Identifying predictors of within-person variance in MRI-based brain volume estimates

Karch, J.D.; Fivelich, E.; Wenger, E.; Lisofski, N.; Becker, M.; Butler, O.; ... ; Kühn, S.

NeuroImage 200 (2019) 575–589

Contents lists available at ScienceDirect

NeuroImage

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

Identifying predictors of within-person variance in MRI-based brain volume estimates

Julian D. Karch [a,b,*], Elisa Filevich [a,c,d], Elisabeth Wenger [a], Nina Lisofsky [e], Maxi Becker [e], Oisin Butler [a], Johan Mårtensson [f], Ulman Lindenberger [a,g], Andreas M. Brandmaier [a,g,1], Simone Kühn [e,h,1]

[a] Center for Lifespan Psychology, Max Planck Institute for Human Development, Lentzeallee 94, 14195, Berlin, Germany
[b] Psychological Institute, Faculty of Social and Behavioral Sciences, Leiden University, Wassenaarseweg 52, 2333, AK Leiden, the Netherlands
[c] Bernstein Center for Computational Neuroscience Berlin, Philippstr. 13, 10115, Berlin, Germany
[d] Institute for Psychology, Humboldt-Universität zu Berlin, Rudower Chaussee 18, 12489, Berlin, Germany
[e] Clinic and Policlinic for Psychiatry and Psychotherapy, University Clinic Hamburg-Eppendorf, Martinistraße 52, 20246, Hamburg, Germany
[f] Department of Clinical Sciences, Lund University, Box 117, 221 00, Lund, Sweden
[g] Max Planck UCL Centre for Computational Psychiatry and Ageing Research, Lentzeallee 94, 14195, Berlin, Germany
[h] Lise Meitner Group for Environmental Neuroscience, Max Planck Institute for Human Development, Lentzeallee 94, 14195 Berlin, Germany

ARTICLE INFO

Keywords:
Structural MRI
Statistical learning
Reliability
Longitudinal change
Time-of-day effects

ABSTRACT

Adequate reliability of measurement is a precondition for investigating individual differences and age-related changes in brain structure. One approach to improve reliability is to identify and control for variables that are predictive of within-person variance. To this end, we applied both classical statistical methods and machine-learning-inspired approaches to structural magnetic resonance imaging (sMRI) data of six participants aged 24–31 years gathered at 40–50 occasions distributed over 6–8 months from the Day2day study. We explored the within-person associations between 21 variables covering physiological, affective, social, and environmental factors and global measures of brain volume estimated by VBM8 and FreeSurfer. Time since the first scan was reliably associated with Freesurfer estimates of grey matter volume and total cortex volume, in line with a rate of annual brain volume shrinkage of about 1 percent. For the same two structural measures, time of day also emerged as a reliable predictor with an estimated diurnal volume decrease of, again, about 1 percent. Furthermore, we found weak predictive evidence for the number of steps taken on the previous day and testosterone levels. The results suggest a need to control for time-of-day effects in sMRI research. In particular, we recommend that researchers interested in assessing longitudinal change in the context of intervention studies or longitudinal panels make sure that, at each measurement occasion, (a) a given participant is measured at the same time of day; (b) all participants are measured at about the same time of day. Furthermore, the potential effects of physical activity, including moderate amounts of aerobic exercise, and testosterone levels on MRI-based measures of brain structure deserve further investigation.

1. Introduction

Brain imaging techniques, in particular, magnetic resonance imaging (MRI), are frequently used to characterize the morphological features or the functioning of the human brain *in vivo*. In structural MRI (sMRI) research, summary measures such as regional volume or cortical thickness are derived from comprehensive raw images. These measures are used to describe geometrical properties (e.g., size or shape) of grey matter structures such as the hippocampus, and the volume, thickness, or surface area of the cerebral cortex. Contemporary research aims at elucidating in how far these measures might be associated with behavioral changes reflecting various brain-related pathologies as well as changes reflecting maturation, learning, and senescence (Benasisch and Urs, 2018; Lindenberger et al., 2006; Lövdén et al., 2013).

* Corresponding author. Psychological Institute, Faculty of Social and Behavioral Sciences, Leiden University, Wassenaarseweg 52, 2333, AK Leiden, the Netherlands.
E-mail address: j.d.karch@fsw.leidenuniv.nl (J.D. Karch).
[1] These authors contributed equally to this manuscript.

https://doi.org/10.1016/j.neuroimage.2019.05.030
Received 16 January 2019; Received in revised form 8 May 2019; Accepted 10 May 2019
Available online 18 May 2019

Researchers interested in longitudinal changes over multiple measurement occasions often use fully or semi-automated pipelines such as voxel-based morphometry (VBM; Ashburner and Friston, 2000) or cortical thickness estimates as derived using FreeSurfer (Fischl, 2012). These longitudinal changes are important, for example, to understand the effects of aging on the brain or to assess the potential of an intervention to elicit brain plasticity (Lövdén et al., 2013). In these cases, methods that accurately measure small differences in brain structure between repeated measurements are crucial.

A critical factor that limits the sensitivity of change detection in longitudinal studies is the reliability of measurements (Brandmaier et al., 2018a,b). Typically, repeated MRI scanning of the same person within a short period does not result in identical images, even when settings of the MRI scanner are held constant (Morey et al., 2010). This may be due to a host of factors, some related to the MRI acquisition itself, such as temperature or humidity in the MRI scanner, some related to the participants' physical or physiological state, such as previous caffeine or water intake. If confounding factors happen to vary systematically across occasions, or across individuals at a given occasion, the observed variation in MR images might give rise to statistically significant differences in estimates of volume or cortical thickness across individuals or occasions even though it reflects short-lived fluctuations of no particular interest, rather than stable individual differences or long-term change. It follows that uncontrolled variation is also relevant for cross-sectional studies, as some of these factors might vary, but go unnoticed, among the individuals or groups of people who are being compared, potentially artificially increasing between-person differences, or masking or inflating group differences. Therefore, it is critically important to assess how much of the within-person variability in measures of brain structure can be explained by confounding factors. Knowledge about the variables influencing within-person variability may allow researchers to increase the reliability of their measures by –if possible— holding these particular factors constant. For example, participants may be asked to avoid certain behaviors before being scanned, or attempts might be made to control these confounds statistically.

The degree to which different measurement characteristics (e.g., session, day, or MR tomograph in multi-site studies) influence reliability of measurement can be identified and estimated in well-designed reliability studies (Brandmaier et al., 2018a,b). Here, we are interested in how much time-varying variables may serve as predictors of within-person variability.

Note that within-person variance may capture differences due to short term variability, long term change, and measurement error (Nesselroade, 1991). What constitutes useful predictors of within-person variance depends on which of the three is in the focus of the analysis. Here, we proceed under the assumption that most of our predictors are not related to true long-term change but merely reflect unsystematic variation that we want to remove from our measurements. This is more likely to be true for variables unrelated to the person such as scanner characteristics or environment variables. In contrast, we cannot exclude the possibility that systematic changes in person-level variables such as in the affective or physiological state are associated with true (short-term or long-term) changes in the outcome of interest. The obvious candidate representing true long-term change is the time elapsed since the first measurement point as it directly codes time and thus also captures long-term change. In sum, the aptitude of predictors as a control measure in future studies ultimately depends on the research question (e.g., is it targeted at short-term variation or long-term trends) and on how much we can assume the predictors' independence from that true change; here, we focus on a purely statistical evaluation of which predictors may explain away within-person variability, and we will discuss our findings in the light of the challenges mentioned earlier. We use publicly available data from the Day2day study (Filevich et al., 2017), in which six participants were scanned between 40 and 50 times over 6–8 months. At each measurement occasion information was recorded on a series of variables that were deemed to be plausible potential modulators of MRI images, according to previous reports or based on anecdotal evidence. This set of potential modulator variables included scanner characteristics, environment-related variables, and participant-specific parameters. The resulting longitudinal data enable us to explore the potential of a number of selected variables (individually and in their interactions) to predict within-person fluctuations in commonly used sMRI estimates of brain structure.

In the following, we carry out exploratory analyses to characterize the ability of the potential modulators to reduce within-person variance. We acknowledge that the large number of relatively arbitrary decisions when setting up analyses of this kind poses a potential threat to the validity and generalizability of the results obtained (Carp, 2012; Simmons et al., 2011). To address and attenuate this problem, we selected a diverse set of analysis strategies originating from different data analysis cultures to obtain a range of solutions to the problem of finding predictors, thus taking a variety of perspectives. Specifically, we applied classical statistical procedures based on the general linear model and supplemented them with statistical learning approaches that have been applied to an increasing range of research fields and problems in recent years. In doing so, we intended to compare the sensitivity of the different statistical techniques in exploring potential predictors of within-person fluctuations. We report and base our conclusions on the pattern of results obtained with the various approaches instead of cherry-picking any particular one post hoc. We emphasize that the present approach is hypothesis-generating (e.g., exploratory) rather than hypothesis-testing (e.g., confirmatory), and can be seen as a first step towards the identification and control of variables for the purpose of increasing the reliability of structural MR measurements.

2. Material & methods

2.1. Data set

The Day2day data set has been extensively described in Filevich et al. (2017). For convenience, we briefly report the study characteristics that are relevant to the present paper. The original data collection was approved by the Ethics Committee of Charité University Clinic, Berlin; see Filevich et al. (2017).

2.2. Participants

Six participants (1 male, mean age 28 years, $SD = 3.06$ years, range: 24–31 years) volunteered to contribute to the data set, for which they were scanned 40–50 times over 6–8 months. In total, 280 measurement points were obtained across all participants. No participant had a diagnosis of a psychiatric disorder or had previously suffered from a mental disease.

Data collection took place between July 2013 and February 2014. In the original study, the investigators aimed at collecting MR images from each participant two to three times a week to capture short-term fluctuations, but each participant was free to arrange a scanning regime that would optimally fit into his or her schedule. Additionally, scanning depended on the availability of the MR scanner. As a result, the MR data were not always collected at regular intervals. The time elapsed between two measurements was 4.24 days on average (SD = 4.52 days, min = 1 days, max = 33 days). Filevich et al. (2017) provide a detailed overview of the temporal distribution of each participant's scanning sessions.

2.3. MRI data

Structural images were collected using a three-dimensional T_1-weighted magnetization prepared gradient-echo sequence (MPRAGE) with the following parameters: TR = 2500 ms, TE = 4.77 ms, TI = 1100 ms, FOV = 256 × 256 × 192 mm^3, flip angle = 7°, bandwidth = 140 Hz/pixel, 1 × 1 × 1 mm^3 voxel size, 9:20 min duration.

2.4. Brain structure measures

Structural data were processed using VBM8 (http://dbm.neuro.uni-jena.de/vbm.html) and SPM8 (http://www.fil.ion.ucl.ac.uk/spm) using default parameters. We only used the cross-sectional processing stream, where each structural image is processed separately. This allowed us to obtain reliability estimates that could be compared to cross-sectional studies, which are currently more common than longitudinal studies. VBM8 involves bias correction, tissue classification, and affine registration. The affine-registered grey matter and white matter segmentations were used to build a customized DARTEL template. We used the measures grey matter volume (VBM-GM), white matter volume (VBM-WM), and grey matter + white matter + cerebrospinal fluid volume (VBM-Total).

Cortical segmentation was performed using the FreeSurfer image analysis suite (http://surfer.nmr.mgh.harvard.edu/). The technical details of these procedures have been described thoroughly elsewhere (Fischl, 2012). All reconstructed data were visually checked for segmentation accuracy at each time point. No manual interventions with the MRI data were performed. Again, we only used the cross-sectional processing scheme, for the same reasons as mentioned above for the VBM analysis. As measures of interest, we extracted total grey matter volume (FS-GM), total cortical volume (FS-Cortex), and total intracranial volume (FS-ICV). Note that FS-ICV represents the total volume covered by the input surface, and therefore includes cortical and subcortical structures, as well as the ventricles. In turn, FS-GM includes both cortical and subcortical structures, whereas FS-Cortex includes only cortical volume.

2.5. Variables

Among those available in the Day2day data set, we selected an *a priori* subset of variables related to scanner status, participant behavior and affect either during the scan or the 24 h before scanning. We restricted our analyses to a list of $p = 21$ predictors that have been shown or are expected to affect measures of brain structure (Table 1). For a detailed description of all available variables, please refer to Filevich et al. (2017).

In addition to the variables available in the Day2day data set, we also included a sMRI data quality measure as a predictor. Specifically, we used Freesurfer's built-in measure of image quality, namely the number of defect holes, which has been suggested recently as a control variable (Rosen et al., 2018).

2.6. Exploratory data analysis

We initially planned to partition the data set into an exploration set and a confirmation set. However, after exploration on the exploration data set, we performed a power analysis, which showed that the confirmation set was not large enough to test the generated hypotheses with adequate power given the expected small effect size. Thus, we decided to use all data for exploration and to label our results as exploratory.

2.7. Within-person consistency

We quantified the within-person consistency in sMRI measures by the intra-class correlation (ICC), as it standardizes the within-person vari-

Table 1
Description and abbreviations for all variables considered.

Variable	Short Label	Comment	Assessment Period	Missing Data Points
General				
Days since the first scan of this person	Days Since First Scan		Scan Session	0
Time of start of the scanning session	Time of Day		Scan Session	11 (3.91%)
Minimum outside temperature on the day of the scan (°C)	Min. Outside Temp.	All weather variables were obtained from the German Weather Service.	Scan Session	0
Maximum outside temperature on the day of the scan (°C)	Max. Outside Temp.		Scan Session	0
Hours of sunshine on the day of the scan	Hours of Sunshine		Scan Session	0
Scanner Characteristics				
MR room temperature (°C)	Room Temperature		Scan Session	2 (0.71%)
MR room Humidity (%)	Room Humidity		Scan Session	2 (0.71%)
MR helium level (%)	Helium Level		Scan Session	7 (2.49%)
Number of Defect Holes	Surface Holes	Measures the sMRI data quality	Scan Session	1 (0.36%)
Physiological Variables				
Caffeine intake in the last 24 h	Caffeine Intake Last 24 h	In an equivalent number of cups of coffee	24 h	0
Caffeine intake in the last 2 h	Caffeine Intake Last 2 h	In an equivalent number of cups of coffee	2 h	6 (2.14%)
Cocoa intake (g) in the last 24 h	Cocoa Intake Last 24 h		24 h	71 (25.27%)
Cocoa intake (g) in the last 2 h	Cocoa Intake Last 2 h		2 h	63 (22.42%)
Weight (kg)	Weight	Participants were weighed without their shoes but otherwise fully dressed.	Scan Session	2 (0.71%)
Alcohol intake in the last 24 h	Alcohol Intake Last 24 h	In the number of alcoholic drinks	24 h	0
Liquid intake in the last 24 h (l)	Liquid Intake Last 24 h		24 h	0
Blood pressure (mmHg, systolic and diastolic)	Blood Pressure Systolic Blood Pressure Diastolic		Scan Session	11 (3.91%), 10 (3.56%)
Estradiol in (pg/mL)	Estradiol	Measured using saliva samples	Scan Session	18 (6.41%)
Testosterone (pg/mL)	Testosterone	Measured using saliva samples	Scan Session	14 (4.98%)
Behavioral and affective variables				
General stress subjective rating	Stress Last 24 h	Subjective rating of the last 24 h on 1–6 Likert scale	24 h	0
Number of steps taken on the day before the scanning	Steps Previous Day	Measured with a FitBit© activity tracker (https://www.fitbit.com)	24 h	14 (4.98%)

Note. Missing Data Points refers to the number (percentage) of missing data. The number represents the number of sessions across all subjects for which this variable is missing.

ance with the total variance, such that if there is no within-person variance (and non-zero between-person variance), the ICC is 1 and if there is only within-person variance the ICC is 0:

$$ICC = 1 - \frac{\sigma_\epsilon^2}{\sigma_b^2 + \sigma_\epsilon^2}$$

The between-person variance σ_b^2 and the within-person variance σ_ϵ^2 were estimated using a random intercept model. Note that with six participants there is relatively little information available to estimate the between-person variance accurately. Indeed, a small simulation study (see supplementary materials) revealed that with 6 participants and 50 measurements per participant the restricted maximum likelihood estimator for the between-person variance σ_b^2 is slightly biased downwards relative to the true value. This, in turn, leads to a slight downward bias of the ICC.

In the context of this study, the ICC should be interpreted with caution. Typically, the ICC is used as an estimate of test-retest reliability (Caceres et al., 2009). This, however, rests on the assumption that the stability of true scores is perfect. To the extent that individuals change in true scores over time, any departure of the ICC from 1.0 may represent a lack of stability, and not a lack of reliability (Brandmaier et al., 2018a,b; Brandmaier et al., 2018a,b). Here, we merely interpret ICC to be the proportion of the total variance that is due to between-person variance.

3. Analysis strategy

We selected a variety of statistical approaches to investigate which variables or combination thereof are predictive of within-person fluctuations. We targeted the most common traditional procedures as well as the most suitable statistical learning procedures to robustly identify predictors of within-person fluctuations. In the following, we describe the methods as well as our rationale for their selection.

3.1. Classical statistical methods

3.1.1. Within-person prediction matrix

In order to investigate to what degree each single variable is predictive of a given brain measure, we employed the strategy proposed by Bland and Altman (1995): First, a baseline model was fitted with only the person as a predictor, that is, a model in which the person-specific intercepts are modeled but no other predictors are included. Then, the respective variable of interest was added as the only predictor (full model). This then resulted in one full model for each predictor-outcome pair. We obtained a regression coefficient and associated *p*-value for each pairing of a predictor and an outcome by performing a model comparison between the baseline model and the corresponding full model using *F*-tests. This is similar to the data-analytic approach taken in standard MRI analysis via statistical parametric mapping (Ashburner, 2012). We call this analysis approach the *within-person prediction matrix*.

3.1.2. Stepwise regression

In stepwise regression, models are iteratively expanded by adding one locally best predictor after the other until a stopping criterion is reached. For our analysis, this translates into the following approach: For a given brain measure, the baseline model with only the participant as a predictor served as the starting point. The most influential variable was then added to the model (i.e., the variable that explains the most within-person variance on its own). As an estimate of explained variance, we chose the adjusted R^2 difference between the model with the added variable and the base model. Once the most influential variable had been added, the process was repeated with the obtained model (with the variable added) as the new base model. That is, the question now became: Which variable explains most within-person variance on top of the already selected variable(s)? This process was repeated until no variable significantly improved model fit, as measured by the *F*-test.

The stepwise regression approach has been widely criticized (Huberty, 1989). Much of this criticism is concerned with the greedy (i.e., locally but not globally optimal) nature of stepwise regression. While we acknowledge these limitations, we nevertheless report results here, as greedy model building is still popular. For example, when constructing structural equation models, the widely recommended approach to start with a so-called null model and to extend the model as long as the model fit improves significantly is also a greedy, non-optimal model construction process (Homburg and Dobratz, 1992). As an alternative, regularized regression is typically regarded as superior to stepwise regression. In Section 3.2.1, we, therefore, report the results of a specific regularization approach, namely LASSO regression, which is commonly recommended as a better procedure.

Beyond the general problems of stepwise regression, a small simulation study (see supplementary materials) revealed that for a data set with the properties of the Day2day data set, the adjusted R^2 difference is an overly optimistic measure of the true R^2. Importantly, however, the Type-I error rate was not inflated. We nevertheless report the adjusted R^2 difference because it is commonly employed and because, despite these valid criticisms, it remains the best alternative available.

3.1.3. Omnibus test

ANOVA is typically used to compare more than two groups with each other. It is generally recommended that an omnibus test should be performed as the first step in order to test the hypothesis that differences between groups exist. Applying this idea of an omnibus test to the issue at hand led us to the following procedure: We first fitted a base model, with the person as the only predictor. Then, all variables of interest were added simultaneously, resulting in the full model. As the final step, these two models were compared via an *F*-test. The major weakness of this method is its low statistical power. The more variables there are in the data set that have no association to the outcome, the lower is the chance to detect a variable with a true association to the outcome. In contrast, LASSO regression (see Section 3.2.1) adds penalties for regression weights and effectively formalizes a prior belief over the regression weights such that most of them are expected to be zero.

In an ANOVA, a post hoc analysis is performed after the omnibus hypothesis has been rejected, in order to identify which pairs of groups differ. In analogy, we performed a post hoc analysis to identify which variables are true predictors of the outcome. In Section 4.2.3, we report the *p*-value for every outcome variable pair.

3.2. Statistical learning methods

In the following, we introduce two approaches commonly taken in statistical learning (also known as machine learning), namely LASSO regression (Tibshirani, 1996) and random forests (Breiman, 2001). In contrast to the general linear model, on which the analyses reported above were based, there is no standard recommendation on the way to apply these methods to repeated-measures data like the Day2day data. We employed the following strategy: We eliminated person-specific effects on the outcomes by subtracting the person-specific mean from them. Thus, the statistical learning methods predicted the within-person fluctuations but not the between-person differences.

To measure the utility of these models, we relied on out-of-sample statistics. Statistical learning models are often so flexible that a perfect in-sample-fit (for example R^2) is not meaningful. As a counter-measure, such models are typically evaluated in terms of their performance on new data, therefore called out-of-sample statistics. We used the out-of-sample R^2, which we calculated by taking the square of the correlation between the predicted and the true within-person fluctuations. The strategy for obtaining the necessary out-of-sample predictions differed between the two approaches and will be explained in the corresponding subsections.

Testing whether an out-of-sample R^2 is higher than chance is problematic because its null distribution is not known. In practice, it would be

possible to derive the null distribution empirically using a permutation approach, as is, for example, done in the machine learning toolbox PRoNTo for neuroimaging (Schrouff et al., 2013). However, in this case, the need for computationally intensive methods (multiple imputation and nested cross-validation) prohibited this option. Instead, we used a heuristic effect-size cut-off of 1% in out-of-sample R^2 to determine whether a statistical learning model performed better than random guessing.

3.2.1. LASSO regression

For our analysis, the number of potential variables was relatively high ($p = 21$) and the number of observations relatively low ($N = 281$). Thus, the standard model-fitting algorithm for the linear model was likely to overfit. To remedy this issue, we also included a penalized linear model. Instead of the regular linear model, which finds the parameter values that maximize the model fit, for example, as quantified by the residual sum of squares (RSS), penalized linear models maximize a trade-off between model fit and a penalty, which is higher for more complex models. Different penalization strategies mostly differ in the employed penalty term and consequently in the quantification of model complexity.

For this analysis, we used the LASSO regression (Tibshirani, 1996). The objective function that is minimized to find regression weights β is:

$$f(\beta) = RSS(\beta) + \lambda||\beta||_1.$$

The non-negative parameter lambda λ (often called a hyper-parameter) is a scalar that determines the relative importance of model fit and simplicity. The higher the value of λ the more model complexity is penalized. Model complexity is quantified as the l^1-norm of the weight vector, which is equivalent to summing the absolute values of all weights. Thus, a penalty of zero is achieved if and only if all regression weights are zero. Compared to other approaches of quantifying model complexity (most notably, the l^2-norm used in ridge regression), the l^1-norm used in LASSO regression has the advantage that many regression weights are set to zero (Tibshirani, 1996). Thus, it can also be used as a feature selection procedure. Indeed, it is commonly suggested as a superior alternative to stepwise regression (Flom and Cassel, 2007).

Following the standard recommendations (Tibshirani, 1996), we used cross-validation with the residual sums of squares as the performance metric to find the hyper-parameter λ. To obtain out-of-sample predictions of the within-person fluctuation based on the resulting model, we again used cross-validation, which results in nested cross-validation (for a detailed description of nested cross-validation, see Karch et al., 2015). For both cross-validation steps, we used regular 10-fold cross-validation.

As the variable importance measure, we report the final estimated weight vector β. Prior to applying LASSO, we standardized all variables such that the absolute value of the coefficients within the weight vector can be interpreted as variable importance.

3.2.2. Random forests

Both the general linear model and LASSO regression only consider linear relationships between the predictor variables and the dependent variables, which means that interactions between variables could not be explored for their predictive potential using these methods. The standard approach taken to examine interactions is to include them as multiplicative terms in the linear model. However, this approach massively increases the number of variables and thereby the risk of overfitting. As an alternative, we employed random forests, which can find non-linear relationships, including interactions, and at the same time implement effective counter-measures against overfitting (Breiman, 2001).

Random forests build on decision trees, a popular statistical learning method that is typically used for classification but can also be used for regression as in our study. The decision tree method is explained in detail elsewhere (e.g., Hastie et al., 2009, Chapter 9.2).

To reiterate, the advantage of decision trees for this study is that they also model non-linear relationships, including interactions among variables. Their disadvantage is that they are susceptible to overfitting. Small changes in the training set, for example, when leaving out a few cases, typically lead to drastic changes in the decision tree. Hence, Breiman (2001) introduced random forests as a counter-measure against overfitting by decision trees. The basic idea is as follows. Instead of growing only one tree, many trees (that is, a forest) are grown. The final prediction is then the average prediction across all trees. To introduce heterogeneity between trees, each tree is grown with a random subset of the data (in our language, a random subset of measurements and variables). For a detailed description of random forests, see Breiman (2001).

Since each tree is grown with a subset of the data only, out-of-sample predictions can be obtained without the need for cross-validation. The out-of-sample prediction for each data point is obtained by using only those trees for the prediction that did not include this data point in the training set. These predictions are also known as out-of-bag predictions.

To obtain an estimate of the importance of each variable in the random forest model, we used random forest variable importance values. Essentially, for each variable, the deterioration of the out-of-sample performance is estimated by random permutation of the values of the respective variables. The performance deterioration is expressed in the percentage of out-of-sample mean squared error increase. For a detailed description of the procedure, see Breiman (2001, Chapter 10).

Random forests also possess hyper-parameters that control how the algorithm grows the forest. For the individual trees, we set the parameters such that each tree was grown to its full depth. We did not employ pruning to avoid overfitting because averaging over many trees is already an effective countermeasure. Indeed, it is well known that for an ensemble method (averaging the predictions of many models) to perform well, substantial diversity across the individual models is required (Kuncheva, 2004, Chapter 10), which speaks against pruning. We averaged across 10,000 trees. The performance of a random forest typically increases asymptotically with the number of trees up to a certain threshold (Oshiro et al., 2012). Thus, the number of trees represents a compromise between performance and computational cost. Our choice of 10,000 trees is well above the heuristic of 128 promoted by Oshiro et al. (2012), and at the same time proved computationally feasible.

3.3. Treatment of missing values

The VBM measures had no missing values. For the FreeSurfer measures, there was one missing value for FS-GM and FS-Cortex and no missing values for intracranial volume. To treat the missing data in the brain volume variables, we employed list-wise deletion.

The missing data information for all predictor variables can be found in Table 1. In summary, 9 out of the 22 predictors had no missing values. Of the remaining predictors, only 3 had more than 5% missing data (Estradiol [6.41%], Cocoa Intake Last 24 h [25.26%], Cocoa Intake Last 2 h 22.42%). Nevertheless, because of the multivariate nature of our analysis, using list-wise deletion, that is, dropping every data point that has a missing value in any of the variables would have resulted in losing more than half of the data.

Therefore, we employed multiple imputation (e.g., Van Buuren, 2012). More specifically, we applied multiple imputation using fully conditional specification as implemented in the R package *mice* (Buuren and Groothuis-Oudshoorn, 2011). We chose an appropriate imputation model for each variable. For the general and the scanner variables, we used predictive mean matching (Van Buuren, 2012, Chapter 3.4). To pick potential predictors for each variable, we did not solely rely on estimates of the correlations but also on prior knowledge. This was to account for the uncertainty of the correlation estimates due to the relatively small data set size. For the remaining, person-specific variables, we added a random intercept for the persons to the imputation model, as provided by the R package *miceadds* (Robitzsch et al., 2018) to account for their nested nature. The selection of the variables of interest again relied on prior knowledge and estimates of the *within-person* correlations.

For every variable, the distribution of the imputed values within each

person was visually checked against the distribution of non-missing values. To aggregate analysis results across the imputed data sets (for example, adjusted R^2 values, variable importance values, and hypothesis test results), we relied on findings presented in Van Buuren (2012, Chapter 6). We standardized all continuous variables to reduce numerical estimation problems.

Before imputation, every variable was thoroughly inspected for outliers. Values that were well outside the reasonable range were set to missing such that multi-imputation could be employed to deal with the resulting uncertainty adequately. The only exception among the variables was Helium Level. Here, a piecewise linear regression model with days since beginning of the study as the predictor proved accurate enough to justify single imputation for the 7 (2.49%) missing values.

4. Results

4.1. Within-person variance

Fig. 1 shows that the estimated ICC was greater than 0.95 for all brain structure measures. Consequently, for all six brain measures, only 5% of the total variance was within-person variance. For four brain measures (VBM-GM, VBM-WM, VBM-Total, and FS-ICV]) it was even higher than 0.98.

4.2. Classical statistical methods

4.2.1. Within-person prediction matrix

Fig. 2 presents the full p-value matrix showing the p-values corresponding to the hypothesis test that variable x linearly predicts within-person fluctuations of brain measure y. Like a statistical parametric mapping analysis, our analysis also had to account for multiple comparisons. As a first step, we took the conventional 0.05 p-value threshold and corrected for multiple comparisons using the conservative Bonferroni correction (this results in a $p = 0.05/(6 \times 21) \approx 0.0004$ threshold for the individual comparisons). Using this strategy, none of the combinations between a brain measure and a variable achieved statistical significance. As a next exploratory step, we used a more liberal p-value cutoff ($p < 0.01$ for the individual comparisons). This resulted in ten significant predictor-outcome pairs (see Fig. 1). Six of these ten pairs involved the two FreeSurfer measures FS-GM and FS-Cortex. These are also the brain measures with the lowest ICC and thus, with proportionally the highest within-person variance to explain away. Therefore, we focused on analyzing FS-GM and FS-Cortex only. However, the same data-analytic approach could be taken for any other brain measure.

The variables that we identified as significant predictors (at the 0.01 level) were: Time of Day, Minimum Outside Temperature, and Maximum Outside Temperature for both brain measures. Days Since First Scan and

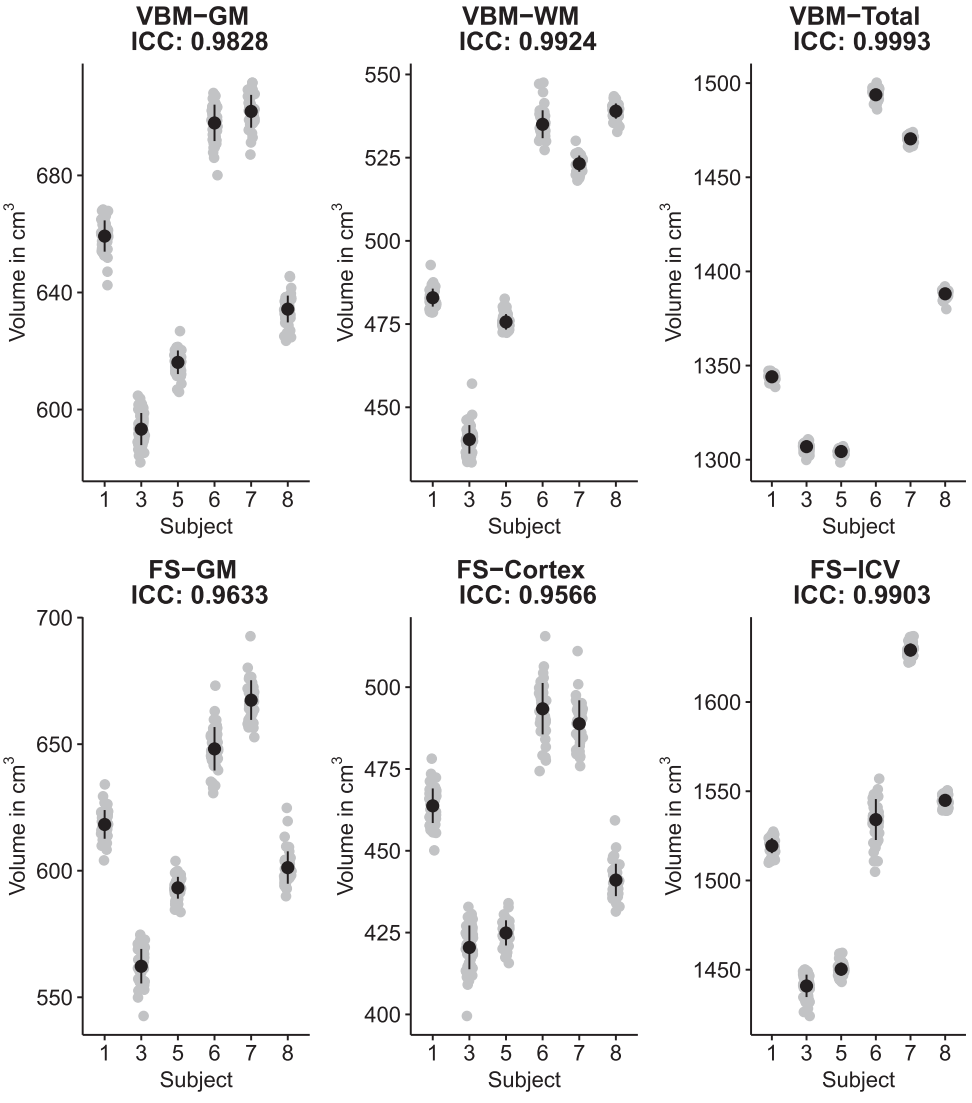

Fig. 1. Visualization of the between- and within-person variance. Each grey dot represents the brain measure for a given person at a given time point. The black dots depict the mean brain measure for each person. The black bars represent the intra-individual standard deviation. FS-GM: FreeSurfer grey matter volume, FS-Cortex: FreeSurfer cortex volume (FS-Cortex), FS-ICV: FreeSurfer intracranial volume, VBM-GM: VBM grey matter volume as assessed, VBM-WM: VBM white matter volume, VBM-Total: grey matter + white matter + cerebrospinal fluid volume.

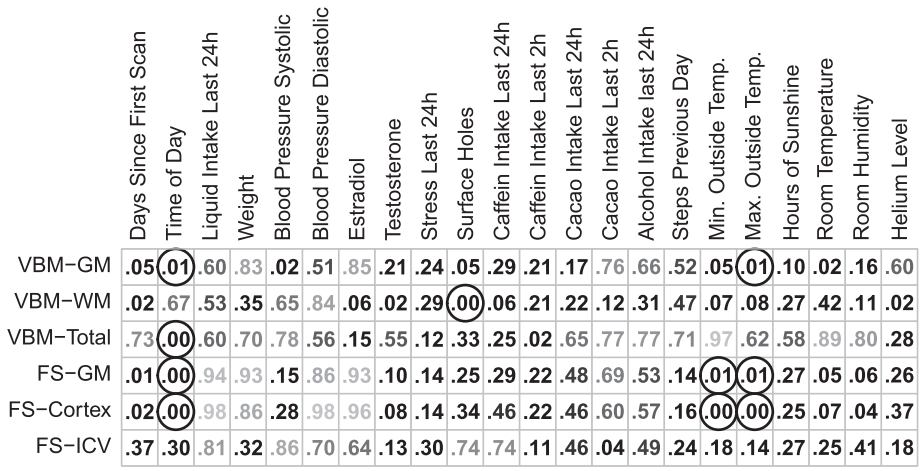

Fig. 2. *p*-values corresponding to the hypothesis test that variable x (column) is a predictor of brain measure y (row). The color of the number (black to grey) corresponds to the magnitude of the *p*-value (low to high). The *p*-values are not corrected for multiple comparisons. The circles denote variable brain measure combinations for which the corresponding *p*-value is smaller than 0.01. FS-GM: FreeSurfer grey matter volume, FS-Cortex: FreeSurfer cortex volume (FS-Cortex), FS-ICV: FreeSurfer intracranial volume, VBM-GM: VBM grey matter volume as assessed, VBM-WM: VBM white matter volume, VBM-Total: grey matter + white matter + cerebrospinal fluid volume.

the two temperature variables were highly correlated (see Fig. 3). We address the issue of collinearity of these variables in more detail in the following.

4.2.2. *Stepwise regression*

As we explained in Section 3.2.1, the best predictors are added step by step in stepwise regression. We summarize the results of this analysis in Fig. 4.

In the first step, Days Since First Scan was among the top predictors for both brain measures, even though it was not the variable with the highest adjusted R^2 improvement. Among the other top three variables, two are highly collinear with Days Since First Scan, namely Maximum Outside Temperature and Minimum Outside Temperature, which is a result of the seasonal change that occurred during the study. Because Days Since First Scan is essentially a marker of a participant's progressing age, and given the vast literature documenting a reduction in cortical thickness and brain volume with healthy aging (see Lindenberger, 2014, for an overview), we chose this variable rather than the variable with the highest adjusted R^2 improvement, as is traditionally done. This allowed us to search for predictors of fluctuations after controlling for the putative effects of healthy aging (note that the directions of the observed effects and their size are addressed below). The adjusted R^2 improvement by including Days Since First Scan in the model was around 0.80%.

After controlling for Days Since First Scan, the effect of Temperature

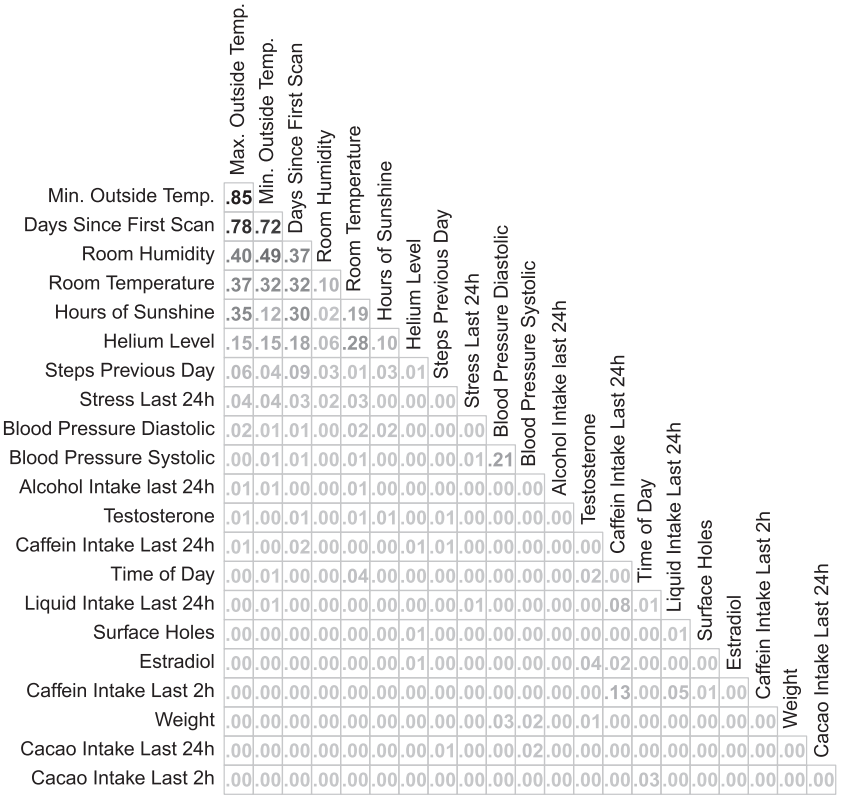

Fig. 3. Strength of the linear relationships between the variables. The numbers express the amount of within-person variance of variable y that is explained by variable x. This relationship is symmetric. The greyscale value of the number corresponds to the amount of variance explained (low to high).

FS−GM

	Time of Day	Min. Outside Temp.	Max. Outside Temp.	Days Since First Scan	Testosterone	Room Temperature	Room Humidity	Steps Previous Day	Blood Pressure Systolic	Stress Last 24h	Caffein Intake Last 2h	Surface Holes	Helium Level	Hours of Sunshine	Caffein Intake Last 24h	Cacao Intake Last 24h	Alcohol Intake last 24h	Cacao Intake Last 2h	Liquid Intake Last 24h	Blood Pressure Diastolic	Weight	Estradiol
Baseline	(12)	(9)	(9)	(7)	4	4	4	2	2	2	1	0	0	0	0	−1	−1	−1	−1	−1	−1	−2
Days	(11)	1	1	0	1	−1	−1	(5)	0	0	1	1	−1	−1	1	−1	−1	−1	−1	−1	−1	−1
Time of Scan	0	0	0	0	3	2	−1	(6)	0	0	1	1	−1	−1	1	−1	−1	0	−1	−1	−1	−1
Steps	0	0	0	0	4	2	−1	0	0	−1	1	1	−1	−1	0	0	−1	0	−1	−1	−1	−1

FS−Cortex

	Time of Day	Max. Outside Temp.	Min. Outside Temp.	Days Since First Scan	Room Humidity	Testosterone	Room Temperature	Stress Last 24h	Steps Previous Day	Caffein Intake Last 2h	Hours of Sunshine	Blood Pressure Systolic	Surface Holes	Helium Level	Caffein Intake Last 24h	Cacao Intake Last 24h	Alcohol Intake last 24h	Cacao Intake Last 2h	Weight	Liquid Intake Last 24h	Blood Pressure Diastolic	Estradiol
Baseline	(13)	(11)	(11)	(8)	(5)	5	4	2	2	1	0	0	0	0	−1	−1	−1	−1	−2	−2	−2	−2
Days	(12)	2	2	0	0	2	−1	0	(5)	1	−2	−1	0	−2	0	−1	−1	−1	−1	−1	−1	−1
Time of Scan	0	2	1	0	−1	4	2	0	(6)	1	−2	−1	0	−2	0	−1	−1	0	−2	−1	−1	−1
Steps	0	2	0	0	−1	5	2	−1	0	1	−2	−1	0	−2	−1	0	−1	0	−1	−1	−1	−1

Fig. 4. Results of the stepwise regression procedure, (a) for FreeSurfer grey matter volume (FS-GM) and (b) for FreeSurfer cortex volume (FS-Cortex). In each cell, the improvement in adjusted R^2 by adding this variable is represented in basis points (one-hundredth percent). In each row, the improvement is measured in comparison to a different base model. In the respective first row, the comparison is against the "baseline" model including only person as a predictor. In the following rows, the base model is extended by the variable labeling the corresponding row. Circles indicate significant adjusted R^2 improvements.

vanished, as expected because of the high collinearity between Days Since First Scan and Temperature. For both brain measures, the strongest additional predictor was Time of Day. Pending upon the direction of the effect, this result might be in line with a recent paper showing that brain volume decreases across the course of the day (Nakamura et al., 2015). The adjusted R^2 improvement was around 0.12%.

After controlling for Days Since First Scan and Time of Day, the strongest predictor was Steps Previous Day. The adjusted R^2 improvement was around 0.06%. While we are not aware of any previous studies linking steps taken on the day before scanning and brain volume, there is a wealth of literature documenting effects of physical exercise on brain volume (for reviews see, Hillman et al., 2008; Voss et al., 2013).

After controlling for Days Since First Scan, Time of Day, and Steps Previous Day, the strongest predictor was Testosterone. The adjusted improvement was around 0.02% and not significant. We, therefore, stopped the stepwise inclusion of predictors at this point.

4.2.3. Omnibus test

The omnibus test of a null hypothesis stating no predictive effect of any of the variables was statistically significant for both brain measures ($p = 0.0018$ for FS-GM and $p = 0.0020$ for FS-Cortex). This shows that one or more variables could significantly explain some amount of variability, although this test does not reveal the identities of these variables.

Fig. 5 shows the p-values for the individual hypothesis tests that a particular variable is a predictor of a brain measure. The p-values are calculated by comparing the full model to the full model without the respective variable. To avoid biases due to collinearities in the variables, we did not include the variables that correlated with Days Since First Scan in the full model (i.e., Min. Outside Temperature, Max. Outside Temperature, Hours of Sunshine, Room Temperature, Room Humidity, and Helium Level).

For both brain measures, Days since First Scan, Time of Day, Testosterone, and Steps Previous Day were deemed predictors of within-person brain fluctuations by the post hoc strategy.

4.3. Statistical learning methods

For both brain measures, LASSO slightly outperformed the random forest method. In Fig. 6, we display the out-of-sample R^2 of the within-person fluctuations for both statistical learning methods. Both methods achieved an R^2 of at least 2%. As a consequence, we investigated the variable importance values of both approaches.

4.3.1. LASSO

In Fig. 7 , we display the standardized coefficients[2] from the LASSO regression. We explain their interpretation using an example. The

[2] We removed the variables highly correlated with Days Since First Scan to ease interpretation. Before we checked that removing them did not influence the predictive accuracy substantially.

	Days Since First Scan	Time of Day	Liquid Intake Last 24h	Weight	Blood Pressure Systolic	Blood Pressure Diastolic	Estradiol	Testosterone	Stress Last 24h	Surface Holes	Caffein Intake Last 24h	Caffein Intake Last 2h	Cacao Intake Last 24h	Cacao Intake Last 2h	Alcohol Intake last 24h	Steps Previous Day
FS−GM	.01	.00	.99	.83	.36	.46	.78	.03	.75	.23	.59	.42	.44	.18	.59	.02
FS−Cortex	.01	.00	.87	.72	.53	.45	.62	.02	.70	.31	.84	.38	.38	.15	.65	.02

Fig. 5. Visualization of the *post hoc* tests, showing the uncorrected *p*-values for the hypothesis that variable x (column) is a predictor of brain measure y (row).

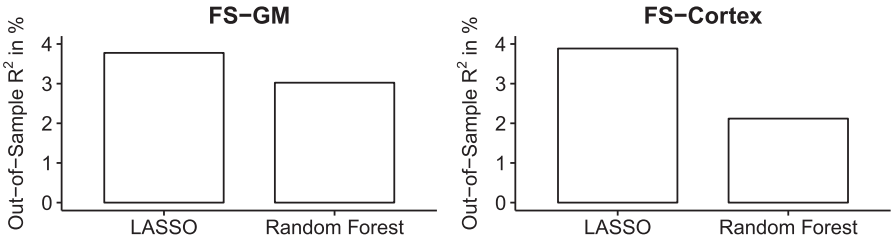

Fig. 6. Out-of-sample R^2 of the within-person fluctuations for FreeSurfer grey matter volume (FS-GM) and FreeSurfer cortex volume (FS-Cortex) as determined by the LASSO and the random forest method.

coefficient of −.07 for Time of Day should be interpreted as the prediction of FS-GM being decreased by .07 standard deviations (of FS-GM) if Time of Day increases by one standard deviation (of Time of Day). The absolute values of the standardized coefficients of all predictor variables were relatively low. The order of importance was the same for both structural measures. The coefficients were even equal up to a precision of two digits. Time of Day was the most important variable and Days Since First scan the fourth most important. Interestingly, the coefficient of Steps Previous Day is exactly 0. Thus, it was not selected as a predictor by LASSO. While the coefficient for Testosterone was also low, it was nonzero (−0.0027 for FS-GM, and −0.0005 for FS-Cortex) and thus selected as a predictor by LASSO. LASSO also deemed some predictors important that were not identified by any of the previous methods. Most notably, the number of surface holes and systolic blood pressure.

4.3.2. *Random forests*

In Fig. 8 , we display the variable importance values[3] derived from the random forest. With an increase of at most 0.57%, the importance of every single variable was relatively low. The order of importance values was identical across FS-GM and FS-Cortex. In terms of relative importance, the top five most important variables were all correlated with Days Since First Scan. Time of Day was only deemed the seventh most important value, and Testosterone follows as the eighth.

4.4. *Summary*

As expected, each analysis strategy led to slightly different conclusions: This is due to the different assumptions made about the data generating processes by each of the approaches. As is often the case when multiple alternatives are possible, no single one of them can be said to be optimal. Instead, each alternative has strengths and weaknesses that should be considered in relation to the desired analysis. Having performed multiple different analyses, however, allows us to triangulate which finding is robust with regard to the chosen analysis strategy. Specifically, we wanted to demonstrate possible approaches to tackle the problem at hand either from a classical statistical inference or a statistical learning framework. We summarize the results of the different analysis strategies in Table 2.

The finding that Time of Day is predictive of within-person variability in structural estimates is the most robust, as all feature selection strategies yielded this as a significant factor. Also, it proved to have the fourth highest random forest variable importance value.

The finding that Days Since First Scan is predictive of within-person variability in structural estimates is also relatively robust. All analysis strategies except the prediction matrix concluded that it is a true predictor. For the prediction matrix, we could not reject the hypothesis that it was an unimportant predictor, but we found a statistical trend. Also, Days Since First Scan reached the highest random forest variable importance value.

While our results also suggest that the two environmental temperature variables (Min. Outside Temperature and Max. Outside Temperature) are useful predictors, it needs to be considered that they correlate highly with Days Since First Scan, as a result of the seasonal change during the study. Therefore, the temperature variables essentially represent a recoding of the latter. While some analysis strategies allowed us to control for this adequately, it was not possible in all of them. Importantly, the stepwise regression analysis, in which we adequately controlled for this, suggests that the temperature variables do not possess any predictive power beyond Days Since First Scan. For these reasons, we excluded the temperature variables in Table 2.

Steps Previous Day was selected by two feature selection strategies and was the eighth most important variable in the random forest analysis. All analysis strategies estimated its effect to be weaker than the effects of Days Since First Scan and Time of Day.

Testosterone was only selected by the post hoc test after the omnibus test. It was deemed the seventh most important variable by the random forest analysis.

4.5. *Size and direction of robust effects*

In a final analysis, we estimated the strengths and directions of the

[3] The number of surface holes was removed from the predictive model because including it severely reduced the out-of-bag accuracy. Consequently, its variable importance was negative.

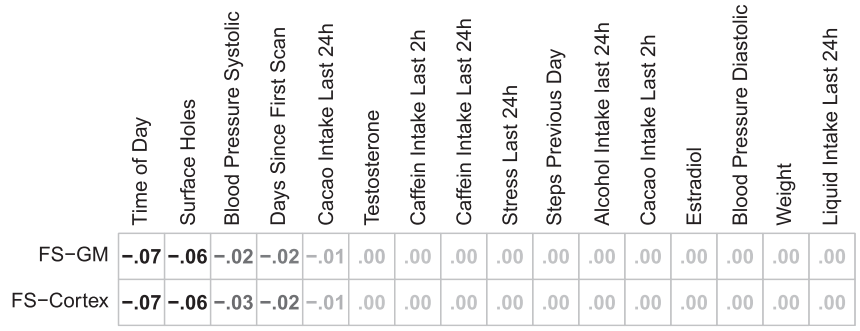

Fig. 7. Visualization of LASSO coefficients. The numbers represent the standardized coefficient values.

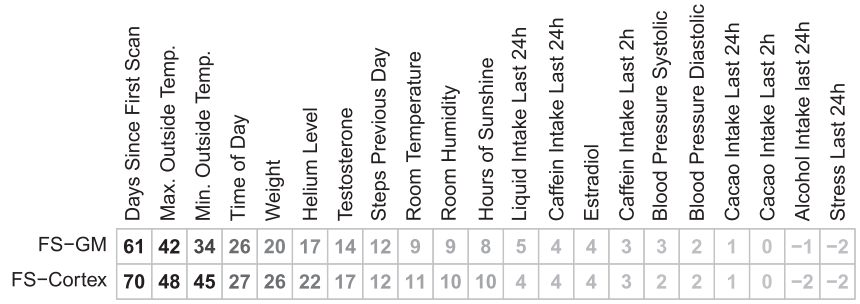

Fig. 8. Visualization of the random forest variable importance values. The numbers represent the increase of the mean squared error when randomly permuting the given variable in basis points. The variables are sorted by importance value.

Table 2
Summary of the analysis results. For simplicity, we report whether a variable is selected (or not) for all analysis strategies but random forests. Selection is denoted by a grid pattern and rejection by white. For random forests, we report the ranking of all variables that had previously been selected by the other strategies.

	Prediction Matrix	Stepwise	Post hoc Omnibus	Random Forest	LASSO
Time of Day				4	1
Days Since First Scan				1	4
Steps Previous Day				8	Not used
Testosterone				7	6

effect sizes of the associations that we deemed robust (Days Since First Scan and Time of Day).

ICC values were relatively high to begin with, so one may ask why searching for predictors of within-variance is worthwhile at all. First of all, it is important to note that ICC is linked to the precision of and statistical power to detect between-person differences in a cross-sectional analysis. It follows that researchers who have greater reliability for individual measures will also have a larger chance to detect correlations between them. Consequently, for cross-sectional correlation studies, the high ICCs we observed directly translate into a high statistical power to detect correlations. Decreasing within-variance is always beneficial as it improves the power and precision of point estimates, and hence also allows to maintain power and precision while affording cheaper designs (e.g., fewer people, fewer measurement occasions; see Brandmaier et al., 2015).

Further note that, in general, measures with larger ICC are not necessarily better measures. For example, in an experimental setting, in which we may be interested in group differences between conditions, the total variance $\sigma_T^2 = \sigma_b^2 + \sigma_e^2$ is directly related to the size of the standard errors and consequently the power to detect group differences. Therefore, larger individual differences (that lead to larger ICCs) usually dilute our measurements of mean group differences, and the power to detect a given experimental mean difference may actually be smaller in a population with a higher ICC (see Brandmaier et al., 2018a,b). We can conclude that in those settings, despite high ICC values, it may still be imperative to reduce within-person variance to achieve adequate levels of precision and statistical power.

Furthermore, to detect individual differences in within-person change in longitudinal settings, ICC is less informative as it does not relate to the magnitude of individual differences of true change (see Brandmaier et al., 2018a,b for an extension of this idea to the reliability of change). A measure that is highly reliable to detect individual differences at one point in time may be entirely unreliable in detecting differences in within-person change if these changes are relatively small. Brandmaier, von Oertzen et al. (2018) have argued that within-person variability in itself is a useful, unstandardized measure to convey precision of a

measurement instrument. Consequently, for detecting the within-person change in longitudinal settings, the observed high ICC values may again be misleading.

To summarize, ICC is a standardized measure of within-person consistency, and, as such, it is useful descriptive statistic. Even when ICC is high, it might still be useful to increase precision and statistical power further by identifying sources of within-person variance that then can be brought under experimental or statistical control.

In the stepwise regression analysis, the within-person variance could be decreased by 1.8% for both outcomes when including Days Since First Scan as a predictor to the model that only uses subject IDs as predictors. Similarly, when adding Time of Day to the model including Days Since first Scan the within-person variance could be further decreased by 3.0% for FS-GM and by 2.70% for FS-Cortex. Taken together these two variables decreased the within-person variance by 4.8% for FS-GM and 4.4% for FS-Cortex. These results are in line with the random forests and LASSO results, according to which up to around 4% of the residual variance can be explained using all predictors.

Within-subject predictors may not only increase power but also correct for bias. To estimate the biasing effect that the within-subject predictors might have, we estimated the size of the regression coefficients using the linear model including person-specific intercepts and the two robust variables. This also enables us to compare our results to previous research (Nakamura et al., 2015; Raz et al., 2005), which quantified the effect size of within-subject predictors using regression coefficients. The confidence intervals for the coefficients for the two predictors are shown in Table 3.

For both predictors the effect size estimate is negative, that is, brain volume declines over a day as well as over a year. Compared to the overall size of areas (FS-GM: roughly 550–700 cm^3, FS-Cortex roughly 400–500 cm^3) the estimated decrease is relatively mild. Interestingly, the estimated decline over a year was approximately the same order of magnitude as the decline over a day. Also, the effects were very similar across the two chosen FreeSurfer brain measures.

In Fig. 9, we visualize the regression coefficients using partial residual plots that show that the implied relationships are rather uniform across the sample and that there are no serious violations of the model assumptions.

5. Discussion

We aimed at finding predictors of within-person variance in structural MRI measures. We selected a set of global MRI brain measures and based our analysis on a unique longitudinal MRI data set with roughly 50 observations per person. As analysis strategy, we chose to report the results of a variety of different statistical approaches to triangulate the problem of finding the best predictors of within-person variation in sMRI. This can also be regarded as a sensitivity analysis and as a demonstration of the approaches that may be useful to explore potential predictors of within-person variance, both from a more traditional statistical modeling perspective and a statistical learning perspective. Our analyses revealed two robust predictors of within-person variance in sMRI: Days Since First Scan and Time of Day.

5.1. Effect of Days Since First Scan

Days Since First Scan can be regarded as a marker of each participant's age-related changes since the inception of the study and is as such a very likely candidate for a predictor rather capturing true change than measurement error (unless measurement error increased over the total time of the study). As such, and because there is widespread agreement in the literature that the brain shrinks during the process of aging (Lindenberger, 2014), it is in principle not surprising that this was a robust predictor. However, most studies reporting structural changes with aging have considered changes over years (Bäckman et al., 2006; Fjell et al., 2009a,b; Fjell et al., 2009a,b; Persson et al., 2016; Raz and Kennedy, 2009; Raz and Rodrigue, 2006), whereas our analyses are rather in line with previously observed short-term changes in control groups of intervention studies (Lövdén et al., 2012). Similar to these studies our analyses were sensitive enough to detect changes in the range of a few months, presumably thanks to the large number of scans of every single individual. This finding highlights the qualitative advantage of collecting a large number of data points for individuals, as was the case for the Day2day study.

Concerning the size of the effect, Raz et al. (2005) studied a sample of participants ranging from 20 to 77 years at two time points approximately five years apart. In this sample, longitudinal annual percentage change varied from −0.1% to −0.9% across brain regions. In our analysis, we modeled the data such that rate of change is modeled as fixed across all participants. In Fig. 10, we translate the exact confidence intervals displayed in Table 3 into approximate percent change values.

With an approximate range of −2.05% to −0.17% change in a year, the estimated size of the effect is in line with previous findings, especially considering the relatively high uncertainty in our estimates. The findings reported by Raz et al. are based on a sample with a much broader adult age range than the one included in Day2day. However, the statistical analyses of Raz et al. and the inspection of individual longitudinal trajectories (e.g., see Figs. 6 and 7 of that paper) both indicate that volume shrinkage is not restricted to late adulthood. Clearly, the proposition that longitudinal volume shrinkage is detectable in early adulthood warrants further investigation, as it might be relevant for both clinical and basic research questions. We also note that in intervention studies, the effects of aging are usually addressed by including an age-matched control group (e.g., Kühn et al., 2014; Wenger et al., 2017). We believe that our results highlight the need to do this.

Table 3
95% confidence intervals for the effect sizes (in cm^3) of the two robust predictors, namely Days Since First Scan and Time of Day.

	FS-GM	FS-Cortex
In a Year	[− 10.28, − 0.89]	[− 9.29, − 0.75]
From 8am to 8pm	[− 8.96, − 1.91]	[− 7.88, − 1.48]

Note. FS-GM: FreeSurfer grey matter volume, FS-Cortex: FreeSurfer cortex volume.

5.2. Effect of time of day

Nakamura et al. (2015) also found that the Time of Day is a predictor of within-person variance in sMRI. In line with our analysis, they observed a decline in brain volume over the day. In their analysis, the size of the effect meant a decline of −0.221% to 0.090% per 12 h. Unfortunately, it is unclear how accurate these estimates are as the authors did not report the uncertainty in these estimates. Approximately translating our results to percent change yields similar rates of decline for both brain measures, FS-GM and FS-Cortex, with a 95% confidence interval of roughly −1.72% to −0.3% change per day).

The underlying causal mechanisms of the effect of Time of Day on brain volume are not yet known. Nakamura et al. (2015) offer some speculative explanations. First, they suggest that this effect might be due to fluid redistribution during the day that is counteracted by long supine periods during the night, thus returning brain volume to normal the next morning. A second alternative is that volume changes reflect hydration status, in turn, caused by diuretic factors. Finally, Nakamura et al. (2015) proposed that factors external to participants, such as heating up of the MR scanner coil due to use throughout the day, could have led to apparent volume changes.

Beyond global structural parameters, effects of Time of Day have also been reported at the functional level. Anderson et al. (2014) measured the BOLD signal during a 1-back task. They found that older adults tested at the peak time of alertness had better performance in the behavioral task than in the afternoon, and that brain activity during the morning (but not afternoon) was comparable to that of younger adults. This

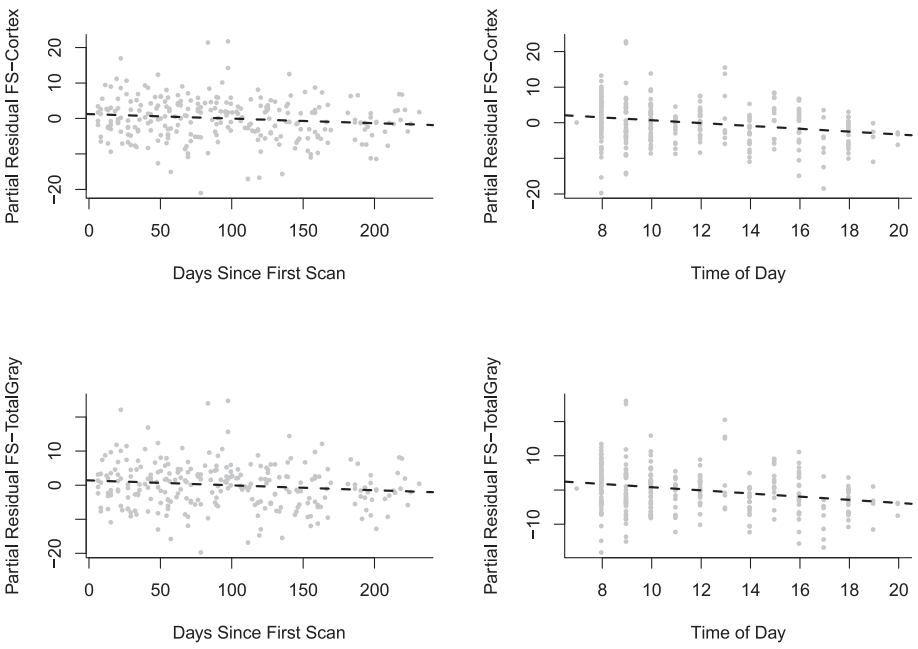

Fig. 9. Visualization of the effects using partial residual plots. Each dot represents a partial residual. The dashed line represents the best regression line between the corresponding variable and the partial residuals. The slope of this line is equivalent to the corresponding regression coefficient. For more details on partial regression plots, see Larsen & McCleary (1972).

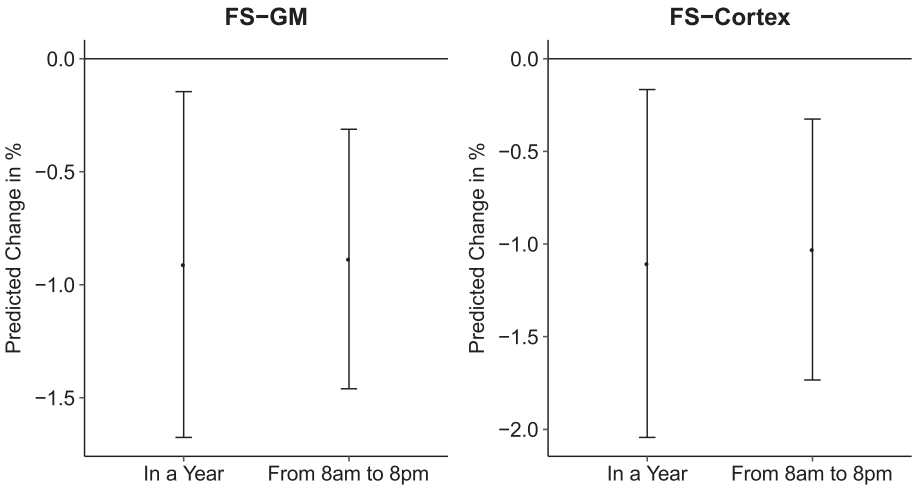

Fig. 10. Visualization of the estimated effect of the two robust predictors, namely Days Since First Scan and Time of Day. The dots denote the maximum likelihood estimate, and the error bars indicate the 95% confidence intervals. Here, they are displayed in approximate percent change. FS-GM: FreeSurfer grey matter volume, FS-Cortex: FreeSurfer cortex volume.

finding complements our results by suggesting that the relevant parameter is not only the time of day but perhaps also its interaction with peak alertness time.

While the Day2day data set was not explicitly designed to address the questions posed here and thus cannot offer a clear answer, we can nevertheless contribute to the refinement of these speculations. We found no effect of liquid intake during the day, suggesting that hydration status may not be a plausible explanation of the effect. Additionally, because Nakamura and colleagues found the diurnal variation in populations of older adults, some of them with multiple sclerosis, mild cognitive impairment, or Alzheimer's disease, they were not able to exclude the possibility that a daily regime of medications (perhaps with diuretic effects) could have affected brain volume. Our results, taken from healthy young adults, show that this cannot be the only explanation, as the Day2day study participants took no medication during the time of scanning. However, more work is needed to understand the causal mechanisms.

When considering whether to counteract this effect when performing longitudinal studies, power and bias have to be considered. In terms of power, our results suggest that controlling for Time of Day decreases the within-variance by roughly 3%, which translates into a roughly 1.75% smaller standard error. To get an intuition for the practical relevance of this decrease, the same decrease in standard error can be achieved by increasing the sample size by 1.75%.

The size of the bias depends on how unequal the to-be-compared groups are in terms of their scanning time. Our results suggest that if one group is always measured around 8 a.m. and another one around 8 p.m., then one can expect to find roughly a 1% difference in average brain volume between these groups even if there are no meaningful differences between the groups. Given that volume increases typically

found in interventional longitudinal studies, for example, as a response to learning juggling, are roughly of the same magnitude (Zatorre et al., 2012), any such systematic group differences in scan time must be avoided.

Possible strategies for controlling for the confounding effect, include restricting or randomizing the time of scanning across persons, and appropriately controlling for it in the statistical analysis (see also, Nakamura et al., 2015). Randomizing or restricting the time of scanning has the advantage that it is a viable solution even if no model for the effect of Time of Day on the measurements is available. However, they complicate data collection as they pose additional organizational constraint for longitudinal studies. The choice between randomization and restricting should take the goal of the study into account. For longitudinal studies, fixing seems most appropriate as it does not only control for potential bias but also maximizes the within-person correlation and thus the power to detect within-person changes. In contrast, when statistically controlling for the effect, data collection can be performed without timing constraints and reliability can be increased (although only mildly). However, it is important to note that a correct model of the effects of Time of Day on the measurement is required.

5.3. Weak evidence for steps taken on previous day and testosterone

We also found weak evidence that steps taken on the day before scanning and testosterone were predictors of within-person variance in sMRI. These results are not as robust as the effects related to time since first scan, and to time of day, and there also is less prior evidence on the effect of testosterone and physical exercise within the last 24 h on sMRI measurements. However, there is a vast literature documenting the long-term effects of physical exercise on brain volume (for reviews see, Hillman et al., 2008; Voss et al., 2013). Some brain structures appear to be more susceptible to change than others. In particular, relationships between hippocampal volume and physical exercise have been extensively reported, particularly in older adults (Chaddock et al., 2010; Duzel et al., 2016; Erickson et al., 2011; Kleemeyer et al., 2016; Maass et al., 2015).

Most previous studies have focused on the long-term effects of performing various elaborate physical exercise programs (Lövdén et al., 2013). In contrast, we found weak evidence for an effect of a short-term variation in everyday movement, namely the number of steps taken on the day before the scan. This effect needs to be replicated with a larger sample in a more targeted study. Given the presence of diurnal fluctuations (see above), we contend that small variations in everyday physical activity may be associated with variations in brain volume, for reasons that need to be identified in subsequent work including animal models.

Sex hormones have also been shown to affect the adult human brain (Chaddock et al., 2010; Duzel et al., 2016; Erickson et al., 2011; Kleemeyer et al., 2016; Maass et al., 2015). Even small and short-term fluctuations of hormones, for example, during the menstrual cycle, are associated with structural brain changes (Comasco and Sundström-Poromaa, 2015; Lisofsky et al., 2015a,b; Peper, van den Heuvel, Mandl, Hulshoff Pol and van Honk, 2011; Toffoletto et al., 2014). Adult testosterone levels also fluctuate, showing a diurnal and seasonal cycle, and are influenced by exercise, for instance (Dabbs, 1990; Zitzmann and Nieschlag, 2001). While changes in functional brain measures have been observed following exogenous testosterone administration (e.g., Bos et al., 2013), these natural short-term fluctuations have not been studied systematically in relation to human brain structure. Animal studies, however, have shown that testosterone induces microstructural changes in grey and white matter, and, for example, influences the survival of new hippocampal neurons in adult male rats (e.g., Garcia-Segura and Melcangi, 2006; Spritzer and Galea, 2007; Sumner and Fink, 1998). The present findings highlight the need to further study the effects of natural hormonal variation on the adult human brain.

5.4. Other variables

Concerning all other variables, our analysis suggests that none of them may be noteworthy predictors of the within-person variance of brain volume as measured by sMRI. While LASSO selected the number of surface holes and the systolic blood pressure as predictors, this result is unique to LASSO and not confirmed by any of the other methods. For the remaining variables, no method suggested that it might be a noteworthy predictor. However, these other variables might have effects that are too small to detect given our data set. Nevertheless, based on these findings, we see no necessity to either experimentally or statistically control for any of the remaining variables included in this analysis, at least within the range of variability investigated in the present study.

5.5. Relationship to previous studies

That brain function and structure may vary spontaneously and over days has been recognized only recently. In particular, the “myConnectome” project (Poldrack et al., 2015) spearheaded the approach of collecting a dense sample of neural, physiological, and psychological data from a single individual over more than a year. The first analyses of this dataset focused on functional and structural connectivity during resting state (Laumann et al., 2015; Poldrack et al., 2015) and revealed, for example, effects of caffeine intake on functional connectivity networks. Here, we did not find enough evidence to support an effect of caffeine on whole-brain structural parameters. We speculate that caffeine might have more “fine-grained” effects, perhaps also with faster dynamics than those that we measure at a global brain structural level.

5.6. Conclusions

The present study yields one clear result: Based on our estimates, average day-to-day fluctuations in Freesurfer estimates of grey matter and overall cortical volume are not reliably smaller than the average negative linear change within one year, and both are reliably different from zero. This important result is in full agreement with early pleas of lifespan psychologists to distinguish between short-term variability and long-term change (Laumann et al., 2015; Poldrack et al., 2015). Since then, behavioral researchers have introduced research design to capture both short-term variability and long-term change within the same study (Hofer and Sliwinski, 2001), and to examine how they are related (e.g., Lövdén et al., 2007). The present results underscore the need to introduce similar considerations and designs when studying changes in brain structure across the lifespan. At the same time, it needs to be kept in mind that our results were obtained in a very small sample of healthy young adults. It is possible that the signal of annual percent change increases relative to the amount of diurnal fluctuation with advancing adult age. Clearly, the generalizability of the present findings needs to be assessed in future studies.

Nevertheless, we dare recommending, based on the present findings, that researchers interested in longitudinal change experimentally control for time of day when planning to carry out a study investigating long-term changes in brain volume, be it in the context of an intervention study or in the context of a longitudinal panel study. In particular, we recommend that researchers interested in assessing long-term longitudinal change make sure that (a) a given participant is measured at precisely the same time of day at each measurement occasion, thereby minimizing the influence of within-person diurnal variability relative to the influence of long-term change; (b) all participants are measured at about the same time of day to minimize the contribution of between-person differences in time of day to individual differences in estimates of brain structure.

The effects of physical exercise on the day before scanning and

testosterone levels are inconclusive yet and require further investigation. For all other assessed variables, our analyses yielded no evidence that they are predictive of within-person variance and therefore do not justify a recommendation that they should be controlled for in future studies. Future research with larger data sets to corroborate and extend our conclusions is warranted.

Appendix A. Supplementary material

The data and code used for this paper can be found at https://doi.org/10.24433/CO.3688518.v1.

References

Anderson, J.A.E., Campbell, K.L., Amer, T., Grady, C.L., Hasher, L., 2014. Timing is everything: age differences in the cognitive control network are modulated by time of day. Psychol. Aging 29 (3), 648–657. https://doi.org/10.1037/a0037243.

Ashburner, J., 2012. SPM: a history. Neuroimage 62 (2), 791–800. https://doi.org/10.1016/j.neuroimage.2011.10.025.

Ashburner, J., Friston, K.J., 2000. Voxel-based morphometry: the methods. Neuroimage 11 (6), 805–821. https://doi.org/10.1006/nimg.2000.0582.

Bäckman, L., Nyberg, L., Lindenberger, U., Li, S.-C., Farde, L., 2006. The correlative triad among aging, dopamine, and cognition: current status and future prospects. Neurosci. Biobehav. Rev. 30 (6), 791–807. https://doi.org/10.1016/j.neubiorev.2006.06.005.

Benasisch, A.A., Urs, R. (Eds.), 2018. Emergent Brain Dynamics - Prebirth to Adolescence, vol. 25. MIT Press, Cambridge, MA.

Bland, J.M., Altman, D.G., 1995. Statistics notes: calculating correlation coefficients with repeated observations: Part 1—correlation within subjects. Br. Med. J. 310 (6977), 446. https://doi.org/10.1136/bmj.310.6977.446.

Bos, P.A., van Honk, J., Ramsey, N.F., Stein, D.J., Hermans, E.J., 2013. Testosterone administration in women increases amygdala responses to fearful and happy faces. Psychoneuroendocrinology 38 (6), 808–817. https://doi.org/10.1016/j.psyneuen.2012.09.005.

Brandmaier, A.M., von Oertzen, T., Ghisletta, P., Lindenberger, U., Hertzog, C., 2018a. Precision, reliability, and effect size of slope variance in latent growth curve models: implications for statistical power analysis. Front. Psychol. 9, 294. https://doi.org/10.3389/fpsyg.2018.00294.

Brandmaier, A.M., von Oertzen, T., Ghisletta, P., Hertzog, C., Lindenberger, U., 2015. LIFESPAN: A tool for the computer-aided design of longitudinal studies. Front. Psychol. 6, 272.

Brandmaier, A.M., Wenger, E., Bodammer, N.C., Kühn, S., Raz, N., Lindenberger, U., 2018b. Assessing reliability in neuroimaging research through intra-class effect decomposition (ICED). eLife 7, e35718. https://doi.org/10.7554/eLife.35718.

Breiman, L., 2001. Random forests. Mach. Learn. 45 (1), 5–32. https://doi.org/10.1023/A:1010933404324.

Buuren, S., Groothuis-Oudshoorn, K., 2011. mice: multivariate imputation by chained equations in R. J. Stat. Softw. 45 (3). https://doi.org/10.18637/jss.v045.i03.

Caceres, A., Hall, D.L., Zelaya, F.O., Williams, S.C.R., Mehta, M.A., 2009. Measuring fMRI reliability with the intra-class correlation coefficient. Neuroimage 45 (3), 758–768. https://doi.org/10.1016/j.neuroimage.2008.12.035.

Carp, J., 2012. On the plurality of (methodological) worlds: estimating the analytic flexibility of FMRI experiments. Front. Neurosci. 6, 149. https://doi.org/10.3389/fnins.2012.00149.

Chaddock, L., Erickson, K.I., Prakash, R.S., VanPatter, M., Voss, M.W., Pontifex, M.B., et al., 2010. Basal ganglia volume is associated with aerobic fitness in preadolescent children. Dev. Neurosci. 32 (3), 249–256. https://doi.org/10.1159/000316648.

Comasco, E., Sundström-Poromaa, I., 2015. Neuroimaging the menstrual cycle and premenstrual dysphoric disorder. Curr. Psychiatr. Rep. 17 (10), 77. https://doi.org/10.1007/s11920-015-0619-4.

Dabbs, J.M., 1990. Salivary testosterone measurements: reliability across hours, days, and weeks. Physiol. Behav. 48 (1), 83–86. https://doi.org/10.1016/0031-9384(90)90265-6.

Duzel, E., van Praag, H., Sendtner, M., 2016. Can physical exercise in old age improve memory and hippocampal function? Brain 139 (Pt 3), 662–673. https://doi.org/10.1093/brain/awv407.

Erickson, K.I., Voss, M.W., Prakash, R.S., Basak, C., Szabo, A., Chaddock, L., et al., 2011. Exercise training increases size of hippocampus and improves memory. Proc. Natl. Acad. Sci. U.S.A. 108 (7), 3017–3022. https://doi.org/10.1073/pnas.1015950108.

Filevich, E., Lisofsky, N., Becker, M., Butler, O., Lochstet, M., Martensson, J., et al., 2017. Day2day: investigating daily variability of magnetic resonance imaging measures over half a year. BMC Neurosci. 18, 65. https://doi.org/10.1186/s12868-017-0383-y.

Fischl, B., 2012. FreeSurfer. Neuroimage 62 (2), 774–781. https://doi.org/10.1016/j.neuroimage.2012.01.021.

Fjell, A.M., Walhovd, K.B., Fennema-Notestine, C., McEvoy, L.K., Hagler, D.J., Holland, D., et al., 2009a. One year brain atrophy evident in healthy aging. J. Neurosci. 29 (48), 15223–15231. https://doi.org/10.1523/JNEUROSCI.3252-09.2009.

Fjell, A.M., Westlye, L.T., Amlien, I., Espeseth, T., Reinvang, I., Raz, N., et al., 2009b. High consistency of regional cortical thinning in aging across multiple samples. Cerebr. Cortex 19 (9), 2001–2012. https://doi.org/10.1093/cercor/bhn232.

Flom, P.L., Cassel, D.L., 2007. Stopping stepwise: why stepwise and similar selection methods are bad, and what you should use. In: Northeast SAS Users Group 2007 Proceedings. Baltimore, Maryland. Retrieved from. http://www.lexjansen.com/nesug/nesug07/sa/sa07.pdf.

Garcia-Segura, L.M., Melcangi, R.C., 2006. Steroids and glial cell function. Glia 54 (6), 485–498. https://doi.org/10.1002/glia.20404.

Hastie, T., Tibshirani, R., Friedman, J.H., 2009. The Elements of Statistical Learning, second ed., vol. 1. Springer, New York.

Hillman, C.H., Erickson, K.I., Kramer, A.F., 2008. Be smart, exercise your heart: exercise effects on brain and cognition. Nat. Rev. Neurosci. 9 (1), 58–65. https://doi.org/10.1038/nrn2298.

Hofer, S.M., Sliwinski, M.J., 2001. Understanding Ageing. An evaluation of research designs for assessing the interdependence of ageing-related changes. Gerontology 47 (6), 341–352. https://doi.org/10.1159/000052825.

Homburg, C., Dobratz, A., 1992. Covariance structure analysis via specification searches. Stat. Pap. 33 (1), 119. https://doi.org/10.1007/BF02925318.

Huberty, C.J., 1989. Problems with stepwise methods—better alternatives. Adv. Soc. Sci. Methodol. 1, 43–70.

Karch, J.D., Sander, M.C., von Oertzen, T., Brandmaier, A.M., Werkle-Bergner, M., 2015. Using within-subject pattern classification to understand lifespan age differences in oscillatory mechanisms of working memory selection and maintenance. Neuroimage 118, 538–552. https://doi.org/10.1016/j.neuroimage.2015.04.038.

Kleemeyer, M.M., Kühn, S., Prindle, J., Bodammer, N.C., Brechtel, L., Garthe, A., et al., 2016. Changes in fitness are associated with changes in hippocampal microstructure and hippocampal volume among older adults. Neuroimage 131, 155–161. https://doi.org/10.1016/j.neuroimage.2015.11.026.

Kühn, S., Gleich, T., Lorenz, R.C., Lindenberger, U., Gallinat, J., 2014. Playing Super Mario induces structural brain plasticity: gray matter changes resulting from training with a commercial video game. Mol. Psychiatry 19 (2), 265–271. https://doi.org/10.1038/mp.2013.120.

Kuncheva, L.I., 2004. Combining Pattern Classifiers: Methods and Algorithms, first ed. Wiley-Interscience, Hoboken, NJ.

Larsen, W.A., McCleary, S.J., 1972. The use of partial residual plots in regression analysis. Technometrics 14 (3), 781–790. https://doi.org/10.1080/00401706.1972.10488966.

Laumann, T.O., Gordon, E.M., Adeyemo, B., Snyder, A.Z., Joo, S.J., Chen, M.-Y., et al., 2015. Functional system and areal organization of a highly sampled individual human brain. Neuron 87 (3), 657–670. https://doi.org/10.1016/j.neuron.2015.06.037.

Lindenberger, U., 2014. Human cognitive aging: *Corriger la fortune?* Science 346 (6209), 572–578. https://doi.org/10.1126/science.1254403.

Lindenberger, U., Li, S.-C., Bäckman, L., 2006. Delineating brain–behavior mappings across the lifespan: substantive and methodological advances in developmental neuroscience. Neurosci. Biobehav. Rev. 30 (6), 713–717. https://doi.org/10.1016/j.neubiorev.2006.06.006.

Lisofsky, N., Lindenberger, U., Kühn, S., 2015a. Amygdala/hippocampal activation during the menstrual cycle: evidence for lateralization of effects across different tasks. Neuropsychologia 67, 55–62. https://doi.org/10.1016/j.neuropsychologia.2014.12.005.

Lisofsky, N., Mårtensson, J., Eckert, A., Lindenberger, U., Gallinat, J., Kühn, S., 2015b. Hippocampal volume and functional connectivity changes during the female menstrual cycle. Neuroimage 118, 154–162. https://doi.org/10.1016/j.neuroimage.2015.06.012.

Lövdén, M., Li, S.-C., Shing, Y.L., Lindenberger, U., 2007. Within-person trial-to-trial variability precedes and predicts cognitive decline in old and very old age: longitudinal data from the Berlin Aging Study. Neuropsychologia 45 (12), 2827–2838. https://doi.org/10.1016/j.neuropsychologia.2007.05.005.

Lövdén, M., Schaefer, S., Noack, H., Bodammer, N.C., Kühn, S., Heinze, H.-J., et al., 2012. Spatial navigation training protects the hippocampus against age-related changes during early and late adulthood. Neurobiol. Aging 33 (3), 620.e9-620.e22. https://doi.org/10.1016/j.neurobiolaging.2011.02.013.

Lövdén, M., Wenger, E., Mårtensson, J., Lindenberger, U., Bäckman, L., 2013. Structural brain plasticity in adult learning and development. Neurosci. Biobehav. Rev. 37 (9), 2296–2310. https://doi.org/10.1016/j.neubiorev.2013.02.014.

Maass, A., Düzel, S., Goerke, M., Becke, A., Sobieray, U., Neumann, K., et al., 2015. Vascular hippocampal plasticity after aerobic exercise in older adults. Mol. Psychiatry 20 (5), 585–593. https://doi.org/10.1038/mp.2014.114.

Morey, R.A., Selgrade, E.S., Wagner, H.R., Huettel, S.A., Wang, L., McCarthy, G., 2010. Scan-rescan reliability of subcortical brain volumes derived from automated segmentation. Hum. Brain Mapp. 31 (11), 1751–1762. https://doi.org/10.1002/hbm.20973.

Nakamura, K., Brown, R.A., Narayanan, S., Collins, D.L., Arnold, D.L., 2015. Diurnal fluctuations in brain volume: statistical analyses of MRI from large populations. Neuroimage 118, 126–132. https://doi.org/10.1016/j.neuroimage.2015.05.077.

Nesselroade, J.R., 1991. The warp and the woof of the developmental fabric. In: Visions of Aesthetics, the Environment & Development: the Legacy of Joachim F. Wohlwill. Lawrence Erlbaum Associates, Inc, Hillsdale, NJ, US, pp. 213–240.

Oshiro, T.M., Perez, P.S., Baranauskas, J.A., 2012. How many trees in a random forest?. In: Machine Learning and Data Mining in Pattern Recognition. Springer, Berlin, pp. 154–168. https://doi.org/10.1007/978-3-642-31537-4_13.

Peper, J.S., van den Heuvel, M.P., Mandl, R.C.W., Hulshoff Pol, H.E., van Honk, J., 2011. Sex steroids and connectivity in the human brain: a review of neuroimaging studies. Psychoneuroendocrinology 36 (8), 1101–1113. https://doi.org/10.1016/j.psyneuen.2011.05.004.

Persson, N., Ghisletta, P., Dahle, C.L., Bender, A.R., Yang, Y., Yuan, P., et al., 2016. Regional brain shrinkage and change in cognitive performance over two years: the

bidirectional influences of the brain and cognitive reserve factors. Neuroimage 126, 15–26. https://doi.org/10.1016/j.neuroimage.2015.11.028.
Poldrack, R.A., Laumann, T.O., Koyejo, O., Gregory, B., Hover, A., Chen, M.-Y., et al., 2015. Long-term neural and physiological phenotyping of a single human. Nat. Commun. 6, 8885. https://doi.org/10.1038/ncomms9885.
Raz, N., Kennedy, K.M., 2009. A systems approach to the aging brain: neuroanatomic changes, their Modifiers, and cognitive correlates. In: Jagust, W., D'Esposito, M. (Eds.), Imaging the Aging Brain. Oxford University Press, pp. 43–70. https://doi.org/10.1093/acprof:oso/9780195328875.003.0004.
Raz, N., Lindenberger, U., Rodrigue, K.M., Kennedy, K.M., Head, D., Williamson, A., et al., 2005. Regional brain changes in aging healthy adults: general trends, individual differences and modifiers. Cerebr. Cortex 15 (11), 1676–1689. https://doi.org/10.1093/cercor/bhi044.
Raz, N., Rodrigue, K.M., 2006. Differential aging of the brain: patterns, cognitive correlates and modifiers. Neurosci. Biobehav. Rev. 30 (6), 730–748. https://doi.org/10.1016/j.neubiorev.2006.07.001.
Robitzsch, A., Grund, S., Henke, T., 2018. M iceadds: Some Additional Multiple Imputation Functions, Especially for "Mice" (R package version 3.0-16.). Retrieved from. https://CRAN.R-project.org/package=miceadds.
Rosen, A.F.G., Roalf, D.R., Ruparel, K., Blake, J., Seelaus, K., Villa, L.P., et al., 2018. Quantitative assessment of structural image quality. Neuroimage 169, 407–418. https://doi.org/10.1016/j.neuroimage.2017.12.059.
Schrouff, J., Rosa, M.J., Rondina, J.M., Marquand, A.F., Chu, C., Ashburner, J., et al., 2013. PRoNTo: pattern recognition for neuroimaging toolbox. Neuroinformatics 11 (3), 319–337. https://doi.org/10.1007/s12021-013-9178-1.
Simmons, J.P., Nelson, L.D., Simonsohn, U., 2011. False-positive psychology: undisclosed flexibility in data collection and analysis allows presenting anything as significant. Psychol. Sci. 22 (11), 1359–1366. https://doi.org/10.1177/0956797611417632.
Spritzer, M.D., Galea, L.A.M., 2007. Testosterone and dihydrotestosterone, but not estradiol, enhance survival of new hippocampal neurons in adult male rats. Dev. Neurobiol. 67 (10), 1321–1333. https://doi.org/10.1002/dneu.20457.
Sumner, B.E., Fink, G., 1998. Testosterone as well as estrogen increases serotonin2A receptor mRNA and binding site densities in the male rat brain. Mol. Brain Res. 59 (2), 205–214. https://doi.org/10.1016/S0169-328X(98)00148-X.
Tibshirani, R., 1996. Regression shrinkage and selection via the lasso. J. R. Stat. Soc. Ser. B 58 (1), 267–288.
Toffoletto, S., Lanzenberger, R., Gingnell, M., Sundström-Poromaa, I., Comasco, E., 2014. Emotional and cognitive functional imaging of estrogen and progesterone effects in the female human brain: a systematic review. Psychoneuroendocrinology 50, 28–52. https://doi.org/10.1016/j.psyneuen.2014.07.025.
Van Buuren, S., 2012. Flexible Imputation of Missing Data. CRC press, Boca Raton, FL.
Voss, M.W., Vivar, C., Kramer, A.F., van Praag, H., 2013. Bridging animal and human models of exercise-induced brain plasticity. Trends Cognit. Sci. 17 (10), 525–544. https://doi.org/10.1016/j.tics.2013.08.001.
Wenger, E., Kühn, S., Verrel, J., Mårtensson, J., Bodammer, N.C., Lindenberger, U., Lövdén, M., 2017. Repeated structural imaging reveals nonlinear progression of experience-dependent volume changes in human motor cortex. Cerebr. Cortex 27 (5), 2911–2925. https://doi.org/10.1093/cercor/bhw141.
Zatorre, R.J., Fields, R.D., Johansen-Berg, H., 2012. Plasticity in gray and white. Nat. Neurosci. 15 (4), 528–536. https://doi.org/10.1038/nn.3045.
Zitzmann, M., Nieschlag, E., 2001. Testosterone levels in healthy men and the relation to behavioural and physical characteristics: facts and constructs. Eur. J. Endocrinol. 144 (3), 183–197. https://doi.org/10.1530/eje.0.1440183.