NeuroImage 224 (2021) 117416

Contents lists available at ScienceDirect

NeuroImage

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

Meta-analysis of generalized additive models in neuroimaging studies

Øystein Sørensen[a,*], Andreas M. Brandmaier[d,e], Dídac Macià[c], Klaus Ebmeier[f], Paolo Ghisletta[g,h,i], Rogier A. Kievit[j], Athanasia M. Mowinckel[a], Kristine B. Walhovd[a,b], Rene Westerhausen[a], Anders Fjell[a,b]

[a] Center for Lifespan Changes in Brain and Cognition, University of Oslo, Pb. 1094 Blindern, Oslo 0317, Norway
[b] Department of Radiology and Nuclear Medicine, Oslo University Hospital, Norway
[c] Departament de Medicina, Facultat de Medicina i Ciències de la Salut, Universitat de Barcelona, and Institut de Neurociències, Universitat de Barcelona, Spain
[d] Center for Lifespan Psychology, Max Planck Institute for Human Development, Berlin, Germany
[e] Max Planck UCL Centre for Computational Psychiatry and Ageing Research, Berlin, Germany
[f] Department of Psychiatry, University of Oxford, UK
[g] Faculty of Psychology and Educational Sciences, University of Geneva, Switzerland
[h] Swiss Distance University Institute, Switzerland
[i] Swiss National Centre of Competence in Research LIVES, University of Geneva, Switzerland
[j] MRC Cognition and Brain Sciences Unit, University of Cambridge, UK

ARTICLE INFO

Keywords:
Data protection
Distributed learning
Generalized additive mixed models
Generalized additive models
Meta-analysis
Privacy

ABSTRACT

Analyzing data from multiple neuroimaging studies has great potential in terms of increasing statistical power, enabling detection of effects of smaller magnitude than would be possible when analyzing each study separately and also allowing to systematically investigate between-study differences. Restrictions due to privacy or proprietary data as well as more practical concerns can make it hard to share neuroimaging datasets, such that analyzing all data in a common location might be impractical or impossible. Meta-analytic methods provide a way to overcome this issue, by combining aggregated quantities like model parameters or risk ratios. Most meta-analytic tools focus on parametric statistical models, and methods for meta-analyzing semi-parametric models like generalized additive models have not been well developed. Parametric models are often not appropriate in neuroimaging, where for instance age-brain relationships may take forms that are difficult to accurately describe using such models. In this paper we introduce meta-GAM, a method for meta-analysis of generalized additive models which does not require individual participant data, and hence is suitable for increasing statistical power while upholding privacy and other regulatory concerns. We extend previous works by enabling the analysis of multiple model terms as well as multivariate smooth functions. In addition, we show how meta-analytic *p*-values can be computed for smooth terms. The proposed methods are shown to perform well in simulation experiments, and are demonstrated in a real data analysis on hippocampal volume and self-reported sleep quality data from the Lifebrain consortium. We argue that application of meta-GAM is especially beneficial in lifespan neuroscience and imaging genetics. The methods are implemented in an accompanying R package `metagam`, which is also demonstrated.

1. Introduction

Combining brain imaging data across studies has great potential in terms of increasing statistical power, enabling discoveries of effects that might not be detectable in any single dataset. Due to regulatory and practical concerns, privacy in particular, it may not be possible to analyze all data in a single place. It may also sometimes be beneficial to analyze data from multiple studies in two stages, even when the data are available at a single location, e.g., when data do not fit in computer memory or runtime is nonlinear in the number of participants (Riley et al., 2010).

Meta-analytic techniques offer one way to increase statistical power without sharing raw data. By estimating the relationships under study separately in each data location, pooled estimates are obtained by combining the estimates without sharing the underlying data. With some exceptions, meta-analytic methods have been developed for combining parameters from parametric statistical models or for effect measures like relative risks (Hedges and Olkin, 1985; Sutton and Higgins, 2008).

* Corresponding author.
E-mail address: oystein.sorensen@psykologi.uio.no (Ø. Sørensen).

https://doi.org/10.1016/j.neuroimage.2020.117416
Received 6 February 2020; Received in revised form 23 September 2020; Accepted 25 September 2020
Available online 2 October 2020

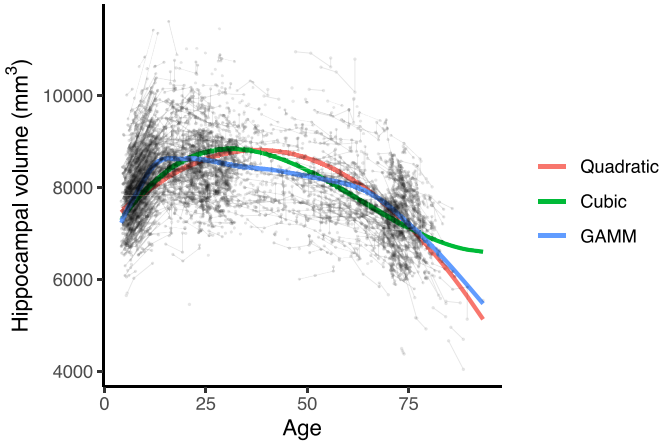

Fig. 1. Modeling lifespan trajectories. Example of modeling lifespan hippocampal volume with longitudinal data using linear mixed models with quadratic and cubic terms for age, as well as a generalized additive model. The black dots show individual observations and the black lines connect subsequent observations from the same individual. The GAMM was fitted with 20 cubic regression splines and a random intercept term for each individual, and the optimal smoothing parameter estimated with restricted maximum likelihood.

However, there are important cases in which it is impractical and suboptimal to enforce a parametric representation of the association under investigation, e.g., when an appropriate parametric model to approximate the data is not known, or its interpretability is not clear, as with high-degree polynomials. Examples include lifetime trajectories of brain development (Fjell et al., 2010), air quality measures (Gasparrini and Armstrong, 2010), and ecological phenomena (Borchers et al., 1997; Pedersen et al., 2019). Generalized additive models (GAMs) (Hastie and Tibshirani, 1986; Wood, 2017) are attractive for studying such relationships, and can easily be extended to longitudinal or other forms of clustered data via generalized additive mixed models (GAMMs), which, in addition to GAMs, can also estimate random effects.

Fig. 1 illustrates modeling lifespan trajectories of hippocampal volume changes using linear mixed models (LMMs) with quadratic and cubic polynomials for the age term, and a GAMM with a smooth term for age.[1] The data were taken from 4364 observations of 2023 healthy participants (age 4–93 years, 1–8 measurements per participant) from the Center for Lifespan Changes in Brain and Cognition (LCBC) longitudinal studies (Fjell et al., 2017; Walhovd et al., 2016). Detailed sample characteristics are presented in Supplementary Material I. The quadratic fit is not flexible enough to capture the steep increase during adolescence - moreover, it estimates the hippocampal volume to increase until the age of around 40. The cubic fit captures the volume growth during adolescence better than the quadratic fit, but fails to capture the decline that occurs after the age of around 70. The GAMM fit, on the other hand, is flexible enough to both capture the steep increase during adolescence, a period of moderate decline during adulthood, and finally a steeper decline at older age.[2]

As the methods for meta-analysis of GAMs and GAMMs are identical, we will refer to both as GAMs in the rest of this paper, unless distinction is necessary. For reasons that we will explain below, in this paper we will not discuss meta-analysis of the underlying parametric functions across GAMs. Rather, we present methods for combining GAM fits for neuroimaging data by pointwise meta-analysis of the fitted values. Although developed for use in meta-analytic neuroimaging studies, the methods can of course be applied to other types of data as well. The models under study can include any number of terms, including multivariate smooth functions. In order to employ these techniques, models should be fit separately for each cohort, with basis functions and knot placement chosen independently. Related previous works include meta-analysis of locally weighted regression fits (Schwartz and Zanobetti, 2000) and meta-analytic estimation of nonlinear dose-response relationships using individual participant data (Crippa et al., 2018; Sauerbrei and Royston, 2011).

The main applications we have in mind are meta-analysis of published results where the effects of interest are represented by functional relationships rather than single parameters, and multi-center studies in which it is impractical or not possible to analyze all brain imaging data in a single location. An example of the latter is the Enhancing Neuro Imaging Genetics through Meta Analysis project (ENIGMA: http://enigma.ini.usc.edu/), where meta-analysis of individual site summary statistics is the commonly applied strategy (e.g., Dennis et al., 2018; van Erp et al., 2018). The methods developed require that some model relating an outcome of interest to a set of explanatory variables has been fitted on data from each cohort, and that the model estimates can be shared across cohorts such that the expected response and their standard errors at new values of the explanatory variables can be computed. We provide a companion R package named `metagam` (Sørensen et al., 2020) containing functions for removing all individual participant data from GAMs fitted with the `mgcv` and `gamm4` packages (Wood and Scheipl, 2017; Wood, 2017), such that the resulting model object only contains aggregate measures which can easily be shared. The package also contains methods for combining the fits and analyzing the results, and will be demonstrated in Section 5.1. The comprehensive review of meta-analysis packages in R by Polanin et al. (2017) does not mention any existing packages for conducting this type of pointwise meta-analysis, so to the best of our knowledge, `metagam` is the first R package to provide this functionality.

The methods presented in this paper were motivated by a project in the Lifebrain consortium (http://www.lifebrain.uio.no/) (Walhovd et al., 2018). The goal was to study the relationship between self-reported sleep and hippocampal volume across six Lifebrain cohorts, and GAMMs were a natural model choice due to the expected non-linear age-relationships for self-reported sleep parameters and hippocampal volume. In this case a safe common data store was in place, but we initially hypothesized that it might be easier to have each cohort fit a model locally and share the overall result rather than analyzing all data in a single place, leading to the development of the methods presented here.

2. Background

2.1. Meta-analysis of parametric models

Consider a situation in which M cohorts $m = 1, \dots, M$ each have a dataset D_m with n_m participants. The response variable of interest is denoted y and there are p explanatory variables represented by the vector $\mathbf{x}$. If subject i in cohort m has been measured n_{mi} times, the data are $D_m = \{(y_{ij}, \mathbf{x}_{ij}), \text{for } i = 1, \dots, n_m, j = 1, \dots, n_{mi}\}$. Notably, this includes the case of individually varying numbers of assessments and time intervals between assessments. In practice, some of the explanatory variables will be time-varying, while others will be time-invariant. Purely cross-sectional data correspond to $n_{mi} = 1$ for all m and i.

Our interest concerns statistical inference on data from all studies, in the case where data cannot be analyzed jointly. When the relationship under study can be represented by a parametric model, well established methods exist for obtaining meta-analytic estimates of the model parameters. For example, if an LMM is used for longitudinal data (Laird and Ware, 1982), parameter estimates from each study can be combined using parametric meta-analysis (DerSimonian and Laird, 1986; Gasparrini et al., 2012). The same applies to related approaches based on structural

[1] The LMMs were fitted using R (Team, 2019) package `nlme` (Pinheiro et al., 2019) and the GAMM was fitted using `mgcv` (Wood, 2017), all with a random intercept term.

[2] Fig. 1 and all other figures in this paper were created using `ggplot2` (Wickham, 2016).

Table 1
Spline coefficients for models described in Section 2.4. The coefficient γ_2 was not possible to determine for Barcelona and Whitehall-II. In addition, γ_1 for Barcelona and Whitehall-II, γ_3 for Whitehall-II, and γ_8 for BASE-II are severe outliers.

Study	γ_1	γ_2	γ_3	γ_4	γ_5	γ_6	γ_7	γ_8
Barcelona	28142	–	6195	7719	7629	7421	7190	6310
BASE-II	4694	8374	6274	7919	7770	7213	7297	−17182
Betula	9605	8481	8380	8072	7840	7389	6994	7734
Cam-CAN	8298	8452	8397	8040	7916	7468	7291	6375
LCBC	8408	8479	8324	7689	7401	7468	7202	5819
Whitehall-II	1625151	–	−120033	7580	7528	7353	6935	6084

equation modeling (e.g., Brandmaier et al., 2018; Kievit et al., 2018) or generalized linear models (McCullagh and Nelder, 1989).

2.2. Generalized additive models

In many applications, assuming that the response [3] y is a smooth function of the explanatory variables, rather than following a model that is linear in its parameters (e.g., polynomial), may lead to better statistical fit, cf. Fig. 1. Generalized additive models (GAMs) (Hastie and Tibshirani, 1986) take this approach. Letting $\mathcal{X}_s$ denote the set of explanatory variables used by smooth function $f_s(\cdot)$, a GAM with S smooth terms can be written on the form

$$y = \beta_0 + \sum_{s=1}^{S} f_s(\mathcal{X}_s) + \epsilon, \tag{1}$$

where β_0 denotes the intercept and ϵ is a normally distributed residual. Constraints necessary for model identification are discussed in Appendix A. Each smooth function is a linear combination of K_s basis functions $b_{ks}(\cdot)$ with weights γ_{ks}, $k = 1, \ldots, K_s$,

$$f_s(\mathcal{X}_s) = \sum_{k=1}^{K_s} b_{ks}(\mathcal{X}_s)\gamma_{ks}. \tag{2}$$

Typically, each basis function is nonzero over a small part of the range of its variables, as defined by its knot locations. A linear parametric term for x_j is given by the special case $\mathcal{X}_s = \{x_j\}$, $K_s = 1$, $b_{1s}(x_j) = x_j$, and hence $f_s(\mathcal{X}_s) = \gamma_{1s}x_j$. Examples are provided in Supplementary Material II.

2.3. Smoothing

Least squares estimation of model (1) with a large number of basis functions for each term typically leads to wiggly estimates which overfit the data. Smoothing is thus necessary, and a popular and efficient solution involves penalizing the second derivatives of the smooth functions, while making sure the number of basis functions is sufficiently large to represent a wide range of functional forms (Wood, 2017). In the context of meta-analysis, smoothing is performed independently for each study. Supplementary Material II presents further details and a visualization of the effect of smoothing.

2.4. Limitations of parametric meta-analysis of generalized additive models

If each study used identical basis functions, a meta-analytic fit could be obtained by treating their weights as linear regression parameters (Gasparrini et al., 2012). However, as also noted by Crippa et al. (2018), if the range of some variable x_j differs between cohorts, enforcing the same knot placement is suboptimal and the model may not even be identified.

As an example, we consider modeling of lifespan trajectories of hippocampal volumes from six European cohorts. The data are further described in Section 5. As shown in Fig. 6 (top), these studies have widely varying age distributions. We fit GAMs relating baseline age to hippocampal volume for each cohort, but enforced the same knot location for all models, placed at eight equally spaced quantiles of the full data sample. Table 1 shows the corresponding spline coefficients. While these coefficients are not directly interpretable, outliers for a given sample indicate that its fit is highly different from the others, and a missing value indicates that the fit for the sample was not identified. As can be seen, Barcelona and Whitehall-II have missing values (-) for spline coefficient γ_2. In addition, there are extreme outliers: Barcelona has a severely outlying value for γ_1, BASE-II has an outlying value for γ_8, and Whitehall-II has outlying values for γ_1 and γ_3. This lack of identification and unstable coefficients is caused by using knot locations which, because they are forced to be equal across cohorts, are not suitable for the actual age distributions.

3. Pointwise meta-analysis of generalized additive models

3.1. Estimation of overall fits in pointwise meta-analysis

We now propose a model for meta-analysis of GAMs. We assume that a GAM has been fitted to the data from each cohort m separately, and that the vector $\mathbf{x}$ represents values of the explanatory variables for which a meta-analytic estimate of the regression function is sought. The expected response in cohort m is then given by

$$\hat{y}_m = \hat{f}_m(\mathbf{x}) = \hat{\beta}_{0,m} + \sum_{s=1}^{S} \hat{f}_{s,m}(\mathcal{X}_s), \; m = 1, \ldots, M. \tag{3}$$

Importantly, the basis functions and knot placements for a given smooth term $\hat{f}_{s,m}(\mathcal{X}_s)$ will in general vary across cohorts m. Each model term has a corresponding estimated standard deviation $\hat{\sigma}_{s,m}(\mathcal{X}_s)$, and the overall fit has estimated standard deviation $\hat{\sigma}_m(\mathbf{x})$.

We illustrate our methods by considering meta-analytic estimation of each single term separately, but note that inference on any combination of smooth terms, including the overall function, is readily obtained with the same methods. Some additional details related to identification of smooth terms are discussed in Appendix A. For ease of notation, we omit the dependency on $\mathcal{X}_s$ and $\mathbf{x}$ in the rest of this section. For example, $f_{s,m}$ means $f_{s,m}(\mathcal{X}_s)$ and $\sigma_{s,m}$ means $\sigma_{s,m}(\mathcal{X}_s)$.

The meta-analytic estimate of smooth term s is the weighted mean

$$\hat{f}_s = \frac{\sum_{m=1}^{M} \hat{f}_{s,m}\left(\hat{\sigma}_{s,m}^2 + \hat{\sigma}_s^2\right)^{-1}}{\sum_{m=1}^{M}\left(\hat{\sigma}_{s,m}^2 + \hat{\sigma}_s^2\right)^{-1}} \tag{4}$$

with standard error

$$\mathrm{se}_{\hat{f}_s} = \left\{ \sum_{m=1}^{M} \hat{\sigma}_{s,m}^2 + \hat{\sigma}_s^2 \right\}^{-1/2}. \tag{5}$$

[3] For ease of presentation, we assume a continuous outcome with normally distributed residuals, corresponding to an identity link function in a generalized additive model. The methods developed extend directly to other outcomes (e.g., binomial or count) by introducing a linear predictor $\eta = \beta_0 + \sum_{s=1}^{S} f_s(\mathcal{X}_s)$ with link function $g(\cdot)$ satisfying $g(y) = \eta$.

The term $\hat{\sigma}_s^2$ represents the estimated between-study variance, and fixed effects meta-analysis corresponds to the special case $\hat{\sigma}_s^2 = 0$. The DerSimonian-Laird estimator for between-sample variance (DerSimonian and Laird, 1986),

$$\hat{\sigma}_s^2 = \max\left\{0, \frac{\sum_{m=1}^{M} \hat{\sigma}_{s,m}^{-2}\left(\hat{f}_{s,m} - \frac{\sum_{m=1}^{M}\hat{\sigma}_{s,m}^{-2}\hat{f}_{s,m}}{\sum_{m=1}^{M}\hat{\sigma}_{s,m}^{-2}}\right) - (M-1)}{\sum_{m=1}^{M}\hat{\sigma}_{s,m}^{-2} - \frac{\sum_{m=1}^{M}\hat{\sigma}_{s,m}^{-4}}{\sum_{m=1}^{M}\hat{\sigma}_{s,m}^{-2}}}\right\}, \tag{6}$$

is computationally efficient as it does not require iteration, making it attractive in pointwise meta-analysis in which a separate estimate is required over a large number of grid points. However, iterative methods may give more accurate estimates (Veroniki et al., 2016). We refer to, e.g. Viechtbauer (2005) and Viechtbauer et al. (2015) for an overview of estimators of between-sample variance, all of which can be used with the methods presented.

Eqs. (4) and(5) are the familiar weighted means formulas used in meta-analysis, and have been used by Sauerbrei and Royston (2011) in a similar setting, focusing on meta-analysis of univariate functions estimated by fractional polynomials. In the fixed effects case, $\hat{f}_s$ is the estimated mean conditional on randomly pooling from the populations of the observed cohorts alone. Random effects analysis, on the other hand, estimates the marginal population effect f_s across all potential studies. See Viechtbauer (2010, Section. 2.3) for an excellent discussion of the interpretation of fixed vs. random effects meta-analyses. Confidence bands with level $(1-\alpha)$ are readily obtained for either estimates as

$$\left[\hat{f}_s + z_{\alpha/2}\mathrm{se}_{\hat{f}_s}, \hat{f}_s + z_{1-\alpha/2}\mathrm{se}_{\hat{f}_s}\right], \tag{7}$$

where z_q denotes the qth quantile of the standard normal distribution.

Pointwise meta-analysis requires software for computing predictions and standard errors for the models fitted in each study. In the case of GAMs, this requires knowledge of the basis functions along with the estimates and covariance matrices of spline weights, quantities which are readily available from software for fitting GAMs, like `mgcv` (Wood, 2017) or `pyGAM` (Servén and Brummitt, 2018). Importantly, individual participant data are not required for computing such predictions from already fitted models.

3.2. Inference for smoothing terms in pointwise meta-analysis

Tests for statistical significance of smooth terms can be performed by combining the p-values from each separate fit using methods for meta-analytic combination of p-values as summarized, e.g., in Becker (1994) or Loughin (2004). In particular, let $p_{s,m}$ denote the p-value obtained in cohort m for the hypothesis $H_{0,m} : f_s(\mathcal{X}_s) = 0$ that the smooth term s is zero over the whole range of explanatory variables $\mathcal{X}_s$ in cohort m, and let $H_{A,m} : f_s(\mathcal{X}_s) \neq 0$ denote the alternative hypothesis. Such p-values can be computed using the methods in Wood (2012). The meta-analytic null hypothesis then states that all p-values are uniformly distributed between 0 and 1, i.e., H_0: $p_{s,m} \sim U(0, 1)$, $m = 1, \ldots, M$, while the meta-analytic alternative hypothesis H_A states that all p-values have the same unknown non-uniform density which is non-increasing in the test statistic (Birnbaum, 1954). A large number of methods exist for computing the combined p-values. For example, Stouffer's sum of z method (Stouffer et al., 1949) uses the Z-score

$$Z_s = \frac{\sum_{m=1}^{M} w_m \Phi^{-1}(1 - p_{s,m})}{\sqrt{\sum_{m=1}^{M} w_m^2}}, \tag{8}$$

where Φ is the standard normal distribution and Φ^{-1} its quantile function, and $w_m, m = 1, \ldots, M$ are meta-analytic weights. Zaykin (2011) suggests defining the weights as the square root of the sample size, $w_m = \sqrt{n_m}$. The combined p-value is then defined by $p_s = 1 - \Phi(Z_s)$.

4. Simulation studies

Simulation studies were conducted in order to compare the performance of the pointwise meta-analysis approach presented in Section 3 to the ideal mega-analysis (McArdle and Horn, 1985) case, in which all data can be analyzed jointly. Section 4.1 reports simulation results comparing estimation of smooth terms, and Section 4.2 reports simulation results comparing statistical inference performance.

4.1. Function estimation

The first set of simulations compared pointwise meta-analysis to mega-analysis in terms of their ability to accurately estimate nonlinear functional forms and to quantify uncertainty with confidence bands. Data were generated from the model

$$y = f_0(x_0) + f_1(x_1) + f_2(x_2) + f_3(x_3) + \epsilon,$$

with all explanatory variables independently uniformly distributed in [0, 1] and $\epsilon \sim N(0, \sigma^2)$. The functional forms assumed were similar to those used by Marra and Wood (2012), and are shown as dashed black lines in Fig. 2.

Datasets with 4000 observations of (x_0, x_1, x_2, x_3, y) were independently sampled 1000 times. For each dataset, the following four cases were considered:

- In the mega-analysis case, all 4000 observations were analyzed jointly. This served as a gold standard, yielding the model that would be fit if all data were available to analyze with a single model.
- In the equal sample size case, the dataset was split into 5 "cohorts" of 800 observations each. Each cohort was analyzed independently, and the meta-analytic fit computed as outlined above.
- In the unequal sample size case, the dataset was split into 5 "cohorts" with 300, 500, 800, 1,000, and 1400 observations each.
- In the unequal range and sample size case, a first "cohort" was created by sampling 300 observations with $x_2 < 0.5$ from the full dataset, the second cohort by sampling 500 observations with $x_2 \geq 0.5$ from the remaining observations, the third cohort by sampling 800 observations with $x_1 < 0.5$ from the remaining observations, the fourth cohort by sampling 1000 observations with $x_1 \geq 0.5$ from the remaining observations, and the fifth cohort contained the remaining 1400 observations. Hence, this case has the same sample sizes as the unequal sample size case, but the ranges of x_1 and x_2 vary between cohorts.

In the latter three cases, fixed effects meta-analysis was conducted. Univariate smooth terms were estimated using cubic regression splines with 20, 10, 30, and 5 basis functions for $f_0(x_0)$, $f_1(x_1)$, $f_2(x_2)$, and $f_3(x_3)$, respectively. Knot placement was determined independently for each cohort, based on the quantiles of the explanatory variables. Second derivative smoothing was performed using generalized cross-validation, and standard error computations for each term included the uncertainty about the overall intercept as described in Marra and Wood (2012). For identifiability, the smooth terms were subject to sum-to-zero constraints over [0,1], cf. Appendix A. In the case study reported in Section 5, with a GAM regressing hippocampal volume on age and sleep quality, the mega-analysis case had an adjusted R squared value $R^2_{adj} = 0.37$ (cf. Supplementary Material IV, p. 13). Setting $\sigma = 1.0$ in the simulations gave $R^2_{adj} \approx 0.40$, thus close to a realistic noise level in neuroimaging studies, while $\sigma = 1.6$ corresponds to a high noise case with $R^2_{adj} \approx 0.20$. All simulations were repeated with each of these noise levels. Computations were performed in R version 3.6.2 (Team, 2019) with the package `mgcv` (Wood, 2017).

Figure 2 shows the average fits over all simulations. One can hypothesize that splitting a dataset into smaller parts and performing smoothing separately might lead to oversmoothing compared to analyzing all data in a single model. Considering Fig. 2 we see that this was the case for estimating $f_2(x_2)$ in the case with $\sigma = 1.0$, in which all meta-analysis

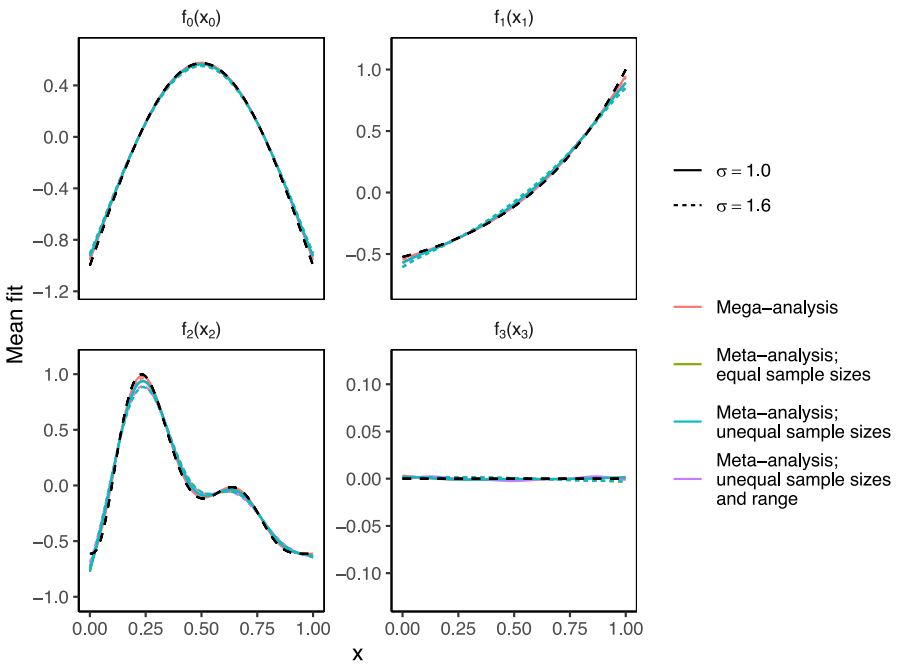

Fig. 2. Simulation estimates overlaid on true functions. Dashed black lines show true functions. The colored lines show mean fits averaged over 1,000 simulations as described in Section 4.1.

Table 2
Mean root-mean-square error of fitted terms in the case of equal sample sizes, unequal sample sizes, and mega-analysis, with residual standard deviation $\sigma = 1.0$ or $\sigma = 1.6$. Standard deviations across simulations are shown in parentheses.

Term	σ	Equal sample size	Unequal sample size	Unequal range and sample size	Mega-analysis
$f_0(x_0)$	1.00	0.037 (0.011)	0.037 (0.011)	0.038 (0.012)	0.035 (0.013)
$f_1(x_1)$	1.00	0.037 (0.014)	0.037 (0.014)	0.040 (0.013)	0.031 (0.011)
$f_2(x_2)$	1.00	0.060 (0.011)	0.060 (0.011)	0.062 (0.012)	0.054 (0.010)
$f_3(x_3)$	1.00	0.016 (0.009)	0.017 (0.009)	0.021 (0.010)	0.017 (0.012)
$f_0(x_0)$	1.60	0.054 (0.019)	0.053 (0.019)	0.055 (0.019)	0.052 (0.022)
$f_1(x_1)$	1.60	0.057 (0.020)	0.056 (0.020)	0.055 (0.018)	0.046 (0.020)
$f_2(x_2)$	1.60	0.089 (0.018)	0.089 (0.019)	0.089 (0.019)	0.079 (0.018)
$f_3(x_3)$	1.60	0.027 (0.015)	0.027 (0.015)	0.030 (0.016)	0.029 (0.020)

Table 3
Mean coverage of 95 % confidence intervals for fitted terms in the case of equal sample sizes, unequal sample sizes, and mega-analysis, with residual standard deviation $\sigma = 1.0$ or $\sigma = 1.6$. Standard deviations across simulations are shown in parentheses.

Term	σ	Equal sample size	Unequal sample size	Unequal range and sample size	Mega-analysis
$f_0(x_0)$	1.00	0.95 (0.21)	0.95 (0.21)	0.95 (0.23)	0.97 (0.16)
$f_1(x_1)$	1.00	0.89 (0.31)	0.90 (0.30)	0.89 (0.31)	0.97 (0.17)
$f_2(x_2)$	1.00	0.88 (0.32)	0.89 (0.31)	0.87 (0.33)	0.96 (0.20)
$f_3(x_3)$	1.00	0.99 (0.10)	0.99 (0.10)	0.96 (0.19)	0.99 (0.11)
$f_0(x_0)$	1.60	0.96 (0.20)	0.96 (0.19)	0.96 (0.20)	0.98 (0.15)
$f_1(x_1)$	1.60	0.86 (0.34)	0.88 (0.33)	0.91 (0.28)	0.97 (0.17)
$f_2(x_2)$	1.60	0.87 (0.33)	0.87 (0.34)	0.87 (0.34)	0.96 (0.20)
$f_3(x_3)$	1.60	0.99 (0.11)	0.99 (0.11)	0.98 (0.15)	0.98 (0.13)

cases slightly underestimated the two peaks of the true term. For the three other terms, the $\sigma = 1.0$ case had very low bias. In the high noise case, with $\sigma = 1.6$, oversmoothing can also be seen in the estimates of $f_1(x_1)$. The two meta-analyses with unequal sample size, also had somewhat too smooth estimates of $f_1(x_1)$ in the $\sigma = 1.0$ case. Overall, however, the average fits in the meta-analysis cases were very close to the true curves.

Table 2 shows the root-mean-square error (RMSE) of the fitted terms over the range [0, 1]. In both noise settings, the meta-analyses with equal and unequal sample size had only slightly higher RMSE than the mega-analytic estimates, and there did not seem to be any systematic difference between them. The meta-analysis with unequal range and sample size had RMSE very close to the two other meta-analytic cases.

Table 3 shows the average coverage across [0, 1] of 95 % confidence intervals computed with (7). The coverage of the confidence intervals of the mega-analytic estimates were close to 95 %, as expected from Marra and Wood (2012), and always conservative. All three meta-analytic cases had very similar coverage, varying between 86 % and 99 %. In particular for $f_1(x_1)$ and $f_2(x_2)$ the confidence intervals were somewhat too narrow, whereas for $f_0(x_0)$ and $f_3(x_3)$ the confidence intervals were slightly conservative.

4.2. Hypothesis testing and power

A second set of simulation experiments was conducted with the goal of comparing the statistical inference performance of meta-analysis to

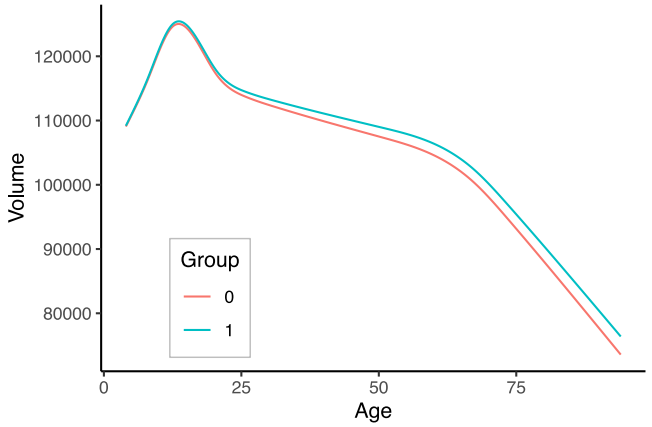

Fig. 3. Lifespan trajectories with group interaction. Functional forms assumed for lifespan trajectories in Section 4.2. Subjects are assumed to belong to either group 0 or 1, whose mean lifespan trajectories differ as shown by the two curves.

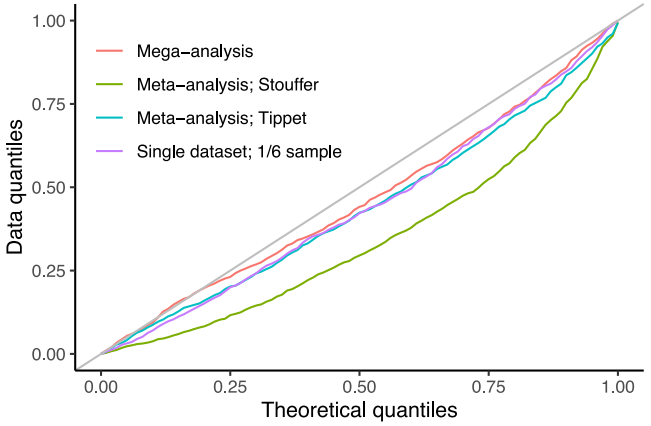

Fig. 4. P-value distribution under the null hypothesis. Quantile-quantile plot of p-values under the null hypothesis as described in Section 4.2, for the case of residual standard deviation equal to 3,500 and total sample size 3,000. Meta-analytic p-values were computed using both Stouffer's and Tippet's method, as shown by the legend.

mega-analysis. Two issues are of particular interest in this regard; first, whether the distribution of p-values is close to uniform when the null hypothesis is true (e.g., Murdoch et al. (2008)), and second, the power to reject a false null hypothesis. A nonlinear functional form approximating the lifespan trajectory of cerebellum cortex volume was estimated with the LCBC data (Fjell et al., 2017; Walhovd et al., 2016), as shown in Fig. 3. For the power analysis, it was assumed that a dichotomous group variable interacted with the lifespan trajectory, leading to slightly higher atrophy for members of the baseline group, especially in advanced ages. For analysis of the null distribution of p-values, the two groups had identical lifespan trajectories. Analyzing this type of smooth interactions is relevant, e.g., when investigating the impact of a given genetic variation on lifespan trajectories of brain measures (Walhovd et al., 2019).

Cross-sectional measurements were simulated with age uniformly distributed between 4 and 94 years, and group memberships randomly allocated to 0 or 1 with equal probabilities. For the mega-analysis, all measurements were analyzed in a single GAM, while for the meta-analysis, the data were first split into 6 datasets and analyzed separately, before a meta-analytic p-value was computed. For reference, the power obtained when using a single dataset of size 1/6th of the total dataset was also computed. A total of 1000 Monte Carlo samples were analyzed for each parameter setting. For the case of a nonzero group interaction, statistical power was computed as the fraction of the 1000 random simulations in which the group interaction was significant at a 5 % level. In the first set of simulations, the total sample size was fixed at 3000 while the residual standard deviation varied between 1000 and 15,000. In the second set of simulations, the residual standard deviation was fixed at 3500, and the total sample size varied between 900 and 3000. In all cases, "cohort fits" were computed by randomly splitting the dataset into 6 equally sized parts. The GAMs used to analyze the data in each sample were of the form

$$y = \beta_{0,m} + f_{1,m}(x_1) + f_{2,m}(x_1)x_2 + \beta_{2,m}x_2 + \epsilon,\ m = 1, \ldots, M,$$

where x_1 is age, $x_2 \in \{0, 1\}$ is an indicator for group membership, and ϵ is a normally distributed residual. The parameter $\beta_{0,m}$ represents the intercept, $\beta_{2,m}$ is the offset effect of membership in group 1, the smooth term $f_{1,m}(x_1)$ represents the age trajectory of subjects in group 0, and $f_{2,m}(x_1)$ represents the difference between the smooth term of subjects in group 1 and subjects in group 0. Hence, subjects in group 1 have age trajectory given by $f_{1,m}(x_1) + f_{2,m}(x_1)$. GAMs were fitted with the `gam` function in `mgcv` (Wood, 2017), using cubic regression splines to construct the smooth terms and generalized cross-validation for smoothing. Knot placement was determined independently for each study. The null hypothesis states that there is no difference between the lifespan trajectories across groups, and the p-values corresponding to this null hypothesis in each sample were directly obtained from the model fit, which uses the methods described in Wood (2012). For the meta-analysis, we compared several different methods for combining p-values: Wilkinson's maximum p (Wilkinson, 1951), Tippet's minimum p (Tippet, 1931), the logit-p method (Becker, 1994), Fisher's sum of logs (Fisher, 1925), Edgington's sum of p (Edgington, 1972), and Stouffer's sum of z (Stouffer et al., 1949), using the implementations in the R package `metap` (Dewey, 2019). As all samples in the meta-analysis were of equal size, equal meta-analytic weights were used in Stouffer's sum of z (8). The other methods do not use weights. Tippet's minimum p method gave p-values closest to uniform under the null hypothesis under most parameter settings, while Stouffer's sum of z method typically gave highest power. The p-values resulting from these two methods are hence shown in the results in this section, while complete results for all methods can be found in Supplementary Material III.

Fig. 4 shows quantile-quantile plots of the p-values obtained by meta-analysis, mega-analysis, and a fit of a single dataset in the case of no actual interaction between the group variable and the lifespan trajectories in the case with sample size 3000 and residual standard deviation 3500. Results for other values of these parameters were similar, and are shown in Supplementary Material III. The gray line shows the ideal reference line. All methods yielded p-values which deviated to some degree from the uniform distribution. Meta-analytic p-values computed using Tippet's minimum p method were close to the p-values obtained either in the mega-analysis or in the single data fit. p-values computed using Stouffer's sum of z, on the other hand, were considerably further from being uniformly distributed. As Fig. 4 shows, the p-values of the mega-analysis were not perfectly uniformly distributed. This is due to the approximate nature of the algorithms used to compute p-values in GAMMs, which need to take into account the overall uncertainty in the smoothing parameter (Wood, 2017, Sec. 6.12).

Fig. 5 (left) shows power curves for varying residual standard errors, and Fig. 5 (right) shows power curves over a range of sample sizes. In both cases, the mega-analytic approach outperforms the meta-analytic approaches. Stouffer's sum of z method obtained power closest to the mega-analysis, while Tippet's minimum p method had lower power. Analyzing a single dataset, representing 1/6th of the total data, gave much lower power than either of the other two approaches. This highlights the benefit of pointwise meta-analysis compared to separate analyses by each center, when data cannot be shared.

To summarize, meta-analysis using Stouffer's sum of z method had power fairly close to that of a mega-analysis, at an increased risk of falsely rejecting true null hypotheses. On the other hand, meta-analysis

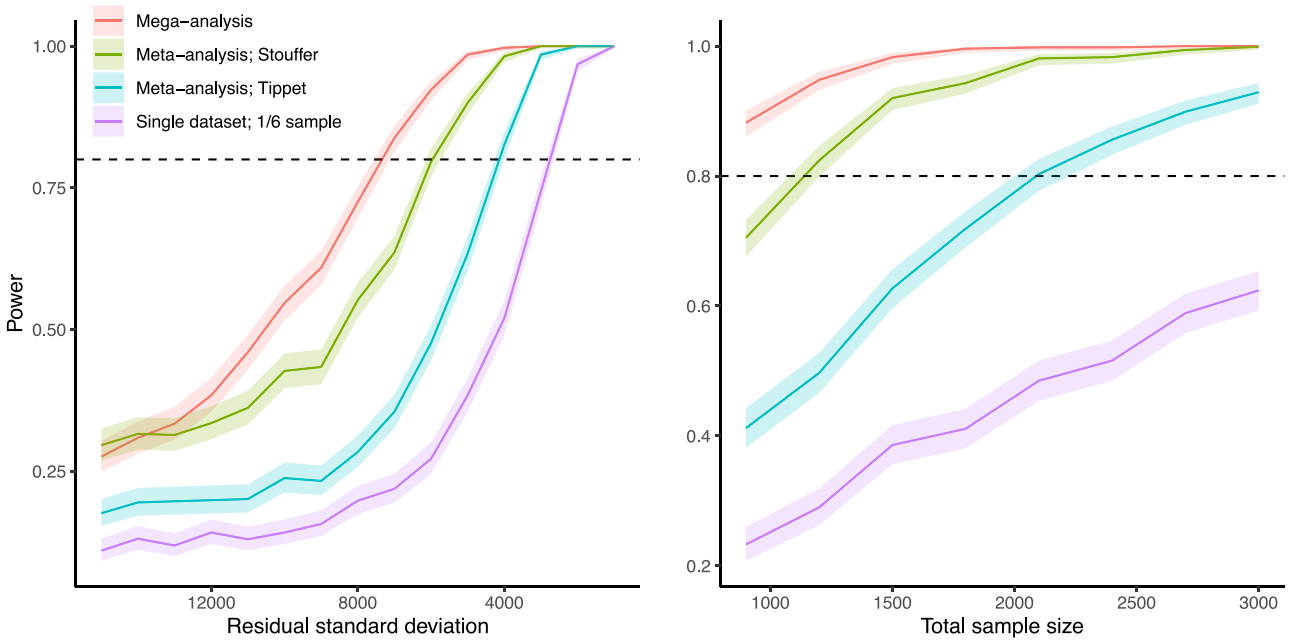

Fig. 5. Statistical power to detect interaction. Results of statistical power simulations desribed in Section 4.2. Left: fixed total sample size 3,000 and varying noise level. Right: fixed noise level σ=3,500 and varying total sample size. Shaded areas around curves show 95 % confidence intervals computed using the R package `Hmisc` (Harrell, 2019). Meta-analytic p-values were computed using both Stouffer's and Tippet's method, as shown by the legend.

using Tippet's minimum p method had risk of falsely rejecting a true null hypothesis close to that of a mega-analysis, at the cost of lower power. The other methods for combining p-values were somewhere inbetween these extremes, as shown in Supplementary Material III.

5. Case study

We will now illustrate the proposed methods on brain imaging data from six European cohorts analyzed by Fjell et al. (2019). The datasets contained measurements of sleep quality and hippocampal volume from the Berlin Study of Aging-II (BASE-II) (Bertram et al., 2013; Gerstorf et al., 2016), the Betula project (Nilsson et al., 1997), the Cambridge Centre for Ageing and Neuroscience study (Cam-CAN) (Taylor et al., 2017), Center for Lifespan Changes in Brain and Cognition longitudinal (LCBC) studies (Fjell et al., 2017; Walhovd et al., 2016), Whitehall-II (Filippini et al., 2014), and University of Barcelona brain studies (Abellaneda-Pérez et al., 2019; neiro et al., 2014; Rajaram et al., 2017). Self-reported sleep and hippocampal volume data from 2843 participants (18–90 years) were included. Longitudinal information on hippocampal volume was available for 1,065 participants, yielding a total of 4621 observations. Mean interval from first to last examination was 3.8 years (range 0.2–11.0 years). Participants were screened to be cognitively healthy and in general not suffer from conditions known to affect brain function, such as dementia, major stroke, multiple sclerosis, etc. Exact screening criteria were not identical across subsamples. Detailed sample characteristics are presented in the Supplementary Material I.

In Fjell et al. (2019), the data were analyzed jointly using GAMMs in a mega-analysis, taking into account both the clustering of repeated measurements within the same subject, and of subjects within a given cohort. However, the methods proposed in this paper enable this type of multi-cohort analysis also when the data cannot be shared. In this particular example we examine how hippocampal volume is related to age and to sleep quality as measured by the global score on the Pittsburgh Sleep Quality Index (PSQI) (Buysse et al., 1989). A low value of the PSQI variable indicates good sleep.

The following model was first fit to data from each study separately:

$$y_{ij} = \beta_0 + f_1(x_{ij,1}) + f_2(x_{ij,1})x_{i,2} + \beta_3 x_{i,3} + b_i + \epsilon_{ij}. \quad (9)$$

y_{ij} denotes hippocampal volume of subject i at timepoint j, $x_{ij,1}$ is the age of subject i at timepoint j, $x_{i,2}$ is the global PSQI score, and $x_{i,3}$ is the sex of subject i. $b_i \sim N(0, \sigma_b^2)$ is the random intercept of subject i and $\epsilon_{ij} \sim N(0, \sigma^2)$ is the residual. The main effect of age is represented by $f_1(x_1)$. $f_2(x_1)x_2$ is a varying-coefficient term (Hastie and Tibshirani, 1993), in which $f_2(x_1)$ is a regression coefficient for sleep quality which varies smoothly with age. Restricted maximum likelihood was used both for smoothing and estimation of random effect terms, and cubic regression splines were used as basis functions. The range of the age variable differed considerably between studies, as shown in the top part of Fig. 6. Hence, both the knot placement and the number of knots used to fit $f_1(x_1)$ and $f_2(x_1)$ was determined for each cohort separately. The simulation procedure described in Wood (2017, Ch. 5.9) was used to ensure that the number of knots was large enough to allow sufficient flexibility for the shapes of the smooth terms. The sleep quality scores were similarly distributed across cohorts, as shown in the bottom part of Fig. 6. Betula differs somewhat in shape from the others, due to a transformation that had to be applied to these data (Fjell et al., 2019). Fig. 7 shows the fits of the term $\beta_0 + f_1(x_1)$ in (9) relating age to hippocampal volume, over the range of ages in each cohort.

For the meta-analysis, we will focus on the effect of age on hippocampal volume including the intercept term, $\beta_0 + f_1(x_1)$, and the age-dependent effect of sleep quality on hippocampal volume, $f_2(x_1)$. To this end, we set up a grid over which to compute the estimates, containing the range of ages from 20 to 90 equally spaced by 0.1 year, and the value of the sleep quality score set to $x_2 = 1$, such that $f_2(x_1)x_2 = f_2(x_1)$, representing the main effect of sleep as a function of age. Random effects meta-analysis was used, with between-study variance estimated with the DerSimonian-Laird estimator shown in Eq. (6).

Fig. 8 shows the meta-analytic fits compared to the full data case. The estimated effects of age on hippocampal volume are very similar between the two approaches, although the meta-analytic fit lies somewhat above the mega-analytic fit for ages below 60 and has somewhat narrower confidence bands at low ages and wider confidence bands at high ages. A possible reason for the narrow confidence bands of the meta-analytic estimate of $f_1(x_1)$ for ages in the range from 30 to 55 years is that this age range is dominated by LCBC and Cam-CAN (Fig. 9), which have very similar estimated functional forms (Fig. 7). As shown in

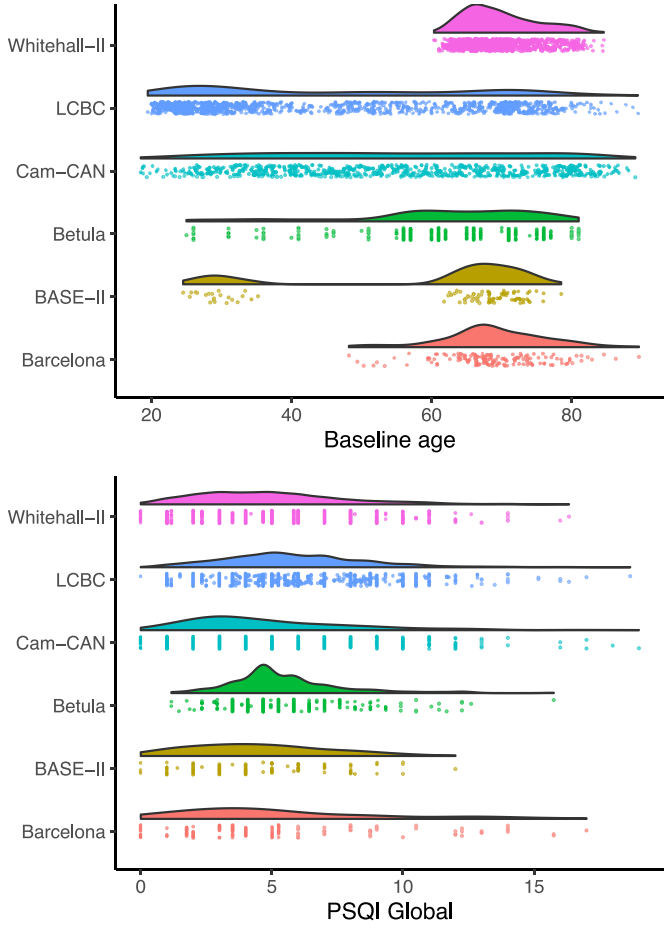

Fig. 6. Empirical distribution of explanatory variables. Raincloud plots (Allen et al., 2019) showing the distribution of baseline age (top) and global PSQI score (bottom) in the data from each study in Section 5.

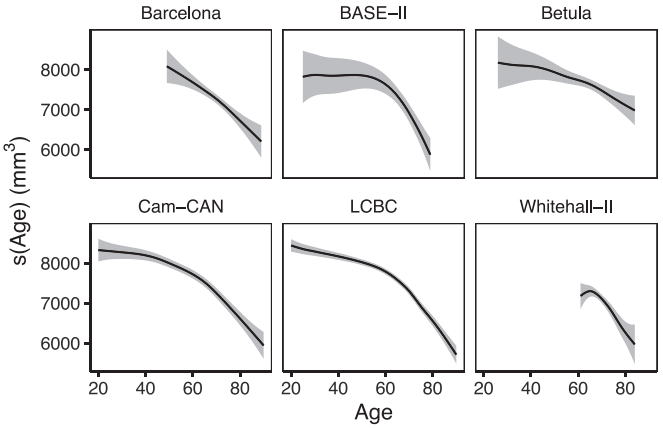

Fig. 7. Age trajectories for each cohort. Estimates of $\beta_0 + f_1(x_1)$ in (9), showing how age predicts hippocampal volume in each cohort. Gray shaded areas are 95 % confidence intervals.

Supplementary Material IV (p. 16), the estimated between-sample variance is even identically zero over part of this range. Since the standard error of the meta-analytic fit is estimated independently at each age (cf. Eq. (5)), the confidence bands hence become narrow, in contrast to the mega-analytic fit, for which the global smoothness assumption and the utilization of repeated measurements contribute to confidence bands whose width has little variation in the interior of the age range.

As in Fjell et al. (2019), there seems to be no effect of global PSQI score on hippocampal volume at any age, as can be seen by the confidence intervals covering zero in both cases (Fig. 8, right). In the meta-analytic case, the estimated curve has a peak at around 70 years, as opposed to the straight line estimated by the full data analysis. However, the confidence bands obtained with the two methods are highly overlapping. We note that while the mega-analysis estimates a linear varying-coefficient term $f_2(x_1)$, the meta-analytic estimate is nonlinear. As shown in Supplementary Material IV, all the individual cohort fits except Betula were very close to linearity. However, pointwise meta-analytic fits are nonlinear by construction, so even if all individual cohort fits estimated a linear effect, the meta-analytic estimate would in general be nonlinear. This can be seen by the fact that $\hat{f}_s(\mathbf{x})$ depends nonlinearly on the covariates $\mathbf{x}$ in Eq. (4), through the products of the estimated smooth terms with the meta-analytic weights. In contrast, the mega-analysis shrinks the total estimate towards a linear function through the second-derivative penalty. As a result, the mega-analytic estimate will be linear when the data do provide sufficient evidence of a nonlinear effect.

In order to quantify how much each study contributes to the meta-analytic fit at each value of an explanatory variable, we propose using dominance plots, visualizing $\hat{\sigma}^2_{s,m}/\mathrm{se}^2_{\hat{f}_s}$ for $m = 1, \ldots, M$. Fig. 9 (left) shows that LCBC and Cam-CAN are the main contributors to the meta-analytic fit for the main effect of age on hippocampal volume for ages up to around 50 years, after which the relative influence of the other studies starts increasing. Furthermore, the heterogeneity of the models fit in each study can be analyzed by computing Cochran's Q statistic (Cochran, 1954) over an explanatory variable, thus comparing $\hat{f}_{s,m}$ for $m = 1, \ldots, M$ independently at each value of the explanatory variable. Fig. 9 (right) shows a heterogeneity plot comparing the main effects of

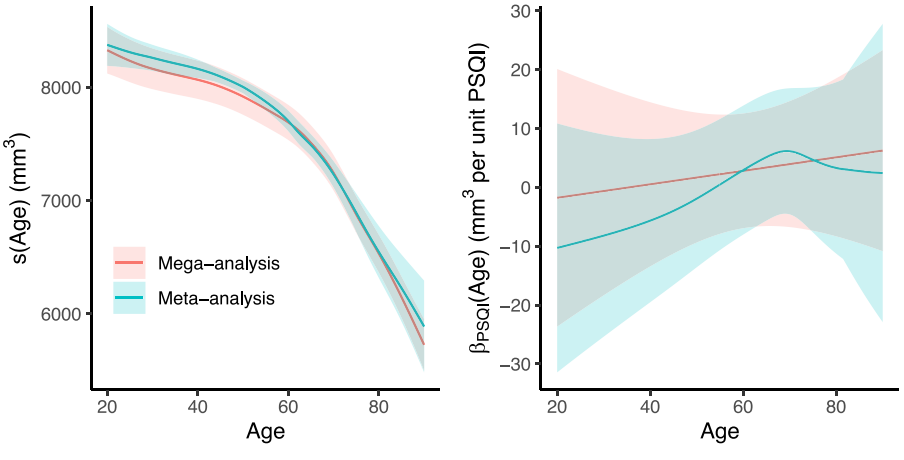

Fig. 8. Comparison of meta-analytic and mega-analytic estimates. Meta-analytic fits obtained as described in Section 5, compared to the corresponding fit obtained with full data. Left: effect of age on hippocampal volume, including the overall intercept. Right: effect of PSQI global score on hippocampal volume as a function of age.

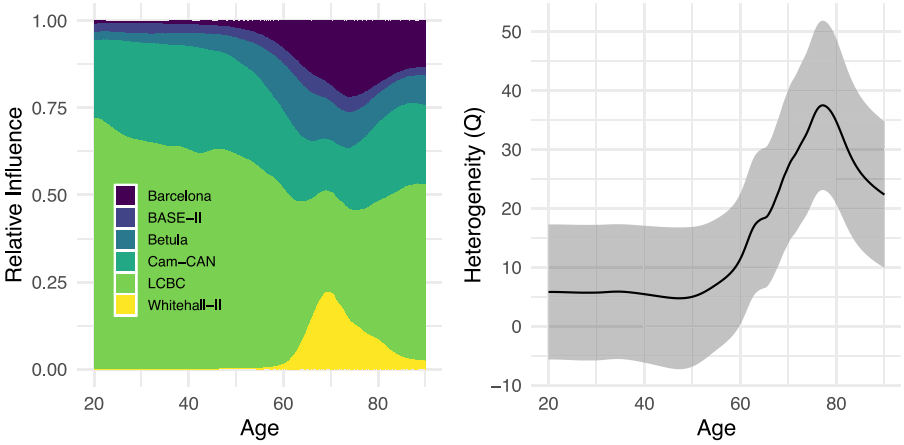

Fig. 9. Dominance and heterogeneity plots. Dominance and heterogeneity plots for $\beta_0 + f_1(x_1)$ in Eq. (9). Left: the relative contribution from each study to the meta-analytic fit over age. Right: Cochran's Q statistic for heterogeneity over age. Shaded areas represent 95 % confidence intervals.

age in each study, with 95 % confidence intervals represented by the shaded gray areas. The confidence interval in the heterogeneity plot does not contain zero for ages above 60, indicating that there is evidence of systematic differences across cohorts in the effect of age on hippocampal volume after the age of 60.

5.1. Pointwise meta-analysis in R with the 'metagam' package

This section shows how the meta-analysis described above can be conducted in R using the `metagam` package, which implements the methods presented in this paper. Some details are omitted for clarity, and are shown in Supplementary Material IV.

First, the following code fits a GAMM to the data for each study using the `mgcv` package (Wood, 2017).

```
library(mgcv)
# Fit GAMM in cohort 1
cohort_gam1 <- gamm(
  Hippocampus ~ s(Age) + s(Age, by = PSQI_Global) + Sex,
  data = cohort_data1, random = list(ID =~ 1), method = "REML")
```

The fitted model objects returned by `gamm()` contain the original data used to fit the model, as well as the responses. The `strip_rawdata()` function from `metagam` removes all individual participant data from each model fit, returning an object containing only aggregate quantities that can be shared without any individual data. The following lines attach the `metagam` package and then create an object `cohort_fit1`, which does not contain any individual-specific data.

```
library(metagam)
cohort_fit1 <- strip_rawdata(cohort_gam1)
```

Assuming each cohort has followed the two steps above, the following code gathers the model fits from each of the six cohorts in a list, creates a grid over which to predict, and finally uses the `metagam()` function to compute the meta-analytic fits.

```
# Combine fits from each cohort in a list
cohort_fits <- list(cohort_fit1, cohort_fit2, cohort_fit3,
                    cohort_fit4, cohort_fit5, cohort_fit6)

# Create a grid over which to compute meta-analytic fits
grid <- data.frame(
  Age = seq(from = 20, to = 90, by = 0.1),
  Sex = factor("Female", levels = c("Female", "Male")),
  PSQI_Global = 1)

# Smooth function of x_1, including overall intercept
metafit_age <- metagam(cohort_fits, grid, terms = "s(Age)",
                       method = "DL", intercept = TRUE)

# Age-varying slope of x_2, not including overall intercept
metafit_psqi <- metagam(cohort_fits, grid,
                        terms = "s(Age):PSQI_Global",
                        method = "DL", intercept = FALSE)
```

The argument `method = ''DL''` specifies that random effects meta-analysis should be used, with the DerSimonian-Laird estimator (DerSimonian and Laird, 1986). The `metafor` package (Viechtbauer, 2010) performs the actual estimation, and all estimators available in `metafor` may be used. By default, predictions from each model are computed over the whole supplied grid, thus extrapolating the estimates from cohorts whose data cover only a subset of the grid. Arguments can be specified in order to compute the predictions from each model only within the range of variables used to fit it. In practice, this latter option does not have much impact, since the standard errors are large outside of the range of the variables used in the fit, and hence the corresponding predictions get a very low weight at these points.

Finally, the dominance and heterogeneity plots shown in Fig. 9 are obtained with the commands:

```
plot_dominance(metafit_age)
plot_heterogeneity(metafit_age)
```

6. Discussion

We have proposed and illustrated a flexible way to obtain meta-analytic fits of GAMs in neuroimaging studies where individual participant data cannot be shared across cohorts. In the simulation studies, the meta-analytic procedure showed estimation performance close to that obtained in the ideal case, in which all data were analyzed in a single model, except that the meta-analytic estimates tended to have somewhat too narrow confidence intervals. Furthermore, the simulations showed that when testing for an interaction between a smooth function and a categorical variable, the distribution of p-values under the null hypothesis of no interaction, and the power to detect an actual interaction, were highly dependent on the chosen method for combining p-values, offering a trade-off between power and the probability of making false rejections. The proposed method is particularly useful when the variables under study have different ranges across cohorts, such that enforcing the same knot placement is suboptimal and might lead to nonidentified models. This is the case in many multi-cohort and consortium studies using neuroimaging data, where for instance age-range or patient distribution across a clinical indicator may vary considerably across samples. Differing variable ranges and knot placement are also inevitable across independent studies using GAMs to estimate some effect of interest in different study populations.

A case study illustrating the use of pointwise meta-analysis was considered in Section 5, in which the effect of sleep quality and age on hippocampal volume was estimated for six European cohorts. Due to the nonlinear lifespan relationship between age and hippocampal volume, GAMMs were preferable to LMMs when analyzing these data. However, the highly varying age distributions (Fig. 6) lead to nonidentified models when the same knot location was enforced across cohorts (cf.

Table 1). Meta-analysis of GAMMs by combining spline weights at each knot (Gasparrini et al., 2012) could hence not be used. The pointwise meta-analysis developed in this paper alleviated these issues, and allowed computing meta-analytic estimates of both the effect of age on hippocampal volume and the age-varying effect of sleep quality on hippocampal volume. Since the full data were available in a single location in this case, the meta-analytic estimates could be directly compared to a mega-analysis in which all data were analyzed jointly. The meta-analytic estimate of the effect of age on hippocampal volume was very close to the mega-analytic estimate (Fig. 8, left), although it had slightly narrower confidence bands for the middle age ranges. The meta-analytic estimate of the effect of sleep was also close to the mega-analytic estimate, both being almost zero over the full age range. A notable difference in the latter case was that while the mega-analysis estimated the effect of sleep to vary linearly with age, the meta-analytic estimate was nonlinear, as it will be by construction. An interesting topic for further study, which would enable a meta-analytic estimate to be linear when the smooth terms from each cohort are close to linear, involves imposing additional constraints on the meta-analytic fit, by using the degrees of freedom of the estimate from each cohort to inform the shape of the overall meta-analytic estimate. Dominance and heterogeneity plots (Fig. 9) were also introduced as additional tools for analyzing the relative impact of each dataset on the meta-analytic fit, and the heterogeneity of the estimated effects, respectively, both as functions of age.

One particular area of application for meta-GAM is imaging genetics. The need for very large sample sizes has long been recognized (Thompson et al., 2014), which imposes challenges due to privacy and data protection as well as practical issues regarding transfer, storage and processing of large amounts of neuroimaging data. These challenges have successfully been overcome in initiatives such as ENIGMA (Bearden and Thompson, 2017; Thompson et al., 2017) using a meta-analytic approach to gene discovery. Classic meta-analytic techniques are often inappropriate in situations where genetic effects are studied in interaction with other variables, such as age in a lifespan study. To test whether effects of genetic variants on a neuroimaging outcome measure vary as a function of age, or whether the lifespan trajectories of a neuroimaging outcome variable differ as a function of genetic variation (Piers, 2018; Walhovd et al., 2019), more complex modeling is needed. This functionality is provided by meta-GAM. As shown in Fig. 8, this meta-analytic approach yielded superior power to detect effects in such situations compared to single studies, although not completely reaching the same statistical power as mega-analyses in cases of total sample size less than 2000. Other examples of situations where meta-GAM would be applicable are when testing whether an effect varies as a function of another continuous variable, such as blood pressure, BMI or sleep duration. In all of these cases, the neuroanatomical outcome variable is expected to show a more complex relationship to the predictor variable than what can be captured by a parametric model. In these cases, meta-analytic GAM will be a powerful strategy to test genetic effects. Thus, we believe the present strategy may be a useful tool in neuroimaging genetics.

An alternative to the pointwise meta-analysis approach presented in this paper is to treat the fitted smooth functions from each cohort as samples from a Gaussian process (Murphy, 2012, Ch. 15). A meta-analytic fit could then be obtained by using these samples to estimate the parameters of a common smoothing kernel. This approach has been taken by Salimi-Khorshidi et al. (2011) for meta-analysis of neuroimaging data. Another alternative is using multiple imputation methods to generate synthetic data in each cohort with the same distributional properties as the original data, which can then be shared and analyzed in a mega-analysis (Little, 1993; Nowok et al., 2016; Rubin, 1993). Other possible extensions include accommodating potential correlation between the pointwise estimates in a given cohort using the robust variance estimation methods developed by Hedges et al. (2010), and to model the effect of cohort-specific covariates using multivariate meta-regression (Berkey et al., 1998). The latter may be used to account for systematic differences between trajectories across cohorts (cf. Fig. 9, right), and hence reduce potential bias in the meta-analytic estimates (Hofer and Piccinin, 2009). Also, deriving meta-analytic weights to use when combining p-values (Rosenthal, 1978) as in Section 4.2 could potentially yield p-values closer to those of the mega-analysis.

Although we have focused on the case in which data are not available in a single location, the proposed methods can also be useful in two-stage analysis with GAMs. In two-stage analysis, models are fitted separately for each cohort as described here, and then fit using meta-analytic techniques (Burke et al., 2016). This approach seems to be somewhat less efficient than analyzing the data jointly in a one-stage model (Boedhoe et al., 2019; Kontopantelis, 2018), but is useful when combining the data is impractical due to storage requirements or harmonization challenges (Sung et al., 2014). Finally, use of meta-GAM as a research synthesis method requires estimates and covariance matrices of spline weights as well as knot placement and basis functions to be properly reported by the studies to be combined in the meta-analysis. The `metagam` package easily allows extraction of such parameters from GAMs, creating model objects which can be made publicly available in repositories like the Open Science Framework (Foster and Deardorff, 2017), https://osf.io/.

7. Conclusion

Here we propose and demonstrate an approach to meta-analysis of neuroimaging results in situations where parametric models might not be appropriate, such as is often the case, e.g., in lifespan research. Parametric models might not be able accurately to capture lifespan trajectories of most neuroanatomical volumes, here as demonstrated for hippocampus. We show how such data can be analyzed using meta-analysis of generalized additive (mixed) models, and demonstrate that this is a powerful approach using simulated as well as real multi-cohort longitudinal data from the Lifebrain consortium. We believe this approach can be successfully applied in a range of settings where neuroimaging variables are used as outcome, especially within lifespan and neuroimaging genetics research, and beyond.

Data and code availability statement

The R scripts used to conduct the simulation studies in Section 4 are available in Supplementary Material V, and the R script used in the case study in Section 5 are available in Supplementary Material IV. The R package `metagam` implementing the methods developed in this paper is available from the Comprehensive R Archive Network (https://cran.r-project.org/package=metagam).

The data supporting the results of the current study are available from the corresponding author on reasonable request, given appropriate ethical and data protection approvals. Requests for data included in the Lifebrain meta-analysis can be submitted to the relevant principal investigators of each study. Contact information can be obtained from the corresponding author.

Ethics statement

The Lifebrain project is approved by the Regional Committee for Medical and Health Research Ethics of South Norway. Each sub-study was approved by the relevant ethical review board in the respective country (see Supplementary Material I).

Declaration of Competing Interest

The authors declare that they have no competing interests.

CRediT authorship contribution statement

Øystein Sørensen: Conceptualization, Methodology, Software, Formal analysis, Writing - original draft, Writing - review & editing, Visu-

alization. **Andreas M. Brandmaier:** Methodology, Software, Writing - original draft, Writing - review & editing, Visualization. **Dídac Macià:** Writing - original draft, Writing - review & editing. **Klaus Ebmeier:** Writing - original draft, Writing - review & editing. **Paolo Ghisletta:** Writing - original draft, Writing - review & editing. **Rogier A. Kievit:** Writing - original draft, Writing - review & editing. **Athanasia M. Mowinckel:** Software, Writing - original draft, Writing - review & editing, Visualization. **Kristine B. Walhovd:** Writing - original draft, Writing - review & editing. **Rene Westerhausen:** Writing - original draft, Writing - review & editing. **Anders Fjell:** Conceptualization, Methodology, Writing - original draft, Writing - review & editing, Visualization, Supervision.

Acknowledgement

The Lifebrain project is funded by the EU Horizon 2020 Grant: 'Healthy minds 0-100 years: Optimising the use of European brain imaging cohorts ("Lifebrain"). Grant agreement number: 732592. Call: Societal challenges: Health, demographic change and well-being. In addition, the different sub-studies are supported by different sources: LCBC: The European Research Council under grant agreements 283634, 725025 (to A.M.F.) and 313440 (to K.B.W.), as well as the Norwegian Research Council (to A.M.F., K.B.W.). Betula: a scholar grant from the Knut and Alice Wallenberg (KAW) foundation to L.N. University of Barcelona: Partially supported by a Spanish Ministry of Economy and Competitiveness (MINECO) grant to D-BF [grant number PSI2015-64227-R (AEI/FEDER, UE)]; by the Walnuts and Healthy Aging study (http://www.clinicaltrials.gov; Grant NCT01634841) funded by the California Walnut Commission, Sacramento, California. BASE-II has been supported by the German Federal Ministry of Education and Research under grant numbers 16SV5537/ 16SV5837/ 16SV5538/ 16SV5536K/ 01UW0808/ 01UW0706/ 01GL1716A/ 01GL1716B. CamCAN: Initial funding from the Biotechnology and Biological Sciences Research Council (BBSRC), followed by support from the Medical Research Council (MRC) Cognition & Brain Sciences Unit (CBU). Work on the Whitehall II Imaging Substudy was mainly funded by Lifelong Health and Wellbeing Programme Grant G1001354 from the UK Medical Research Council ("Predicting MRI Abnormalities with Longitudinal Data of the Whitehall II Substudy") to Dr Ebmeier. The Wellcome Centre for Integrative Neuroimaging is supported by core funding from award 203139/Z/16/Z from the Wellcome Trust.

Appendix A. Identifiability constraints on smooth terms

The smooth terms in the GAM (1) are only uniquely determined up to some additive constant. In order to compute the model fit, constraints have to be imposed on the smooth terms, effectively fixing $f_s(0)$ to some constant value. The default in the R package `mgcv` is to let each smooth term $f_s(\mathcal{X}_s)$ sum to zero over the observed data $\mathcal{X}_s$. This means requiring that the smooth term estimated from data in each cohort satisfy

$$\sum_{\mathbf{x}\in\mathcal{X}_{s,m}} f_{s,m}(\mathbf{x}) = 0,\ m = 1,\dots,M, \tag{A.1}$$

where we let $\mathcal{X}_{s,m}$ denote the actual values of $\mathcal{X}_s$ in cohort m. Using this approach the smooth term in each cohort has been constrained to sum to zero over its own data, and hence the terms are not directly comparable without correcting for this difference in offset. This is particularly important when the values of $\mathcal{X}_{s,m}$ cover different ranges across cohorts, as in Fig. 6.

One solution is to note that the smooth plus its intercept are comparable across cohorts, since the difference between the constraints is captured by the intercept term. To be precise, assume a GAM with a single smooth term f_1 is fit to data in cohorts m_1 and m_2, where the smooth term is constrained according to the data in cohort m_1, i.e.,

$$\sum_{\mathbf{x}\in\mathcal{X}_{s,m_1}} f_{s,m}(\mathbf{x}) = 0,\ m = m_1, m_2.$$

This yields estimates $\tilde{\beta}_{0,m} + \tilde{f}_{s,m}$ for $m = m_1, m_2$, and the terms $\tilde{f}_{s,m_1}$ and $\tilde{f}_{s,m_2}$ would be directly comparable. Instead constraining f_{s,m_2} over its own data would lead to a shift $\Delta\tilde{\beta}_{0,m_2}$ in the intercept estimated in cohort m_2, i.e.,

$$\sum_{\mathbf{x}\in\mathcal{X}_{s,m_1}} f_{s,m_2}(\mathbf{x}) = \Delta\tilde{\beta}_{0,m_2} + \sum_{\mathbf{x}\in\mathcal{X}_{s,m_2}} f_{s,m_2}(\mathbf{x}) = 0.$$

The estimated intercept in cohort m_2 would now be $\hat{\beta}_{0,m_2} = \tilde{\beta}_{0,m_2} + \Delta\tilde{\beta}_{0,m_2}$, where $\Delta\tilde{\beta}_{0,m_2}$ takes into account the difference between the sum-to-zero constraint in cohort m_1 and in cohort m_2. This argument generalizes to any number of cohorts and smooth terms. Hence, sum-to-zero constraints of the form (A.1) for each smooth can be imposed independently in each cohort fit, as long as the estimated intercept β_0 is added to each smooth term before combining. This implies replacing $\hat{f}_{s,m}$ with $\hat{\beta}_{0,m} + \hat{f}_{s,m}$ in equation (4). An important point for interpretation is that when using this option, the meta-analytic estimate of $\beta_0 + f_s$ incorporates both differences between estimated intercepts and differences between estimated smooth terms across cohorts.

Another way to resolve this issue is by imposing a constraint for each smooth term, specifying a point at which it should be exactly zero (Wood, 2017, Ch. 5.4.1). If the same point constraints have been applied when fitting the GAM to the data from each cohort, the smooth terms are all on the same scale and can be combined meta-analytically as described in Section 3. This approach hence replaces (A.1) by

$$f_{s,m}\left(\mathcal{X}_s^{pc}\right) = 0,\ m = 1,\dots,M, \tag{A.2}$$

for some point $\mathcal{X}_s^{pc}$ which is identical across cohorts. An advantage of this approach is that it does not require the intercept to be included in the meta-analysis; hence the meta-analytic estimate $\hat{f}_s$ contains only the smooth term. On the other hand, point constraint may lead to wider confidence bands for the smooth terms (Wood, 2017, Ch. 5.4.1). Also, this approach requires that point constraints are specified as part of the model to be fit to the data from each cohort. Note that the confidence interval for a smooth term subject to point constraint (A.2) does not need to have zero width at the constraint point $\mathcal{X}_s^{pc}$. The methods for constructing confidence intervals developed by Marra and Wood (2012) based on the work by Nychka (1988), take into account the uncertainty about the overall intercept as well as the uncertainty about the smooth term, and these typically yield better coverage properties than confidence intervals which only model the uncertainty of the smooth term.

Supplementary material

Supplementary material associated with this article can be found, in the online version, at 10.1016/j.neuroimage.2020.117416

References

Abellaneda-Pérez, K., Vaqué-Alcázar, L., Vidal-Piñeiro, D., Jannati, A., Solana, E., Bargalló, N., Santarnecchi, E., Pascual-Leone, A., Bartrés-Faz, D., 2019. Age-related differences in default-mode network connectivity in response to intermittent theta-burst stimulation and its relationships with maintained cognition and brain integrity in healthy aging. NeuroImage 188, 794–806. doi:10.1016/j.neuroimage.2018.11.036.

Allen, M., Poggiali, D., Whitaker, K., Marshall, T.R., Kievit, R.A., 2019. Raincloud plots: a multi-platform tool for robust data visualization. Wellcome Open Res. 4 (63). doi:10.12688/wellcomeopenres.15191.1.

Bearden, C.E., Thompson, P.M., 2017. Emerging global initiatives in neurogenetics: The enhancing neuroimaging genetics through meta-analysis (ENIGMA) consortium. Neuron 94 (2), 232–236. doi:10.1016/j.neuron.2017.03.033.

Becker, B.J., 1994. Combining significance levels. In: The handbook of research synthesis. Russell Sage Foundation, New York, NY, US, pp. 215–230. ISBN 0-87154-226-9 (Hardcover)

Berkey, C.S., Hoaglin, D.C., Antczak-Bouckoms, A., Mosteller, F., Colditz, G.A., 1998. Meta-analysis of multiple outcomes by regression with random effects. Statistics in Medicine 17 (22), 2537–2550. 10.1002/(sici)1097-0258(19981130)17:22<2537::aid-sim953>3.0.co;2-c

Bertram, L., Böckenhoff, A., Demuth, I., Düzel, S., Eckardt, R., Li, S.-C., Lindenberger, U., Pawelec, G., Siedler, T., Wagner, G.G., Steinhagen-Thiessen, E., 2013. Cohort profile: the Berlin aging study II (BASE-II)†. Int. J. Epidemiol. 43 (3), 703–712. doi:10.1093/ije/dyt018.

Birnbaum, A., 1954. Combining independent tests of significance*. J. Am. Stat. Assoc. 49 (267), 559–574. doi:10.1080/01621459.1954.10483521.

Boedhoe, P.S.W., Heymans, M.W., Schmaal, L., Abe, Y., Alonso, P., Ameis, S.H., Anticevic, A., Arnold, P.D., Batistuzzo, M.C., Benedetti, F., Beucke, J.C., Bollettini, I., Bose, A., Brem, S., Calvo, A., Calvo, R., Cheng, Y., Cho, K.I.K., Ciullo, V., Dallaspezia, S., Denys, D., Feusner, J.D., Fitzgerald, K.D., Fouche, J.-P., Fridgeirsson, E.A., Gruner, P., Hanna, G.L., Hibar, D.P., Hoexter, M.Q., Hu, H., Huyser, C., Jahanshad, N., James, A., Kathmann, N., Kaufmann, C., Koch, K., Kwon, J.S., Lazaro, L., Lochner, C., Marsh, R., Martínez-Zalacaín, I., Mataix-Cols, D., Menchón, J.M., Minuzzi, L., Morer, A., Nakamae, T., Nakao, T., Narayanaswamy, J.C., Nishida, S., Nurmi, E.L., O'Neill, J., Piacentini, J., Piras, F., Piras, F., Reddy, Y.C.J., Reess, T.J., Sakai, Y., Sato, J.R., Simpson, H.B., Soreni, N., Soriano-Mas, C., Spalletta, G., Stevens, M.C., Szeszko, P.R., Tolin, D.F., van Wingen, G.A., Venkatasubramanian, G., Walitza, S., Wang, Z., Yun, J.-Y., Working-Group, E.-O., Thompson, P.M., Stein, D.J., van den Heuvel, O.A., Twisk, J.W.R., 2019. An empirical comparison of meta- and mega-analysis with data from the enigma obsessive-compulsive disorder working group. Frontiers in Neuroinformatics 12 (102). doi:10.3389/fninf.2018.00102. ISSN 1662-5196

Borchers, D.L., Buckland, S.T., Priede, I.G., Ahmadi, S., 1997. Improving the precision of the daily egg production method using generalized additive models. Can. J. Fish. Aquat. Sci. 54 (12), 2727–2742. doi:10.1139/f97-134.

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

Burke, D.L., Ensor, J., Riley, R.D., 2016. Meta-analysis using individual participant data: one-stage and two-stage approaches, and why they may differ. Stat. Med. 36 (5), 855–875. doi:10.1002/sim.7141.

Buysse, D.J., Reynolds, C.F., Monk, T.H., Berman, S.R., Kupfer, D.J., 1989. The Pittsburgh sleep quality index: a new instrument for psychiatric practice and research. Psychiatry Res. 28 (2), 193–213. doi:10.1016/0165-1781(89)90047-4.

Cochran, W.G., 1954. The combination of estimates from different experiments. Biometrics 10 (1), 101. doi:10.2307/3001666.

Crippa, A., Thomas, I., Orsini, N., 2018. A pointwise approach to dose-response meta-analysis of aggregated data. Int. J. Stat. Med.Res. 7, 25–32. doi:10.6000/1929-6029.2018.07.02.1.

Dennis, E.L., Wilde, E.A., Newsome, M.R., Scheibel, R.S., Troyanskaya, M., Velez, C., Wade, B.S.C., Drennon, A.M., York, G.E., Bigler, E.D., Abildskov, T.J., Taylor, B.A., Jaramillo, C.A., Eapen, B., Belanger, H., Gupta, V., Morey, R., Haswell, C., Levin, H.S., Hinds, S.R., Walker, W.C., Thompson, P.M., Tate, D.F., 2018. ENIGMA military brain injury: a coordinated meta-analysis of diffusion MRI from multiple cohorts. 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI2018). IEEE doi:10.1109/isbi.2018.8363830.

DerSimonian, R., Laird, N., 1986. Meta-analysis in clinical trials. Controlled Clinical Trials 7 (3), 177–188. doi:10.1016/0197-2456(86)90046-2. ISSN 0197-2456

Dewey, M., 2019. metap: meta-analysis of significance valuesR package version 1.2.

Edgington, E.S., 1972. An additive method for combining probability values from independent experiments. J. Psychol. 80 (2), 351–363. doi:10.1080/00223980.1972.9924813.

van Erp, Theo, G.M., Walton, E., Hibar, D.P., Schmaal, L., Jiang, W., Glahn, D.C., Pearlson, G.D., Yao, N., Fukunaga, M., Hashimoto, R., Okada, N., Yamamori, H., Bustillo, J.R., Clark, V.P., Agartz, I., Mueller, B.A., Cahn, W., de Zwarte, S.M.C., Pol, H., Hulshoff, E., Kahn, R.S., Ophoff, R.A., van Haren, Neeltje, E.M., Andreassen, O.A., Dale, A.M., Doan, N.T., Gurholt, T.P., Hartberg, C.B., Haukvik, U.K., Jørgensen, K.N., Lagerberg, T.V., Melle, I., Westlye, L.T., Gruber, O., Kraemer, B., Richter, A., Zilles, D., Calhoun, V.D., Crespo-Facorro, B., Roiz-Santiañez, R., Tordesillas-Gutiérrez, D., Loughland, C., Carr, V.J., Catts, S., Cropley, V.L., Fullerton, J.M., Green, M.J., Henskens, F.A., Jablensky, A., Lenroot, R.K., Mowry, B.J., Michie, P.T., Pantelis, C., Quidé, Y., Schall, U., Scott, R.J., Cairns, M.J., Seal, M., Tooney, P.A., Rasser, P.E., Cooper, G., Weickert, C.S., Weickert, T.W., Morris, D.W., Hong, E., Kochunov, P., Beard, L.M., Gur, R.E., Gur, R.C., Satterthwaite, T.D., Wolf, D.H., Belger, A., Brown, G.G., Ford, J.M., Macciardi, F., Mathalon, D.H., O'Leary, D.S., Potkin, S.G., Preda, A., Voyvodic, J., Lim, K.O., McEwen, S., Yang, F., Tan, Y., Tan, S., Wang, Z., Fan, F., Chen, J., Xiang, H., Tang, S., Guo, H., Wan, P., Wei, D., Bockholt, H.J., Ehrlich, S., Wolthusen, R.P.F., King, M.D., Shoemaker, J.M., Sponheim, S.R., Haan, L.D., Koenders, L., Machielsen, M.W., van Amelsvoort, T., Veltman, D.J., Assogna, F., Banaj, N., de Rossi, P., Iorio, M., Piras, F., Spalletta, G., McKenna, P.J., Pomarol-Clotet, E., Salvador, R., Corvin, A., Donohoe, G., Kelly, S., Whelan, C.D., Dickie, E.W., Rotenberg, D., Voineskos, A.N., Ciufolini, S., Radua, J., Dazzan, P., Murray, R., Marques, T.R., Simmons, A., Borgwardt, S., Egloff, L., Harrisberger, F., Riecher-Rössler, A., Smieskova, R., Alpert, K.I., Wang, L., Jönsson, E.G., Koops, S., Sommer, I.E.C., Bertolino, A., Bonvino, A., Giorgio, A.D., Neilson, E., Mayer, A.R., Stephen, J.M., Kwon, J.S., Yun, J.-Y., Cannon, D.M., McDonald, C., Lebedeva, I., Tomyshev, A.S., Akhadov, T., Kaleda, V., Fatouros-Bergman, H., Flyckt, L., Busatto, G.F., Rosa, P.G.P., Serpa, M.H., Zanetti, M.V., Hoschl, C., Skoch, A., Spaniel, F., Tomecek, D., Hagenaars, S.P., McIntosh, A.M., Whalley, H.C., Lawrie, S.M., Knöchel, C., Oertel-Knöchel, V., Stäblein, M., Howells, F.M., Stein, D.J., Temmingh, H.S., Uhlmann, A., Lopez-Jaramillo, C., Dima, D., McMahon, A., Faskowitz, J.I., Gutman, B.A., Jahanshad, N., Thompson, P.M., Turner, J.A., Farde, L., Flyckt, L., Engberg, G., Erhardt, S., Fatouros-Bergman, H., Cervenka, S., Schwieler, L., Piehl, F., Agartz, I., Collste, K., Victorsson, P., Malmqvist, A., Hedberg, M., Orhan, F., 2018. Cortical brain abnormalities in 4474 individuals with schizophrenia and 5098 control subjects via the enhancing neuro imaging genetics through meta analysis (ENIGMA) consortium. Biol. Psychiatry 84 (9), 644–654. doi:10.1016/j.biopsych.2018.04.023.

Filippini, N., Zsoldos, E., Haapakoski, R., Sexton, C.E., Mahmood, A., Allan, C.L., Topiwala, A., Valkanova, V., Brunner, E.J., Shipley, M.J., Auerbach, E., Moeller, S., Uğurbil, K., Xu, J., Yacoub, E., Andersson, J., Bijsterbosch, J., Clare, S., Griffanti, L., Hess, A.T., Jenkinson, M., Miller, K.L., Salimi-Khorshidi, G., Sotiropoulos, S.N., Voets, N.L., Smith, S.M., Geddes, J.R., Singh-Manoux, A., Mackay, C.E., Kivimäki, M., Ebmeier, K.P., 2014. Study protocol: the Whitehall II imaging sub-study. BMC Psychiatry 14 (1). doi:10.1186/1471-244x-14-159.

Fisher, R.A., 1925. Statistical Methods for Research Workers. Oliver and Boyd, Edinburgh.

Fjell, A.M., Idland, A.-V., Sala-Llonch, R., Watne, L.O., Borza, T., Brækhus, A., Lona, T., Zetterberg, H., Blennow, K., Wyller, T.B., Walhovd, K.B., 2017. Neuroinflammation and tau interact with amyloid in predicting sleep problems in aging independently of atrophy. Cereb. Cortex 28 (8), 2775–2785. doi:10.1093/cercor/bhx157.

Fjell, A.M., Sørensen, Ø., Amlien, I.K., Bartrés-Faz, D., Bros, D.M., Buchmann, N., Demuth, I., Drevon, C.A., Düzel, S., Ebmeier, K.P., Idland, A.-V., Kietzmann, T.C., Kievit, R., Kühn, S., Lindenberger, U., Mowinckel, A.M., Nyberg, L., Price, D., Sexton, C.E., Solé-Padullés, C., Pudas, S., Sederevicius, D., Suri, S., Wagner, G., Watne, L.O., Westerhausen, R., Zsoldos, E., Walhovd, K.B., 2019. Self-reported sleep relates to hippocampal atrophy across the adult lifespan – results from the lifebrain consortium. Sleep doi:10.1093/sleep/zsz280.

Fjell, A.M., Walhovd, K.B., Westlye, L.T., Østby, Y., Tamnes, C.K., Jernigan, T.L., Gamst, A., Dale, A.M., 2010. When does brain aging accelerate? dangers of quadratic fits in cross-sectional studies. NeuroImage 50 (4), 1376–1383. doi:10.1016/j.neuroimage.2010.01.061. ISSN 1053-8119

Foster, E.D., Deardorff, A., 2017. Open science framework (OSF). J. Med. Lib. Assoc. 105 (2). doi:10.5195/jmla.2017.88.

Gasparrini, A., Armstrong, B., 2010. Time series analysis on the health effects of temperature: Advancements and limitations. Environmental Research 110 (6), 633–638. doi:10.1016/j.envres.2010.06.005. ISSN 0013-9351

Gasparrini, A., Armstrong, B., Kenward, M.G., 2012. Multivariate meta-analysis for non-linear and other multi-parameter associations. Stat. Med. 31 (29), 3821–3839. doi:10.1002/sim.5471.

Gerstorf, D., Bertram, L., Lindenberger, U., Pawelec, G., Demuth, I., Steinhagen-Thiessen, E., Wagner, G.G., 2016. Editorial. Gerontology 62 (3), 311–315. doi:10.1159/000441495.

Harrell, F. E., 2019. Hmisc: Harrell MiscellaneousR package version 4.3-0. https://CRAN.R-project.org/package=Hmisc.

Hastie, T., Tibshirani, R., 1986. Generalized additive models. Statist. Sci. 1 (3), 297–310. doi:10.1214/ss/1177013604. 08

Hastie, T., Tibshirani, R., 1993. Varying-coefficient models. J. R. Stat. Soc. 55 (4), 757–779. doi:10.1111/j.2517-6161.1993.tb01939.x.

Hedges, L.V., Olkin, I., 1985. Statistical Methods for Meta-analysis. Academic Press, Orland, FL.

Hedges, L.V., Tipton, E., Johnson, M.C., 2010. Robust variance estimation in meta-regression with dependent effect size estimates. Research Synthesis Methods 1 (1), 39–65. doi:10.1002/jrsm.5. 1

Hofer, S.M., Piccinin, A.M., 2009. Integrative data analysis through coordination of measurement and analysis protocol across independent longitudinal studies. Psychol. Methods 14 (2), 150–164. doi:10.1037/a0015566.

Kievit, R.A., Brandmaier, A.M., Ziegler, G., van Harmelen, A.-L., de Mooij, S.M.M., Moutoussis, M., Goodyer, I.M., Bullmore, E., Jones, P.B., Fonagy, P., Lindenberger, U., Dolan, R.J., 2018. Developmental cognitive neuroscience using latent change score models: a tutorial and applications. Dev. Cognit. Neurosci. 33, 99–117. doi:10.1016/j.dcn.2017.11.007.

Kontopantelis, E., 2018. A comparison of one-stage vs two-stage individual patient data meta-analysis methods: a simulation study. Res. Synth. Methods doi:10.1002/jrsm.1303.

Laird, N.M., Ware, J.H., 1982. Random-effects models for longitudinal data. Biometrics 38 (4), 963–974. ISSN 0006341X, 15410420

Little, R.J.A., 1993. Statistical analysis of masked data. Journal of Official Statistics 9 (2), 407. https://search.proquest.com/docview/1266808565?accountid=14699

Loughin, T.M., 2004. A systematic comparison of methods for combining p-values from independent tests. Computational Statistics & Data Analysis 47 (3), 467–485. doi:10.1016/j.csda.2003.11.020. ISSN 0167-9473

Marra, G., Wood, S.N., 2012. Coverage properties of confidence intervals for generalized additive model components. Scand. J. Stat. 39 (1), 53–74. doi:10.1111/j.1467-9469.2011.00760.x.

McArdle, J.J., Horn, J.L., 1985. Mega Analyses of the Wais: Structural and Dynamic Models of Adult Intellectual Ability. National institute of aging grant report. University of Viginia.

McCullagh, P., Nelder, J.A., 1989. Generalized Linear Models. Chapman & Hall / CRC, London.

Murdoch, D.J., Tsai, Y.-L., Adcock, J., 2008. P-values are random variables. Am. Stat. 62 (3), 242–245. doi:10.1198/000313008x332421.

Murphy, K.P., 2012. Machine Learning: A Probabilistic Perspective. The MIT Press. ISBN 0262018020, 9780262018029

neiro, V.-P., Martin-Trias, D., Arenaza-Urquijo, P., Eider, M., Sala-Llonch, R., Clemente, I.C., Mena-Sánchez, I., Bargalló, N., Falcón, C., Pascual-Leone, A., Bartrés-Faz, D., 2014. Task-dependent activity and connectivity predict episodic memory network-based responses to brain stimulation in healthy aging. Brain Stimul. 7 (2), 287–296. doi:10.1016/j.brs.2013.12.016.

Nilsson, L.-G., Bäckman, L., Erngrund, K., Nyberg, L., Adolfsson, R., Bucht, G., Karlsson, S., Widing, M., Winblad, B., 1997. The Betula prospective cohort study: Memory, health, and aging. Aging Neuropsychol. Cognit. 4 (1), 1–32. doi:10.1080/13825589708256633.

Nowok, B., Raab, G.M., Dibben, C., 2016. synthpop: bespoke creation of synthetic data in R. J. Stat. Softw. 74 (11). doi:10.18637/jss.v074.i11.

Nychka, D., 1988. Bayesian confidence intervals for smoothing splines. J. Am. Stat. Assoc. 83 (404), 1134–1143. doi:10.1080/01621459.1988.10478711.

Pedersen, E.J., Miller, D.L., Simpson, G.L., Ross, N., 2019. Hierarchical generalized additive models in ecology: an introduction with mgcv. PeerJ 7 (e6876). doi:10.7717/peerj.6876.

Piers, R.J., 2018. Structural brain volume differences between cognitively intact ApoE4 carriers and non-carriers across the lifespan. Neural Regener. Res. 13 (8), 1309. doi:10.4103/1673-5374.235408.

Pinheiro, J., Bates, D., DebRoy, S., Sarkar, D., 2019. R core team, nlme: linear and nonlinear mixed effects modelsURL https://CRAN.R-project.org/package=nlme. R package version 3.1-143.

Polanin, J.R., Hennessy, E.A., Tanner-Smith, E.E., 2017. A review of meta-analysis packages in R. J. Educ. Behav. Stat. 42 (2), 206–242. doi:10.3102/1076998616674315.

Rajaram, S., Valls-Pedret, C., Cofán, M., Sabaté, J., Serra-Mir, M., Pérez-Heras, A.M., Arechiga, A., Casaroli-Marano, R.P., Alforja, S., Sala-Vila, A., Doménech, M., Roth, I., Freitas-Simoes, T.M., Calvo, C., López-Illamola, A., Haddad, E., Bitok, E., Kazzi, N., Huey, L., Fan, J., Ros, E., 2017. The walnuts and healthy aging study (WAHA): protocol for a nutritional intervention trial with walnuts on brain aging. Front. Aging Neurosci. 8. doi:10.3389/fnagi.2016.00333.

Riley, R.D., Lambert, P.C., Abo-Zaid, G., 2010. Meta-analysis of individual participant data: rationale, conduct, and reporting. BMJ 340. doi:10.1136/bmj.c221. ISSN 0959-8138

Rosenthal, R., 1978. Combining results of independent studies. Psychol. Bull. 85 (1), 185–193. doi:10.1037/0033-2909.85.1.185.

Rubin, D.B., 1993. Discussion statistical disclosure limitation. Journal of Official Statistics 9 (2), 461. 06

Salimi-Khorshidi, G., Nichols, T.E., Smith, S.M., Woolrich, M.W., 2011. Using Gaussian-process regression for meta-analytic neuroimaging inference based on sparse observations. IEEE Trans. Med. Imaging 30 (7), 1401–1416. doi:10.1109/tmi.2011.2122341.

Sauerbrei, W., Royston, P., 2011. A new strategy for meta-analysis of continuous covariates in observational studies. Stat. Med. 30 (28), 3341–3360. doi:10.1002/sim.4333.

Schwartz, J., Zanobetti, A., 2000. Using meta-smoothing to estimate dose-response trends across multiple studies, with application to air pollution and daily death. Epidemiology 11 (6), 666–672. ISSN 10443983. http://www.jstor.org/stable/3703820

Servén, D., Brummitt, C., 2018. pygam: generalized additive models in python. Doi: 10.5281/zenodo.1208723.

Sørensen, Ø., Brandmaier, A. M., Mowinckel, A. M., 2020. metagam: meta-analysis of generalized additive models. https://CRAN.R-project.org/package=metagam. R package version 0.1.2.

Stouffer, S.A., Suchman, E.A., DeVinney, L.C., Star, S.A., Williams, R.M.J., 1949. The American Soldier, vol 1: Adjustment During Army Life. Princeton University Press, Princeton.

Sung, Y.J., Schwander, K., Arnett, D.K., Kardia, S.L.R., Rankinen, T., Bouchard, C., Boerwinkle, E., Hunt, S.C., Rao, D.C., 2014. An empirical comparison of meta-analysis and mega-analysis of individual participant data for identifying gene-environment interactions. Genet. Epidemiol. 38 (4), 369–378. doi:10.1002/gepi.21800.

Sutton, A.J., Higgins, J.P.T., 2008. Recent developments in meta-analysis. Stat. Med. 27 (5), 625–650. doi:10.1002/sim.2934.

Taylor, J.R., Williams, N., Cusack, R., Auer, T., Shafto, M.A., Dixon, M., Tyler, L.K., Cam-CAN, Henson, R.N., 2017. The cambridge centre for ageing and neuroscience (cam-CAN) data repository: structural and functional MRI, MEG, and cognitive data from a cross-sectional adult lifespan sample. NeuroImage 144, 262–269. doi:10.1016/j.neuroimage.2015.09.018.

Team, R.C., 2019. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/

Thompson, P.M., Andreassen, O.A., Arias-Vasquez, A., Bearden, C.E., Boedhoe, P.S., Brouwer, R.M., Buckner, R.L., Buitelaar, J.K., Bulayeva, K.B., Cannon, D.M., Cohen, R.A., Conrod, P.J., Dale, A.M., Deary, I.J., Dennis, E.L., de Reus, M.A., Desrivieres, S., Dima, D., Donohoe, G., Fisher, S.E., Fouche, J.-P., Francks, C., Frangou, S., Franke, B., Ganjgahi, H., Garavan, H., Glahn, D.C., Grabe, H.J., Guadalupe, T., Gutman, B.A., Hashimoto, R., Hibar, D.P., Holland, D., Hoogman, M., Pol, H., Hulshoff, E., Hosten, N., Jahanshad, N., Kelly, S., Kochunov, P., Kremen, W.S., Lee, P.H., Mackey, S., Martin, N.G., Mazoyer, B., McDonald, C., Medland, S.E., Morey, R.A., Nichols, T.E., Paus, T., Pausova, Z., Schmaal, L., Schumann, G., Shen, L., Sisodiya, S.M., Smit, D.J.A., Smoller, J.W., Stein, D.J., Stein, J.L., Toro, R., Turner, J.A., van den, H., Martijn, P., van den, H., Odile, L., van Erp, Theo, G.M., van Rooij, D., Veltman, D.J., Walter, H., Wang, Y., Wardlaw, J.M., Whelan, C.D., Wright, M.J., Ye, J., 2017. ENIGMA and the individual: predicting factors that affect the brain in 35 countries worldwide. NeuroImage 145, 389–408. doi:10.1016/j.neuroimage.2015.11.057.

Thompson, P.M., Stein, J.L., Medland, S.E., Hibar, D.P., Vasquez, A.A., Renteria, M.E., Toro, R., Jahanshad, N., Schumann, G., Franke, B., Wright, M.J., Martin, N.G., Agartz, I., Alda, M., Alhusaini, S., Almasy, L., Almeida, J., Alpert, K., Andreasen, N.C., Andreassen, O.A., Apostolova, L.G., Appel, K., Armstrong, N.J., Aribisala, B., Bastin, M.E., Bauer, M., Bearden, C.E., Bergmann, O., Binder, E.B., Blangero, J., Bockholt, H.J., Bøen, E., Bois, C., Boomsma, D.I., Booth, T., Bowman, I.J., Bralten, J., Brouwer, R.M., Brunner, H.G., Brohawn, D.G., Buckner, R.L., Buitelaar, J., Bulayeva, K., Bustillo, J.R., Calhoun, V.D., Cannon, D.M., Cantor, R.M., Carless, M.A., Caseras, X., Cavalleri, G.L., Chakravarty, M.M., Chang, K.D., Ching, C.R.K., Christoforou, A., Cichon, S., Clark, V.P., Conrod, P., Coppola, G., Crespo-Facorro, B., Curran, J.E., Czisch, M., Deary, I.J., de Geus, E.J.C., den Braber, A., Delvecchio, G., Depondt, C., de Haan, L., de Zubicaray, G.I., Dima, D., Dimitrova, R., Djurovic, S., Dong, H., Donohoe, G., Duggirala, R., Dyer, T.D., Ehrlich, S., Ekman, C.J., Elvsåshagen, T., Emsell, L., Erk, S., Espeseth, T., Fagerness, J., Fears, S., Fedko, I., Fernández, G., Fisher, S.E., Foroud, T., Fox, P.T., Francks, C., Frangou, S., Frey, E.M., Frodl, T., Frouin, V., Garavan, H., Giddaluru, S., Glahn, D.C., Godlewska, B., Goldstein, R.Z., Gollub, R.L., Grabe, H.J., Grimm, O., Gruber, O., Guadalupe, T., Gur, R.E., Gur, R.C., Göring, H.H.H., Hagenaars, S., Hajek, T., Hall, G.B., Hall, J., Hardy, J., Hartman, C.A., Hass, J., Hatton, S.N., Haukvik, U.K., Hegenscheid, K., Heinz, A., Hickie, I.B., Ho, B.-C., Hoehn, D., Hoekstra, P.J., Hollinshead, M., Holmes, A.J., Homuth, G., Hoogman, M., Hong, L.E., Hosten, N., Hottenga, J.-J., Pol, H., Hulshoff, E., Hwang, K.S., Jack, C.R., Jenkinson, M., Johnston, C., Jönsson, E.G., Kahn, R.S., Kasperaviciute, D., Kelly, S., Kim, S., Kochunov, P., Koenders, L., Krämer, B., Kwok, J.B.J., Lagopoulos, J., Laje, G., Landen, M., Landman, B.A., Lauriello, J., Lawrie, S.M., Lee, P.H., Hellard, S.L., Lemaître, H., Leonardo, C.D., shan Li, C., Liberg, B., Liewald, D.C., Liu, X., Lopez, L.M., Loth, E., Lourdusamy, A., Luciano, M., Macciardi, F., Machielsen, M.W.J., MacQueen, G.M., Malt, U.F., Mandl, R., Manoach, D.S., Martinot, J.-L., Matarin, M., Mather, K.A., Mattheisen, M., Mattingsdal, M., Meyer-Lindenberg, A., McDonald, C., McIntosh, A.M., McMahon, F.J., McMahon, K.L., Meisenzahl, E., Melle, I., Milaneschi, Y., Mohnke, S., Montgomery, G.W., Morris, D.W., Moses, E.K., Mueller, B.A., Maniega, S.M.n., Mühleisen, T.W., Müller-Myhsok, B., Mwangi, B., Nauck, M., Nho, K., Nichols, T.E., Nilsson, L.-G., Nugent, A.C., Nyberg, L., Olvera, R.L., Oosterlaan, J., Ophoff, R.A., Pandolfo, M., Papalampropoulou-Tsiridou, M., Papmeyer, M., Paus, T., Pausova, Z., Pearlson, G.D., Penninx, B.W., Peterson, C.P., Pfennig, A., Phillips, M., Pike, G.B., Poline, J.-B., Potkin, S.G., Pütz, B., Ramasamy, A., Rasmussen, J., Rietschel, M., Rijpkema, M., Risacher, S.L., Roffman, J.L., Roiz-Santiañez, R., Romanczuk-Seiferth, N., Rose, E.J., Royle, N.A., Rujescu, D., Ryten, M., Sachdev, P.S., Salami, A., Satterthwaite, T.D., Savitz, J., Saykin, A.J., Scanlon, C., Schmaal, L., Schnack, H.G., Schork, A.J., Schulz, S.C., Schür, R., Seidman, L., Shen, L., Shoemaker, J.M., Simmons, A., Sisodiya, S.M., Smith, C., Smoller, J.W., Soares, J.C., Sponheim, S.R., Sprooten, E., Starr, J.M., Steen, V.M., Strakowski, S., Strike, L., Sussmann, J., Sämann, P.G., Teumer, A., Toga, A.W., Tordesillas-Gutierrez, D., Trabzuni, D., Trost, S., Turner, J., den Heuvel, M.V., van der, W., Nic, J., van Eijk, K., van Erp, Theo, G.M., van Haren, Neeltje, E.M., vant Ent, D., van Tol, M.-J., Hernández, M., Valdés, C., Veltman, D.J., Versace, A., Völzke, H., Walker, R., Walter, H., Wang, L., Wardlaw, J.M., Weale, M.E., Weiner, M.W., Wen, W., Westlye, L.T., Whalley, H.C., Whelan, C.D., White, T., Winkler, A.M., Wittfeld, K., Woldehawariat, G., Wolf, C., Zilles, D., Zwiers, M.P., Thalamuthu, A., Schofield, P.R., Freimer, N.B., Lawrence, N.S., Drevets, W., 2014. The ENIGMA consortium: large-scale collaborative analyses of neuroimaging and genetic data. Brain Imaging Behav. doi:10.1007/s11682-013-9269-5.

Tippet, L., 1931. The Methods of Statistics. Williams and Norgate, London.

Veroniki, A.A., Jackson, D., Viechtbauer, W., Bender, R., Bowden, J., Knapp, G., Kuss, O., Higgins, J.P.T., Langan, D., Salanti, G., 2016. Methods to estimate the between-study variance and its uncertainty in meta-analysis. Res. Synth. Methods 7 (1), 55–79. doi:10.1002/jrsm.1164.

Viechtbauer, W., 2005. Bias and efficiency of meta-analytic variance estimators in the random-effects model. J. Educ. Behav. Stat. 30 (3), 261–293. doi:10.3102/10769986030003261.

Viechtbauer, W., 2010. Conducting meta-analyses in R with the metafor package. Journal of Statistical Software, Articles 36 (3), 1–48. doi:10.18637/jss.v036.i03. ISSN 1548-7660

Viechtbauer, W., López-López, J.A., Sánchez-Meca, J., Marín-Martínez, F., 2015. A comparison of procedures to test for moderators in mixed-effects meta-regression models. Psychol. Methods 20 (3), 360–374. doi:10.1037/met0000023.

Walhovd, K.B., Fjell, A.M., Sørensen, Ø., Mowinckel, A.M., Reinbold, C.S., Idland, A.-V., Watne, L.O., Franke, A., Dobricic, V., Kilpert, F., Bertram, L., Wang, Y., 2019. Genetic risk for Alzheimer's disease predicts hippocampal volume through the lifespan. bioRxiv doi:10.1101/711689.

Walhovd, K.B., Fjell, A.M., Westerhausen, R., Nyberg, L., Ebmeier, K.P., Lindenberger, U., Bartres-Faz, D., Baare, W.F.C., Siebner, H.R., Henson, R., 2018. Healthy minds 0-100 years: optimising the use of European brain imaging cohorts ("Lifebrain"). Eur. Psychiatry 47 (76–77). doi:10.1016/j.eurpsy.2017.10.005.

Walhovd, K.B., Krogsrud, S.K., Amlien, I.K., Bartsch, H., Bjørnerud, A., Due-Tønnessen, P., Grydeland, H., Hagler, D.J., Håberg, A.K., Kremen, W.S., Ferschmann, L., Nyberg, L., Panizzon, M.S., Rohani, D.A., Skranes, J., Storsve, A.B., Sølsnes, A.E., Tamnes, C.K., Thompson, W.K., Reuter, C., Dale, A.M., Fjell, A.M., 2016. Neurodevelopmental origins of lifespan changes in brain and cognition. Proc. Natl. Acad. Sci. 113 (33), 9357–9362. doi:10.1073/pnas.1524259113.

Wickham, H., 2016. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag, New York. ISBN 978-3-319-24277-4. URL https://ggplot2.tidyverse.org

Wilkinson, B., 1951. A statistical consideration in psychological research. Psychol. Bull. 48 (2), 156–158. doi:10.1037/h0059111.

Wood, S., Scheipl, F., 2017. gamm4: Generalized additive mixed models using 'mgcv' and 'lme4'. URL https://CRAN.R-project.org/package=gamm4. R package version 0.2-5.

Wood, S.N., 2012. On p-values for smooth components of an extended generalized additive model. Biometrika 100 (1), 221–228. doi:10.1093/biomet/ass048. 10

Wood, S.N., 2017. Generalized Additive Models: an Introduction with R, seccond ed. Chapman and Hall/CRC.

Zaykin, D.V., 2011. Optimally weighted z-test is a powerful method for combining probabilities in meta-analysis. J. Evol. Biol. 24 (8), 1836–1841. doi:10.1111/j.1420-9101.2011.02297.x.