]'s estimate of site variation is based on a partial pooling approach, in which the variation between site-specific parameters is bound by a shared prior. The approach showed better performance when evaluated with respect to metrics accounting for the quality of the predictive mean and variance compared to a complete pooling of site parameters and to ComBat harmonization, and similar performance to a no-pooling approach, with the benefit of reduced risk of over-fitting due to the shared site variance. Moreover, [Kia et al., 2020] also showed that the posterior distribution of site parameters from the training set can also be used as an informed prior to make predictions in an unseen, new test set, outperforming predictions from complete pooling and uninformed priors, and overcoming a weakness of ComBat. The method was also able to display heterogeneity between individuals with varying clinical diagnoses in associated brain regions of 1017 clinical patients of the study.

The present paper is a replication and extension of the approach by [Kia et al., 2020]. Based on several successful attempts of using Gaussian Process Regression to map non-linearity in normative models [Kia and Marquand, 2018, Marquand et al., 2016, Marquand et al., 2014], we extend the normative model with the capacity to account for site effects by adding a Gaussian process to model non-linear effects between age and the brain structure. In addition, our model is fully Bayesian and entails a hierarchical structure, including priors and hyper priors for each parameter. We use data from the ABIDE (autism brain imaging data exchange, `http://preprocessed-connectomes-project.org/abide/`) data set to compare a non-linear, Gaussian version of the model, to a linear hierarchical Bayesian version accounting for site effects that does not include the Gaussian Process term. We show that the hierarchical Bayesian models including a site parameter perform better than existing methods for dealing with additive and multiplicative site effects, including *ComBat* and regressing out site. We discuss the normative hierarchical Bayesian methods with regard to their implications for neuroimaging data-sharing initiatives and their use as general technique to correct for site effects.

2 Methods

In this section we will introduce the data used in this study and the pre-processing steps applied, followed by a conceptual and mathematical description of our approach to include site as predictor in a normative hierarchical Bayesian model. We will also illustrate other methods (than including site as predictor) to accommodate for site effects that will be used to validate our approach against. Lastly, we will outline which measures will be used for model comparison.

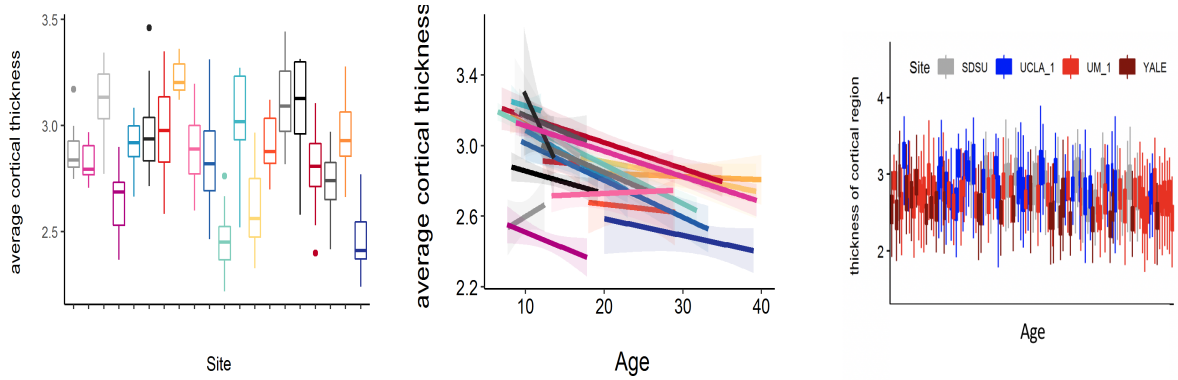
2.1 Data

The following sub-section aims to give a description of the ABIDE data set, including a study on the scope of site effects in the data.

2.1.1 ABIDE data set

The ABIDE consortium (`http://preprocessed-connectomes-project.org/abide/`) was founded to facilitate research and collaboration on autism spectrum disorders by data aggregation and sharing. The consortium provides a publicly available structural magnetic resonance imaging (MRI) data set and corresponding phenotypic information of 539 individuals with autism spectrum disorder and 573 age-matched typical controls. For this study, only data from healthy individuals were included. As those healthy controls are meant to be complementary to the autism branch in the data set, 403 out of 539 subjects in this study were male. The data was processed using a standardized protocol [Craddock et al., 2013] of the FreeSurfer standard pipeline (Desikan-Kiliany Atlas) as part of the *Preprocessed Connectomes Project* [Craddock et al., 2013] and has been made available for download on the *preprocessed* section of the ABIDE initiative. For the current study we focused on cortical thickness measures of the 34 bilateral regions of the Desikian Killiany atlas parcellation [Desikan et al., 2006] as a part of the FreeSurfer [Fischl et al., 2004] output and the average cortical thickness across all 34 regions. We chose to include cortical thickness measures since they show a strong (negative) association with age (unlike measures of surface area, which remain more stable across the life span [Storsve et al., 2014]).

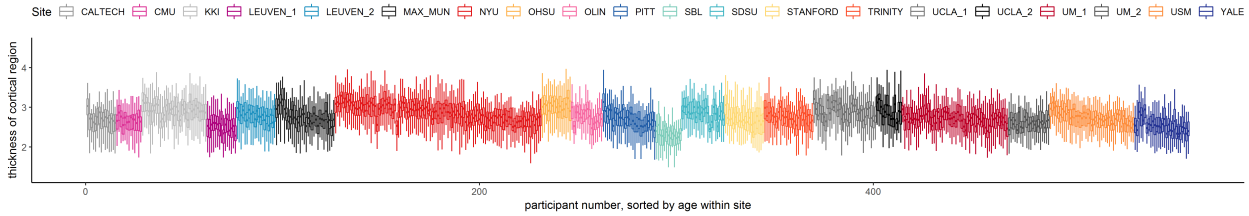
bioRxiv preprint doi: https://doi.org/10.1101/2021.02.09.430363; this version posted February 11, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.



(a) Distribution of average cortical thickness measures of 573 individuals, grouped by the 20 acquisition sites the data were collected at (each boxplot describes the distribution of one site).

(b) Average cortical thickness of 573 individuals regressed onto age, grouped by site (each regression line describes one site).

(c) Thickness measures of all 34 cortical regions average cortical thickness grouped by individual, colored by site, sorted by age (each boxplot represents one individual). Displayed are 4 out of 20 sites from the ABIDE data set



(d) Distribution of all 34 cortical regions average cortcial thickness per individual, summarized as boxplot (each boxplot represents one individual). Boxplots are coloured by site and ordered by age within site.

Figure 1: Site effects in 573 healthy individuals from the ABIDE data set.

2.1.2 Site effects in the ABIDE data set

The ABIDE data set has been obtained by aggregating data from 20 independent samples collected at 17 different scanning locations [Di Martino et al., 2014]. Although all data has been collected with 3 Tesla scanners and preprocessed in a harmonized way [Craddock et al., 2013], sequence parameters for anatomical and functional data, as well as type of scanner varied across sites [Di Martino et al., 2014]. In addition, sites differ in distribution of age and sex and in sample size. An overview of site-specific data is provided in Table 1 and in [Di Martino et al., 2014].

The ABIDE data set is affected by site specific effects that are unlikely to be explained by biological variation. They manifest as linear and non-linear interactions between *scanning site*, covariates (for example age and sex), and cortical measures. Similar to *batch effects* in genomics [Leek et al., 2010], those effects lead to a clustering of the data caused by external factors related to the scanning- and analysis process. With the aim to estimate to which extent the ABIDE data set is affected by site effects, we calculated an ANCOVA with age as covariate. It revealed that average cortical thickness differed between site (main effect site: $F(19, 516) = 4.4$, $p < 0.1 \times 10^{-8}$, sum contrast). In addition we tested for differences in variance between sites. Bartlett's sphericity test [Bartlett, 1937] showed a difference in variance between sites even after regressing out variance that could be explained by age and sex ($p < 0.001$). The site effects in the ABIDE data set are visualized in Fig. 1.

2.2 Splitting the ABIDE data set into training and test sets

To evaluate the performance of the models, we split the data into a training set (70% of data) and a test set (30% of data) using the R package *caret* and *splitstackshape*, while the distribution of age, sex and site was preserved between sets. Thus, training and test sets contained individuals from the same sites ("within-site-split"). An overview of the distribution of age and sex for the training and test sets can be found in Fig. 2. Subsequently, the training and test sets were standardized region-wise based on location and scale parameters of the training set. For the model estimation process, only complete pairs of observations (per region) were used.

bioRxiv preprint doi: https://doi.org/10.1101/2021.02.09.430363; this version posted February 11, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

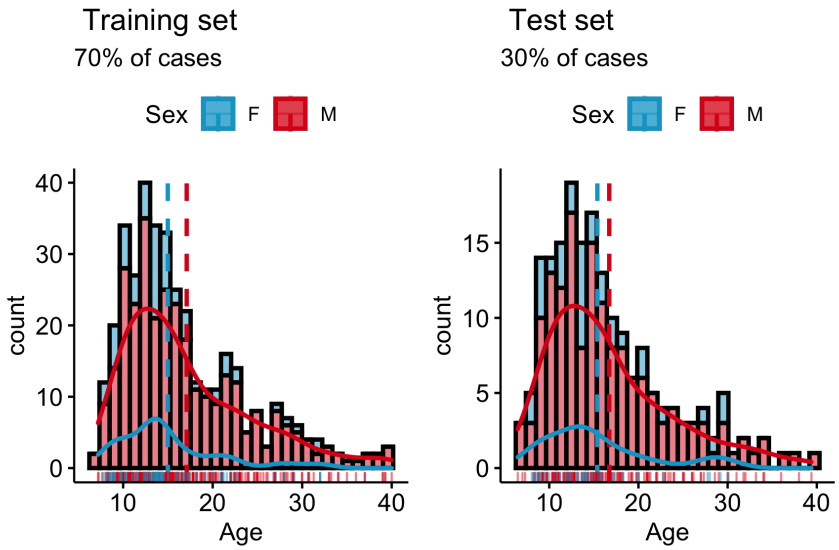


Figure 2: Overview over phenotypic information in the ABIDE data set. // Age male subjects: M = 17.5. SD = 8.3. Age female subjects: M = 15.6, SD = 7.0. Range = 6.5-40

2.3 site as a predictor in a Hierarchical Bayesian Model

With the aim to create reliable normative models in multi-site neuroimaging data, we developed and compared two versions of a hierarchical Bayesian models that include site as a predictor. In a hierarchical linear version of the model, site is modeled hierarchically, resulting in a random effect for site (Hierarchical Bayesian Linear Model, HBLM). In a non-linear version of the model, a Gaussian Process for age is added to test whether performance is increased if the model is also able to capture non-linear effects between age and *thickness of the cortical region* ("Hierarchical Bayesian Gaussian Process Model, HBGPM"). Both Hierarchical Bayesian models were trained and tested in a within site split (see section 2.2 on splitting the multi-site ABIDE data set.)

2.4 Comparison models

To get a better understanding of the performance of our approach, we performed a second analysis, comparing the hierarchical Bayesian approach with site as predictor to predictions made from a data that other methods managing site effects had been applied on. In the following, those alternative models will be summarized under the term *comparison models*.

Of note, the approach used to accommodate for site effects in the comparison models is fundamentally different from the approach used in the hierarchical Bayesian models. In the hierarchical Bayesian approach, multi-level modeling is used to account for site-variance without removing it, whereas different methods of *harmonization* are used on the data to remove variance related to site as part of the comparison models approach.

In detail, the comparison model approach entailed a two-step procedure, in which site effects are first *harmonized* by three different common models of site harmonization, and then a simple Bayesian linear algorithm, with an additive term for age and sex, but without site as a predictor is used to make predictions in Stan [Stan Development Team, 2020b]. The harmonization procedures include i) regressing out site effect from the cortical thickness measures using linear regression and using the residuals as input to the simple Bayesian linear model (thus, removing additive variant components of site), ii) using ComBat [Johnson et al., 2007, Fortin et al., 2017] to clear the data from site effects (thus, harmonizing for additive and multiplicative effects of site, and iii) using ComBat as above, but explicitly preserving the variance associated with sex and age; an approach which will be referred to as *modified ComBat* in the following. Predictions made from raw data (thus, without any treatment of site effects) were used as a baseline model. An overview over all pipelines for all models can be found in Fig. 3.

2.5 Performance measures

2.5.1 Measures of model performance

Model performance is assessed using several common performance metrics. The Pearson's correlation coefficient ρ indicates the linear association between true and predicted value of cortical thickness measures. However, correlations

bioRxiv preprint doi: https://doi.org/10.1101/2021.02.09.430363; this version posted February 11, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

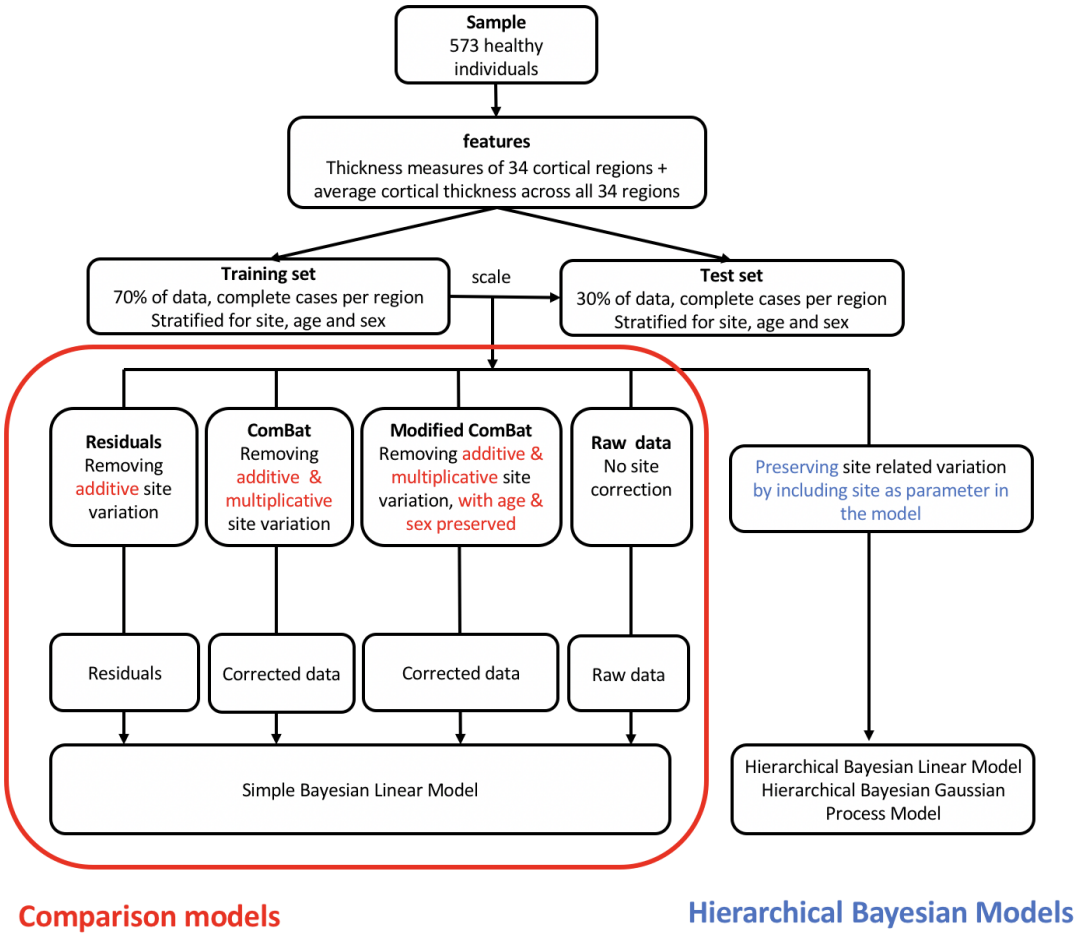


Figure 3: Pipelines for hierarchical Bayesian and comparison models

are not a sensitive error measure and cannot capture the "miss" between true and predicted value. Hence, we also calculate the standardized version of the root mean squared error (SRMSE) and the point-wise log-likelihood at each data point in the test set as a metric indicating deviance from the true value. However, these measures only take into account the estimate of the mean, and do not account for variations in the estimate of the variance. Thus we also compute the proportion of variance explained (EV) by the predicted values and a standardized version of the log-loss (mean standardized log-loss, MSLL [Rasmussen and Williams, 2006]). The latter does not only take into account the variance of the test set, but also standardizes it by the variance of the training set, making a comparison between the models possible. This step is necessary as various methods of correcting for site might also have an impact on the variance remaining in the data.

2.5.2 Measures of goodness of the simulation in Stan

Parameters indicating the goodness of the model simulation process in Stan itself, like convergence, effective sample size, and trace plots can be found in the supplementary material.

2.6 Model specification

In this section we show how normative models describing the association between age, and sex, and cortical thickness measures can be modeled on data comprising site effects using a hierarchical Bayesian linear mixed model with a Gaussian Process term, which allows to model non-linear association between age and cortical thickness measures.

bioRxiv preprint doi: https://doi.org/10.1101/2021.02.09.430363; this version posted February 11, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Following the notations of [Gelman, 2008, Rasmussen and Williams, 2006], we model a target vector $\mathbf{y} \in \mathbb{R}^{n\times 1}$ containing the the individual responses y_i for each subject $i = 1, \ldots, n$ and each region, using a latent function $\mathbf{f} = f(\mathbf{x})$. $f_i = f(x_i)$ is the evaluation of the latent function for an input vector $\mathbf{x}_i$ containing all p input variables of subject i, and is considered to differ from the true response variables by additive noise ϵ_i with the variance η_i and $N(0, \sigma^2)$ along the diagonal, with $\mathbf{I}$ being a $n \times n$ identity matrix:

$$\mathbf{y} \sim \mathcal{N}(\mathbf{f}, \sigma^2\mathbf{I}), \tag{1}$$

or, for the individual case:

$$y_i = f(x_i) + \epsilon_i, \tag{2}$$

with:

$$\epsilon_i \sim \mathcal{N}\,(0, \sigma^2).$$

The ability of the model to deal with site effects is obtained by introducing a random effect for site $s = 1, 2, \ldots, q$ so that the prediction for the ith subject is a combination of fixed and varying effects:

$$\mathbf{f} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \boldsymbol{\gamma}, \tag{3}$$

where $\boldsymbol{\gamma}$ is an additional non-linear component (defined in (5) below) and the estimate for one particular subject i is calculated the following:

$$f_i = \sum_{j=1}^{p} x_{ij}\beta_j + \sum_{s=1}^{q} z_{is}u_s + \gamma_i \tag{4}$$

with

$$\boldsymbol{\beta} \sim \mathcal{N}(0, \Sigma_j)$$

$$\mathbf{u} \sim \mathcal{N}(0, \Sigma_s).$$

Here, $\boldsymbol{\beta}$ is a $1 \times p$ vector containing the fixed regression weights corresponding to an $n \times p$ input matrix $\mathbf{X}$ with columns $j = 1, \ldots, p$. In case of non-centralized data one column of ones for an intercept offset has to be added. Similarly, $\mathbf{u}$ is a $1 \times q$ vector containing the weights for random effects across subjects, corresponding to a dummy coded $n \times q$ matrix $\mathbf{Z}$ modeling site. For all linear models, in (3) we assume γ_i = 0.

For the non-linear models we assume γ is a Gaussian Process with mean function $m(x)$ and covariance function $k(x, x')$ to allow for non-linear dependencies between the predictors and the target variable:

$$\boldsymbol{\gamma} \sim GP(m(x), k(x, x')). \tag{5}$$

In our case, we set $m(x) = 0$ and define $k(x, x')$ as the additional non-linear component in the following squared-exponential form:

$$k(x, x') = \sigma_f^2 exp(-\frac{1}{2l_j^2}(x, x'))^2, \tag{6}$$

with free parameters for the signal variance term σ_f^2 and the length scale l. Note this allows to specify two sources of variance: The signal variance σ_f^2 and the noise variance σ^2 as modeled in (1).

From a hierarchical Bayesian point of view, random effects are equal to a hierarchical structure of sources of variation. For modeling site effects, introducing a hierarchical structure has the benefit that it allows to include structural dependencies between sites via partial pooling. Thus, instead of modeling site effects as an effect shared between sites or independently from each other, a semi-independent association between sites can be obtained via assuming that all site parameters originate from a shared first-order prior distribution. This concept has been used elsewhere [Kia et al., 2020, Gelman et al., 2013, Mathys et al., 2012].

bioRxiv preprint doi: https://doi.org/10.1101/2021.02.09.430363; this version posted February 11, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

We hence induce shared priors and hyper priors $\boldsymbol{\theta_0}$ for site s, i.e. $\forall s, u_s \sim Inv\Gamma(2,2)$, and a uniform prior for the length scale $l \sim U(1,8)$. We use Stan [Carpenter et al., 2017, Stan Development Team, 2020b] to estimate all free parameters $\boldsymbol{\theta} = (\boldsymbol{\beta^T}, \mathbf{u^T}, \mathbf{l^T}, \sigma, \sigma_f)$ performing Bayesian inference:

$$p(\boldsymbol{\theta}|\boldsymbol{X}, \boldsymbol{y}, \boldsymbol{\theta_0}) = \frac{p(\boldsymbol{\theta}, \boldsymbol{X}, \boldsymbol{y}, \boldsymbol{\theta_0})}{p(\boldsymbol{X}, \boldsymbol{y}, \boldsymbol{\theta_0})} = \frac{p(\boldsymbol{\theta}|\boldsymbol{X}, \boldsymbol{y}, \boldsymbol{\theta_0})}{p(\boldsymbol{X}, \boldsymbol{y}, \boldsymbol{\theta_0})} \tag{7}$$

where $p(\boldsymbol{X}, \boldsymbol{y}, \boldsymbol{\theta_0}) = \int p(\boldsymbol{\theta}) p(\boldsymbol{X}, \boldsymbol{y}, \boldsymbol{\theta_0}|\boldsymbol{\theta})\, d\boldsymbol{\theta}$.

2.6.1 Posterior predictive distribution

We obtain the posterior predictive distribution y_* for a new sample x_* via:

$$\begin{aligned} p(y_*|\boldsymbol{y}) &= \int p(y_*, \boldsymbol{\theta}|x_*, \boldsymbol{X}, \boldsymbol{y}, \boldsymbol{\theta_0})\, d\boldsymbol{\theta} \\ &= \int p(y_*|\boldsymbol{\theta}, x_*, \boldsymbol{X}, \boldsymbol{y}, \boldsymbol{\theta_0})\, p(\boldsymbol{\theta}|\boldsymbol{X}, \boldsymbol{y}, \boldsymbol{\theta_0})\, d\boldsymbol{\theta} \\ &= \int p(y_*|\boldsymbol{\theta})\, p(\boldsymbol{\theta}|\boldsymbol{X}, \boldsymbol{y}, \boldsymbol{\theta_0})\, d\boldsymbol{\theta} \end{aligned} \tag{8}$$

as $\boldsymbol{y}$ and y_* are considered to be conditionally independent given $\boldsymbol{\theta}$ [Gelman et al., 2013].

Further, the predictive distribution can be computed exactly, writing the joint distribution of the known data $\boldsymbol{y}$, $\boldsymbol{X}$ and the new sample x_*, with the variance being determined by sample variance σ^2 and the Gaussian kernel $k(x, x')$:

$$k(x, x') = \begin{bmatrix} \boldsymbol{K} + \sigma^2 \boldsymbol{I} & k_* \\ k_*^T & k_{**} \end{bmatrix} \tag{9}$$

Here, $\boldsymbol{K}$ is an $n \times n$ covariance matrix of training data, k_{**} denotes the variance at the test sample points and k_* is the covariance between y_* the known data.

2.6.2 Comparison models

We compare the hierarchical Bayesian attempt to normative modeling to commonly used harmonization techniques in which site is controlled for by subtracting an estimate of the site effect from the data prior to fitting the normative model. These methods included: i) removing additive effects of site, by regressing out site effects via linear regression and using the residuals as input for the simple Bayesian linear model to obtain the normative scores, ii) harmonizing for additive and multiplicative effects of site using ComBat [Johnson et al., 2007, Fortin et al., 2017], iii) *modified ComBat*, thus, using ComBat as before, but preserving biological variance of interest i.e., sex and age. All these methods involve removing site effects prior to estimating the normative scores in contrast to our method in which we explicitly model site within the normative modeling framework. These harmonized data, obtained as output from the harmonization techniques, are subsequently used for normative modeling in a simple Bayesian linear model that does neither take into account site effects nor non-linear dependencies between age and measures of cortical thickness. Thus, equation (3) is reduced to $\boldsymbol{f} = \boldsymbol{X}\boldsymbol{\beta}$ with $\boldsymbol{\beta} \sim \mathcal{N}(0, \sum_j)$. In addition we use this simple Bayesian linear model to make one set of predictions for each regions from data that was not in any way harmonized for site (raw data model).

R [R Core Team, 2020] was used for preprocessing of all data and to create the data set where site was regressed out, and for preprocessing the data with ComBat [Johnson et al., 2007, Fortin et al., 2017].

2.6.3 Implementation: Normative modeling in Stan

Both the hierarchical Bayesian and the comparison model version of the normative models were implemented in Stan [Carpenter et al., 2017, Stan Development Team, 2020b], a probabilistic C++ based programming language to perform Bayesian Inference, and analyzed in R [R Core Team, 2020] using the package rstan [Stan Development Team, 2020a]. Stan allows to directly compute the log posterior density of a model given the known variables x and y. It uses the No-U-Turn Sampler (NUTS) [Hoffman and Gelman, 2014], a variation of Hamiltonian Monte Carlo Sampling [Duane et al., 1987, Neal et al., 2011, Neal, 1994] to generate representative samples from the posterior distribution of parameters and hyper parameters θ, each of which has the marginal distribution $p(\theta \mid y, x)$. This is achieved by first approximating the distribution of the data to a defined threshold in a warm up period and then randomly sampling

bioRxiv preprint doi: https://doi.org/10.1101/2021.02.09.430363; this version posted February 11, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

from the model, generating new draws of parameters for each iteration and calculating the response of the model. This approach of sampling instead of fitting allows for the simulation of complex models for which the derivation of an analytical solution of the posterior is computationally costly or not possible.

The Bayesian framework provides access to the full posterior distribution and to the distribution of all parameters. This allows to deduce the a variance estimate of each parameter, leading to a parameter estimate that is not only described by its mean, but also by the (un)-certainty around the mean estimation, providing information on its accuracy and reliability. Moreover, we can use the posterior distribution of each site-specific parameter from the training set as prior for the test set, allowing to make predictions for unfamiliar sites.

The Stan code for the HBLM, the HBGPM and the simple Bayesian linear model without site as predictor can be found at `https://github.com/likeajumprope/Bayesian_normative_models`.

2.7 Model simulation process in Stan

Parameters indicating the goodness of the model simulation process in Stan [Carpenter et al., 2017, Stan Development Team, 2020b] itself, like convergence, effective sample size and trace plots can be found in the supplementary material.

3 Results

Both the HBLM and the HBGPM outperformed all other comparison models with respect to all performance measures considered in this study. In detail, the HBLM and the HBGPM showed higher average values of the Pearson's correlation coefficient ρ (Table 2), lower average SRMSEs (Table 3), smaller average LL (Table 4) and higher average proportions of EV (Table 5) than all comparison models ($p < 0.001$ for all comparisons). For none of these comparisons did the non-linear HBGPM outperform the linear HBLM. In addition to the mean comparisons reported in Table 2 - 5, the distribution of all performance measures across all 34 regions and for average cortical thickness across the entire cortex per model can be found in Fig. 4. A detailed comparison of all models with respect to to ρ, SRMSE, EV and LL can be found in the supplementary material.

3.1 Mean standardized log loss

To also account for the second order statistics of the posterior distributions created by each model, we calculated the mean standardized log loss (MSLL). This measure can only be calculated for the test set, as it is the log loss standardized by the mean loss of the training data set [Rasmussen and Williams, 2006]. Hence, the MSLL gives an indication of whether a model is able to predict the data better than the mean of the training set (with more negative values being better). An overview of the MSLL for all cortical thickness measures of all regions for all models is given in Fig. 5a. The only models that perform better for most regions than the mean of the training data set are the Hierarchical Bayesian models ($MSLL_{HBGPM} < 0$ for all regions; $MSLL_{HBLM} < 0$ for all but one region), in contrast to prediction from the *residuals* and the *ComBat* model, where none of the predictions perform better than the mean of the training data set ($MSLL_{residuals} > 0$ for all regions; $MSLL_{ComBat} > 0$ for all regions, see Fig. 5a. The MSLL for the *modified ComBat* model and *raw data* model were region-dependent, with 45 % regions (16 out of 35) for the *modified ComBat* model and 17% of regions (six out of 35) for the *raw data* model performing better than predictions from the mean of the training set. It should also be mentioned that for some individual regions the comparison models performed very poorly (max $MSLL_{ComBat}$ = 356, max $MSLL_{mod.ComBat}$ = 138, max $MSLL_{raw}$ = 1252; max $MSLL_{residuals}$ = 517) and show measures that exceeded the plotted range of Fig. 5a. In contrast, the maximum MSLL for the hierarchical Bayesian models was max -0.056 for the *HBGPM* and max 0.08 for the *HBLM*.

3.2 Predictive Variance

We also observed that the models differ in the variance of predicted values, as visualized in Fig. 5b for average cortical thickness. For the *ComBat*, the *raw data* and the *residuals* model the range of predicted values was severely restricted (range predicted values *raw data*, test set: [2.60 - 3.03], range predicted values *residuals*, test set [2.64 - 3.00]; range predicted values *ComBat*, test set: [2.73 - 2.97]. These intervals cover 9.2 %, 7.9 % and 8.0 % of the original test set variance, respectively. The *modified ComBat* model retained 29.0% of the original test set variance (range predicted value *modified ComBat* [2.55 = 3.01]. In other words, all harmonization techniques had a reduced predictive variance and were instead biased toward predicting the mean, sometimes severely. In contrast, this bias was substantially reduced

bioRxiv preprint doi: https://doi.org/10.1101/2021.02.09.430363; this version posted February 11, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Site (Abbreviation)	Manufacturer	Platform	Voxel Size [mm]	TR [ms]	TE [ms]	n	males [%]	age range [years]
California I. of Technology (Caltech)	SIEMENS	TIM TRIO	1.0×1.0×1.0	1590	2.73	15	0.73	17-39
Carnegie Mellon U. (CMU)	SIEMENS	TIM TRIO	1.0×1.0×1.0	1870	2.48	13	0.77	20-40
Kennedy Krieger I. (KKI)	Philips	Achieva	1.0×1.0×1.0	8	3.7	33	0.73	8-12
Ludwig Maximilians U. Munich (LMU)	SIEMENS	VERIO	1.0×1.0×1.0	1800	3.06	15	1	18-29
NYU Langone Medical Center (NYU)	SIEMENS	ALLERGRA	1.3×1.0×1.3	2530	3.25	20	0.75	12-16
Olin I. of Living at Hartford Hospital (OLIN)	SIEMENS	ALLEGRA	1.0×1.0×1.0	2500	2.74	30	0.9	7-35
Oregon Health and Science U. (OHSU)	SIEMENS	TIM TRIO	1.0×1.0×1.0	2300	3.58	105	0.75	6-31
San Diego State U. (SDSU)	GE	MR750	1.0×1.0×1.0	NA	NA	15	1	8 12
Social Brain Lab (SBL)	Philips	INTERA	1.0×1.0×1.0	9	3.5	16	0.875	10-23
Stanford U. (STANFORD)	GE	SIGNA	0.859 × 1.500 × 0.859	8.4	1.8	27	0.85	9 33
Trinity Centre for Health Sciences (TRINITY)	Philips	Achieva	1.0 ×1.0 ×1.0	3.9	8.5	13	1	20-39
U. of California Los Angeles 1 ($UCLA_1$)	SIEMENS	TIM TRIO	1.0 × 1.0 × 1.2	2300	2.84	22	0.73	8-16
U. of California Los Angeles 2 ($UCLA_2$)	SIEMENS	TIM TRIO	1.0 × 1.0 × 1.2	2300	2.84	20	0.8	7-12
U. of Leuven 1 ($LEUVEN_1$)	Philips	Achieva	0.98× 0.98 × 1.20	9.6	4.6	25	1	12-25
U. of Leuven 2 ($LEUVEN_2$)	Philips	Achieva	0.98× 0.98×1.20	9.6	4.6	32	0.88	9-17
U. of Michigan 1 (UM_1)	GE	SIGNA	NA	250	5.7	13	0.85	9-13
U. of Michigan 2 (UM_2)	GE	SIGNA	NA	250	5.7	54	0.69	8-19
U. of PittsburghSchool of Medicine (PITT)	SIEMENS	ALLERGRA	1.1×1.1×1.1	2100	3.93	21	0.95	13 28
U. of Utah School of Medicine (USM)	SIEMENS	TIM TRIO	1.0×1.0×1.2	2300	2.91	43	1	8-39
Yale Child Study Center (YALE)	SIEMENS	TIM TRIO	1.0×1.0×1.2	1230	1.73	28	0.71	7-17

Table 1: The scanner parameters and sample specifications of the ABIDE data set.

bioRxiv preprint doi: https://doi.org/10.1101/2021.02.09.430363; this version posted February 11, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

in the hierarchical Bayesian models, which retained 57.0 % (*HBLM*) and 65.0 % (*HBGPM*) of the original test variance (range predicted values *HBLM*, test set: [2.43 - 3.23]; range predicted values *HBGPM*, test set: [2.38 - 3.28]).

	Mean Correlation (STD)		Post-hoc comparison					
ρ	training set	test set	HBLM	HBGPM	mod. ComBat	ComBat	residuals	raw data
HBLM	0.734 (0.06)	0.694 (0.06)		ns.	***	***	***	***
HBGPM	0.752 (0.05)	0.705 (0.06)	ns.		***	***	***	***
mod. ComBat	0.541 (0.15)	0.568 (0.16)	***	***		***	***	***
ComBat	0.289 (0.09)	0.343 (0.11)	***	***	***		ns.	***
residuals	0.267 (0.08)	0.329 (0.12)	***	***	***	ns		***
raw data	0.435 (0.14)	0.435 (0.16)	***	***	***	*	**	

Table 2: Post-hoc tests of correlations between true and predicted values. Cell values indicate post-hoc comparison significance values (adjusted by tukey method for a comparing a family of 6 estimates). Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1 ns. blue: test set. yellow: training set.

	Mean SRMSE (STD)		Post-hoc comparison					
SRMSE	training set	test set	HBLM	HBGPM	mod. ComBat	ComBat	residuals	raw data
HBLM	0.0608 (0.006)	0.066 (0.005)		n.s	***	***	***	***
HBGPM	0.0587 (0.006)	0.064 (0.006)	ns.		***	***	***	***
mod. ComBat	0.0763 (0.007)	0.075 (0.008)	***	***		***	n.s	ns.
ComBat	0.0872 (0.003)	0.085 (0.005)	***	***	***		***	***
residuals	0.0865 (0.003)	0.085 (0.004)	***	***	ns.	***		n.s
raw data	0.0808 (0.006)	0.085 (0.008)	***	***	***	***	***	

Table 3: Post-hoc tests of SRMSE between true and predicted values. Cell values indicate post-hoc comparison significance values (adjusted by tukey method for a comparing a family of 6 estimates). Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1 ns. blue: test set. yellow: training set.

LL	training set	test set
HBLM	-1.050	-1.121
HBGPM	-1.020	-1.109
ComBat mod.	-1.225	-1.193
ComBat	-1.374	-1.336
residuals	-1.381	-1.394
raw	-1.299	-1.335

Table 4: Averaged log loss for training and test set.

EV	training set	test set
HBLM	0.5674	0.5
HBGPM	0.5397	0.485
ComBat mod.	0.3146	0.338
ComBat	0.0918	0.122
residuals	0.0778	0.114
raw	0.2091	0.208

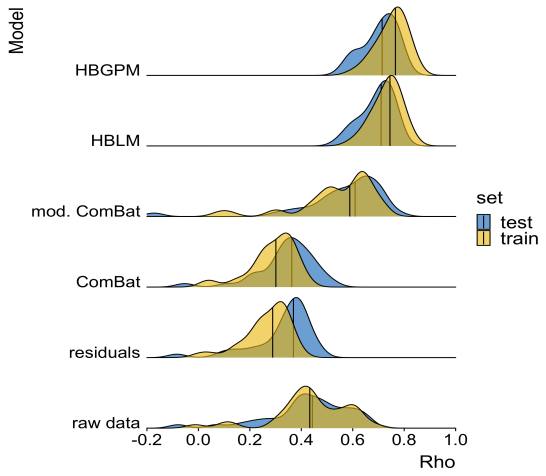
Table 5: Averaged explained variance for training and test set.

4 Discussion

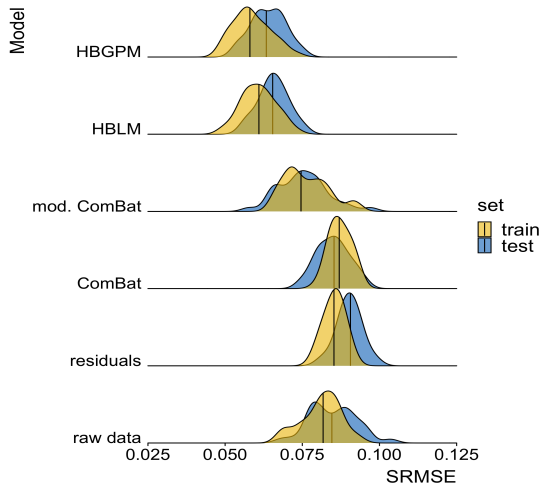
In this work, we aim to provide a method that allows the application of normative modeling to neuroimaging data sets that are affected by site effects resulting from pooling data between sites. In contrast to other methods of harmonizing for additive and multiplicative site effects in the data prior to the normative modeling (e.g., regressing out site effects, harmonization with *ComBat*), our approach is based on modeling site as predictor within the normative modeling framework. The benefit of this approach is that it does not entail removing variance and thus cannot lead to an overestimation of site variance and accidental removal of meaningful variation in case the latter is confounded with site variation. Using a hierarchical Bayesian approach, we propose two versions of normative models that were able to control for site effects. In both versions, site is modeled via a random intercept offset, but one version only models linear effects of age on cortical thickness (Hierarchical Bayesian Linear Model, HBLM), whereas the other version also includes a Gaussian process term in order to allow potential non-linear relationships between age and cortical thickness measures (Hierarchical Bayesian Gaussian Process Model; HBGPM).

The normative models are trained on a training set consisting of healthy individuals from the ABIDE data set (70% of the data from 20 different sites, within-site split, preserving the distribution of age and sex across training and

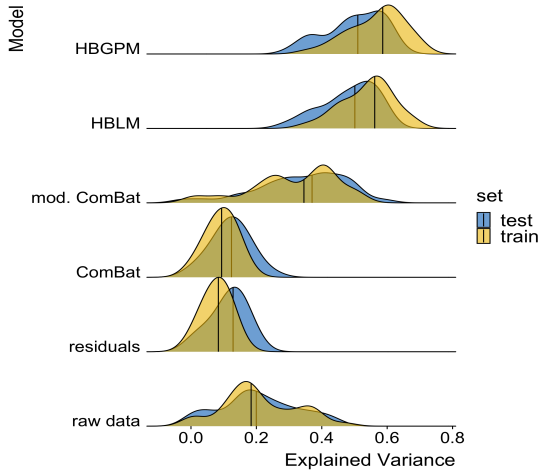
bioRxiv preprint doi: https://doi.org/10.1101/2021.02.09.430363; this version posted February 11, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.



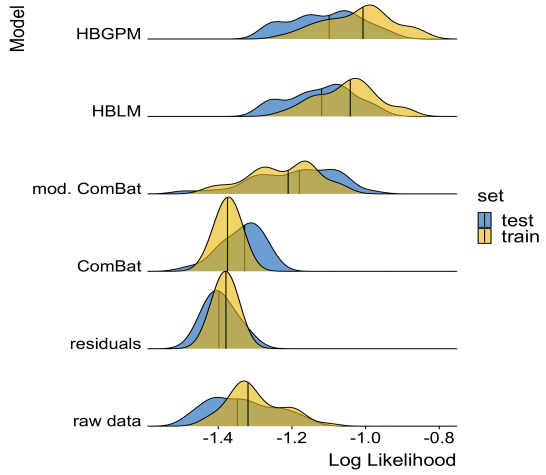
(a) Distribution of Pearson's correlation coefficient ρ for 35 cortical regions, indicating the correlation between true and predicted values, training and test set.



(b) SRMSE for 35 cortical regions, indicating the deviation true and predicted values of six different models for the training and the test set.



(c) Explained variance for 35 cortical regions, training and test set.



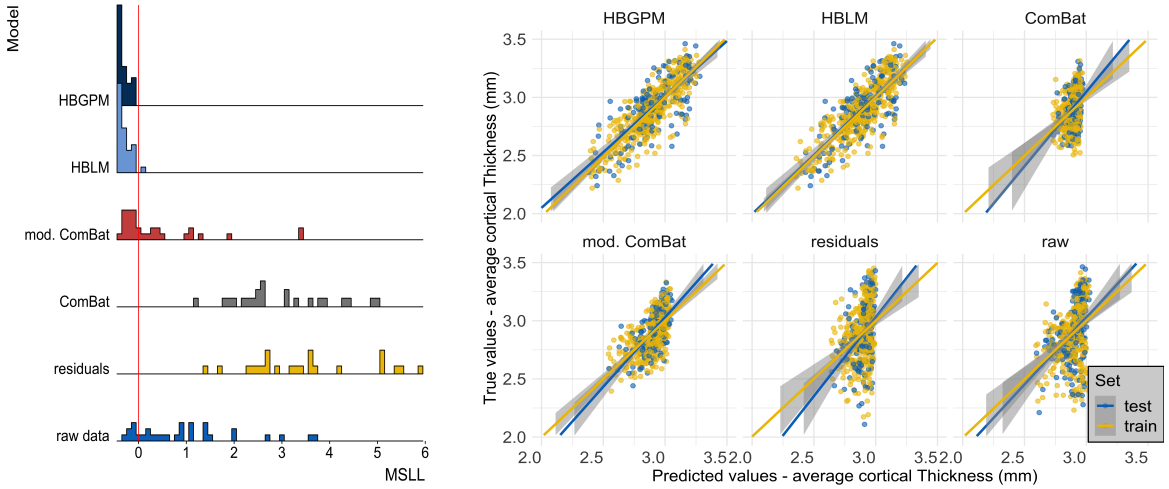
(d) Log likelihood distribution for 35 cortical regions, training and test set.

Figure 4: Performance measures

test set) and we present results from generalization to a test set (the remaining 30% of the data from the same sites). We compare the performance of our hierarchical Bayesian normative models explicitly modeling site effects applied to cortical thickness measures derived from FreeSurfer [Fischl et al., 2004]) to other commonly used methods to deal with site effects. These alternative methods included: i) regressing out site via linear regression and using the residuals, removing additive site variation, ii) applying *ComBat* [Fortin et al., 2017, Fortin et al., 2018] to harmonize additive and multiplicative site effects in the data, and iii) *modified ComBat*, hence applying ComBat while preserving age and sex effects in the data. Cortical thickness measures cleared from site effects using these alternative methods are used as dependent variables in a normative model with age and sex as predictors but excluding site. For comparison reasons, we also include a fourth model where we made predictions from raw data uncorrected for any site effects.

We report three main findings: (1) Our normative hierarchical Bayesian models (both the linear *HBLM* and non-linear *HBGPM* version), explicitly modeling site effects within the normative modeling framework, outperform all alternative harmonization models with respect to model fit, including correlations between true and predicted values (ρ), standardized root mean square error (SRMSE), explained variance scores (EV), log-likelihood (LL) and the mean standardized log loss (MSLL); (2) the non-linear model did not significantly improve prediction of cortical thickness

bioRxiv preprint doi: https://doi.org/10.1101/2021.02.09.430363; this version posted February 11, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.



(a) MSLL distribution for 35 cortical regions, test set.

(b) Predicted variance vs. actual variance for average cortical thickness for each model derived from predictions of 573 individuals.

Figure 5: Mean standardized log loss and predicted variance for 35 cortical regions

based on age, sex and site compared to the linear model; (3) all methods, but in particular the harmonization methods lead to an undesirable shrinking of the variance in the predictions.

We showed that when using neuroimaging structural data sets pooled across different sites and scanners for estimating normative models, better predictive performance can be achieved by including site as a predictor than using a two-step approach of first harmonizing the data with respect to site and subsequently creating a normative model using these "cleared" data. This conclusion is based on results showing that the hierarchical Bayesian models outperformed the harmonizing comparison models on all of the performance metrics we examined. This includes the predictions derived from data that was cleared from site effects by a version of ComBat [Fortin et al., 2017, Fortin et al., 2018] in which variation associated with age and sex was preserved, which was the best performing method across all harmonizing models. We observed a higher correlation between true and predicted values and LL values closer to zero for our hierarchical Bayesian models explicitly modeling site effects with a random intercept offset, indicating better model fit. As a key factor of normative models is that they are not only able to estimate the predictive mean, but also give an estimate of the predictive variance and variation around the mean [Marquand et al., 2019, Marquand et al., 2016], we also included explained variance scores and the MSLL as performance metrics. Our HBLM and HBGPM models showed higher explained variance than the alternative models. In addition, the HBLM and HBGPM showed a negative MSLL in the test set; a metric which contrasts the log loss between the true and predicted values by the loss that would be achieved using the mean and the variance of the training set [Rasmussen and Williams, 2006], thus capturing differences in variance in the data sets. This benefit in performance for the hierarchical Bayesian models is in line with previous literature using a similar paradigm [Kia et al., 2020]. [Kia et al., 2020] showed that a hierarchical Bayesian regression approach using site as a batch effect lead to a better performance than complete pooling, no pooling and ComBat. In detail, our findings match [Kia et al., 2020]'s findings with respect to the comparison between a normative model created from hierarchical Bayesian regression (HBR) and a *modified ComBat* version in a data set with the same sites in training and test set. Their findings are in line with ours with respect to ρ ([Kia et al., 2020]: *HBR* range: 0.4 - 0.9, *modified ComBat* range: 0.2 - 0.8), SMSE: ([Kia et al., 2020]: *HBR* range: 0.2 - 0.9, *modified ComBat* range: 0.4 - 1.0) and MSLL ([Kia et al., 2020]: *HBR* range: -0.7 - -1.0, *modified ComBat* range: -0.04 - 0.0), except that the MSLL for the *modified Combat* model was worse in our study (see Figs. 4a, 4b, 5a). Therefore, our findings replicate the findings of [Kia et al., 2020] using an independent data set and separate implementation and extend that method to model non-linear functions using a Gaussian process term.

We anticipated that the non-linear version of the normative model, which included a Gaussian Process for age, would perform better than the linear version, as studies have shown that the association between age and regions of cortical thickness can be non-linear, especially for older age ranges [Storsve et al., 2014]. However, our results showed similar performance in predicting cortical thickness based on age, sex and site for both linear and non-linear models. This might be due to the fact that the the age range in our sample was restricted, ranging from 6-40 years, thus likely capturing an age range where the association between age and cortical thickness is still mostly linear [Wierenga et al., 2014]. As a consequence, the non-linear version of the model was not able to improve the overall

bioRxiv preprint doi: https://doi.org/10.1101/2021.02.09.430363; this version posted February 11, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

performance. Nonetheless, since other structural brain measures, including sub-cortical volumes and cortical surface area [Wierenga et al., 2014, Raznahan et al., 2011], have shown stronger non-linear associations with age, non-linear normative models may outperform a linear model for other types of structural brain imaging measures.

Despite an overall good performance of our models, it should also be mentioned that the performance showed substantial variation between regions, as reflected in the variation in ρ values, SRMSE, EV, LL and MSLL within models. We assume that this due to the fact that, although *average* cortical thickness shows a strong association with age, different cortical brain regions differ in their association with age and the magnitude of this correlation also changes across the life span ([Storsve et al., 2014]).

All models, but in particular the comparison models, have a significant shrinkage effect on the variance of the predicted values, indicating that harmonization techniques remove variance that is useful in predicting the response variable. This is most extreme for regressing out site effects and leads to poor performance across all performance metrics. We also observe that the performance of the residuals model is similar to the *ComBat* model without the preservation of age and sex, which is particularly reflected in the similarities of predicted variance in Fig. 5b and in the SRMSE. Both models suffer a loss of more than 90% of their original test variance. In contrast, the performance improves when variables like age and sex are preserved, as demonstrated by an increase in performance measures when using the version of *ComBat* in which variation associated with age and sex was preserved. We argue that the similarity in performance between *ComBat* and the *residuals* model is an indicator of the same underlying process, showing a weakness of the harmonization approach: merely regressing out site effects leads to the removal of meaningful variation correlated with the predictors of interest (in this case age and sex), especially when these predictors of interest are correlated with the site effects, which subsequently led to worse predictions of cortical thickness based on age and sex. This can be partially prevented by preserving important sources of variation when regressing out site effects, as shown for the *modified ComBat* model, where specified sources of variance were preserved when regressing out site effects. However, our results show two additional flaws of the harmonization approach: 1) as already pointed out by [Kia et al., 2020], in order to specify sources of variance that should be retained, all those sources of variance have to be known, which is not always the case; 2) even with age and sex preserved the modified ComBat model only retains 40% of the original variance. Our hierarchical Bayesian models including the prediction-based approach, in contrast, preserves known and unknown interactions between site and biological covariates by specifically modeling site, thus overcoming this requirement. The result is reflected in larger proportions of variance retained (see Fig. 5b. The advantage of the hierarchical Bayesian approach becomes particularly clear when considering that the scores derived from normative models are relative scores describing the deviation from a predicted normative mean. Thus, the normative deviation score is not affected by the absolute value of the predicted mean, and the number of predictors in the model does not influence the normative score.

Previous attempts to estimate the centiles of normative models have included polynomial regression [Kessler et al., 2016], support vector regression [Erus et al., 2015], quantile regression [Huizinga et al., 2018, Lv et al., 2020] and Gaussian process regression [Wolfers et al., 2018b], providing different degrees of the ability to separate between sources of variances and making individual predictions (for an overview see [Marquand et al., 2019]). We chose a hierarchical Bayesian framework for the implementation of our normative model as it has several advantages. The distribution-based structure based on posteriors allows for the separation and integration of different sources of variances, including epistemic (uncertainty in the model parameters), aleatoric (inherent variability in data) and prior variation, which are all considered when predicting cortical thickness based on age, sex and site. This allows for both the integration of already known information in the form of priors into the predictions, and for an adjustment of the precision of the estimate based on the uncertainty at each data point. In addition, the Bayesian framework, as implemented in software packages like Stan [Carpenter et al., 2017, Stan Development Team, 2020b], allows to draw samples from the full posterior distribution at the level of individual participants, which leads to an exact estimate of all parameters instead of an approximation. In particular in comparison to quantile regression, the distributional assumption entailed in the hierarchical Bayesian approach also allows to get more precise estimates of the underlying centiles, particularly in the outer centiles, which are usually of primary interest and where the data are sparsest. The proposed Bayesian framework also offers an elegant way to integrate site effects into normative models. site effects can be modeled via a hierarchical random effect structure, in which different sites are modeled semi-independently, sharing variation via a combined prior of higher order. This approach, also known as partial pooling, allows for including site-specific variance into the prediction for site, while at the same time constraining the amount of between-site variation to a maximum.

Whilst the primary aim of this study was to develop a novel method for dealing with site effects specifically within a normative modeling framework, the method can be used as general approach to clear neuroimaging data from site, age and sex effects. This is due to the fact that a normative score describes an individual's cortical thickness *in relation to* the variance explained by the predictor variables in the normative model (age, sex and site). Hence, they can be seen as "cleaned" cortical thickness measures that can be the basis for further analysis, for example to establish the

bioRxiv preprint doi: https://doi.org/10.1101/2021.02.09.430363; this version posted February 11, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

association between cortical thickness measures and clinical or demographic information. Another potential clinical use of a normative model based on healthy controls could be, that, once established, it can be used to derive individualized deviation scores from individuals with a psychiatric or neurological disorder. Their individual deviation scores can be considered as the degree of deviation from the normative variation and be used for further analysis, for example to predict clinically useful information.

Our proposed method has two potential disadvantages. The first one is related to the computational cost associated with estimating the covariance matrix within the Gaussian Process for the non-linear models, which in our analysis amounted to 25 hours per model per region and could only be mastered via parallel processing on a cluster. This is due to the fact that using the non-linear Gaussian Process term becomes very time and memory expensive with growing *n* ($O(n^2)$). Thus, in cases in which the relationship between the predictor and the outcome is estimated to be close to linear, the need for the more complex non-linear model should be carefully considered. Secondly, the between-site split and the model at its current state only allow generalizations to a test set which includes individuals from the same sites as the training set, thus where the site variation is known. However, especially in clinical settings, generalizing the model and making predictions in data from new sites is an important additional goal. Despite the fact that we cannot use the posterior distribution of one particular site as a prior when applying the model to a new, unknown site, the hierarchical Bayesian framework still allows using the posterior parameter distributions of *all* sites as derived from the training data set as priors for site parameters when applying the model to a new site. This approach has already been successfully demonstrated in [Kia et al., 2020] where the posterior parameter distribution of site derived from the training data was fed as a informative prior for the site predictor in a normative model applied to the test data consisting of new (unknown) sites. This use of a so called informed priors leads to more accurate and precise predictions than the broad, unspecific prior that would have to be used in cases where the distribution of the data is unknown [Kia et al., 2020]. Thus, despite some loss in precision, the Bayesian framework can, in contrast to all other methods examined in this paper, be adapted to make predictions to new, unknown sites.

5 Conclusion

We proposed an extended version of a normative modeling approach that is able to accommodate for site effects in neuroimaging data. The method is superior to previous approaches, including regressing out site and versions of ComBat [Fortin et al., 2017, Johnson et al., 2007] and facilitates the estimation of normative models based on neuroimaging data pooled across many different scan sites. A further extension of the model to make generalizations to new sites and the application to clinical data will be the objective of future work.

6 Online material

The supplementary material and the Stan code for the HBLM, HBGPM and simple Bayesian linear model can be found at `https://github.com/likeajumprope/Bayesian_normative_models`.

bioRxiv preprint doi: https://doi.org/10.1101/2021.02.09.430363; this version posted February 11, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

References

[Bartlett, 1937] Bartlett, M. S. (1937). Properties of sufficiency and statistical tests. *Proceedings of the Royal Society of London. Series A-Mathematical and Physical Sciences*, 160(901):268–282.

[Bethlehem et al., 2018] Bethlehem, R., Seidlitz, J., Romero-Garcia, R., and Lombardo, M. (2018). Using normative age modelling to isolate subsets of individuals with autism expressing highly age-atypical cortical thickness features. *bioRxiv*, page 252593.

[Carpenter et al., 2017] Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., and Riddell, A. (2017). Stan: A probabilistic programming language. *Journal of statistical software*, 76(1).

[Chen et al., 2014] Chen, J., Liu, J., Calhoun, V. D., Vasquez, A. A., Zwiers, M. P., Gupta, C. N., Frannke, B., and Turner, J. A. (2014). Exploration of scanning effects in multi-site structural MRI studies. *Journal of Neuroscience Methods*, 23(15):37–50.

[Craddock et al., 2013] Craddock, C., Benhajali, Y., Chu, C., Chouinard, F., Evans, A., Jakab, A., Khundrakpam, B. S., Lewis, J. D., Li, Q., Milham, M., et al. (2013). The neuro bureau preprocessing initiative: open sharing of preprocessed neuroimaging data and derivatives. *Neuroinformatics*, 4.

[Desikan et al., 2006] Desikan, R. S., Ségonne, F., Fischl, B., Quinn, B. T., Dickerson, B. C., Blacker, D., Buckner, R. L., Dale, A. M., Maguire, R. P., Hyman, B. T., Albert, M. S., and Killiany, R. J. (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.

[Di Martino et al., 2014] Di Martino, A., Yan, C. G., Li, Q., Denio, E., Castellanos, F. X., Alaerts, K., Anderson, J. S., Assaf, M., Bookheimer, S. Y., Dapretto, M., Deen, B., Delmonte, S., Dinstein, I., Ertl-Wagner, B., Fair, D. A., Gallagher, L., Kennedy, D. P., Keown, C. L., Keysers, C., Lainhart, J. E., Lord, C., Luna, B., Menon, V., Minshew, N. J., Monk, C. S., Mueller, S., Müller, R. A., Nebel, M. B., Nigg, J. T., O'Hearn, K., Pelphrey, K. A., Peltier, S. J., Rudie, J. D., Sunaert, S., Thioux, M., Tyszka, J. M., Uddin, L. Q., Verhoeven, J. S., Wenderoth, N., Wiggins, J. L., Mostofsky, S. H., and Milham, M. P. (2014). The autism brain imaging data exchange: Towards a large-scale evaluation of the intrinsic brain architecture in autism. *Molecular Psychiatry*, 19(6):659–667.

[Duane et al., 1987] Duane, S., Kennedy, A. D., Pendleton, B. J., and Roweth, D. (1987). Hybrid monte carlo. *Physics letters B*, 195(2):216–222.

[Erus et al., 2015] Erus, G., Battapady, H., Satterthwaite, T. D., Hakonarson, H., Gur, R. E., Davatzikos, C., and Gur, R. C. (2015). Imaging Patterns of Brain Development and their Relationship to Cognition. *Cerebral Cortex*, 25(6):1676–1684.

[Feczko et al., 2019] Feczko, E., Miranda-Dominguez, O., Marr, M., Graham, A. M., Nigg, J. T., and Fair, D. A. (2019). The Heterogeneity Problem: Approaches to Identify Psychiatric Subtypes. *Trends in Cognitive Sciences*, 23(7):584–601.

[Fischl et al., 2004] Fischl, B., Van Der Kouwe, A., Destrieux, C., Halgren, E., Ségonne, F., Salat, D. H., Busa, E., Seidman, L. J., Goldstein, J., Kennedy, D., Caviness, V., Makris, N., Rosen, B., and Dale, A. M. (2004). Automatically Parcellating the Human Cerebral Cortex. *Cerebral Cortex*, 14(1):11–22.

[Fortin et al., 2018] Fortin, J. P., Cullen, N., Sheline, Y. I., Taylor, W. D., Aselcioglu, I., Cook, P. A., Adams, P., Cooper, C., Fava, M., McGrath, P. J., McInnis, M., Phillips, M. L., Trivedi, M. H., Weissman, M. M., and Shinohara, R. T. (2018). Harmonization of cortical thickness measurements across scanners and sites. *NeuroImage*, 167(June 2017):104–120.

[Fortin et al., 2017] Fortin, J. P., Parker, D., Tunç, B., Watanabe, T., Elliott, M. A., Ruparel, K., Roalf, D. R., Satterthwaite, T. D., Gur, R. C., Gur, R. E., Schultz, R. T., Verma, R., and Shinohara, R. T. (2017). Harmonization of multi-site diffusion tensor imaging data. *NeuroImage*, 161:149–170.

[Foulkes and Blakemore, 2018] Foulkes, L. and Blakemore, S.-J. (2018). Studying individual differences in human adolescent brain development. *Nature neuroscience*, 21(3):315–323.

[Fried, 2017] Fried, E. (2017). Moving forward: how depression heterogeneity hinders progress in treatment and research. *Expert Review of Neurotherapeutics*, 17(5):423–425.

[Gelman, 2008] Gelman, A. (2008). *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge University Press.

[Gelman et al., 2013] Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., and Rubin, D. B. (2013). *Bayesian data analysis*. CRC press.

bioRxiv preprint doi: https://doi.org/10.1101/2021.02.09.430363; this version posted February 11, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

[Gronenschild et al., 2012] Gronenschild, E. H. B. M., Habets, P., Jacobs, H. I. L., Mengelers, R., Rozendaal, N., van Os, J., and Marcelis, M. (2012). The Effects of FreeSurfer Version, Workstation Type, and Macintosh Operating System Version on Anatomical Volume and Cortical Thickness Measurements. *PLoS ONE*, 7(6):e38234.

[Han et al., 2006] Han, X., Jovicich, J., Salat, D., van der Kouwe, A., Quinn, B., Czanner, S., Busa, E., Pacheco, J., Albert, M., Killiany, R., Maguire, P., Rosas, D., Makris, N., Dale, A., Dickerson, B., and Fischl, B. (2006). Reliability of MRI-derived measurements of human cerebral cortical thickness: The effects of field strength, scanner upgrade and manufacturer. *NeuroImage*, 32(1):180–194.

[Hoffman and Gelman, 2014] Hoffman, M. D. and Gelman, A. (2014). The no-u-turn sampler: adaptively setting path lengths in hamiltonian monte carlo. *J. Mach. Learn. Res.*, 15(1):1593–1623.

[Huizinga et al., 2018] Huizinga, W., Poot, D., Vernooij, M., Roshchupkin, G., Bron, E., Ikram, M., Rueckert, D., Niessen, W., and Klein, S. (2018). A spatio-temporal reference model of the aging brain. *NeuroImage*, 169:11–22.

[Insel et al., 2010] Insel, T., Cuthbert, B., Garvey, M., Heinssen, R., Pine, D. S., Quinn, K., Sanislow, C., and Wang, P. (2010). Research domain criteria (rdoc): toward a new classification framework for research on mental disorders.

[Insel, 2014] Insel, T. R. (2014). The NIMH Research Domain Criteria (RDoC) Project: Precision Medicine for Psychiatry. *American Journal of Psychiatry*, 171(4):395–397.

[Johnson et al., 2007] Johnson, W. E., Li, C., and Rabinovic, A. (2007). Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics*, 8(1):118–127.

[Kessler et al., 2016] Kessler, D., Angstadt, M., and Sripada, C. (2016). Growth charting of brain connectivity networks and the identification of attention impairment in youth. *JAMA psychiatry*, 73(5):481–489.

[Kia et al., 2020] Kia, S. M., Huijsdens, H., Dinga, R., Wolfers, T., Mennes, M., Andreassen, O. A., Westlye, L. T., Beckmann, C. F., and Marquand, A. F. (2020). Hierarchical bayesian regression for multi-site normative modeling of neuroimaging data. In Martel, A. L., Abolmaesumi, P., Stoyanov, D., Mateus, D., Zuluaga, M. A., Zhou, S. K., Racoceanu, D., and Joskowicz, L., editors, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2020*, pages 699–709, Cham. Springer International Publishing.

[Kia and Marquand, 2018] Kia, S. M. and 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)*, volume 11072 LNCS, pages 127–135.

[Leek et al., 2010] Leek, J. T., Scharpf, R. B., Bravo, H. C., Simcha, D., Langmead, B., Johnson, W. E., Geman, D., Baggerly, K., and Irizarry, R. A. (2010). Tackling the widespread and critical impact of batch effects in high-throughput data. *Nature Reviews Genetics*, 11(10):733–739.

[Lv et al., 2020] Lv, J., Biase, M. D., Cash, R. F., Cocchi, L., Cropley, V., Klauser, P., Tian, Y., Bayer, J., Schmaal, L., Cetin-Karayumak, S., Rathi, Y., Pasternak, O., Bousman, C., Pantelis, C., Calamante, F., and Zalesky, A. (2020). Individual deviations from normative models of brain structure in a large cross-sectional schizophrenia cohort. *bioRxiv*, page 2020.01.17.911032.

[Marquand et al., 2014] Marquand, A. F., Brammer, M., Williams, S. C., and Doyle, O. M. (2014). Bayesian multi-task learning for decoding multi-subject neuroimaging data. *NeuroImage*, 92:298–311.

[Marquand et al., 2019] Marquand, A. F., Kia, S. M., Zabihi, M., Wolfers, T., Buitelaar, J. K., and Beckmann, C. F. (2019). Conceptualizing mental disorders as deviations from normative functioning. *Molecular Psychiatry*, 24(10):1415–1424.

[Marquand et al., 2016] Marquand, A. F., Rezek, I., Buitelaar, J., and Beckmann, C. F. (2016). Understanding Heterogeneity in Clinical Cohorts Using Normative Models: Beyond Case-Control Studies. *Biological Psychiatry*, 80(7):552–561.

[Mathys et al., 2012] Mathys, C. D., Prüssmann, K., Stephan, K. E., and Behrens, T. (2012). HIERARCHICAL GAUSSIAN FILTERING Construction and variational inversion of a generic Bayesian model of individual learning under uncertainty.

[Miller et al., 2016] 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., et al. (2016). Multimodal population brain imaging in the uk biobank prospective epidemiological study. *Nature neuroscience*, 19(11):1523.

[Mirnezami et al., 2012] Mirnezami, R., Nicholson, J., and Darzi, A. (2012). Preparing for Precision Medicine. *New England Journal of Medicine*, 366(6):489–491.

[Mueller et al., 2005] Mueller, S. G., Weiner, M. W., Thal, L. J., Petersen, R. C., Jack, C. R., Jagust, W., Trojanowski, J. Q., Toga, A. W., and Beckett, L. (2005). Ways toward an early diagnosis in alzheimer's disease: the alzheimer's disease neuroimaging initiative (adni). *Alzheimer's & Dementia*, 1(1):55–66.

bioRxiv preprint doi: https://doi.org/10.1101/2021.02.09.430363; this version posted February 11, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

[Neal, 1994] Neal, R. M. (1994). An improved acceptance procedure for the hybrid monte carlo algorithm. *Journal of Computational Physics*, 111(1):194–203.

[Neal et al., 2011] Neal, R. M. et al. (2011). Mcmc using hamiltonian dynamics. *Handbook of markov chain monte carlo*, 2(11):2.

[R Core Team, 2020] R Core Team (2020). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

[Rasmussen and Williams, 2006] Rasmussen, C. E. and Williams, C. K. I. (2006). *GAUSSIAN PROCESSES FOR MACHINE LEARNING*. MIT Press, Cambridge.

[Raznahan et al., 2011] Raznahan, A., Shaw, P., Lalonde, F., Stockman, M., Wallace, G. L., Greenstein, D., Clasen, L., Gogtay, N., and Giedd, J. N. (2011). How Does Your Cortex Grow? *Journal of Neuroscience*, 31(19):7174–7177.

[Stan Development Team, 2020a] Stan Development Team (2020a). RStan: the R interface to Stan. R package version 2.21.2.

[Stan Development Team, 2020b] Stan Development Team (2020b). Stan modeling language users guide and reference manual, version 2.25.

[Storsve et al., 2014] Storsve, A. B., Fjell, A. M., Tamnes, C. K., Westlye, L. T., Overbye, K., Aasland, H. W., and Walhovd, K. B. (2014). Differential Longitudinal Changes in Cortical Thickness, Surface Area and Volume across the Adult Life Span: Regions of Accelerating and Decelerating Change. *Journal of Neuroscience*, 34(25):8488–8498.

[Thompson et al., 2020] Thompson, P. M., Jahanshad, N., Ching, C. R. K., Salminen, L. E., Thomopoulos, S. I., Bright, J., Baune, B. T., Bertolín, S., Bralten, J., Bruin, W. B., Bülow, R., Chen, J., Chye, Y., Dannlowski, U., de Kovel, C. G. F., Donohoe, G., Eyler, L. T., Faraone, S. V., Favre, P., Filippi, C. A., Frodl, T., Garijo, D., Gil, Y., Grabe, H. J., Grasby, K. L., Hajek, T., Han, L. K. M., Hatton, S. N., Hilbert, K., Ho, T. C., Holleran, L., Homuth, G., Hosten, N., Houenou, J., Ivanov, I., Jia, T., Kelly, S., Klein, M., Kwon, J. S., Laansma, M. A., Leerssen, J., Lueken, U., Nunes, A., Neill, J. O., Opel, N., Piras, F., Piras, F., Postema, M. C., Pozzi, E., Shatokhina, N., Soriano-Mas, C., Spalletta, G., Sun, D., Teumer, A., Tilot, A. K., Tozzi, L., van der Merwe, C., Van Someren, E. J. W., van Wingen, G. A., Völzke, H., Walton, E., Wang, L., Winkler, A. M., Wittfeld, K., Wright, M. J., Yun, J.-Y., Zhang, G., Zhang-James, Y., Adhikari, B. M., Agartz, I., Aghajani, M., Aleman, A., Althoff, R. R., Altmann, A., Andreassen, O. A., Baron, D. A., Bartnik-Olson, B. L., Marie Bas-Hoogendam, J., Baskin-Sommers, A. R., Bearden, C. E., Berner, L. A., Boedhoe, P. S. W., Brouwer, R. M., Buitelaar, J. K., Caeyenberghs, K., Cecil, C. A. M., Cohen, R. A., Cole, J. H., Conrod, P. J., De Brito, S. A., de Zwarte, S. M. C., Dennis, E. L., Desrivieres, S., Dima, D., Ehrlich, S., Esopenko, C., Fairchild, G., Fisher, S. E., Fouche, J.-P., Francks, C., Frangou, S., Franke, B., Garavan, H. P., Glahn, D. C., Groenewold, N. A., Gurholt, T. P., Gutman, B. A., Hahn, T., Harding, I. H., Hernaus, D., Hibar, D. P., Hillary, F. G., Hoogman, M., Hulshoff Pol, H. E., Jalbrzikowski, M., Karkashadze, G. A., Klapwijk, E. T., Knickmeyer, R. C., Kochunov, P., Koerte, I. K., Kong, X.-Z., Liew, S.-L., Lin, A. P., Logue, M. W., Luders, E., Macciardi, F., Mackey, S., Mayer, A. R., McDonald, C. R., McMahon, A. B., Medland, S. E., Modinos, G., Morey, R. A., Mueller, S. C., Mukherjee, P., Namazova-Baranova, L., Nir, T. M., Olsen, A., Paschou, P., Pine, D. S., Pizzagalli, F., Rentería, M. E., Rohrer, J. D., Sämann, P. G., Schmaal, L., Schumann, G., Shiroishi, M. S., Sisodiya, S. M., Smit, D. J. A., Sønderby, I. E., Stein, D. J., Stein, J. L., Tahmasian, M., Tate, D. F., Turner, J. A., van den Heuvel, O. A., van der Wee, N. J. A., van der Werf, Y. D., van Erp, T. G. M., van Haren, N. E. M., van Rooij, D., van Velzen, L. S., Veer, I. M., Veltman, D. J., Villalon-Reina, J. E., Walter, H., Whelan, C. D., Wilde, E. A., Zarei, M., and Zelman, V. (2020). ENIGMA and global neuroscience: A decade of large-scale studies of the brain in health and disease across more than 40 countries. *Translational Psychiatry*, 10(1):100.

[Volkow et al., 2018] Volkow, N. D., Koob, G. F., Croyle, R. T., Bianchi, D. W., Gordon, J. A., Koroshetz, W. J., Pérez-Stable, E. J., Riley, W. T., Bloch, M. H., Conway, K., et al. (2018). The conception of the abcd study: From substance use to a broad nih collaboration. *Developmental cognitive neuroscience*, 32:4–7.

[Wierenga et al., 2014] Wierenga, L. M., Langen, M., Oranje, B., and Durston, S. (2014). Unique developmental trajectories of cortical thickness and surface area. *NeuroImage*, 87:120–126.

[Wolfers et al., 2019] Wolfers, T., Beckmann, C. F., Hoogman, M., Buitelaar, J. K., Franke, B., and Marquand, A. F. (2019). Individual differences v. the average patient: mapping the heterogeneity in ADHD using normative models. *Psychological Medicine*, pages 1–10.

[Wolfers et al., 2020] Wolfers, T., Beckmann, C. F., Hoogman, M., Buitelaar, J. K., Franke, B., and Marquand, A. F. (2020). Individual differences v. the average patient: mapping the heterogeneity in adhd using normative models. *Psychological Medicine*, 50(2):314–323.

[Wolfers et al., 2018a] 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., and Marquand, A. F. (2018a).

bioRxiv preprint doi: https://doi.org/10.1101/2021.02.09.430363; this version posted February 11, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Mapping the Heterogeneous Phenotype of Schizophrenia and Bipolar Disorder Using Normative Models. *JAMA Psychiatry*, 75(11):1146.

[Wolfers et al., 2018b] Wolfers, T., Doan, N. T., Kaufmann, T., Alnæs, D., Moberget, T., Agartz, I., Buitelaar, J. K., Ueland, T., Melle, I., Franke, B., et al. (2018b). Mapping the heterogeneous phenotype of schizophrenia and bipolar disorder using normative models. *JAMA psychiatry*, 75(11):1146–1155.

[Zabihi et al., 2019] Zabihi, M., Oldehinkel, M., Wolfers, T., Frouin, V., Goyard, D., Loth, E., Charman, T., Tillmann, J., Banaschewski, T., Dumas, G., et al. (2019). Dissecting the heterogeneous cortical anatomy of autism spectrum disorder using normative models. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*, 4(6):567–578.