DEPARTMENT OF
STATISTICS

Advances in statistical methods for large-scale binary-valued neuroimaging data

A thesis submitted for the degree
Doctor of Philosophy

November 2021

Petya Kindalova
St Peter's College

Department of Statistics
University of Oxford

Contents

Acknowledgements

I would like to thank my supervisors Professor Thomas Nichols and Professor Ioannis Kosmidis for their endless support, guidance and advice throughout my DPhil. I will be forever thankful to both of them for our regular meetings and the thorough feedback they provided to every written output of our research. They gave me hope and inspiration at difficult times and for that I would always be extremely grateful. I would also like to thank the Department of Statistics at the University of Oxford and EPSRC for the opportunity to have had a fully funded position. Thanks should also be extended to the admin staff at the department who have always brightened my day with their smiles and positivity.

I must extend my gratitude to two academics whom I met during my Integrated Masters studies at the Department of Mathematics and Statistics at the University of Glasgow. First, special thanks to Professor Duncan Lee who passionately supported my idea to pursue doctoral studies. Second, I will always be immensely grateful to Dr Ludger Evers who encouraged me to apply to the University of Oxford and let me have dreams I never thought could become reality.

I would also like to thank all my friends who supported me throughout the roller-coaster of emotions over the last five years. Namely, I would like to thank my friend Aleksandar Kolev, who did a PhD at UCL at roughly the same time as me, since sharing every step of the way with a close friend gave me the extra courage and confidence I needed.

Last but not least, I would like to thank my partner Krasimir Kremakov and my family. I couldn't have gone through this journey without their endless support, cheerful messages, and continuous belief in my success.

Abstract

White matter lesions are common in the ageing brain and their size, location, and evolution have been shown to be informative of diagnosis, treatment, or prevention of neurological conditions such as multiple sclerosis. The work presented in this thesis explores the use of voxel-based approaches for modelling binary lesion data obtained from brain Magnetic Resonance Imaging scans. We seek to develop methods that are computationally efficient and stable for massive datasets with very low lesion incidence, and demonstrate the value of these methods with a real dataset to disentangle contributing risk factors of lesion incidence.

Our contributions are spread across three main chapters including two published articles and one preprint, and they could be summarised as

Chapter 2 Kindalova et al. (2021a) explore whether the potential gains in estimator accuracy justify the use of a more computationally intensive spatial modelling approach as opposed to a mass-univariate approach to modelling voxel-wise binary lesion data. A method comparison of three cross-sectional lesion mapping approaches is facilitated through the development of a novel simulation framework of artificial lesion masks, which mimics features of real lesion masks.

Chapter 3 Veldsman et al. (2020) use data on 13,680 healthy ageing UK Biobank participants at one time point to explore the effects of the individual cerebrovascular risk factors (e.g. waist-to-hip ratio and smoking) on lesion load and on lesion probability, which has been an obstacle in the literature so far due to the dominating effect of hypertension and the presence of comorbidities.

Chapter 4 Kindalova et al. (2021b) adopt a generalized estimating equations approach to modelling longitudinal binary-valued outcomes. By adding a Jeffreys-prior penalty to log-link generalized estimating equations for relative risk regression, finiteness of the estimates along with superior convergence rate are demonstrated in an extensive simulation study as well as in a UK Biobank application on 1,578 participants with data from 2 visits.

Chapter 1

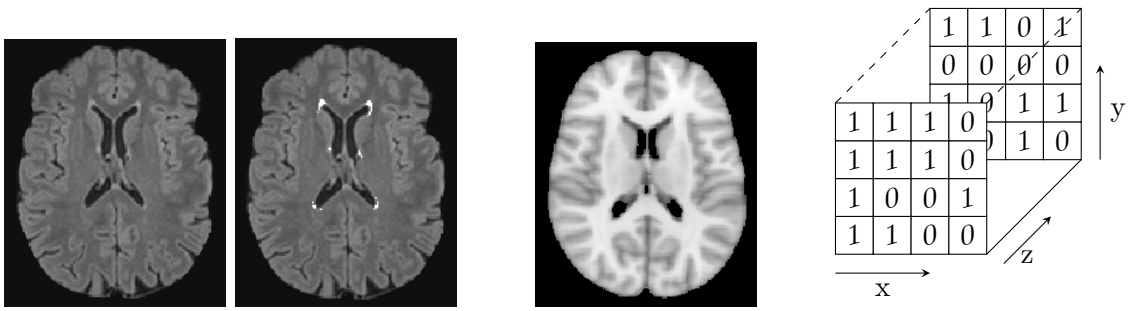
Opening remarks

1.1 Introduction

White matter hyperintensities (WMHs) of presumed vascular origin (Wardlaw et al., 2013) are common in the ageing brain and are associated with progressive cognitive impairment (Debette and Markus, 2010). Early pathology studies suggest that WMHs indicate brain tissue damage due to myelin loss or axonal degeneration (Fazekas et al., 1993). However, modern technology allows for the detection of subtle white matter changes through Magnetic Resonance Imaging (MRI). Structural MRI scans come in many variants, but two of the most common are T1-weighted and T2-weighted. T1-weighted images show white matter as most intense, gray matter darker and cerebral spinal fluid darkest. T2-weighted images show cerebral spinal fluid as most intense, gray matter darker and white matter darkest. T1-weighted images are mostly used to discriminate gray from white matter, and T2-weighted images - to characterize brain structure. For each participant scanned in a study, the structural MRI data take the form of a 3D image, i.e. a 3D snapshot of the entire brain in high resolution.

WMHs, broadly referred to as lesions, are evident as hyperintensities on T2-weighted, fluid attenuated inversion recovery (FLAIR) images, and the degree of signal abnormality (or "whiteness") reflects the amount of tissue damage. For an illustration of a T2 FLAIR image, see the axial slice (horizontal plane) in Figure 1.1(a), where the whiter areas capping the ventricles indicate the presence of lesions.

Pre-processing of the raw MRI images, such as the example in Figure 1.1(a), is required to enable any statistical analysis at the group level. The pre-processing generally



(a) T2 FLAIR (b) Segmentation (c) Mapping (d) Binary matrix

Figure 1.1: From brain MRI raw data to binary matrices. (a) One axial slice of a T2 FLAIR image in native space, (b) Lesion segmentation step, with voxels highlighted in white being identified as lesions, (c) An axial slice of the MNI standard anatomical T1 image shown for reference, (d) A schematic illustration of 3D binary lesion data in matrix form ready for statistical analysis.

includes (i) brain extraction (removal of non-brain tissues from the image), (ii) spatial normalisation, linear alignment (Jenkinson et al., 2002a) to a standard brain template, e.g. the Montreal Neurological Institute (MNI) brain atlas, and non-linear warping to this template (Andersson et al., 2007) to maximize correspondence across individuals in light of significant cross-subject variation in brain structure, and (iii) segmentation of lesions, creating a binary mask that marks voxels (volumetric pixels) as being a part of a lesion (1) or not (0).

Whether created manually, or by an automated segmentation procedure such as BIANCA (Griffanti et al., 2016), lesion maps are in the "subject space" (also known as native space), i.e. each map has unique dimensions and the 3D coordinates have no particular correspondence to any other subjects; see Figure 1.1(b). Then the estimated spatial normalisation parameters can be applied to the binary lesion maps to map them to a common space with the standard anatomical T1 image shown in Figure 1.1(c) for reference. The resulting images are binary matrices (Figure 1.1(d)) representing the brain of each individual in a common space, where 1's mark the exact locations of the lesions.

1.1.1 Motivation

In brain imaging, there are multiple types of lesions depending on the factor influencing their development. For example, multiple sclerosis (MS), the most common neurological disorder in young adults, is an autoimmune disease of the central nervous system. MS

causes inflammation that damages myelin further resulting in lesions throughout the brain. The dynamics of lesion development in space and time form the MS diagnostic criteria (Polman et al., 2011) and lesion location is shown to be a prognostic factor of disease progression (Dalton et al., 2012). Simply focusing on the total lesion burden (e.g. volume and number of lesions) could result in limited findings about MS outcomes, also called the clinoco-radiological paradox (Barkhof, 2002). Another example from a recent study by Ghaznawi et al. (2021) shows that geometric and topographic features of lesions could reveal increased risk of stroke. Also, cognitive decline has been shown to be associated with periventricular WMHs (Prins et al., 2005; Godin et al., 2010). Irrespective of the type of lesions, clearly lesion location, size and evolution are important prognostic factors of MS disease progression, cognitive decline in healthy ageing, stroke and dementia, and many more.

The binary lesion images could be used as the prognostic factors at the individual level instead of aggregating the data up to brain region level or to the absolute total lesion volume. Adopting a voxel-based approach, known as *mass-univariate*, means that a model is fitted at each voxel independently. Such analysis can reveal patterns of current/new lesions and their association with risk factors, or expose the localized effect of a treatment.

1.1.2 Thesis outline

The remainder of this chapter provides background to the research completed and briefly outlines the main contributions of the thesis. There are three main chapters, two of which are motivated by cross-sectional binary-valued neuroimaging data and one by longitudinal.

In Chapter 2, we develop a novel simulation framework of artificial lesion maps in order to compare three approaches for modelling binary lesion masks, and we then apply one of the methods to a large-scale neuroimaging data set to estimate the effect of systolic blood pressure on lesion probability. The chapter is published in *NeuroImage* (Kindalova et al., 2021a).

Chapter 3 includes a large sample study exploring the association between WMHs and cerebrovascular risk factors such as diabetes and smoking, both in terms of total lesion load and voxel-wise, as well as exploring the impact of WMHs on cognition. The

chapter is published in *NeuroImage: Clinical* (Veldsman et al., 2020) and involves a joint first author Michele Veldsman and Petya Kindalova.

In Chapter 4, we adopt a log-link generalized estimating equations approach to modelling repeated measures binary data to account for the potential correlation between visits within each subject and to ensure the direct interpretability of the estimated effects as relative risks. We propose practical adjustments to the generalized estimating equations approach to ensure stable estimates when dealing with either low, or high incidence events. The proposed method's performance is demonstrated through an extensive simulation study along with a real data neuroimaging application. The preprint (Kindalova et al., 2021b) is under consideration for publication.

Chapter 5 contains a summary of the completed research along with future directions.

1.2 Cross-sectional modelling of binary lesion masks

Lesion location has proven to be important since the first visual scales quantifying lesion burden (Fazekas et al., 1987), where the distinction is made between periventricular and deep white matter regions. Periventricular and deep WMHs have been shown to have different relationships with cerebrovascular risk factors (Griffanti et al., 2018), and potentially different associations with cognition (Mortamais et al., 2013). Beyond this classification, there may be important information hidden in the spatial distribution of lesions, which could be revealed by voxel-based analysis of the binary lesion masks.

1.2.1 Overview

Given mapped to a common space binary lesion masks for a number of subjects at one time point (cross-sectional data), lesion probability maps (LPMs) that show the empirical lesion rate at each voxel can be created. Some studies look for significant differences between LPMs, where subjects are grouped by a visual rating scale (Enzinger et al., 2006), MS disease subtypes (Kincses et al., 2011; Filli et al., 2012), level of cognitive impairment in MS patients (Rossi et al., 2012), etc. All those studies are based on a linear approach (general linear model) combined with permutation-based inference (Nichols and Holmes, 2002). Those modelling approaches are known as *mass-univariate* since they fit a model at each voxel independently and do not explicitly

account for the local spatial dependence in the brain. However, using standard linear regression methods with lesion incidence as the outcome variable could be considered ill-advised since it ignores the binary distribution of the data.

1.2.1.1 Generalized linear model

To respect the binary nature of the lesion data (binary matrices as in Figure 1.1(d)), logistic or probit functions could be directly fitted to the lesion probability through maximum likelihood estimation. Such generalized linear models (GLMs) have been used in the literature, for example Holland et al. (2008) performed logistic regression and found no consistent difference in the WMH distributions across patients with Alzheimer disease/mild cognitive impairment or amyloid angiopathy vs. healthy ageing controls in a sample of 102 individuals. Other voxel-wise analyses based on a logistic regression approach include the work of Rostrup et al. (2012) and Lampe et al. (2019b). Rostrup et al. (2012) mapped the lesion probability as a function of clinical risk factors such as hypertension and alcohol consumption in a sample of 605 participants. Lampe et al. (2019b) tested the association between lesion topography and obesity in a sample of 1,825 individuals. However, the limitations of logistic regression arising due to small sample size and/or low lesion incidence have not been discussed in any of the works mentioned above, as far as we know. The two potential issues we discuss below are 'data separation' and finite-sample bias, which have been investigated in depth for binary-response models.

Each subject i out of N comes with a binary lesion mask in common space $Y_i \subset \mathbb{R}^3$, where we consider M volumetric pixels as a discretization of the brain as visualized in Figure 1.1(d). $Y_i(s_j)$ represents a Bernoulli random variable with lesion probability $p_i(s_j)$ for subject i at voxel s_j, $j = 1, \ldots, M$. A generalized linear model can be written as

$$[Y_i(s_j) \mid p_i(s_j)] \sim \text{Bernoulli}(p_i(s_j)) \qquad \text{(stochastic component)} \tag{1.1}$$

$$g(p_i(s_j)) = \eta_i(s_j) \qquad \text{(link function)} \tag{1.2}$$

$$\eta_i(s_j) = \boldsymbol{x}_i^\top \boldsymbol{\beta}(s_j) \qquad \text{(deterministic component)}\,, \tag{1.3}$$

where $\boldsymbol{x}_i$ is a P-vector of subject-specific covariates (e.g. age or systolic blood pressure) for subject i and $\boldsymbol{\beta}(s_j)$ is a P-vector of parameters at each voxel s_j. The function $g(.)$

is assumed to be monotonic and it links the expectation of the stochastic outcome to the deterministic part. Note that under the mass-univariate modelling framework we assume that $Y_1(s_1), \ldots, Y_1(s_M), \ldots, Y_N(s_1), \ldots, Y_N(s_M)$ are independent random variables given $p_i(s_j)$. To obtain the maximum likelihood estimators (MLEs) $\hat{\boldsymbol{\beta}}(s_j)$, we need to maximize the log-likelihood

$$l(\boldsymbol{\beta}(s_j)) = \sum_{i=1}^{N} \log f(y_i(s_j) \mid \boldsymbol{x}_i;\ \boldsymbol{\beta}(s_j))\,,$$

where $y_i(s_j)$ is a realization of the random variable $Y_i(s_j)$ and $f(y_i(s_j) \mid \boldsymbol{x}_i;\ \boldsymbol{\beta}(s_j))$ is the associated probability mass function. So we need to solve the score equations of the form

$$U(\boldsymbol{\beta}(s_j)) = \left(\frac{\partial l(\boldsymbol{\beta}(s_j))}{\partial \beta_1(s_j)}, \ldots, \frac{\partial l(\boldsymbol{\beta}(s_j))}{\partial \beta_P(s_j)}\right)^\top = \mathbf{0}\,,$$

where $U(\boldsymbol{\beta}(s_j))$ is the score P-vector at voxel s_j. The estimated standard errors of the MLEs are calculated as $\sqrt{\{I(\boldsymbol{\beta}(s_j))^{-1}\}_{pp}}$ using the square roots of the diagonal elements of the inverse of the Fisher information matrix $I(\boldsymbol{\beta}(s_j))$, which is defined as the expected value of the negative Hessian matrix of the log-likelihood.

1.2.1.2 Data separation and finite-sample bias

Maximum likelihood (ML) estimation of a GLM typically requires fitting an iterative procedure, such as iteratively reweighted least squares (IRLS) (Green, 1984). When data separation occurs, the iterative optimization fails to converge and the resulting ML estimate has one or more infinite-valued components. In logistic regression, data separation occurs when a covariate (or a linear combination of covariates) perfectly predicts the outcome; for a formal definition see Albert and Anderson (1984). Awareness of this issue is crucial in clinical applications, where dealing with rare outcomes or small sample sizes is common, since software packages deal with separation differently and might as well report a finite estimate due to early stopping of the iterative algorithm.

ML is the most commonly used estimation method due to the desirable asymptotic properties of the ML estimator. However, the ML estimator can suffer from considerable bias in the small sample size setting due to finite sample properties being far from what is expected asymptotically, further having severe impact on inference. On a side

note, the bias of an estimator is defined as the difference of its expected value from the parameter to be estimated $B(\boldsymbol{\beta}) = \mathrm{E}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})$, and the bias being small means that if we repeat the experiment infinitely many times, the average of the resulting estimates would get close to the true value.

There are numerous methods to reduce bias in parameter estimation (for a detailed review, see Kosmidis 2014), broadly divided into explicit and implicit methods. Explicit methods rely on a one-step procedure, where the bias is estimated and subtracted from the MLE, examples being jackknife (Quenouille, 1956), bootstrap (Hall and Martin, 1988) and asymptotic bias correction approximations of the bias function (Cordeiro and McCullagh, 1991). However, those explicit methods rely on the existence of the MLE, i.e. in the case of data separation, the bias-corrected estimate inherits the instabilities of the MLE.

One implicit method introduced by Firth (1993) addresses this issue for logistic regression by correcting for the first-order bias in maximum likelihood. Note that the first-order bias is the first-order term in the asymptotic expansion of the bias of the ML estimator:

$$B(\boldsymbol{\beta}) = \frac{b_1(\boldsymbol{\beta})}{N} + \frac{b_2(\boldsymbol{\beta})}{N^2} + O(N^{-3}),$$

for an appropriate set of functions $b_1(\boldsymbol{\beta}), b_2(\boldsymbol{\beta}), \ldots$, which are $O(1)$ as $N \to \infty$. This approach was further developed for exponential family GLMs (Kosmidis and Firth, 2009; Kosmidis et al., 2020). Those implicit methods do not depend on the MLE. Instead, a penalty $A(\boldsymbol{\beta})$ is added to the score functions in order to get an estimator with asymptotically smaller bias:

$$U^*(\boldsymbol{\beta}) = U(\boldsymbol{\beta}) \; + \; A(\boldsymbol{\beta}) = \mathbf{0},$$

where $A(\boldsymbol{\beta})$ is a P-vector penalty term expressed as the product $-b_1(\boldsymbol{\beta})I(\boldsymbol{\beta})/N$. The general form of the first-order bias is derived by Kosmidis and Firth (2009) for exponential family non-linear models. Note that this approach guarantees finite-valued estimates, as shown for logistic regression by Kosmidis and Firth (2021), and achieves first-order unbiased estimates, thus addressing both limitations of the MLE. Iterative optimization is required to obtain the bias-reduced estimates and an example is the iter-

ative first-order bias adjustments procedure (Kosmidis and Firth, 2010); note that this adjustment adds very little to the computational cost of obtaining the MLE. However, this approach to bias reduction is only possible when b_1/N, U and I are in closed form; for example generalized linear mixed models have intractable first-order bias term. In addition, note that caution needs to be taken when adjusting for bias in parametric estimation since not all approaches keep the invariance properties of the MLE under reparameterization.

1.2.1.3 Spatial dependence

A mass-univariate voxel-based approach to modelling binary lesions masks does not explicitly take into account the spatial dependence between nearby voxels, and amounts to a massive multiple testing problem. The simplest approach to inference at the brain level is to apply a fixed threshold to the test statistic image (e.g. z-statistics resulting from voxel-wise logistic regression) to give inference at some significance level, e.g. $\alpha = 0.001$, with no formal control for multiple testing. Alternatively, the familywise error (FWE) rate α_{FWE}, the chance of one or more false positive voxels, can be controlled with the Bonferroni method, using a significance level of α_{FWE}/M, where M is the number of voxels in the analysis. However, the Bonferroni correction is very conservative when there is dependence among the tests, as is the case for MRI data like we consider. An alternative is to control the false discovery rate (FDR) α_{FDR}, the expected proportion of false positives among detected voxels. The Benjamini-Hochberg method (Benjamini and Hochberg, 1995) for controlling FDR is valid under both independence and positive dependence (Genovese et al., 2002), making it ideally suited for imaging data. Alternatively, a more computationally demanding permutation-based approach could be used for FWE control, where the null distribution of the maximum test statistic (e.g. t-statistic) is built over permutations of the input data and by testing the achieved test statistic against it provides corrected p-values (Nichols and Holmes, 2002); note, however, special adaptations are needed for binary data (Hemerik et al., 2020).

While FDR and permutation adapt to spatial dependence, alternative approaches can directly use the spatial structure of the signal as a basis of inference. A cluster size test uses a primary threshold to create spatial contiguous suprathreshold regions (known as clusters), and then assesses statistical significance on the basis of cluster size. While

the parametric null distributions for cluster size are available for Gaussian data (Friston et al., 1994), permutation would be required for discrete data. Another approach to directly incorporate spatial information is the threshold-free cluster enhancement (TFCE) method introduced by Smith and Nichols (2009). TFCE amounts to running many cluster size tests with primary thresholds, defining a family of cluster test results, which are then integrated into a single TFCE map. There is no parametric null distribution available for TFCE, so permutation is required to obtain FWE-corrected p-values. TFCE is directly implemented in the *randomise* permutation-based inference FSL tool, which makes it easy to use and a preferred choice in clinical applications (Rostrup et al., 2012; Lampe et al., 2019b).

Alternatively, we can directly include the local spatial dependence in the brain in the modelling stage. An example is a Bayesian spatial generalized linear mixed model (BSGLMM) introduced by Ge et al. (2014). The model is based on the GLM defined in Equations (1.1), (1.2), and (1.3), with the difference that the deterministic components $\eta_i(s_j)$ are explicitly defined as functions of space. This is achieved through the inclusion of spatially varying coefficients $\boldsymbol{\beta}(s_j)$, which are latent spatial processes modelled jointly using a multivariate pairwise difference prior model of the form

$$[\boldsymbol{\beta}(s_j) \mid \boldsymbol{\beta}(-s_j), \boldsymbol{\Sigma}] \sim \text{MVN}\left[\frac{\sum_{s \in S_k} \boldsymbol{\beta}(s)}{N(s_j)}, \frac{\boldsymbol{\Sigma}}{N(s_j)}\right],$$

where $\boldsymbol{\beta}^T = [\boldsymbol{\beta}^T(s_1), \ldots, \boldsymbol{\beta}^T(s_M)]$ is a $P \times M$ column vector and $\boldsymbol{\beta}(-s_j)$ includes all the elements of $\boldsymbol{\beta}$ except the coefficients at voxel s_j. Also, S_k denotes the set of neighboring locations for voxel s_j, where voxels are assumed to be neighbours if they share a face, and the cardinality of this set is denoted as $N(s_j)$. $\boldsymbol{\Sigma}$ is a $P \times P$ symmetric positive definite matrix.

1.2.2 Contributions

Before outlining the method developments and novel data applications included in this thesis, we must introduce briefly the large-scale real data set used to support our research. The UK Biobank[1] (UKB) is a large prospective study which collected extraordinarily rich baseline data and samples for over 500,000 participants between 2006 and

[1] `http://www.ukbiobank.ac.uk/`

2010. The data already collected and to be collected in the longitudinal follow-up assessments include phenotypic and genotypic details collected through questionnaires, physical measures, multimodal imaging, genome-wide genotyping, etc. (Sudlow et al., 2015). The population level imaging data used in our research is part of the imaging extension of the UKB study funded in 2016, which includes MRI brain, heart and abdomen scans of 100,000 participants from the existing cohort. The brain imaging data description and example analytic approaches are presented after the first 5,000 participants' data release (Miller et al., 2016) and the automated processing and quality control pipeline for the first 10,000 imaged subjects is described by Alfaro-Almagro et al. (2018). After preprocessing of the T2-weighted images[2], lesion masks in common space are available for about 20,000 subjects from the first imaging visit. Our work is entirely based on voxel size of $2 \times 2 \times 2$ mm^3, which implies the data to model consist of about 20,000 binary lesion masks of dimension $91 \times 109 \times 91$ voxels (i.e. the x, y, and z dimensions of binary matrix as in Figure 1.1(d)).

1.2.2.1 Method comparison with a novel simulation framework

Given that more biobank-scale data sets are available, the scalability and computational efficiency become increasingly important along with the accuracy of the modelling methods used for lesion mapping. The work presented in Chapter 2 is motivated by the lack of agreement in the literature on whether a mass-univariate voxel-based modelling approach is better for binary lesion data than a more computationally demanding spatial modelling approach that accounts for local spatial dependence. In particular, it is not clear whether the small sample size and typically low lesion incidence result in diverging or highly biased estimates when the default maximum likelihood estimation method is used. On the other hand, Ge et al. (2014) demonstrate the benefits of spatial regularization, but the authors use a limited simulation study of 2D images with lesion patterns not resembling realistic lesion masks. In addition, other simulation approaches introduce homogeneous lesion patterns (Chard et al., 2010) or smooth the simulated lesion masks (Sundaresan et al., 2019), both of which introduce stronger spatial dependencies.

Thus, to allow for a fair comparison between the alternative lesion mapping methods,

[2]The UKB T2-weighted protocol uses a FLAIR contrast with the 3D SPACE optimized readout, which shows strong contrast for WMHs.

we design a novel simulation framework that produces realistic binary lesion data that is calibrated to a given generalized linear model. The realism is achieved through matching real features of lesion masks (total lesion volume, average lesion size, and lesion count) across a reference data set and the simulated data sets. Briefly, the simulation steps are as follows

Step 1 *Learn parameters from reference data.* The reference data set consists of 13,680 healthy ageing individuals from the UK Biobank and estimated maps for intercept and age effects are obtained voxel-wise.

Step 2 *Construct simulation design.* Sample plausible age values for N^* subjects, and using the estimates from Step 1, create the linear predictor for each simulated lesion mask out of N^*.

Step 3 *Simulate smooth noise for linear predictor.* Simulate a zero-mean Gaussian Random Field (GRF) independently for each subject. By tuning the parameters of the GRF covariance, we match the desired lesion summaries (e.g. average lesion size) across age groups for the simulated data sets and the reference data set.

Step 4 *Generate binary lesion data.* Transform the sum of the linear predictor (Step 2) and the noise (Step 3) into a probability and threshold to get binary lesion masks.

Under this realistic simulation setting, we compared three lesion modelling approaches including a mass-univariate generalized linear model with either maximum likelihood estimates, or bias-reduced estimates, and a Bayesian spatial model (BSGLMM). A review of the regression approaches, details on the simulation framework and the formal method comparison along with a real data application are included in Chapter 2.

1.2.2.2 Cerebrovascular risk-related lesions

White matter hyperintensities (or lesions) are indicators of poor brain health (Wardlaw et al., 2015) and are associated with cerebrovascular burden. However, the relationship between lesion location and cerebrovascular risk factors is typically explored in clinical

populations of relatively small sample size and often the lesion data is aggregated up to regional level, e.g. brain structures, periventricular vs. deep white matter, or even up to the brain level. Also, the voxel-based analysis we are aware of typically uses maximum likelihood estimation without mentioning its potential instabilities for small sample size and low lesion incidence rates.

Taking advantage of the large data set of 13,680 healthy ageing subjects, a subset of the UKB data set, we performed the following thorough data analyses

- Aggregating the binary lesion data up to the brain level, we explored (i) the relationship between total lesion load and age by levels of cerebrovascular risk factors, e.g. diabetic or non-diabetic participants, using fitted smooth curves, and (ii) the dependence of total lesion load on cerebrovascular risk factors through multiple linear regression. The aim of this brain level analysis was to outline the significant contributors to the cerebrovascular burden related to the presence of lesions.
  The six cerebrovascular risk factors we focused on are hypertension, hypercholesterolemia, diabetes, smoking, waist-to-hip ratio, and apolipoprotein-E (APOE) $\varepsilon 4$ allele status as well as the composite score of six categorical variables representing those six risk factors.

- Voxel-wise analysis was performed to explore the marginal and joint effect of all six cerebrovascular risk factors on lesion probability, using mass-univariate regression modelling with mean bias-reduced estimates. The results of 40,0001 regressions were used to assess the relative importance of the risk factors and the spatial extent of their effect.

- The association between total lesion load and speed of processing (used as a cognition variable) was investigated through multiple linear regression analysis, and mediation analysis was used to determine whether any of the cerebrovascular risk factors explained their relationship.

The large-scale data set of 13,680 UKB participants allowed us to examine the effects of the individual cerebrovascular risk factors on lesion load and lesion probability, which has been an obstacle in the literature so far due to the dominating effect of hypertension, especially in small samples, when hypertension is often comorbid with obesity, diabetes and smoking.

Reproducible research

The `R` code to generate lesion masks using our simulation framework is available through the Open Science Framework website `https://osf.io/h7sxr/`, where a demonstration of the scalable parallel GLMs implementation, used in both cross-sectional data projects, can be found.

The spatial distribution maps produced as part of the Chapter 3 analysis are available at NeuroVault: `https://neurovault.org/collections/AZQTNVUF/`.

1.3 Longitudinal modelling of binary lesion masks

In 1986, it was suggested by Awad et al. (1986) that *"MRI lesions reflect "wear and tear" in the brain parenchyma which accompany aging and chronic cerebrovascular disease"*. At the time, the authors recognised the need for prospective follow-up MRI studies of healthy ageing adults to find the clinical significance of lesion progression. Over the last 35 years, lesion progression across multiple scans per subject over time was first manually classified by visual inspection into grades, e.g. the Austrian stroke prevention study (Schmidt et al., 1999, 2005) or the Rotterdam scan study (Van Dijk et al., 2008). More recently, the focus moved to changes in total white matter lesion load, e.g. the SMART-MR study (Kloppenborg et al., 2012).

In a small study of 51 healthy ageing volunteers scanned 3-years apart, age, sex, and cerebrovascular risk were not found to be significant predictors of WMH progression (Sachdev et al., 2007), where progression was defined as changes in WMH volumes of brain regions of interest, with the strongest predictor being the baseline level of WMH. In a bigger sample of 668 non-demented adults scanned 3 years apart, Van Dijk et al. (2008) found that higher age, female sex, hypertension, and smoking were associated with WMH progression as well as that cognitive function (information processing speed) was associated with periventricular WMH progression. With better quality MRI scans, bigger studies, longer follow-up times, and more regular repeated scans, the risk factors for lesion progression will get better understood, with lesion location further contributing to disentangling their relationship.

1.3.1 Overview

One approach used to model change over time is to use summaries such as binary values arising from manually grading the white matter lesions change from baseline to follow-up as absent or present for each subject and then use logistic regression to examine the association between vascular risk factors and lesion progression (Schmidt et al., 1999). Five years later, with one more time point available (baseline, 3 and 6 years) for the same cohort, Schmidt et al. (2005) used repeated measures analysis of variance (ANOVA) to assess the effect of time on white matter volume and generalized estimating equations to further evaluate the effect of change in white matter lesion volume on cognition. Voxel-based methods have been used in a similar manner to LPMs in cross-sectional settings, but grouping subjects by visit (Sachdev et al., 2007). Other than that, we are not aware of voxel-wise modelling of repeated binary lesion maps, which might be due to the small sample size of longitudinal studies so far.

1.3.1.1 Longitudinal data modelling

Longitudinal studies consist of repeated measurements over time, where some form of natural clustering is present, such as repeated brain scans of the same subject or weekly water samples from a lake, where the individual or the lake are considered as the clusters. The resulting data could be correlated within each cluster and the correlation should be modelled to achieve valid inference. There are a variety of approaches for modelling longitudinal data and the main considerations around the choice of an appropriate model depend on whether the repeated measures outcome is continuous or categorical, what explanatory variables we would like to adjust for, how the correlation of the outcomes is accounted for in the model, and whether the modelling assumptions are reasonable for the data at hand.

The data collected along with the outcome of interest could be time-varying (e.g. weekly temperature of the water of the lake) or time-invariant (e.g. the altitude of the lake) and depending on the modelling approach we select, we might not be able to include time-varying covariates. For continuous outcomes, the simplest modelling approach with limited covariate adjustment would be repeated measures analysis of variance, where time-varying covariates cannot be included in the model but only one categorical

variable (for example, time, as was done by Schmidt et al. 2005), and the continuous response is assumed to be normally distributed. We can also do paired t-tests to compare a continuous outcome between two time points, but apart from the assumption that the paired differences are normally distributed, we cannot really account for any explanatory variables at all. In clinical applications, adjusting for demographic and lifestyle factors is typically used, so the methods mentioned above might be of limited use. Since the outcome of interest in our work is binary, we would focus on methods for binary data modelling for the rest of this section.

When modelling clustered binary outcomes, a naive modelling approach could be to ignore any potential correlation and resort to using generalized linear models (GLMs) under an independence assumption. This approach provides consistent estimates but incorrect standard errors (Zeger and Liang, 1992), which may lead to incorrect conclusions; for such an example, see the Cannon et al. (2001) analysis of a childhood health intervention data collected in Brazil. Another idea often used in econometrics is to account for the between-cluster variability through the inclusion of cluster-specific intercepts to the GLM (Lancaster, 2000). Adding an intercept per cluster means (i) we can no longer estimate the effect of time-invariant covariates, (ii) for large data sets, the incidental parameter problem might occur; incidental parameters here are the cluster-specific intercepts and they are treated as nuisance parameters whose number increases as the number of clusters increases. The 'problem' for binary data models, as introduced by Neyman and Scott (1948) and discussed thoroughly by Lancaster (2000), is the inconsistency of the maximum likelihood estimates of the covariates of interest (referred to as structural parameters by Neyman and Scott 1948). This fixed effects model is favoured by econometricians and some remedies to the incidental parameter problem include orthogonal reparameterization of the fixed effects (Lancaster, 2002), and bias correction methods (Arellano and Hahn, 2010; Fernández-Val and Weidner, 2016).

Beyond the approaches mentioned above, most commonly marginal (also known as population average models) or conditional (such as mixed models) models are used to model longitudinal data. Marginal models typically use generalized estimating equations (GEE) for estimation, and conditional models rely on maximum likelihood (Laird and Ware, 1982). For a detailed overview of these methods, see Gardiner et al. (2009), and for a thoughtful critique "To GEE or not to GEE", see Hubbard et al. (2010) who

ultimately come down on the side of GEE. The two approaches have similarities since they can account for the binary nature of lesion data and the potential within-cluster correlation, they can handle time-invariant and time-varying independent variables, but they differ in the parameter interpretation and in the modelling assumptions. We will review those points of distinction by discussing each approach separately.

In generalized mixed effects model, random effects are added to the model to account for the unobserved cluster heterogeneity. The random effects are assumed to be random variables, with imposed distributional assumptions to account for the within-cluster correlation in longitudinal data. Mixed models allow for both marginal and cluster-specific inference, but they require the distributional assumptions involving the fixed effects and those for the random effects to be correctly specified. If the random effects distribution is misspecified, the power of the tests can be influenced or the Type I error rates can be inflated (Litière et al., 2007). In addition, the random effects are assumed to be uncorrelated with the explanatory variables, and if there is reason to believe that the unobserved heterogeneity is correlated with the explanatory variables, mixed models could result in inconsistent estimates. The latter assumption cannot be validated, which is the reason that the fixed effects approach mentioned above is favoured by econometricians.

A marginal model specifies the conditional mean, i.e. the mean of the outcome given the covariates, and accounts for the within-cluster correlation through the inclusion of a 'working' correlation matrix and robust computation of standard errors. The mean and the variance of the outcome are usually suggested by a distribution in the exponential family. However, the GEE approach does not rely on distributional assumptions, and as long as the conditional mean is correctly specified, we would get consistent estimates even if the working correlation is misspecified (Liang and Zeger, 1986). Modelling the mean directly implies that inference can be made about the average effect of variables on the response in a population. However, no subject-specific effects can be obtained as for mixed models.

Undoubtedly, mixed models are more appropriate if the main interest lies in estimating each subject's individual response to a change in a covariate. They can handle both discrete and continuous outcomes and can account for the dependence in the data through random slopes and intercepts or a parametric correlation model, though GEE

ultimately has more flexibility by accounting for any form of dependence through its robust standard errors. The use of mixed models leads to greater computational complexity than for marginal models as the number of random effects increases, and it could potentially lead to biased estimates if the model assumptions are not met. For those reasons and due to the fact we cannot easily verify whether model assumptions are supported by the data, especially given the vast number of regressions we fit across the brain, we focus on the marginal approach to modelling longitudinal binary lesion data.

1.3.1.2 Generalized estimating equations

We can think of generalized estimating equations (GEE) as a natural extension of GLMs as introduced in Section 1.2.1.1. Before discussing GEE models in more detail, we introduce the notation for longitudinal data. Here we are suppressing the voxel indexing for simplicity, with the model introduced below to be fitted at each voxel independently.

Building up on the cross-sectional data notation, suppose that each subject i out of N comes with a correlated binary response y_{it} at time point t, $t = 1, \ldots, T$, here we assume a balanced data set for notational simplicity, but the methods introduced below accommodate unbalanced data. The marginal mean vector is defined as $\boldsymbol{\mu}_i = \mathrm{E}(\boldsymbol{y}_i|X_i)$, where $\boldsymbol{y}_i = (y_{i1}, \ldots, y_{iT})^\top$ is a T-vector of binary responses and X_i is a $T \times P$ matrix of subject-specific covariates with rows $\boldsymbol{x}_{i1}, \ldots, \boldsymbol{x}_{iP}$. As in the cross-sectional analysis, $g(.)$ is a known link function, where $g(\mu_{it}) = \boldsymbol{x}_{it}^\top \boldsymbol{\beta}$ and $\boldsymbol{\beta}$ is a P-vector of unknown parameters. The full model specification is as follows

$$
\begin{aligned}
\mathrm{E}(\boldsymbol{y}_i|X_i) = \boldsymbol{\mu}_i = (\mu_{i1}, \ldots, \mu_{iT})^\top & \qquad \text{(marginal mean)} \\
V_i = W_i^{1/2} R_i(\boldsymbol{\alpha}) W_i^{1/2}/\phi & \qquad \text{(working covariance)} \\
g(\mu_{it}) = \eta_{it} & \qquad \text{(link function)} \\
\boldsymbol{\eta}_i = X_i \boldsymbol{\beta} & \qquad \text{(deterministic component)}\,,
\end{aligned}
$$

where the inclusion of a working covariance matrix V_i explicitly accounts for the within-cluster correlation (Liang and Zeger, 1986). The covariance matrix depends on (i) a within-cluster correlation matrix $R(\boldsymbol{\alpha})$, with $\boldsymbol{\alpha}$ being an unknown parameter vector to be estimated, (ii) a diagonal matrix W_i with $v_{it} = \mathrm{Var}(\mu_{it})$ on the diagonal, where $\mathrm{Var}(\mu_{it})$ is a known variance function, and (iii) a dispersion parameter ϕ.

The generalized estimating equations, as defined by Liang and Zeger (1986), are then

$$U(\boldsymbol{\beta}, \boldsymbol{\alpha}, \phi) = \sum_{i=1}^{N} D_i^\top V_i^{-1}(\boldsymbol{y}_i - \boldsymbol{\mu}_i) = \mathbf{0}\,, \tag{1.4}$$

where $D_i = \partial \boldsymbol{\mu}_i / \partial \boldsymbol{\beta}$ is a $T \times P$ matrix and U is a P-vector. An iterative procedure is then used to solve the estimating equations, where the one suggested by Liang and Zeger (1986) entails a Newton-Raphson iteration for the estimation of $\boldsymbol{\beta}$ and method of moments estimation for $\boldsymbol{\alpha}$ and ϕ.

As already mentioned, the asymptotic normality of the estimates is robust to misspecification of V_i (Liang and Zeger, 1986) when the number of clusters is sufficiently large, but choosing a dependence structure closer to the real one could result in efficiency gains (Wang and Carey, 2003). We can either rely on previous studies to pick the correlation structure, or we can make a reasonable guess and consider the choice of the correlation matrix as part of the data analysis; for more details on the choice of the working correlation, see Ziegler and Vens (2010) and Westgate and Burchett (2017).

Link function choice and potential issues The logit link function is typically selected when modelling binary data. However, the usage of different link functions to achieve the best interpretability of the estimated coefficients is discussed in the literature for both GLMs (Wacholder, 1986) and GEEs (Lu and Tilley, 2001). Therefore, if the aim is to interpret relative risks, which in medical applications often is, as opposed to odds ratios or absolute risk, the natural choice of a link function is the log link. When the outcome incidence is low (typically below 10%), the odds ratio approaches the risk ratio. However, in cohort studies and randomized control trials it sometimes happens that odds ratios are inappropriately interpreted as relative risks (Knol et al., 2012) since the odds ratio overestimates the risk ratio when the risk ratio is above 1 (Zhang and Yu, 1998). Thus, the typical choice for modelling cross-sectional binary data is log-Binomial regression.

However, the direct interpretability of the estimated effects as relative risks when using the log link function comes at the expense of convergence issues when the outcome incidence is approaching 1 (Knol et al., 2012), leading to numerical issues with iterative estimation procedures. As already mentioned, the mean and variance of the response are typically informed by a distribution in the exponential family and a fix to

the convergence issues proposed by Carter et al. (2005) replaces the Binomial distribution with Poisson distribution. Note there are no distributional assumptions imposed on the outcome data, and the main difference is the use of the log link with identity variance function instead of a quadratic curve. Carter et al. (2005) prove the consistency and the asymptotic normality of the estimators. Similar ideas are suggested by McNutt et al. (2003) and Zou (2004) for cross-sectional data but they do not discuss the theoretical properties of the estimator. For longitudinal data, Yelland et al. (2011b) found empirically through extensive simulations that log-Poisson regression has superior convergence rates when compared to log-Binomial regression (using GEE to account for the within-cluster correlation).

Boundary estimates Even though convergence issues due to high outcome incidence can be alleviated through the use of identity variance function for binary data, there could be further issues with convergence that are often not given full consideration. Yelland et al. (2011b) stated that *"Surprisingly, modified Poisson regression also failed to converge on rare occasions"*, and later Pedroza and Truong (2017) reported lack of convergence for the log-Poisson regression model in small sample size settings. One possible reason could be the presence of boundary estimates, i.e. the estimate falls on the boundary of the parameter space, which is more likely to happen for rare events and/or small samples. As we have discussed in Section 1.2.1, data separation has been thoroughly studied for logistic regression (Albert and Anderson, 1984; Lesaffre and Albert, 1989), and its causes and a variety of solutions have been recently discussed by Mansournia et al. (2017). To our knowledge, the only solution proposed for the logit Binomial GEE is introduced by Mondol and Rahman (2019), where a Jeffreys-prior penalty is used to ensure finite estimates in the presence of separation.

1.3.2 Contributions

The motivation for the work presented in Chapter 4 arises from the population brain imaging UK Biobank data set. A subset of the subjects included in the cross-sectional analyses in Chapters 2 and 3 were invited for a follow-up MRI scan, which leads to a data set on about 1,600 participants (after pre-processing) for two visits about 2 years apart.

To our knowledge, the GEE approach to modelling correlated binary outcomes has not been used to model changes in lesion incidence voxel-wise, and it is known that the GEE approach appropriately accounts for the within-subject correlation. To further ensure clinically relevant interpretatibility of the estimated coefficients, we use the log-link function to get relative risk estimates across the brain, e.g. at each voxel we get an estimated ratio of the probability of a lesion at a particular voxel if a participant is diabetic to the probability of a lesion if a participant is not diabetic, which is adjusted for demographic and lifestyle factors. Additionally, the nature of lesion data implies that high incidence areas are concentrated around the ventricles, but the majority of voxels have lesion incidence below 10%. Aiming to ensure stable estimates in the presence of varying lesion incidence, we use log-link GEE with identity variance and unknown dispersion.

1.3.2.1 Penalized generalized estimating equations

As mentioned in Section 1.3.1, longitudinal models could also suffer from boundary estimates, especially when the data set is small and/or the outcome incidence is low. The log-link GEE specification we have described above does not prevent parameter estimates being infinite, so we propose a penalty to be added to the generalized estimating functions to resolve this issue. The penalty we propose is the gradient of the Jeffreys prior, similarly to the penalty suggested by Mondol and Rahman (2019) for logit-link GEEs. The penalty added to the standard GEE in (1.4) for the p-the regression coefficient is of the form

$$A_p(\boldsymbol{\beta}, \boldsymbol{\alpha}) = \frac{1}{2}\text{trace}\big[I(\boldsymbol{\beta}, \boldsymbol{\alpha}, \phi)^{-1}\frac{\partial}{\partial \beta_p} I(\boldsymbol{\beta}, \boldsymbol{\alpha}, \phi)\big]\,,$$

where $I(\boldsymbol{\beta}, \boldsymbol{\alpha}, \phi) = \text{E}(-\partial U(\boldsymbol{\beta}, \boldsymbol{\alpha}, \phi)/\partial \boldsymbol{\beta})$. In Chapter 4 we derive the form of the penalty and the resulting modified estimating equations, which end up having the form

$$U_p^*(\boldsymbol{\beta}, \boldsymbol{\alpha}, \phi) = \phi \sum_{i=1}^{N} X_{ip}^\top W_i^{1/2} R(\boldsymbol{\alpha})^{-1} W_i^{-1/2}(\boldsymbol{y}_i - \boldsymbol{\mu}_i) + \frac{1}{2}\sum_{i=1}^{N}\sum_{t=1}^{T} h_{it} x_{itp} = 0\,,$$

where X_{ip}, $\boldsymbol{y}_i$, $\boldsymbol{\mu}_i$ are T-column vectors, and h_{it} is the t-th diagonal element of the i-th block of the projection matrix

$$H_i = W_i^{1/2} R(\boldsymbol{\alpha})^{-1} W_i^{1/2} X_i \Big[\sum_{i=1}^{N} X_i^\top W_i^{1/2} R(\boldsymbol{\alpha})^{-1} W_i^{1/2} X_i \Big]^{-1} X_i^\top .$$

We refer to the above penalized estimating equations as relative risk penalized GEE. The inclusion of the penalty ensures the finiteness of the estimates.

We perform an extensive simulation study to compare the performance of the standard log-link GEE and the proposed penalized log-link GEE, both with identity variance function and unknown dispersion. We monitor the frequency of boundary estimates occurrence by observing the behaviour of the diagonal elements of I^{-1} (the inverse of the expected negative Jacobian matrix) across iterations of the fitting procedure, which is shown to diverge as iterations grow (Lesaffre and Albert, 1989) for logistic regression with cross-sectional data.

It is known that cerebrovascular risk factors increase the risk of neurological disorders such as dementia. Therefore, applying the penalized GEE approach to the UK Biobank repeated lesion maps, our interest lies in uncovering the association between the change in the spatial distribution of lesions and cerebrovascular risk factors longitudinally as a continuation of the cross-sectional work completed in Chapter 3.

Reproducible research

The `R` code to fit the standard and the penalized GEE is available at the GitHub repository `https://github.com/petyakindalova/PGEE_Illustration` along with an illustration of the code used for the simulation study. The spatial relative risk maps produced as part of the real data application are available at NeuroVault `https://neurovault.org/collections/SUDHNHAA/`.

Chapter 2

Voxel-wise and spatial modelling of binary lesion masks: Comparison of methods with a realistic simulation framework

The attached paper is published in *NeuroImage* (Kindalova et al., 2021a) and it is based on joint work with Ioannis Kosmidis and Thomas E. Nichols. A statement of contribution is published as part of the paper.

Abstract

Objectives White matter lesions are a very common finding on MRI in older adults and their presence increases the risk of stroke and dementia. Accurate and computationally efficient modelling methods are necessary to map the association of lesion incidence with risk factors, such as hypertension. However, there is no consensus in the brain mapping literature whether a voxel-wise modelling approach is better for binary lesion data than a more computationally intensive spatial modelling approach that accounts for voxel dependence.

Methods We review three regression approaches for modelling binary lesion masks including mass-univariate probit regression modelling with either maximum likelihood estimates, or mean bias-reduced estimates, and spatial Bayesian modelling, where the regression coefficients have a conditional autoregressive model prior to account for local spatial dependence. We design a novel simulation framework of artificial lesion maps to compare the three alternative

lesion mapping methods. The age effect on lesion probability estimated from a reference data set (13,680 individuals from the UK Biobank) is used to simulate a realistic voxel-wise distribution of lesions across age. To mimic the real features of lesion masks, we propose matching brain lesion summaries (total lesion volume, average lesion size and lesion count) across the reference data set and the simulated data sets. Thus, we allow for a fair comparison between the modelling approaches, under a realistic simulation setting.

Results Our findings suggest that bias-reduced estimates for voxel-wise binary-response generalized linear models (GLMs) overcome the drawbacks of infinite and biased maximum likelihood estimates and scale well for large data sets because voxel-wise estimation can be performed in parallel across voxels. Contrary to the assumption of spatial dependence being key in lesion mapping, our results show that voxel-wise bias-reduction and spatial modelling result in largely similar estimates.

Conclusion Bias-reduced estimates for voxel-wise GLMs are not only accurate but also computationally efficient, which will become increasingly important as more biobank-scale neuroimaging data sets become available.

2.1 Introduction

White matter hyperintensities of presumed vascular origin (WMHs), also known as white matter lesions or leukoaraiosis (Wardlaw et al., 2013), are signs of cerebral small vessel disease (SVD) in the brain. Lesions are evident on Magnetic Resonance Imaging (MRI) as hyperintensities on the T2-weighted, fluid attenuated inversion recovery (FLAIR), and proton density-weighted brain images. WMHs are common in the aging brain and are associated with cerebrovascular burden (Rostrup et al., 2012; Griffanti et al., 2018; Lampe et al., 2019b). It is not clear how different contributors to the cerebrovascular burden, such as hypertension or smoking history, relate to the spatial distribution of WMHs in the brain and this has given rise to the exploitation of a variety of statistical methods in the field.

There are other types of lesions in brain imaging, which differ by the factor influencing their development. For example, multiple sclerosis (MS) is an autoimmune disease of the central nervous system, which causes the destruction of myelin further resulting in brain and spinal cord lesions. Another example are stroke lesions, which can have very similar signal intensities as WMHs and are of vascular origin too. However, independent of the type of brain lesions, their size, location, growth, etc., are important for diagnosis, treatment or prevention. While all types of lesion data motivate the current work, going forward we will focus on WMHs of presumed vascular origin.

As originally created, the MRI scans exist in so-called native space, and do not correspond to any other subject's brain. These native space images are used to quantify the severity of lesions either based on a visual scoring system (e.g. Fazekas scale Fazekas et al. 1987), or by segmenting the lesions by producing a binary lesion map indicating lesion presence/absence. Visual scoring as well as manual lesion segmentation are quite common in neurodegenerative diseases such as MS, even though they are expensive, time-consuming and subject to inter-rater variability (Rudick et al., 2012; Hagens et al., 2019). An objective automated segmentation procedure, such as BIANCA (Griffanti et al., 2016), is preferable since it provides a scalable method to obtain reproducible lesion maps on thousands of subjects.

Whether created manually or by an automated method, the native space lesion maps can be transformed to the MNI atlas space, producing aligned binary lesion maps ready

for group analyses. For example, a voxel-wise analysis can compare the distribution of patterns of lesions from different disease subtypes (Filli et al., 2012), or perform voxel-wise linear regressions between lesion probability and different clinical disability scores (Charil et al., 2003; Kincses et al., 2011). Approaches such as the ones mentioned are known as mass-univariate since they fit a model at each voxel independently, ignoring any spatial dependence between nearby voxels which is later accounted for at the inference stage (e.g. using a method like false discovery rate (FDR) correction that allows for positive spatial dependence (Genovese et al., 2002)). While some authors have used a standard linear model with lesion incidence as response (Kincses et al., 2011), this is ill-advised as it ignores the binary and heteroscedastic nature of the data.

Mass-univariate voxel-wise modelling of lesion masks that accounts for the binary nature of the data is done through maximum likelihood estimation of a generalized linear model (GLM), e.g. logistic or probit regression. While the GLM has been used in the voxel-wise brain lesion mapping literature (Lampe et al., 2019b; Rostrup et al., 2012), to our knowledge the limitations of logistic or probit regression with small sample size and/or low incident responses has not been addressed. These issues have been thoroughly investigated in the statistics literature and a short overview is provided here.
Outside of linear models, maximum likelihood estimation typically requires iterative optimization, such as iteratively reweighted least squares (IRLS) (Green, 1984). When a covariate (or a combination of covariates) in a logistic or probit regression model perfectly separates the outcome variable, 'data separation' occurs (Albert and Anderson, 1984) and the maximum likelihood estimates (MLEs) for those covariates are infinite. Hence the iterative procedure for maximum likelihood will diverge or, even worse, stop early, reporting massive in absolute value estimates without any warning that the estimates are in reality infinite. This is more likely to happen when dealing with rare responses or small sample size. For example, with lesion data, it could happen if only subjects older than 60 years of age have a lesion at a particular voxel and no subject younger than 60 does. In such cases, estimated standard errors also diverge to infinity but faster than the estimates. As a result, the commonly used Wald statistics become artificially small in absolute value masking any significance in evidence when testing. In addition, the optimal properties of the ML estimator only hold asymptotically, and finite sample properties may be far from what is expected asymptotically. To address

both limitations of the MLE, the use of a bias-reduction approach (Kosmidis and Firth, 2009; Kosmidis et al., 2020) in mass-univariate voxel-wise modelling is explored. The method guarantees finite-valued estimates (Kosmidis and Firth, 2021) as it corrects for the first-order bias of the ML estimator. Furthermore, bias-reduced estimates are fast to obtain (typically being only slightly more expensive than MLEs) and the voxel-wise modelling allows for parallel implementation, which makes the method feasible for large imaging data sets.

In contrast to the mass-univariate approaches for brain image analysis, Ge et al. (2014) introduce a Bayesian Spatial Generalized Linear Mixed Model (BSGLMM). While still accounting for the binary nature of the data, the main difference between BSGLMM and the classical GLMs is that BSGLMM accounts for the local spatial dependence in the brain through the inclusion of spatially varying coefficients to a Bayesian spatial model. Spatially varying coefficients are latent spatial processes (or fields) and they are modelled jointly using a multivariate pairwise difference prior model, a particular instance of the Multivariate Conditional Autoregressive (MCAR) model. Given that the method estimates an entire brain mask of coefficients for each covariate in a model (e.g. age and sex), there is a considerable computational burden, which is partly alleviated by a parallel graphical processing unit (GPU) implementation (Ge et al., 2014).

The motivation for the present work is the lack of validation for the mass-univariate generalized linear regression model, and the only very limited simulation framework used to evaluate the BSGLMM method. In particular, it is not known whether the sample sizes and typical incident rates found in WMH studies produce highly biased (or even divergent) estimates of regression effects with standard maximum likelihood estimators. With the Bayesian approach, while Ge et al. (2014) provide simulations showing the benefits of spatial regularization, those evaluations used only 2D images with large homogeneous lesion patterns that do not reflect the highly structured and inhomogeneous patterns found in real data.

To gain a better understating of the differences between the alternative lesion mapping methods, in this paper we develop a novel simulation framework of artificial lesion maps. We estimate the effect of age on lesion probability in a reference data set (a subset of the UK Biobank data set Miller et al. 2016) and we use it to simulate a realistic voxel-wise distribution of lesions across age. We use age as a covariate since it is thought to be the

strongest risk factor for the presence of lesions, although the simulation approach could be adapted to utilise effect maps of any risk factor. To mimic the real features of lesion masks we suggest matching brain lesion summaries (total lesion volume, average lesion size and lesion count) across the reference data set and the simulated data sets. In this way, we allow for a more realistic, fairer comparison between the modelling approaches.

In this paper, we compare three alternative approaches for modelling binary lesion masks, two mass-univariate regression methods and the BSGLMM method, with the remainder of the paper organised as follows. In Section 2.2.1 we start by providing the details behind these different methods. We set out the steps of our proposed novel simulation framework in Section 2.2.2, which mimics features of real lesion masks. We then apply the three modelling approaches to simulated data sets and evaluate their performance in terms of a range of estimation accuracy metrics, such as bias and mean squared error, as well as spatial overlap between Wald statistics (reference versus estimated z-scores), false positive control and computational cost (Section 2.3.1). To demonstrate the scalability of one of the methods, we apply it to a subset of the UK Biobank data, where we estimate the effect of systolic blood pressure on lesion probability (Section 2.3.2).

Software to generate lesion masks using our simulation framework is available through the Open Science Framework website[1], which also provides a demonstration of the parallel GLMs implementation.

2.2 Materials and methods

2.2.1 Summary of existing regression methods

Suppose that there are N individuals and that each subject $i\,(i = 1, \ldots, N)$ comes with a binary lesion mask $\boldsymbol{Y}_i \in \mathcal{B} \subset \mathbb{R}^3$. $\mathcal{B}$ is the human brain and we consider M cubic cells (voxels) as a discretization of the 3D brain on a regular rectangular grid, where s_j denotes the jth voxel within the brain $(j = 1, \ldots, M)$. When modelling binary lesions masks voxel-wise, we consider two approaches that ignore spatial dependence and one that explicitly models that dependence.

[1]Project URL: `https://osf.io/h7sxr/`

2.2.1.1 Generalized linear model

A mass-univariate approach fits a model at each voxel marginally, ignoring spatial dependence. A generalized linear model (GLM) is required in order to respect the binary nature of the data. Every GLM has a link function, deterministic and stochastic components, which we write as

$$[Y_i(s_j) \mid p_i(s_j)] \sim \text{Bernoulli}(p_i(s_j)) \qquad \text{(stochastic component)} \qquad (2.1)$$

$$g(p_i(s_j)) = \eta_i(s_j) \qquad \text{(link function)} \qquad (2.2)$$

$$\eta_i(s_j) = \boldsymbol{x}_i^\top \boldsymbol{\beta}(s_j) \qquad \text{(deterministic component)}, \qquad (2.3)$$

where

- $Y_i(s_j)$ denotes a Bernoulli random variable with probability of success $p_i(s_j)$ and probability mass function $f(y_i(s_j) \,|\, \boldsymbol{x}_i;\, \boldsymbol{\beta}(s_j))$, where $y_i(s_j)$ is a realization of random variable $Y_i(s_j)$ that represents the presence ($Y_i(s_j)$=1) or absence of a lesion for subject i at voxel s_j. Note that in this mass-univariate voxel-wise modelling framework, $Y_1(s_1), \ldots, Y_N(s_1), \ldots, Y_1(s_M), \ldots, Y_N(s_M)$ are assumed to be independent random variables given $p_i(s_j)$.

- g denotes the link function, which is a monotonic function that relates the expectation of the stochastic outcome to the deterministic component.

- $\boldsymbol{x}_i$ denotes the P-vector of subject-specific covariates for subject i, where $\boldsymbol{X}$ is the full rank model matrix that collects $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N$ in its rows and has columns $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_P$.

- $\boldsymbol{\beta}(s_j) = (\beta_1(s_j), \ldots, \beta_P(s_j))^\top$ is a P-vector of parameters at each voxel s_j; these are fixed effects.

The GLM outlined in Eqs. (2.1-2.3) is fitted at each voxel s_j independently. We obtain the maximum likelihood estimators (MLEs) $\hat{\boldsymbol{\beta}}(s_j)$ by maximizing the log-likelihood

$$l(\boldsymbol{\beta}(s_j)) = \sum_{i=1}^{N} \log f(y_i(s_j) \mid \boldsymbol{x}_i;\, \boldsymbol{\beta}(s_j))\,, \qquad (2.4)$$

through an iterative optimization procedure, such as IRLS (Green, 1984). The MLE is

typically the default choice of an estimator because of its optimal asymptotic properties (consistency, asymptotic normality and efficiency). If the model assumptions are adequate, then any inferential procedures based on those estimates, such as tests using Wald statistics (also known as standardized coefficients or z-scores) are also asymptotically correct. However, for finite sample size N the estimates can be unstable and biased.

Bias-reduction in parametric estimation has been thoroughly studied in the literature; for a detailed review see Kosmidis (2014). There are many methods, such as bootstrap, which correct for bias, but they rely on the existence of the MLE. However, if data separation occurs, the MLE for one or more covariates is infinite[2], which apart from computational issues also results in invalid Wald-type inference (extremely wide and uninformative Wald-type confidence intervals due to large standard errors). The bias-correction approach which we focus on in this work was first introduced in Firth (1993) for logit link binomial GLMs and then was further developed for exponential families and applied in generalized nonlinear models (Kosmidis and Firth, 2009; Kosmidis et al., 2020). Adjustments to the score equations (partial derivatives of the log-likelihood set to zero) ensure that estimates $\tilde{\boldsymbol{\beta}}(s_j)$ have asymptotically smaller bias than what the MLE typically has; see 2.A.2 for details. Furthermore, obtaining the MeanBR estimates is only a modest addition to the computational complexity for computing the MLEs. The MeanBR method is implemented in the `R` package `bsglm2` (Kosmidis et al., 2020; Kosmidis, 2020) as an extension to the base `R` `glm` tool.

For the current analyses of simulated and real data, we have chosen probit link Φ^{-1}, where Φ indicates the standard normal cumulative distribution function; we use probit link to ensure comparability with the link used in the BSGLMM approach. Finally, at each voxel, we obtain maximum likelihood estimates $\hat{\boldsymbol{\beta}}(s_j)$ and mean bias-reduced estimates $\tilde{\boldsymbol{\beta}}(s_j)$ along with Wald statistics $\hat{\mathbf{z}}(s_j)$ and $\tilde{\mathbf{z}}(s_j)$ based on those estimates, respectively.

2.2.1.2 Bayesian spatial generalized linear mixed model

The spatial generalized linear mixed model (GLMM) is based on the GLM presented above. However, the deterministic components are explicitly defined functions of space.

[2]Software packages handle separation differently depending on their convergence criterion and the user might not be notified.

While the stochastic component and the link function in Eqs. (2.1) and (2.2) are the same, the deterministic component introduced by Ge et al. (2014) is:

$$\eta_i(s_j) = \boldsymbol{x}_i^T(\boldsymbol{\alpha} + \boldsymbol{\beta}(s_j))\,, \tag{2.5}$$

where the key difference is the inclusion of spatially varying coefficients in addition to the fixed effects. In particular,

- $\boldsymbol{\alpha}$ denotes a P-vector of parameters, fixed effects.
- $\boldsymbol{\beta}(s_j)$ denotes a P-vector of mean-zero random effects, one at each voxel s_j. These random effects are spatially varying voxel-specific effects.

The last bit of the model specification is to assign priors to all parameters in order to complete the specification of the hierarchical model. This is done in the following way:

- fixed effects' priors are flat, improper, uninformative, i.e. $\pi(\boldsymbol{\alpha}) \propto \mathbf{1}$.
- random effects (spatially varying coefficients) have Markov random field (multivariate conditional autoregressive (MCAR) model) priors to account for the spatial dependence. Two voxels are considered to be neighbors if they share a face, i.e. a maximum of six neighbors. In terms of notation, S_k denotes the set of neighboring locations for location s_j and the cardinality of this set is denoted as $N(s_j)$. The MCAR prior can be written as

$$[\boldsymbol{\beta}(s_j) \mid \boldsymbol{\beta}(-s_j), \boldsymbol{\Sigma}] \sim \text{MVN}\left[\frac{\sum_{s\in S_k}\boldsymbol{\beta}(s)}{N(s_j)}, \frac{\boldsymbol{\Sigma}}{N(s_j)}\right], \tag{2.6}$$

  where

  – $\boldsymbol{\beta}(-s_j)$ denotes $\boldsymbol{\beta}$ excluding the coefficients at voxel s_j.

  – $\boldsymbol{\beta}^T = [\boldsymbol{\beta}^T(s_1), \ldots, \boldsymbol{\beta}^T(s_M)]$ is a $P \times M$ column vector.

  – $\boldsymbol{\Sigma}$ is a $P \times P$ symmetric positive definite matrix.

  – The inverse of the hyperparameter $\boldsymbol{\Sigma}$ is needed and its inverse is assumed to have Wishart prior, i.e. $\boldsymbol{\Sigma}^{-1} \sim \text{W}(\nu, \boldsymbol{I}_P)$, where ν is set to 0 in Ge et al. (2014) and $\boldsymbol{I}_P$ is a $P \times P$ identity matrix.

The joint distribution of $\boldsymbol{\beta}$ is improper and not identifiable (Ge et al., 2014), but all the full conditional distributions are well-defined but not easy to sample from. A graphical processing unit (GPU) allows for parallel implementation of the Gibbs sampler derived in Ge et al. (2014)[3]. At each voxel, posterior summaries for the spatially varying coefficients $\boldsymbol{\beta}^*(s_j)$ are obtained along with standardized posterior effects or z-scores (posterior mean divided by posterior standard deviation) $\mathbf{z}^*(s_j)$.

2.2.2 Simulations

Simulation of brain lesions is complicated by the need for a generative model that accounts for dependence in the data. The mass-univariate model makes no attempt to model dependence, and while the BSGLMM explicitly models dependence, it does so on the regression parameters not the data itself. That is, the BSGLMM assumes that the binary lesion data $\boldsymbol{Y}_i$ are independent given the (CAR-regularised) regression parameters $\boldsymbol{\beta}$. Thus even if an accurate regression model could be fit everywhere in the brain, simulation of lesion data $\boldsymbol{Y}$ from either the mass-univariate, or a conditionally independent Bayesian model would be characterised by independent "salt and pepper" noise, i.e. random isolated lesions of 1 or 2 voxels, or single voxel omissions from an otherwise large lesion.

Thus in this work we develop a novel simulation approach that generates realistic binary lesion data that is calibrated to a given generalized linear model. We use this approach to compare three alternative methods (see Section 2.2.1) for modelling the spatial distribution of white matter lesions, and to assess their performance in terms of a variety of measures of accuracy, such as mean squared error (MSE).

2.2.2.1 Simulation procedure

We aim to simulate $\boldsymbol{Y}_1^*, \ldots, \boldsymbol{Y}_{N^*}^*$ lesion masks for N^* subjects that follow a generalized linear model for a given map of regression parameters. Given an existing data set (referred to as 'reference' data) of N lesion masks $\boldsymbol{Y} = (\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_N)$ and a vector $\boldsymbol{X}_2 = (X_{12}, \ldots, X_{N2})$ of centered age, artificial binary lesion masks are simulated as follows:

Step 1 Learn parameters from reference data.

[3]Code available at `https://www.nisox.org/Software/BSGLMM/`

For a reference dataset $\boldsymbol{Y}$ and a model matrix $\boldsymbol{X} = [\mathbf{1}_N \quad \boldsymbol{X}_2]$, where $\mathbf{1}_N$ is an N-vector of ones, obtain estimated maps for intercept and age effects $\boldsymbol{\beta}(s_j) = [\beta_1(s_j) \quad \beta_2(s_j)]$, at each s_j $(j = 1, \ldots, M)$. These coefficients $\boldsymbol{\beta}$ are considered as truth going forward.

Step 2 Construct simulation design.

Create the model matrix $\boldsymbol{X}^*$ by simulating age $\boldsymbol{X}_2^*$ for N^* subjects, where $x_{2m}^* \sim \mathrm{U}(\min(\boldsymbol{X}_2), \max(\boldsymbol{X}_2))$, $(m = 1, \ldots, N^*)$, the uniform distribution on the age range in the reference data set. Then the simulation model matrix is $\boldsymbol{X}^* = [\mathbf{1}_{N^*} \quad \boldsymbol{X}_2^*]$.

Step 3 Simulate smooth noise for linear predictor.

Simulate a zero-mean Gaussian Random Field (GRF) with squared exponential covariance function independently for each of the N^* subjects. The `R` package `RandomFields` (Schlather et al., 2020) and its function `RFsimulate()` are used to simulate a GRF with covariance $C(h) = \sigma^2 \exp(-h^2/2\ell^2)$, where h is the distance between voxels, and the two parameters are the variance σ^2 and the scale ℓ. The scale determines the dependence between voxels.

Step 4 Generate binary lesion data.

Create a binary lesion mask for subject m as $Y_m^*(s_j) = \mathbb{I}\{\Phi(\boldsymbol{x}_m^{*\top}\boldsymbol{\beta}(s_j) + \mathrm{GRF}_m(s_j)) > 0.5\}$, where $\boldsymbol{x}_m^*$ is the mth row of the simulated model matrix $\boldsymbol{X}^*$ and $\mathrm{GRF}_m(s_j)$ is the value of the simulated GRF for subject m at voxel s_j. In particular, we first add the true effect and noise and transform the sum into a lesion probability using the cumulative distribution function of the standard normal before thresholding the lesion probabilities at 0.5 to get binary lesion masks. Note that the threshold of 0.5 ensures that we match the lesion incidence found in the reference data $\boldsymbol{Y}$, set via the intercept term since we are using centered age.

In our illustration, the reference data set $\boldsymbol{Y}$ consists of binary lesion masks of 13,680 UK Biobank (UKB) (Miller et al., 2016) participants along with their age at scan date; the data set is described further in Section 2.2.3. Since our binary lesion mask simulator takes effect maps and GRF parameters as inputs, we make the following choices: (i) the

effect maps $\boldsymbol{\beta}(s_j)$ are mean bias-reduced estimates obtained by fitting voxel-wise GLMs with probit link function and age as the only covariate in the model, and (ii) the use of probit link GLMs to model the simulated data means the variance parameter σ^2 should be fixed to 1 to match the standard Normal variance, and thus there is only one free GRF parameter ℓ. Note that we simulate lesion masks of the same resolution as the reference data lesion masks.

2.2.2.2 Tuning of simulation parameters

Aiming to mimic the real features of the data, we tune the scale parameter ℓ of the GRF to minimise the discrepancies between reference and simulated data medians of the following lesion mask summaries: (i) total lesion volume, (ii) lesion count, (iii) average lesion size. Specifically, looking over ten age bins,

(i) total lesion volume is defined as the number of lesion-affected voxels;

(ii) lesion count is determined using the `FSL cluster`[4] function (connectivity 6);

(iii) average lesion size is defined as total lesion volume divided by lesion count;

and the age bins are determined by the deciles of the reference data set age distribution.

We repeat Steps (3–4) on a grid of scale parameters conditionally on the noise component in Step 3 and until a satisfactory match is found between the simulated medians and reference medians across age bins.

2.2.2.3 Measures of accuracy

Once the GRF parameter is tuned, the three regression modelling methods can be applied and their performance compared across R repetitions. First, we repeat Steps (3–4) (Section 2.2.2.1) R times to obtain R simulated data sets. Then, for each simulated data set r ($r = 1, \ldots, R$), we fit N^* lesion masks $\boldsymbol{Y}^{*\,(r)}$ on age $\boldsymbol{X}_2^*$ to obtain ML $\hat{\boldsymbol{\beta}}(s_j)^{(r)}$, MeanBR $\tilde{\boldsymbol{\beta}}(s_j)^{(r)}$ and BSGLMM $\boldsymbol{\beta}^*(s_j)^{(r)}$ intercept and age estimated maps and their associated z-scores.

To compare the performance of the three regression modelling methods, we calculate the following measures of accuracy of MLE $\hat{\boldsymbol{\beta}}(s_j)$, voxel-wise:

[4] `https://fsl.fmrib.ox.ac.uk/fsl/fslwiki/Cluster`

- Bias B: $\mathrm{B}(\hat{\boldsymbol{\beta}}(s_j)) \approx \frac{1}{R}\sum_r \hat{\boldsymbol{\beta}}(s_j)^{(r)} - \boldsymbol{\beta}(s_j)$,
- Mean squared error MSE: $\mathrm{MSE}(\hat{\boldsymbol{\beta}}(s_j)) \approx \frac{1}{R}\sum_{r=1}^{R}(\hat{\boldsymbol{\beta}}(s_j)^{(r)} - \boldsymbol{\beta}(s_j))^2$,
- Probability of underestimation PU: $\mathrm{PU}(\hat{\boldsymbol{\beta}}(s_j)) \approx \frac{1}{R}\sum_{r=1}^{R}\mathbb{I}\{\hat{\boldsymbol{\beta}}(s_j)^{(r)} < \boldsymbol{\beta}(s_j)^{(r)}\}$.

The corresponding summaries are estimated for $\tilde{\boldsymbol{\beta}}(s_j)$ and $\boldsymbol{\beta}^*(s_j)$. We also explore the Pearson correlation coefficient between the estimated coefficients and the reference data coefficients as another measure of estimator accuracy, resulting in one correlation coefficient per realisation r across the three methods.

To make inference about the effect of a covariate on the lesion probability across the brain, z-score maps are typically explored. Given the difference in the sample size N of the reference data set and N^* of the simulated data sets, the power to detect significant age effect varies and use of a fixed z-score threshold (e.g. ± 1.96) to compare maps is not appropriate. Thus, we fix the z-score threshold to a particular percentile of the z-score distribution (in absolute value), such that we select the highest M^* z-scores. We explore the Dice similarity coefficient (DSC) (Dice, 1945) to measure the spatial overlap between a reference result and one of the three methods, e.g. the highest M^* reference age z-scores $\mathbf{z}(s_j)$ and the the highest M^* age z-scores $\hat{\mathbf{z}}(s_j)^{(r)}$ for simulated data set r. DSC results are the mean across R repetitions. We note that in the image validation literature, a DSC greater than 0.7 is interpreted as good overlap (Zijdenbos et al., 1994; Zou et al., 2004).

We also create maps of the lesion incidence across the brain, where $p(s_j)$ denotes the reference data set lesion incidence at voxel s_j and $\hat{p}(s_j)$ denotes the lesion incidence for a simulated data set.

Software to generate lesion masks using our simulation framework is available through the Open Science Framework website[5], which also provides a demonstration of the parallel GLMs implementation.

2.2.3 Application

To demonstrate the scalability of the mass-univariate approaches (ML and MeanBR), we apply them to a subset of the UK Biobank data (Miller et al., 2016). The data set includes 13,680 healthy ageing individuals, for details on the selection criteria see

[5]Project URL: `https://osf.io/h7sxr/`

Veldsman et al. (2020). Voxel-wise analysis is used to investigate the effect of systolic blood pressure (BP) on the spatial distribution of lesions while controlling for confounding (age, sex, age by sex interaction and head size scaling are included as confounding variables; this is the minimal set of confounding variables suggested by Alfaro-Almagro et al. 2020). The mean age of the participants is 62.9 years ($\pm$7.4 years) with 53% being female (7,236 women). Two sequential measurements of systolic BP were taken in each subject (either manual, or automatic measurement) and the average of these two readings is used as our main covariate of interest. Note that blood pressure is known to be a dominant risk factor for the presence of lesions (Debette and Markus, 2010), so it was chosen for illustrative purposes.

To generate the binary lesion masks for these 13,680 subjects, we use the Brain Intensity Abnormality Classification Algorithm (BIANCA) (Griffanti et al., 2016) to segment the lesions. BIANCA's inputs include T1-weighted and T2-weighted FLAIR images (Alfaro-Almagro et al., 2018). The BIANCA output image in native space is thresholded at 0.8 and binarised as part of the segmentation, where the threshold is optimised as part of the BIANCA training on manually segmented masks of subjects from the UKB cohort. Those binary maps in subject space are then registered to 2mm MNI space by applying the estimated spatial normalisation parameters derived as part of the published UKB preprocessing pipeline (Alfaro-Almagro et al., 2018). More specifically, the generation of T2 FLAIR images in MNI space includes T2 FLAIR to T1 linear registration (FLIRT, Jenkinson et al. 2002b) and T1 to MNI non-linear warping (FNIRT, Andersson et al. 2007). The resulting images are binarised with a 0.5 threshold, as interpolation produces non-binary values. It is these 13,680 binary lesion masks (reference data $\boldsymbol{Y}$) that are fit using a mass-univariate approach (i) to define the reference MeanBR estimates for intercept and age used in the simulator (Step 1 of the simulator in Section 2.2.2.1 with age as the only covariate), and (ii) to obtain ML and MeanBR estimates for systolic blood pressure across voxels while accounting for confounding due to age, sex, age by sex and head size scaling.

2.3 Results

2.3.1 Results on the simulated data

The reference data set used to obtain the reference coefficients is the subset of the UKB data set described in Section 2.2.3. We fit 13,680 lesion masks on age to obtain MeanBR voxel-wise estimates for the intercept and age terms. The model includes only age as a covariate and the analysis mask comprises the 72,603 voxels with non-zero lesion incidence.

2.3.1.1 Simulation setting

Illustration of simulation steps To tune the GRF scale parameter, we simulate one data set of N^*=1,000 subjects for various scale values. Figure 2.1 demonstrates the resulting simulated masks $\boldsymbol{Y}^*$ for two scale parameter choices for subjects aged 50 and 70 years. Increasing the scale parameter increases the smoothness of the GRF (lower granularity), i.e. the scale parameter controls the number of lesions and their size; if the variance parameter is fixed, increasing the scale parameter leads to lower count but bigger size of lesions (see Figures 2.2 and 2.B.2). Given that the true age effect suggests higher lesion probability with increasing age, we would expect to see more lesions for an individual aged 70, which is indeed the case in the illustration in Figure 2.1.

Tuning of simulation parameters As described in Section 2.2.2.2, our main goal is to match as closely as possible the reference data (UKB data) lesion summaries to the simulated lesion summaries. Figure 2.2 includes the results for one of the lesion summaries we considered - average lesion size. The top plot represents the median average lesion size across 10 age groups for five simulation settings along with the mean and median for the reference data set (dashed and solid black lines). By visual inspection, we found that the best scale parameter value based on all three summaries (also see Figures 2.B.1 and 2.B.2) is ℓ=1.5. The side by side boxplots of average lesion size in one simulated data set of 1000 subjects and in the UKB data set of 13,680 participants across age groups suggests the chosen simulation setting follows closely the trend in the reference data across age and the variability in the simulated data is lower than the variability in the reference data. Note we repeated the experiment for a second seed to make sure the

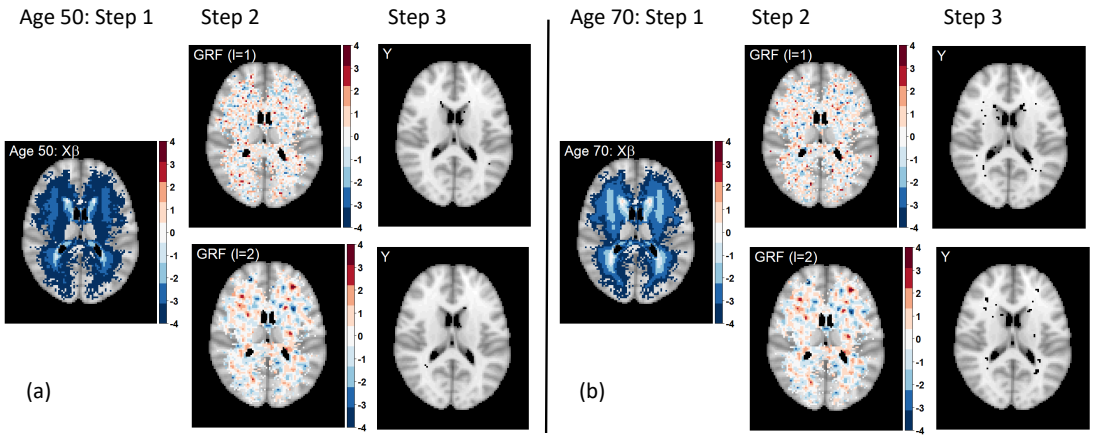


Figure 2.1: Illustration of simulation steps for scale parameters ℓ=1 and ℓ=2 for a subject aged 50 (a) and a subject aged 70 (b). Higher age is associated with higher lesion probability in the periventricular areas (Step 1), as shown by the linear predictor for a 50-year-old and a 70-year-old. Higher scale parameter leads to coarser GRF component (Step 2). The choice of the GRF parameter is crucial for simulating realistic lesion masks. Voxels with zero lesion incidence for the UKB data set ($p(s_j)$=0) are plotted as transparent to show a standard anatomical MRI for reference; axial slice z=45 shown.

lesion summaries do not vary substantially and a plot complementary to Figure 2.2 is included in Figure 2.B.3.

Estimated age effect We have tuned the scale parameter of the GRF and the simulations from now on assume the variance and scale parameters are fixed to 1 and 1.5, respectively. Exploring the achieved lesion probability for a single simulated data set of 1,000 subjects and the UKB lesion probability based on 13,680 participants (Figure 2.3), we observe that the highest lesion probability regions are consistent across the two maps but the simulated data set does not achieve as wide a spatial coverage as the real data set (40,338 non-zero lesion incidence for the simulated data set vs 72,603 for the reference data set, respectively). Note that the UKB data set is about 14 times bigger than the simulated data set, i.e. with a single simulated data set of that size we cannot capture the rarer lesions in the outer white matter. The more limited coverage is also observed for the estimated regression coefficients since we simply do not fit the mass-univariate GLMs at voxels with zero lesion incidence. However, BSGLMM has larger z-scores due to the variance reduction of the smoothness prior and careful inspection suggests possible bleeding of signal into areas where ML and MeanBR do not capture any signal.

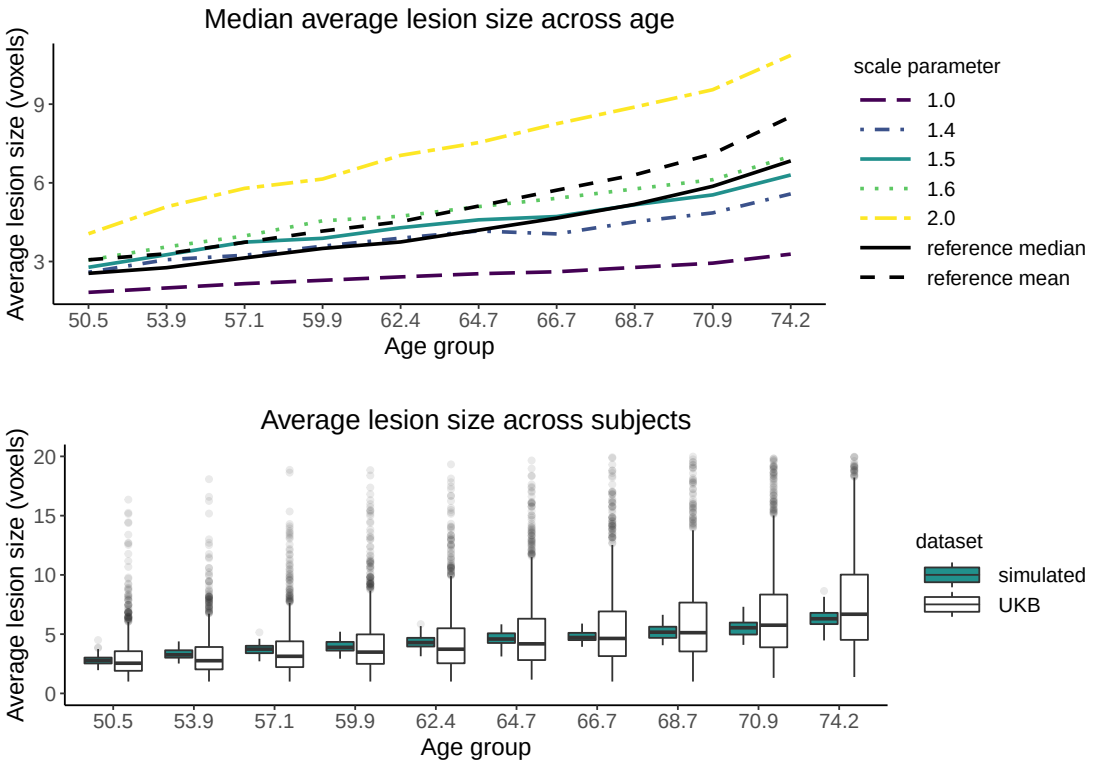


Figure 2.2: Gaussian random field parameter tuning by matching the reference data (UKB) median average lesion size across age bins (black solid line). (top) Plot of median average lesion size across age bins for five simulation settings (five GRF scale parameter values) and reference data values (black lines). Legend values indicate the scale parameter value ℓ used to simulate a GRF for each subject in the simulated sample. (bottom) Boxplots of average lesion size in UKB participants (white) and in one simulated 1000-subject sample with GRF scale parameter ℓ=1.5 (blue) across ten age bins. Note the x-axis labels denote the center of each age bin, the y-axis units are in 2mm^3 voxels, and the variance GRF parameter is fixed to 1 for all simulation settings.

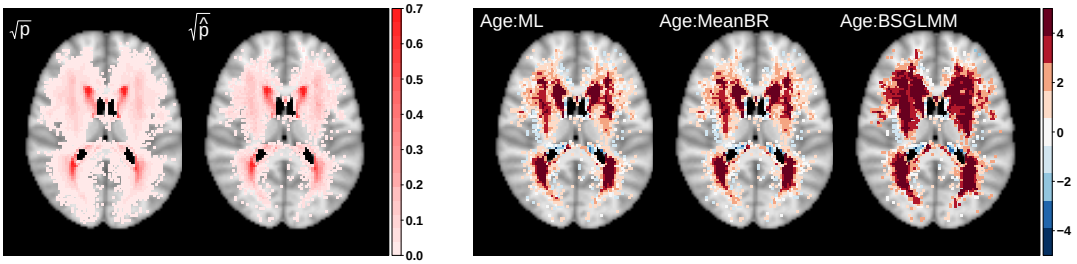


Figure 2.3: Square-root transformed lesion probability based on 13,680 UKB participants $\sqrt{\boldsymbol{p}}$ and lesion probability based on one simulated sample with N^*=1,000 subjects $\sqrt{\hat{\boldsymbol{p}}}$ (left panel) and significance maps (z-scores) for the effect of age across methods (right panel). 72,603 voxels have non-zero lesion probability for the UKB data set and 40,338 for the simulated data set, respectively, which explains the difference in spatial coverage in the left panel.

2.3.1.2 Estimator accuracy

We have visually compared the lesion probability maps and significance maps for one simulated data set against the UKB data set, but in order to quantify the difference between the three modelling approaches, we repeat the experiment R=1,000 times, using the chosen simulation scale parameter for two sample sizes of N^*=250 and N^*=1,000. We estimate $\hat{\boldsymbol{\beta}}(s_j)^{(r)}$, $\tilde{\boldsymbol{\beta}}(s_j)^{(r)}$ and $\boldsymbol{\beta}^*(s_j)^{(r)}$ $(r = 1, \ldots, R)$ and their associated z-scores and measures of accuracy as described in Section 2.2.2.3. We focus on voxels with lesion incidence in the reference data set $p > 0.005$ to ensure the lesion count is not too low in the simulated data sets.

Shrinkage effect The plots of the estimated coefficients across the three methods against the reference coefficients (UKB) for age (Figures 2.4, 2.B.4 and 2.B.5), for one realisation of N^*=1,000 subjects, suggest that MeanBR and BSGLMM estimates are closer to the UKB reference coefficients than ML estimates. The plots highlight the shrinkage effect of the coefficients towards zero, especially for the voxels with the lowest lesion incidence (Figure 2.B.5). This is the result of bias reduction for MeanBR $\tilde{\boldsymbol{\beta}}$ (see Kosmidis and Firth 2021) and the effect of the prior for BSGLMM $\boldsymbol{\beta}^*$.

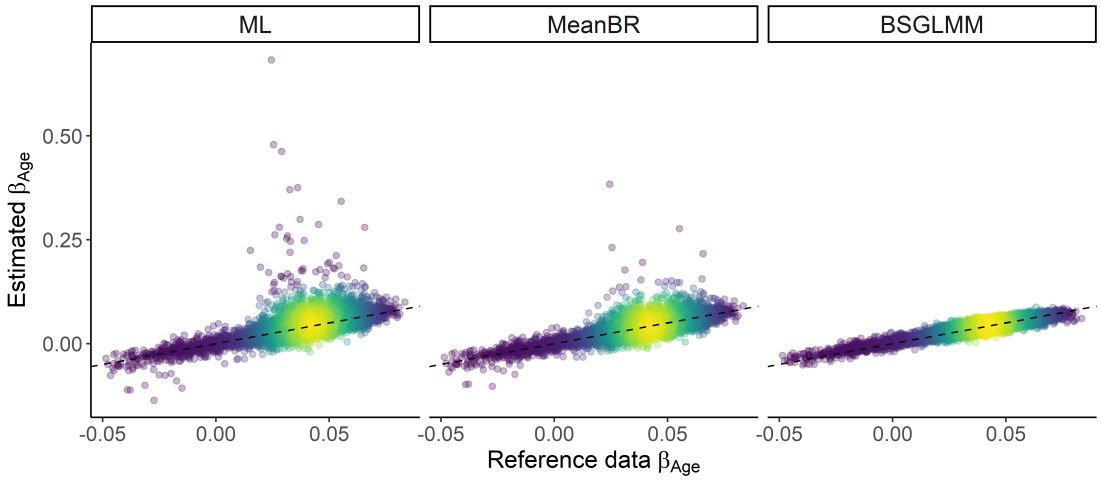


Figure 2.4: Estimated coefficients $\hat{\boldsymbol{\beta}}_{\text{Age}}$ (ML), $\tilde{\boldsymbol{\beta}}_{\text{Age}}$ (MeanBR), $\boldsymbol{\beta}^*_{\text{Age}}$ (BSGLMM) vs. $\boldsymbol{\beta}_{\text{Age}}$ (reference). Each point is coloured according to the density of points on an invisible grid overlaid on the plots (the brighter the colour, the higher the density of the points) and the identity superimposed (dashed black line). Bias reduction and the effect of the prior result in shrinkage of the coefficients towards zero with the Bayesian model following the equality line most closely. One simulated data set of 1,000 subjects used; 11,632 voxels with reference data lesion incidence $p > 0.005$ and finite MLEs plotted.

Accuracy We compare mean squared error (MSE), bias, probability of underestimation (PU) and correlation coefficient across bins of voxels for N^*=250 in Table 2.1 and for N^*=1,000 in Table 2.2. The summaries presented for the estimates across methods are conditional on the MLEs finiteness. The Bayesian method has better performance in terms of MSE and correlation due to the smoothness prior, which reduces estimator variance at the expense of higher bias. Note that Pearson's correlation is sensitive to outliers, thus the poor ML performance for low lesion incidence (e.g. large ML estimates as seen in Figure 2.4). The PU values suggest a slightly positively skewed estimates for the Bayesian method, i.e. tendency for overestimate the estimates. The opposite holds for the mass-univariate approaches, where we tend to underestimate the coefficients. According to standard asymptotic theory, all estimators should converge to a Normal distribution as the sample size increases, having PU of 50%, or equivalently being median unbiased. Reassuringly, increasing the sample size from 250 to 1000 subjects (Table 2.1 vs Table 2.2) gets PU closer to 50%. Overall, BSGLMM performs better for smaller sample size and for low lesion probability, but BSGLMM and MeanBR perform similarly for N^*=1,000.

If we further explore the spatial overlap between the highest M^* voxels, the DSCs across methods suggest good spatial overlap (Table 2.3). If we select a small number of voxels, i.e. voxels with the highest M^*=1,000 z-scores, all methods seem to detect the strongest age effect very well. The Bayesian method has the worst overlap between the three methods. We understand this to be a reflection of the BSGLMM's tendency to "bleed out" stronger effects into weaker effect areas, a problem perhaps more severe at N^*=250.

Table 2.1: Comparing methods across R=1,000 data sets, N^*=250 subjects each. The measures of accuracy are averaged across voxel bins based on the reference data (UKB) lesion incidence p with standard deviation in brackets. BSGLMM is more accurate in terms of MSE and correlation values, but has higher bias than MeanBR. All values are multiplied by 1000 except probability of underestimation (PU) represented in percentage and Pearson correlation ρ with range $(-1, 1)$.

Accuracy metrics	# voxels / Method	$p \in (0.005, 1]$ 11,634	$p \in (0.005, 0.01]$ 4,510	$p \in (0.01, 0.05]$ 4,981	$p \in (0.05, 0.1]$ 935	$p > 0.1$ 1,208
MSE	ML	43.67 (49.94)	93.63 (42.30)	17.15 (20.66)	0.24 (0.08)	0.11 (0.03)
	MeanBR	1.21 (1.08)	1.94 (1.13)	1.01 (0.75)	0.21 (0.05)	0.10 (0.03)
	BSGLMM	0.21 (0.13)	0.33 (0.13)	0.15 (0.05)	0.10 (0.02)	0.07 (0.01)
Bias	ML	32.12 (31.38)	61.92 (27.75)	18.36 (14.22)	2.10 (1.27)	0.81 (0.60)
	MeanBR	-1.07 (2.12)	-2.75 (2.36)	-0.03 (1.12)	0.10 (0.43)	0.02 (0.32)
	BSGLMM	1.62 (4.94)	2.30 (5.98)	1.43 (4.56)	0.92 (3.15)	0.35 (2.03)
PU (%)	ML	43.92 (4.01)	41.25 (3.84)	44.59 (2.94)	47.49 (1.97)	48.36 (1.72)
	MeanBR	57.42 (6.38)	61.30 (7.39)	56.34 (3.97)	52.45 (1.89)	51.25 (1.65)
	BSGLMM	43.97 (16.63)	41.47 (19.47)	44.54 (15.50)	46.75 (12.16)	48.78 (9.53)
ρ	ML	0.12 (0.013)	0.12 (0.018)	0.18 (0.033)	0.79 (0.014)	0.86 (0.009)
	MeanBR	0.46 (0.029)	0.34 (0.034)	0.49 (0.040)	0.80 (0.013)	0.86 (0.009)
	BSGLMM	0.80 (0.006)	0.75 (0.010)	0.81 (0.008)	0.87 (0.009)	0.89 (0.007)

Table 2.2: Comparing methods across R=1,000 data sets, N^*=1,000 subjects each. The measures of accuracy are averaged across voxel bins based on the reference data (UKB) lesion incidence p with standard errors in brackets. All values are multiplied by 1000 except probability of underestimation (PU) represented in percentage and Pearson correlation ρ with range $(-1, 1)$.

Accuracy metrics	# voxels / Method	$p \in (0.005, 1]$ 11,634	$p \in (0.005, 0.01]$ 4,510	$p \in (0.01, 0.05]$ 4,981	$p \in (0.05, 0.1]$ 935	$p > 0.1$ 1,208
MSE	ML	1.61 (14.45)	3.89 (22.76)	0.23 (3.36)	0.05 (0.01)	0.03 (0.01)
	MeanBR	0.22 (0.31)	0.40 (0.44)	0.15 (0.06)	0.05 (0.01)	0.03 (0.01)
	BSGLMM	0.07 (0.03)	0.08 (0.03)	0.06 (0.02)	0.04 (0.01)	0.02 (0.01)
Bias	ML	3.13 (3.32)	5.87 (3.68)	1.85 (1.24)	0.46 (0.35)	0.21 (0.20)
	MeanBR	0.16 (0.50)	0.31 (0.65)	0.08 (0.39)	0.01 (0.22)	0.01 (0.15)
	BSGLMM	0.80 (2.95)	1.13 (3.86)	0.75 (2.49)	0.31 (1.34)	0.13 (0.81)
PU (%)	ML	47.34 (2.23)	46.27 (2.25)	47.61 (1.90)	48.83 (1.74)	49.10 (1.59)
	MeanBR	53.22 (2.68)	54.79 (2.68)	52.83 (2.08)	51.27 (1.61)	50.53 (1.56)
	BSGLMM	46.24 (13.29)	44.75 (16.21)	46.48 (11.96)	48.41 (8.91)	49.15 (6.96)
ρ	ML	0.58 (0.144)	0.45 (0.134)	0.81 (0.029)	0.94 (0.003)	0.96 (0.002)
	MeanBR	0.76 (0.022)	0.65 (0.032)	0.83 (0.006)	0.94 (0.003)	0.96 (0.002)
	BSGLMM	0.90 (0.002)	0.85 (0.004)	0.90 (0.003)	0.95 (0.003)	0.96 (0.002)

Table 2.3: Dice similarity coefficient (DSC) when comparing reference (UKB) and simulation z-scores estimated across the three regression methods. DSCs are obtained across R=1,000 data sets, $N^* \in \{250, 1000\}$ subjects each and the spatial overlap considered is between the highest M^* z-scores, where $M^* \in \{1000, 5000, 10000\}$.

N^*	Method	Dice similarity coefficient		
		M^*=1,000	M^*=5,000	M^*=10,000
$N^* = 250$	ML	0.824	0.774	0.748
	MeanBR	0.821	0.756	0.718
	BSGLMM	0.740	0.704	0.736
$N^* = 1000$	ML	0.907	0.871	0.857
	MeanBR	0.907	0.868	0.848
	BSGLMM	0.888	0.813	0.815

False positive control To further compare the methods in terms of false positive detection, we simulate a single data set with no age effect added to the true effect component in Steps 1 and 2, Section 2.2.2.1 (only reference data (UKB) intercept map used), but we add age $\boldsymbol{X}_2^*$ as a covariate when fitting all three models. We explore the same accuracy metrics as in Tables 2.1 and 2.2 by setting $\boldsymbol{\beta}(s_j) = 0$ for all voxels s_j and the Bayesian method performs best in terms of lowest MSE and the percentage of underestimation is very close to 50%, which is to be expected if the estimates are symmetric around zero (Table 2.4). Interestingly, the effect of the prior in the Bayesian method (shrinkage towards zero) reduces the bias to be smaller or comparable to the MeanBR method since the true effect is set to zero in this case. We observe (Table 2.5) that all methods appear to be conservative in their false positive rate especially for low lesion incidence voxels with the Bayesian method always being most conservative. This is also evident from the quantile–quantile plots (see Figure 2.B.6), where as lesion incidence decreases, the normality deviations increase.

Table 2.4: Comparing methods across one null age effect data set of N^*=1,000 subjects. The measures of accuracy are averaged across voxel bins based on the reference data (UKB) lesion incidence p. All values are multiplied by 1000 except probability of underestimation (PU) represented in percentage.

Accuracy metrics	# voxels / Method	$p \in (0.005, 1]$	$p \in (0.005, 0.01]$	$p \in (0.01, 0.05]$	$p \in (0.05, 0.1]$	$p > 0.1$
		11,596	4,473	4,978	934	1,211
MSE	ML	12.08	31.08	0.19	0.04	0.02
	MeanBR	0.17	0.28	0.13	0.04	0.02
	BSGLMM	0.03	0.03	0.03	0.02	0.01
Bias	ML	-0.62	-1.62	-0.02	0.05	0.07
	MeanBR	0.28	0.59	0.10	0.07	0.08
	BSGLMM	0.12	0.14	0.10	0.09	0.09
PU (%)	ML	49.39	49.06	49.66	49.84	49.17
	MeanBR	49.14	48.66	49.46	49.73	49.17
	BSGLMM	49.14	48.73	49.32	50.48	48.92

Table 2.5: False positive rate evaluation. Number of voxels with z-scores significant at 5% for a two-sided test for age when no age effect is included. Percentage is calculated within each bin (column) based on the reference data (UKB) lesion incidence p. All methods appear to be conservative since we would expect 5% false positives, i.e. about 700 voxels across each row. These results are based on one simulated data set of 1,000 subjects used, 11,596 voxels with lesion incidence in the reference data greater than 0.005 and infinite MLEs discarded. Note, for this one simulated data set a 5% FDR correction found no significant voxels for any method.

$N(\lvert \mathbf{z} \rvert > 1.96)$ / Method	% voxels (# voxels)				
	$p \in (0.005, 1]$	$p \in (0.005, 0.01]$	$p \in (0.01, 0.05]$	$p \in (0.05, 0.1]$	$p > 0.1$
	11,596	4,473	4,978	934	1,211
ML	3.6% (422)	2.1% (95)	4.5% (225)	5.5% (51)	4.2% (51)
MeanBR	3.5% (406)	2.3% (104)	4.1% (205)	5.2% (49)	4.0% (48)
BSGLMM	1.6% (182)	0.4% (17)	1.8% (88)	3.6% (34)	3.6% (43)

2.3.1.3 Computational time and scalability

On average, ML and MeanBR take about 15–20 min for each 250-subject data set and about 50–60 min for each 1,000-subject data set for single-core jobs, and 3–4 min and 10–12 min for parallel jobs (8 cores), respectively. The difference in computational cost between ML and MeanBR is minimal with bias-reduction increasing the computational cost by only a few minutes for a 1000-subject data set. Note that the number of regressions per simulated data set varies depending on the number of non-zero lesion incidence voxels. 23,404 regressions are performed on average per 250-subject data set and 40,286 per 1000-subject data set, respectively, instead of 228,483 (voxels in the brain mask). Our code determines the voxels with non-zero lesion incidence first and creates a matrix of binary values only for those voxels to be used as input to the GLMs. This trick saves computation time, but also allows better RAM management for big UKB-scale data sets since it avoids reading in all lesion masks at once. For the simulated data sets this implementation might not be optimal in terms of speed, but it makes the UKB application possible even without parallel implementation.
BSGLMM takes about 16 min for 100,000 iterations of the Gibbs sampler for a 250-subject data set and about 60 min for a 1,000-subject data set, respectively. BSGLMM is performed on an NVIDIA TESLA K80 GPU card with 12 GB RAM and 2,496 threads. While the BSGLMM run time is comparable to the ML and MeanBR, note that there is a practical upper limit of subjects due to a GPU RAM constraint; the problem arises since the Bayesian method implementation loads all binary masks limiting its application to UKB-scale data.

To summarise, while BSGLMM's GPU implementation is computationally efficient, the ML and MeanBR have more flexibility in how parallelism can be used, making the latter easier to apply at biobank scale.

2.3.2 Results on the real data

We choose to fit the mass-univariate voxel-wise GLM with MeanBR estimates due to its scalability to the UK Biobank data set of 13,680 subjects, but also obtain MLEs to check how often separation occurs. The models we fit include systolic BP as the main effect of interest and age, sex, age by sex interaction and head size scaling as confounders. On

72,603 regressions across the brain (voxels with non-zero lesion incidence), sex MLEs are infinite for 23,330 voxels (32%); separation is more likely to occur for binary covariates and, for example, systolic BP MLEs are infinite for only 260 voxels (0.4%).

Spatial distribution of lesions The lesion incidence across 13,680 UKB participants suggests the areas with the highest probabilities cover the periventricular and deep white matter regions (Figure 2.5). Fitting voxel-wise GLMs with systolic BP as our main covariate of interest, we explore its effect on lesion probability (Figures 2.5 and 2.B.7). Figure 2.5 includes axial slices of z-scores for the effect of systolic BP (right) along with the UKB lesion probability (left); the darker the colour, the stronger the effect of systolic BP on lesion probability. The spatial distribution of lesions mirrors what is well known clinically, that is lesions are classically found capping the ventricles, clustering around the ventricles and within the deep white matter (Fazekas et al., 1987). Hypertension is known to be one of the strongest predictors of the presence of lesions (Dufouil et al., 2001). Consistent with the literature, we find hypertension related lesions distributed in periventricular and deep white matter regions as well as capping the ventricles (Moroni et al., 2018). We get 14,108 voxels with z-scores greater than 1.96 in absolute value (in comparison, 11,251 for z-scores based on MLEs, respectively). Thus, systolic BP has a strong effect on lesion probability as expected based on the existing literature.

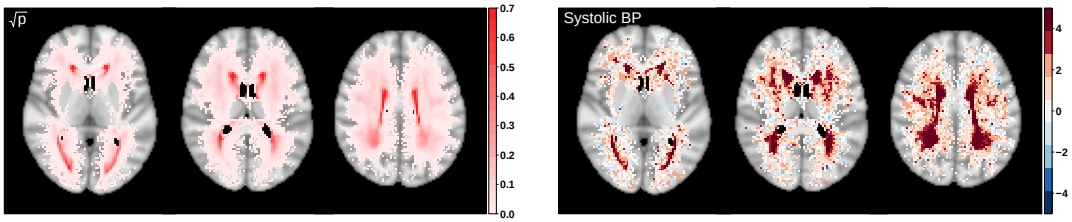


Figure 2.5: Square-root transformed lesion probability based on 13,680 UKB participants $\sqrt{p}$ (left panel) and significance maps (z-scores based on MeanBR estimates) for the effect of systolic BP (right panel); age, sex, age by sex interaction and head size scaling included as confounders. 72,603 voxels have non-zero lesion probability; axial slices $\{40, 45, 50\}$.

Computational time ML and MeanBR take about 3 h and 3.5 h, respectively, utilizing batch jobs run on 8 cores. As mentioned in Section 2.3.1.3, we use the empirical incidence mask to select the non-zero incidence voxels and then run the GLMs for those voxels only, here 72,603 regressions.

2.4 Discussion

Simulation framework Using binary lesion masks of 13,680 healthy aging UK Biobank participants as our reference data set, we develop a binary lesion mask simulator. Age is used as the sole regressor and by using a reference data set, we start building our simulation study by setting the true age coefficient map to the UKB-derived one. In other simulation studies, binary lesion masks or 2D slices are simulated, but the true coefficients are not available (Sundaresan et al., 2019; Ge et al., 2014), which does not allow any comparison between competing methods.

We made the artificial lesion masks as realistic as possible through tuning to make sure the artificial and real lesion masks share important lesion characteristics, such as lesion size, lesion count and lesion volume. This step potentially overcomes the drawbacks of other simulation approaches, where the same count, size and shape lesions are simulated (6 spheroid lesions of size 5 voxels in each dimension Chard et al. 2010) or smoothing of the resulting simulated lesion masks (Sundaresan et al., 2019) is applied, which could introduce stronger spatial dependencies than what is expected from real lesion masks.

Our simulator code is available[6] and ready to use to simulate binary lesion masks for healthy aging individuals. However, if the simulation framework is to be adjusted to any patient reference data set, e.g. Dementia patients, binary lesion masks and age for those patients are needed to obtain the coefficient maps and to tune the simulator as described in Section 2.2.2.2.

Method comparison We compare three alternative regression approaches for modelling of binary lesion masks. Two of them rely on voxel-wise fitting of generalized linear models using maximum likelihood and mean bias-reduction. The other is a Bayesian hierarchical model that takes into account the spatial dependence in the brain though the inclusion of spatially varying coefficients.

The bias and mean squared error of the maximum likelihood and mean bias-reduced coefficients suggest poorer performance of the maximum likelihood estimator, which is in line with the widely dispersed MLEs in the coefficient plots (Figure 2.4). BSGLMM

[6]Project URL: `https://osf.io/h7sxr`

seems to perform slightly better in terms of mean-squared error values, but has higher bias than mean bias-reduced estimates due to the spatial regularization it imposes. When comparing the ability of the methods to detect the voxels with the strongest age effect on lesion probability, all three methods seem to perform similarly well. Null simulations find that for a non-existent age effect all methods are valid but conservative, with false positive rates lowest for low incidence voxels.

The resources required to apply those methods vary significantly since the Bayesian spatial model utilises a GPU implementation to decrease the computational burden. For the size of the simulated data sets, all methods are relatively fast to perform with about an hour run-time for one data set of sample size 1,000 subjects (single core for mass-univariate). However, for the spatial model there is a practical upper limit on the number of subjects due to the GPU RAM constraint since all lesion masks need to be loaded in memory. On the contrary, for the mass-univariate methods, parallel implementation is possible given that the methods are applied independently at each voxel. Thus, mass-univariate approaches are computationally practical for large data sets.

UK Biobank application Reassuringly, the distribution of lesions in the real data reflects the known distribution of lesions associated with age and hypertension (Dufouil et al., 2001). Further work by our group demonstrates the clinical utility of the mass-univariate method (mean bias-reduced estimates) in mapping the spatial distribution of lesions associated with different cerebrovascular risk factors (Veldsman et al., 2020). Application of the mass-univariate methods (ML and MeanBR) to lesion masks on 13,680 subjects demonstrates that total separation occurs quite often for binary covariates (32% of voxels have infinite sex estimates) even in such big data sets, thus mean bias-reduced estimates would be favoured. The run-time of about 3 h suggests that voxel-wise modelling is feasible for large data sets; heavier parallelism (we use a maximum of 8 cores) can reduce run-time substantially.

Limitations Our simulation framework is not adapted for automated tuning, i.e. a grid of scale values for the Gaussian Random Field are explored. An automated procedure could be developed but the merits might not outweigh the computational effort. Further improvement could be introduced by allowing the GRF scale parameter to vary across

age groups to achieve a closer match to the suggested empirical lesion summaries. To match the variability in the reference data better, a more flexible covariance function than the squared exponential (e.g. Matern at the expense of an extra parameter to tune) or a non-stationary GRF might need to be adopted. However, our goal is to provide a simulation framework for the comparison of lesion mapping methods and we believe that matching the median lesion summaries across age groups is sufficient for the fair comparison of the three approaches and any alternatives that may result from future research.

Note that we do not account for any left-right symmetry of lesions, we have not imposed any physiological boundaries or 3D dependence in the entire brain when simulating the lesion masks. However, we do not believe this has any impact on the results presented here since the mass-univariate approaches do not account for the spatial dependence in the brain and the spatial model only accounts for local spatial dependence. We refer to the lesion masks as 'realistic' but this is not meant to imply any clinical realism (given the drawbacks mentioned) and we see the lesion masks as useful in a methods development or methods comparison context.

We generate the lesion masks in MNI space by using outputs from the published UK Biobank pipeline (Alfaro-Almagro et al., 2018). Sensitivity analysis to registration or lesion segmentation approaches could be of future interest but it is out of the scope of the current statistical work since we focus on masks in MNI space for the design of the simulation framework. The proposed lesion mask simulator could be tuned to reflect features of lesions independent of the image resolution, but the method comparison results presented are specific to the sampling resolution of 2mm^3 voxels and we have not performed sensitivity analysis to other voxel sizes.

Note that the lack of scalability of the Bayesian approach is due to the GPU memory constraint and it could be overcome by either a time-consuming CPU implementation of the Gibbs sampler proposed by Ge et al. (2014), or by adopting a divide-and-conquer method for Bayesian inference. The latter involves splitting the data into smaller subsets (computationally manageable), sampling from the posterior distribution on all subsets and then combining the posterior samples to approximate the full data posterior, where possible methods include the ones suggested by Srivastava et al. (2018); Minsker et al. (2017); Li et al. (2017). We have focused our method comparison on the implementation

available instead.

Investigating the effect of systolic blood pressure on lesion probability, we present test statistics at all non-zero lesion incidence voxels to demonstrate the scalability of the method. We could have excluded voxels where the lesion incidence fell too low and then use false discovery rate correction to account for multiple testing (Veldsman et al., 2020) to achieve better inference.

Conclusion The proposed simulation framework mimics real features of the data, which allows for a fair comparison between the lesion mapping methods through realistic experiments. Our findings suggest that bias-reduced estimates for voxel-wise binary-response generalized linear models overcome the instabilities of maximum likelihood estimates, and scale well for large data sets due to parallel implementation. Contrary to the assumption of spatial dependence being key in lesion mapping, our results show that voxel-wise bias-reduction and spatial modelling result in largely similar estimates, but bias-reduction is computationally feasible for biobank-scale neuroimaging data.

Credit authorship contribution statement

Petya Kindalova: Conceptualization, Data curation, Formal analysis, Methodology, Validation, Visualization, Writing - original draft, Writing - review & editing. **Ioannis Kosmidis:** Conceptualization, Funding acquisition, Methodology, Supervision, Writing - review & editing. **Thomas E. Nichols:** Conceptualization, Funding acquisition, Methodology, Supervision, Writing - review & editing.

Acknowledgements

Thank you to Timothy Johnson for his valuable input on the testing of our simulation framework. We would also like to thank Michele Veldsman for her help with the clinical interpretations of the real data analysis.

Appendices

2.A Iterative estimation: maximum likelihood and bias-reduction

2.A.1 Maximum likelihood estimates

The typical iterative algorithm used to find the maximum likelihood estimates (MLEs) for generalized linear models (GLMs) is iteratively reweighted least squares (IRLS) (Green, 1984). IRLS is equivalent to Fisher scoring obtain an iterative solution to the estimating equations (also known as score equations)

$$\frac{\partial l}{\partial \boldsymbol{\beta}} = \left(\frac{\partial l}{\partial \beta_1}, \ldots, \frac{\partial l}{\partial \beta_P}\right)^\top = U(\boldsymbol{\beta}) = \mathbf{0}, \tag{2.A.1}$$

where l is the log-likelihood, $U(\boldsymbol{\beta})$ is the score vector (p-vector). A Taylor series expansion for $\partial l/\partial \boldsymbol{\beta}$ (Eq. (2.A.1)) gives the standard Newton-Raphson method for solving the estimating equations

$$\boldsymbol{\beta}^* \approx \boldsymbol{\beta} + \left[\frac{-\partial^2 l}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^\top}\right]^{-1} U(\boldsymbol{\beta}) = \boldsymbol{\beta} + [J(\boldsymbol{\beta})]^{-1} U(\boldsymbol{\beta}), \tag{2.A.2}$$

where $J(\boldsymbol{\beta}) = -\partial^2 l/\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^\top$ is the observed information matrix, $\boldsymbol{\beta}$ is the initial value of the parameters and $\boldsymbol{\beta}^*$ is the updated value. Evaluation of U and J is repeated until convergence and the resulting estimates are the MLEs we report in the paper denoted as $\hat{\boldsymbol{\beta}}$.

If we replace the observed information $J(\boldsymbol{\beta})$ with the expected information (Fisher information) $I(\boldsymbol{\beta}) = \mathbb{E}(J(\boldsymbol{\beta}))$ in the Newton–Raphson, the Fisher scoring iteration results. Fisher scoring is typically preferred since the Fisher information is useful post-hoc to estimate the asymptotic variance of the parameters. Note that for canonical link (e.g. logit link function for Binomial GLMs), observed and expected information coincide, hence Fisher scoring is equivalent to Newton–Raphson.

2.A.2 Bias-reduced estimates

The bias-correction method we use to obtain mean bias-reduced (MeanBR) estimates $\tilde{\boldsymbol{\beta}}$ was first introduced in Firth (1993) and was then applied and developed further for exponential family models (Kosmidis and Firth, 2009; Kosmidis et al., 2020). The method is known as adjusted score equations, i.e. a penalty $A(\boldsymbol{\beta})$ is added to the score equations in Equation (2.A.1) in order to get estimates with asymptotically smaller bias

$$U^*(\boldsymbol{\beta}) = U(\boldsymbol{\beta}) \; + \; A(\boldsymbol{\beta}) = \mathbf{0}, \tag{2.A.3}$$

where $A(\boldsymbol{\beta})$ is a p-vector based on the expected information matrix $I(\boldsymbol{\beta})$ and on the observed information $J(\boldsymbol{\beta})$. General formulae for the adjusted score equations are derived by Kosmidis and Firth (2009), showing that solving the mean bias-reducing score functions by iterative optimization (e.g. IRLS) results in higher-order mean unbiased estimators. What is interesting is that the general form of the first order bias is of the form

$$\frac{b_1(\boldsymbol{\beta})}{N} = -[I(\boldsymbol{\beta})]^{-1} A(\boldsymbol{\beta}), \tag{2.A.4}$$

where the mean bias function $B(\boldsymbol{\beta})$ of the MLE of $\boldsymbol{\beta}$ can be expanded in decreasing powers of N as

$$B(\boldsymbol{\beta}) = \mathbb{E}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) = \frac{b_1(\boldsymbol{\beta})}{N} + \frac{b_2(\boldsymbol{\beta})}{N^2} + \frac{b_3(\boldsymbol{\beta})}{N^3} + O(N^{-4})$$

for an appropriate set of functions $b_1(\boldsymbol{\beta}), b_2(\boldsymbol{\beta}), \ldots$, which are $O(1)$ as $N \to \infty$. Thus, the adjustment to the score functions $A(\boldsymbol{\beta})$ is a function of the first-order bias and the Fisher information, i.e. iteratively subtracting the first-order bias in the Fisher scoring updates (Kosmidis and Firth, 2010). The iterative procedure from Equation (2.A.2) becomes a quasi Fisher scoring to obtain MeanBR estimates

$$\boldsymbol{\beta}^* \approx \boldsymbol{\beta} \; + \; [I(\boldsymbol{\beta})]^{-1} U^*(\boldsymbol{\beta}) \, . \tag{2.A.5}$$

Here it is 'quasi' since we are using the expectation of the second derivatives of the scores $U(\boldsymbol{\beta})$, instead of the second derivative of the adjusted scores $U^*(\boldsymbol{\beta})$. Note that

the iterated first-order bias adjustment is only possible when $b_1(\boldsymbol{\beta})$ is available in closed-form.

2.B Supplementary figures and tables.

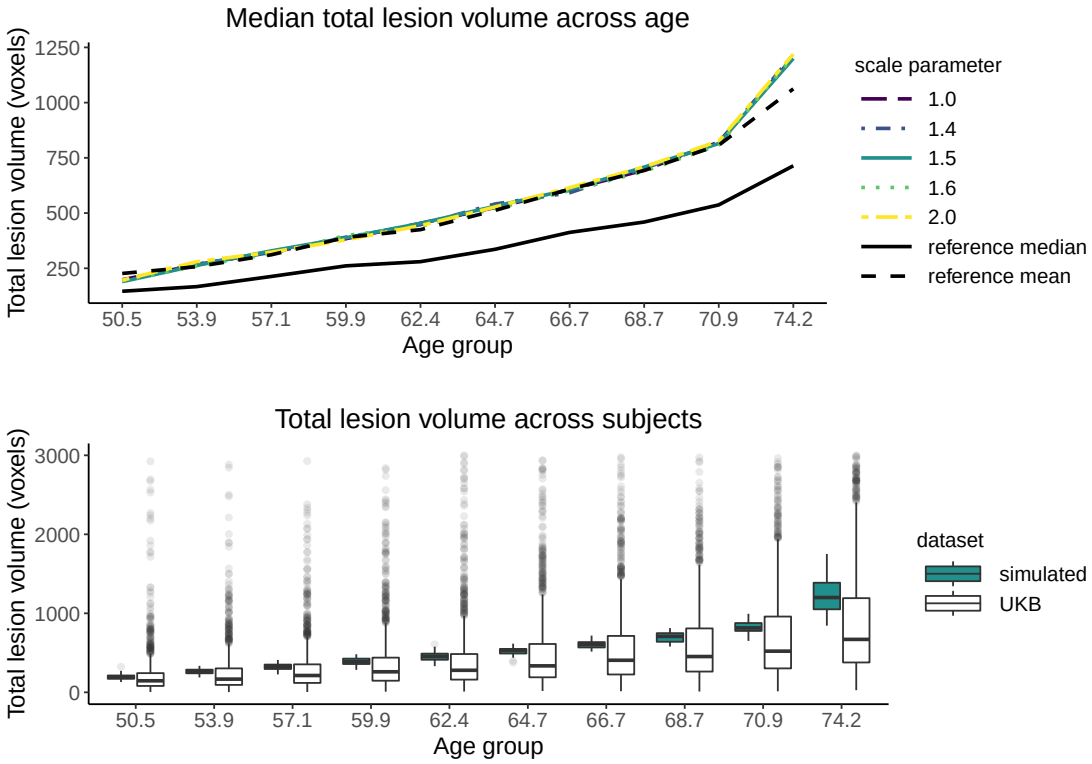


Figure 2.B.1: Gaussian random field parameter tuning by matching the reference data (UKB) median total lesion volume across age bins (black solid line). (Top) Plot of median total lesion volume across age bins for five simulation settings (five GRF scale parameter values) and reference data values (black lines). Legend values indicate the scale parameter value ℓ used to simulate a GRF for each subject in the simulated sample. (Bottom) Boxplots of total lesion volume in UKB participants (white) and in one simulated 1000-subject sample with GRF scale parameter ℓ=1.5 (blue) across ten age bins. Note the x-axis labels denote the center of each age bin, the y-axis units are in 2mm^3 voxels, and the variance GRF parameter is fixed to 1 for all simulation settings.

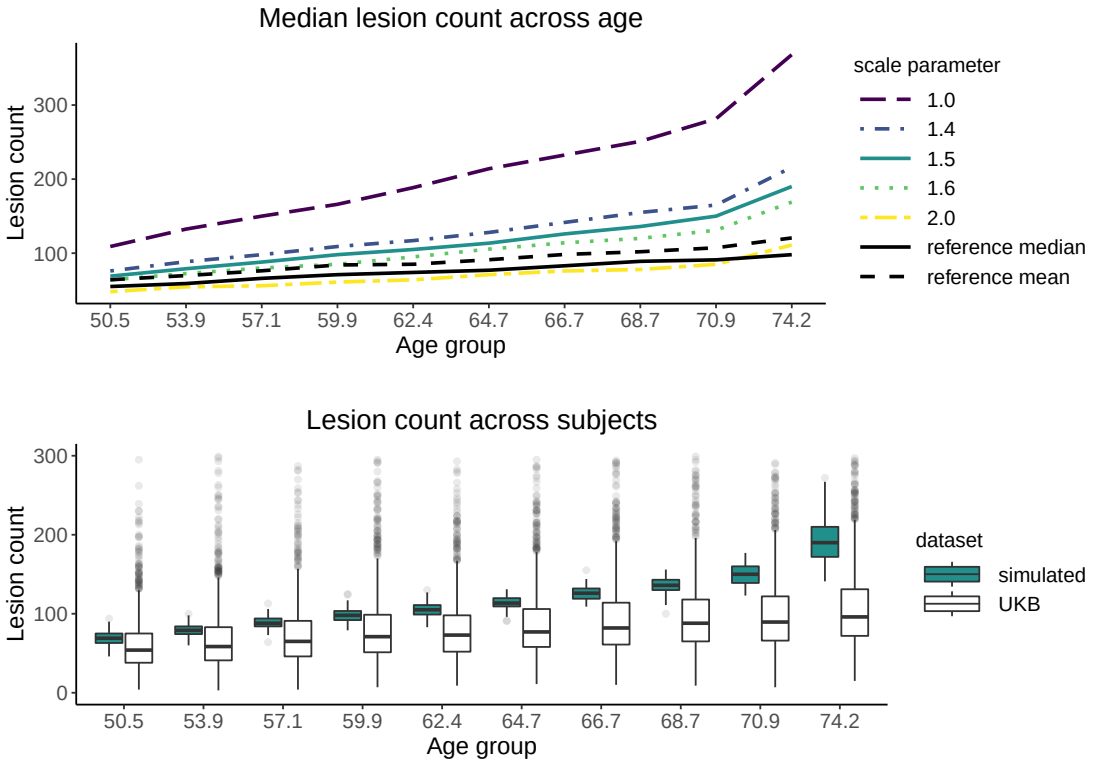


Figure 2.B.2: Gaussian random field parameter tuning by matching the reference data (UKB) median lesion count across age bins (black solid line). (Top) Plot of median lesion count across age bins for five simulation settings (five GRF scale parameter values) and reference data values (black lines). Legend values indicate the scale parameter value ℓ used to simulate a GRF for each subject in the simulated sample. (Bottom) Boxplots of lesion count in UKB participants (white) and in one simulated 1000-subject sample with GRF scale parameter ℓ=1.5 (blue) across ten age bins. Note the x-axis labels denote the center of each age bin, the y-axis units are in number of connected components (lesions), and the variance GRF parameter is fixed to 1 for all simulation settings.

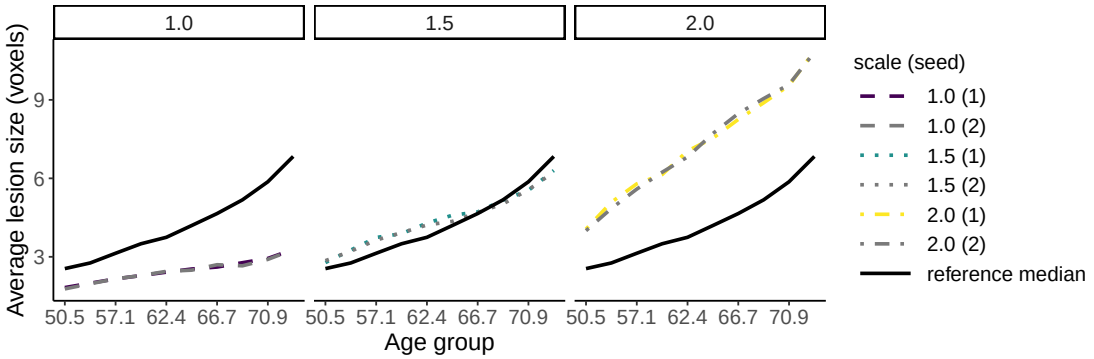


Figure 2.B.3: Gaussian random field parameter tuning by matching the reference data (UKB) median average lesion size across age bins (black solid line) replicated for two seeds. Legend values indicate the scale parameter value ℓ used to simulate a GRF for each subject in the simulated sample and the seed in brackets. The lesion summaries do not vary substantially between seeds. Note the x-axis labels denote the center of the age bins, the y-axis units are in 2mm^3 voxels, and the variance GRF parameter is fixed to 1 for all simulation settings.

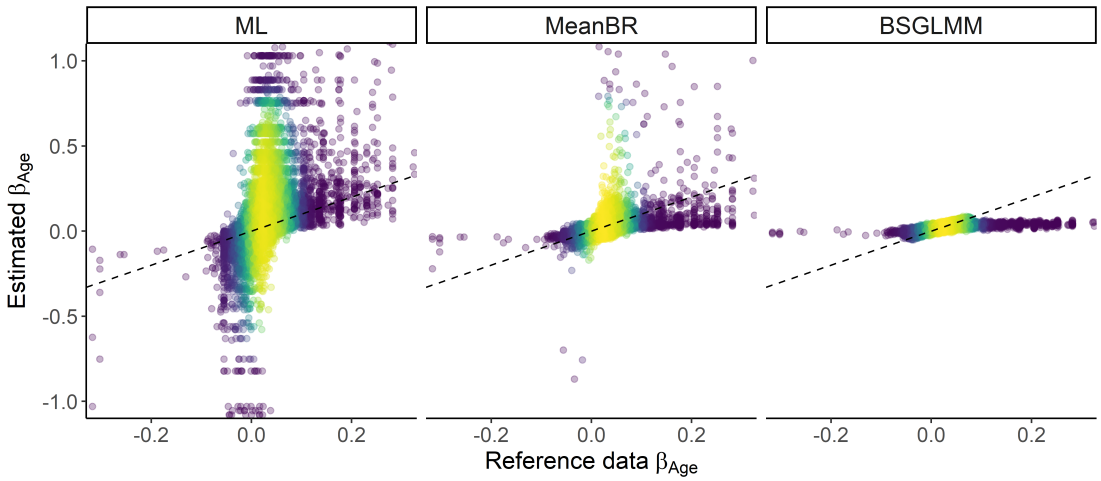


Figure 2.B.4: Estimated coefficients $\hat{\boldsymbol{\beta}}_{\text{Age}}$ (ML), $\tilde{\boldsymbol{\beta}}_{\text{Age}}$ (MeanBR), $\boldsymbol{\beta}^*_{\text{Age}}$ (BSGLMM) vs. $\boldsymbol{\beta}_{\text{Age}}$ (reference). Each point is coloured according to the density of points in an invisible grid overlaid on the plots (the brighter the colour, the higher the density of the points) and the identity superimposed (dashed black line). Bias reduction and the effect of the prior result in shrinkage of the coefficients towards zero with the Bayesian model following the equality line most closely. The 'horizontal effect' observed mostly at the BSGLMM plot (826 voxels have reference data coefficients greater than 0.1 in absolute value) occurs when the lesion incidence is very low. One simulated data set of 1000 subjects used; 40,338 voxels with finite MLEs plotted.

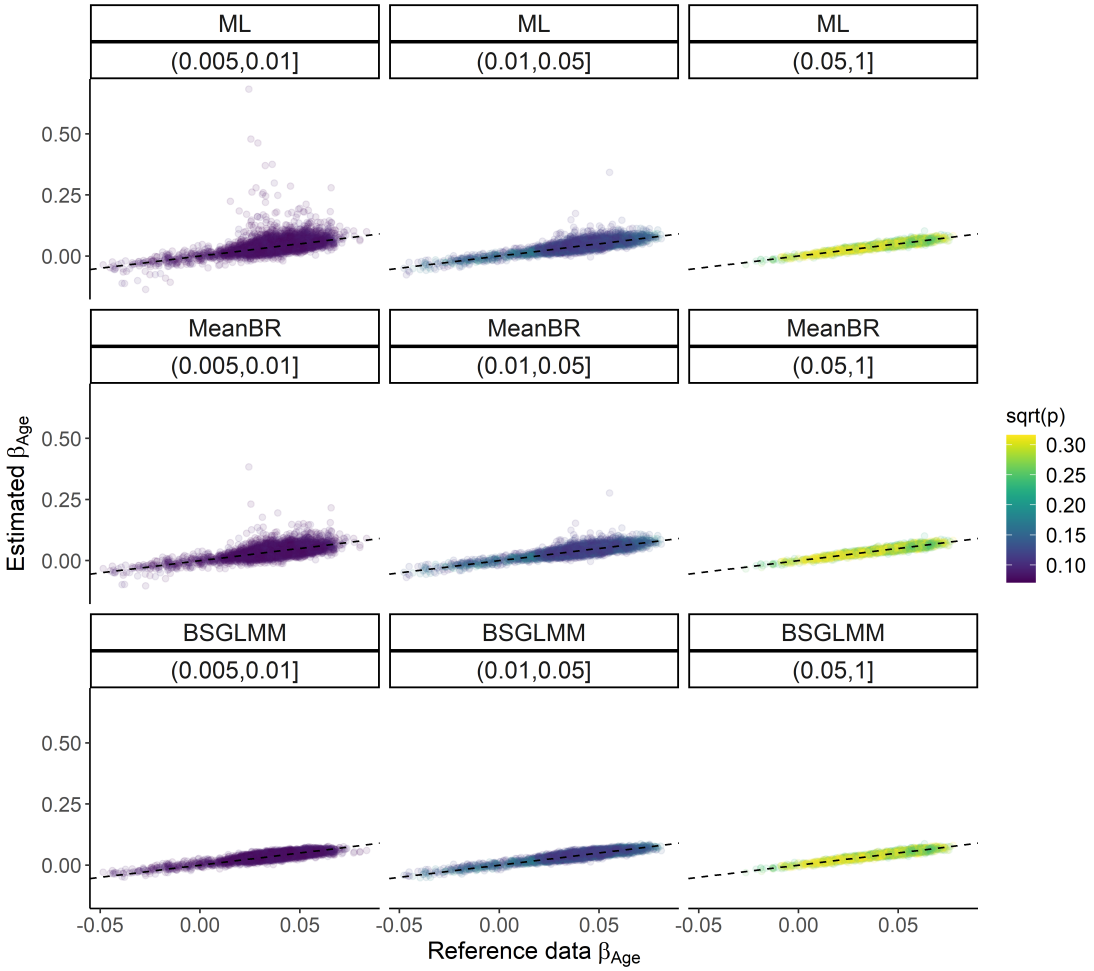


Figure 2.B.5: Estimated coefficients $\hat{\boldsymbol{\beta}}_{\text{Age}}$ (ML), $\tilde{\boldsymbol{\beta}}_{\text{Age}}$ (MeanBR), $\boldsymbol{\beta}^*_{\text{Age}}$ (BSGLMM) vs. $\boldsymbol{\beta}_{\text{Age}}$ (reference) across bins of voxels. Each point is coloured according to the square-root lesion probability $\sqrt{p}$ suggesting shrinkage is observed for voxels with low lesion incidence. One simulated data set of 1000 subjects used; 11,632 voxels with reference data lesion incidence $p > 0.005$ and finite MLEs plotted.

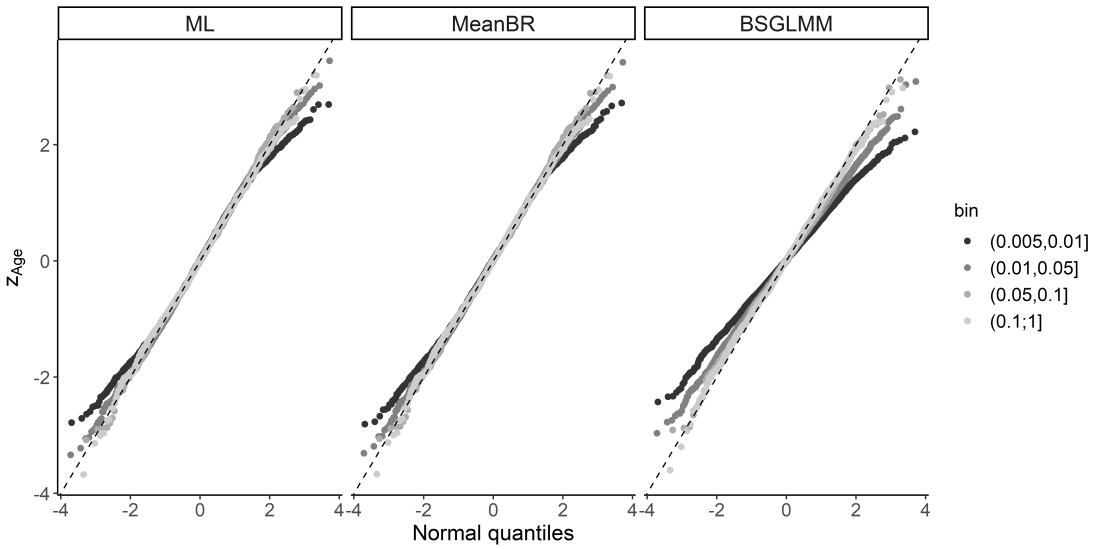


Figure 2.B.6: Quantile–quantile (QQ) plots of the quantiles of the simulated data z-scores across bins of voxels versus the theoretical quantiles from a Normal distribution. The lower the lesion incidence (darker colour), the greater the deviations from a linear trend, i.e. the rarer the lesions, the greater the deviations from normality.

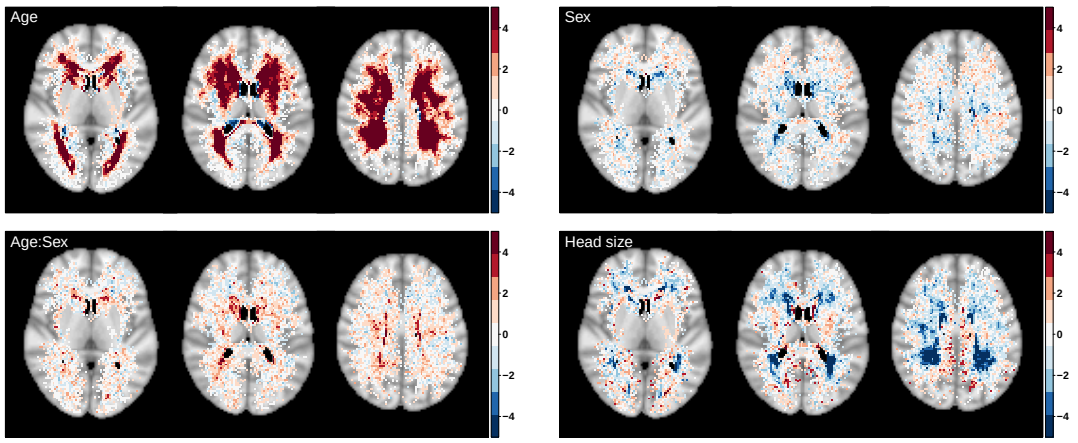


Figure 2.B.7: Significance maps (z-scores based on MeanBR estimates) for the effect of age, sex (baseline men), age by sex interaction and head size scaling to complement Figure 5. Data on 13,680 UK Biobank participants used. 72,603 voxels with non-zero lesion probability shown with zero-lesion incidence voxels plotted as transparent to show anatomical MRI for reference; axial slices $\{40, 45, 50\}$ shown.

Chapter 3

Spatial distribution and cognitive impact of cerebrovascular risk-related white matter hyperintensities

The attached paper is published in *NeuroImage: Clinical* (Veldsman et al., 2020) and it is based on joint work with Michele Veldsman, Masud Husain, Ioannis Kosmidis and Thomas E. Nichols. A statement of contribution is published as part of the paper.

Abstract

Objectives White matter hyperintensities (WMHs) are considered macroscale markers of cerebrovascular burden and are associated with increased risk of vascular cognitive impairment and dementia. However, the spatial location of WMHs has typically been considered in broad categories of periventricular versus deep white matter. The spatial distribution of WHMs associated with individual cerebrovascular risk factors (CVR), controlling for frequently comorbid risk factors, has not been systematically investigated at the population level in a healthy ageing cohort. Furthermore, there is an inconsistent relationship between total white matter hyperintensity load and cognition, which may be due to the confounding of several simultaneous risk factors in models based on smaller cohorts.

Methods We examined trends in individual CVR factors on total WMH burden in 13,680 individuals (aged 45-80) using data from the UK Biobank. We estimated the spatial distribution of white matter hyperintensities associated

with each risk factor and their contribution to explaining total WMH load using voxel-wise probit regression and univariate linear regression. Finally, we explored the impact of CVR-related WMHs on speed of processing using regression and mediation analysis.

Results Contrary to the assumed dominance of hypertension as the biggest predictor of WMH burden, we show associations with a number of risk factors including diabetes, heavy smoking, APOE $\varepsilon 4/\varepsilon 4$ status and high waist-to-hip ratio of similar, or greater magnitude to hypertension. The spatial distribution of WMHs varied considerably with individual cerebrovascular risk factors. There were independent effects of visceral adiposity, as measured by waist-to-hip ratio, and carriage of the APOE $\varepsilon 4$ allele in terms of the unique spatial distribution of CVR-related WMHs. Importantly, the relationship between total WMH load and speed of processing was mediated by waist-to-hip ratio suggesting cognitive consequences to WMHs associated with excessive visceral fat deposition.

Conclusion Waist-to-hip ratio, diabetes, heavy smoking, hypercholesterolemia and homozygous APOE $\varepsilon 4$ status are important risk factors, beyond hypertension, associated with WMH total burden and warrant careful control across ageing. The spatial distribution associated with different risk factors may provide important clues as to the pathogenesis and cognitive consequences of WMHs. High waist-to-hip ratio is a key risk factor associated with slowing in speed of processing. With global obesity levels rising, focused management of visceral adiposity may present a useful strategy for the mitigation of cognitive decline in ageing.

3.1 Introduction

White matter hyperintensities (WMHs) of presumed vascular origin (Wardlaw et al., 2013) are widely recognised as an indicator of poor brain health (Wardlaw et al., 2015). Age remains the strongest predictor for the presence of WMHs. However, the total burden of WMHs is higher in individuals with cerebrovascular risk (CVR) factors, like hypertension or hypercholesterolemia. WMHs triple the risk of stroke and double the risk of dementia suggesting they reflect pathological processes and are not simply a consequence of ageing (Debette and Markus, 2010). A number of studies have examined the relationship between different CVR factors and total WMH burden (Debette and Markus, 2010). Hypertension usually emerges as the dominant risk factor (Debette and Markus, 2010). Beyond this, it is less clear which risk factors are associated with the presence of WMHs in different regions of the brain, and with impaired cognition, when controlling for other risk factors. In other words, there is not sufficient evidence as to which risk factors make an independent contribution to WMH spatial distributions across the brain and associated cognitive impairment. This is important because it may change the focus of clinical management of risk factors beyond control of blood pressure.

Since the first visual scales attempting to quantify WMH burden (Fazekas et al., 1987), it has been recognised that WMHs disproportionately fall within periventricular (PV-WMHs) areas or in deep white matter regions (D-WMHs). Classification in this way has proven useful because deep and periventricular WMHs have different underlying microstructure, different associations with CVR factors (Griffanti et al., 2018) and potentially different relationships to cognition (Mortamais et al., 2013). Although pathology studies are relatively rare compared to imaging studies, there is some evidence of different pathological processes underlying WMHs in different regions (Wardlaw et al., 2015). PV-WMHs are associated with cerebral ischaemia and demyelination of adjacent fibre tracts as well as ependymal loss around the ventricles (Fazekas et al., 1993; Kim et al., 2008). PV-WMHs capping the ventricles are thought to be non-ischaemic in nature and reflect more generalised gliosis (Fazekas et al., 1998). In contrast, D-WMHs are thought to be of more ischaemic origin, the degree of confluence reflecting the degree of ischaemic damage with the most severe being marked loss of fibres and arteriolosclerosis (Fazekas et al., 1993). The spatial variability of WMHs associated with individual

CVR factors has been investigated qualitatively using visual rating scales or broad region of interest approaches (De Leeuw et al., 2001; Strassburger et al., 1997; Van Dijk et al., 2004). There has been much less investigation of WMHs at a whole brain level, specifically looking at the probability of the presence of WMHs for a given risk factor, voxel-wise. Beyond the deep and periventricular classification, there may be important clues in the spatial distribution of WMHs that explain their pathogenesis and contribution to vascular cognitive impairment.

There are conflicting reports over whether hypertension, thought to be the strongest predictor of total WMH burden, is associated with D-WMHs specifically (Moroni et al., 2018; Strassburger et al., 1997) or more diffuse WMHs throughout the brain (Moroni et al., 2018; Wiseman et al., 2004). Diabetes presents a similarly confused picture in the literature. Some reports show no difference in total volume between diabetic patients and non-diabetic controls (De Bresser et al., 2018). One cross-sectional study showed increased D-WMH volume in diabetic patients as well as reduced blood flow. The reduced blood flow may explain the pathogenesis of diabetes related WMHs resulting from ischaemia in deep white matter (Abraham et al., 2016). Diabetes has also been associated with WMH load as a part of metabolic syndrome, showing a strong association with subcortical and periventricular WMHs (Abraham et al., 2016). However, neither body mass index (BMI), nor diabetes appeared to drive this relationship, instead hypertension was the predominant risk factor associated with WMH load. Hypertension frequently dwarfs the effects of the other CVR factors, and sample sizes are usually too low to examine the individual CVR factors - especially because of the high frequency of hypertension as comorbid with high BMI, diabetes and smoking.

The relationship between WMH load and smoking is also inconsistent across studies, but has been shown to be an independent risk factor, when controlling for age, in 1,814 participants of the Framingham Offspring cohort (Jeerakathil et al., 2004). BMI and visceral fat, either measured directly or indexed by waist-to-hip ratio (WHR), have also been associated with higher total WMH loads (Kim et al., 2017; Lampe et al., 2019b) and shown a preference for deep white matter, although this also varies by study (Griffanti et al., 2018; Lampe et al., 2019b). Finally, carriage of the apolipoprotein-E (APOE) $\varepsilon 4$ allele is associated with higher total volume and higher accumulation of WMHs over time (Sudre et al., 2017), but also shows close interdependence with other CVR factors,

such as hypertension (Salvadó et al., 2019). Examining the independent contribution of the APOE $\varepsilon 4/\varepsilon 4$ genotype to the spatial distribution of WMHs is particularly difficult in studies with small cohorts, due to the relatively low prevalence of this allele in the general population (around 13%). Individual risk factors show some interaction with age, for example one large multi-centre stroke study in China showed high cholesterol to be a more important risk factor in older age (Ryu et al., 2014). Finally, sex appears to interact with some of the risk factors, such that the total WMH load is higher in females and the predictive risk factors different to males (Sachdev et al., 2009). Overall, what emerges from the literature is conflicting and complicated associations between individual risk factors and the presence of WMHs beyond those associated with age.

The literature to date has also shown a very mixed picture with regards to the relationship between total WMH load and cognition (Debette and Markus, 2010). Where a relationship has been observed, impairment to speed of processing and executive function are frequently associated with increasing total WMH load. A review of studies between 1990-2013 investigating the relationship between cognition and WMH load in the general population (Mortamais et al., 2013) found equivocal results. Five studies found a significant association between WMH load and global cognition and two reports failed to find an association (Mortamais et al., 2013). In studies that have investigated the spatial distribution of WMHs, cognitive decline was associated with periventricular WMHs (Godin et al., 2010; Prins et al., 2005). It is not clear whether particular risk factors increase the likelihood of cognitive decline associated with WMHs or whether it is just the total burden of global WMHs that is important.

The purpose of this study was to use population level imaging, demographic and lifestyle data from the UK Biobank to answer the following questions. Firstly, what is the cross-sectional relationship between total WMH load and age in the presence and absence of individual risk factors at the population level? This serves to clarify overall trends associated with the different risk factors and to highlight potential interactions between variables such as age and sex. Secondly, our main aim was to investigate whether the spatial distribution of WMHs, estimated voxel-wise across the whole brain, varied for individual CVR factors? Here, we were interested in the contribution of individual risk factors and the spatial distribution of WMHs whilst controlling for all other related CVR factors. We extend the literature in several ways, by demonstrating

quantitative methods to estimate the spatial distribution, voxel-wise, and by taking advantage of a large sample to examine the unique effects of individual risk factors. Finally, we investigated the relationship between individual risk factors and speed of processing. We take advantage of a large dataset that enables systematic examination of individual risk factors, whilst controlling for other risks and a novel method to estimate the probability of CVR-associated WHMs at a voxel-wise level.

3.2 Methods

3.2.1 Participants

The study was conducted under Biobank application number 34077, and imaging data shared within the University of Oxford under application number 8107. UK Biobank participants gave written, informed consent for the study, which was approved by the Research Ethics Committee under application 11/NW/0382.

Participants were selected according to the flow chart in Figure 3.1, starting with 22,292 T1 images. Applying the published UK Biobank automated processing and quality control pipeline (Alfaro-Almagro et al., 2018), 1,985 T1 images were classified as non-usable (for a full list of T1 image imperfections see Table 3, Alfaro-Almagro et al. (2018)). Following the T2 FLAIR pipeline (Alfaro-Almagro et al., 2018), further 1,219 participants with T2 FLAIR images missing (incidents during acquisition, protocol changes) or non-usable (quality control including problems with the acquisition or with the registration to T1) were excluded, i.e. 20,188 participants with available WMH segmented brain images. We excluded data from individuals (590 total) with a current diagnosis or history of neurological or neuropsychiatric disease based on self-reported, non-cancer illness during a verbal interview with a trained nurse. Excluded diagnostic categories were traumatic brain injury, transient ischaemic attack, stroke, haematoma, infection of the nervous system, brain abscess, haemorrhage or skull fracture, encephalitis, meningitis, amyotrophic lateral sclerosis, multiple sclerosis, Parkinson's disease, Alzheimer's disease (AD), epilepsy or alcohol or drug dependency (see Table 3.A.1 for a list of excluded participants by condition). Given known differences in cardiovascular disease and CVR factors between ethnicities (Howard, 2013; Benjamin et al., 2017), and the low proportion of non-white individuals in the Biobank cohort (Fry

et al., 2017), we elected to exclude individuals who self-declared as non-white ethnicity (Figure 3.1). Participants with missing data and one individual with unusually high WMH load (about 12,000 voxels affected by WMHs) were also excluded. As a result, the final dataset for the analysis consisted of 13,680 individuals.

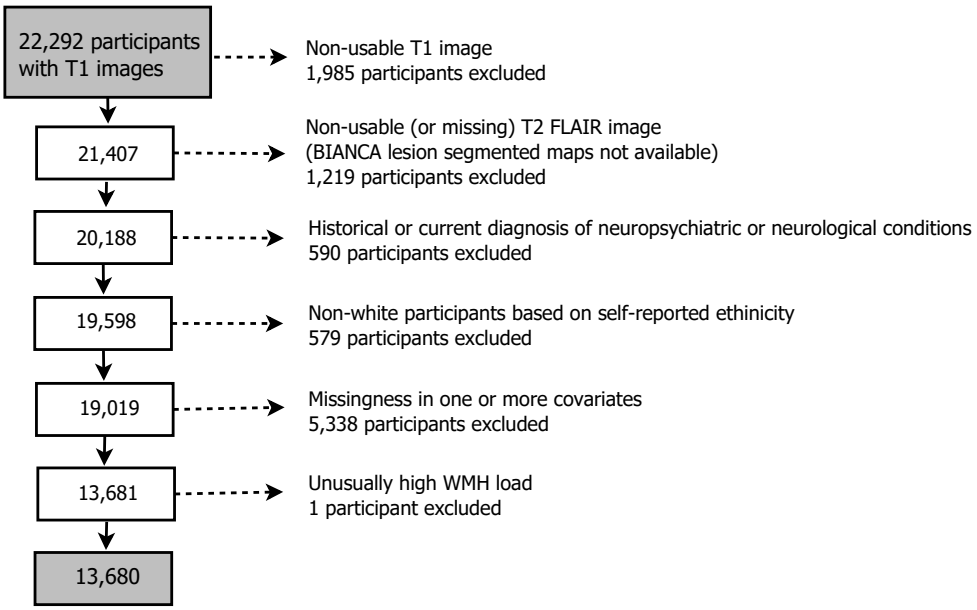


Figure 3.1: Diagram demonstrating the flow of gradually refining participants starting from all UK Biobank participants with available T1 structural brain images.
Most common neuropsychiatric/neurological conditions in decreasing number of participants: stroke (185), transient ischaemic attack (115), epilepsy (77), etc. (see Table 3.A.1 for a full list).
Missingness by risk factor: hypertension risk (2,075), hypercholesterolemia (211), diabetes (165), smoking in pack years (3,146), waist-to-hip ratio (413), and APOE-ε status (514). Some individuals have more than one of the variables missing; for characteristics of individuals excluded due to missingness, see Table 3.A.2.

3.2.2 Cerebrovascular risk factors

We investigated the main CVR factors known to be associated with the presence of WMHs, including hypertension, hypercholesterolemia, smoking, diabetes, waist-to-hip ratio and APOE-ε polymorphism status. Using categorical variables for each risk factor (described below), we also calculated a cumulative CVR score based on the sum of these variables to examine the impact of multiple risk factors on WMH distribution and its relationship to cognition. We did not use an established cardiovascular risk score, such as the Framingham Risk Score (Lloyd-Jones et al., 2004), for two reasons. Firstly, the established scores typically require continuous measures, such as high density lipoprotein

levels, which were not available in UK Biobank at the time of analysis. Secondly, the established scores are typically used to calculate the future risk of a cardiovascular event, such as coronary heart disease. Here, we were interested in the cumulative effects of multiple CVR factors on WMH spatial distribution as well as on cognitive impairment, not on the risk of a specific cerebrovascular event.

Hypertension Blood pressure (BP) was measured using a digital BP monitor (Omron), or a manual sphygmomanometer when the digital monitor was not available. Two readings were taken moments apart and we used the average of these two readings. To increase the reliability of our indicator for hypertension, we included self-reported medication for BP as an additional indicator of high BP. Therefore, our indicator variable 'hypertension risk' had value 1 for anyone (i) with average BP measuring over 140/90mmHg (Boffa et al., 2019) and/or (ii) on medication for high BP; otherwise the indicator had value 0.

Hypercholesterolemia We used medication for cholesterol as an indicator for diagnosed hypercholesterolemia. Participants responded to the question "Do you take any of the following medications?" with the option of "cholesterol lowering medication" as part of a questionnaire presented on touch screen tablets. For those who selected "cholesterol lowering medication", our indicator variable for hypercholesterolemia was assigned a value of 1, and 0 for those who answered this question but did not select this option.

Diabetes Diabetes diagnosis was determined from responses to the question "Has a doctor ever told you that you have diabetes?". This question was part of a questionnaire presented on touch screen tablets. Based on the answer to the question, the indicator variable for diabetes was 1 (answer 'yes') or 0 (answer 'no').

Smoking Current and past smokers, and non-smokers, were divided into groups according to their pack year history. Pack years was calculated as the daily number of cigarettes divided by pack size (20) and multiplied by the number of years smoking. The number of years smoking was the age at stopping smoking, or the age at testing for current smokers, minus the age at which smoking was started. Pack years was appropriately adjusted for those who reported giving up smoking for more than six months. We grouped participants into non-smokers (less than or equal to 10 pack years), smokers (more than 10 and less than or equal to 50 pack years), heavy smokers (more than 50

pack years) (Lubin et al., 2016), which resulted in a discrete 'smoking score' covariate with three levels: non-smoker (0), smoker (1) and heavy smoker (2).

Waist-to-hip ratio Waist circumference is a measure of visceral and subcutaneous fat, while hip circumference is thought to represent subcutaneous fat only. The ratio therefore represents an elevated proportion of intra-abdominal fat (Shuster et al., 2011). Waist and hip circumferences were manually measured in centimeters and used to calculate the WHR. The threshold for high WHR was set according to the WHO guidelines, set for each sex (World Health Organization, 2008) (0.9 for males and 0.85 for females). In the subsequent analysis we used WHR either as a continuous covariate, or as an indicator of high WHR (1) or not (0).

APOE-ε status Carriage of the APOE $\varepsilon 4$ allele only (not carriage of $\varepsilon 3/\varepsilon 3$ or $\varepsilon 2/\varepsilon 2$ or $\varepsilon 2/\varepsilon 3$) was considered a cerebrovascular risk factor based on a substantial body of research for APOE $\varepsilon 4$ as a risk factor for cardiovascular disease (McCarron et al., 1999) and sporadic dementia. The genotyping pipeline is described in full here (Bycroft et al., 2018). Based on the number of $\varepsilon 4$ alleles, a discrete APOE-ε status covariate was created as 0 (no carrier of $\varepsilon 4$ allele), 1 (heterozygous, i.e. $\varepsilon 3/\varepsilon 4$) and 2 (homozygous, i.e. $\varepsilon 4/\varepsilon 4$).

CVR score The CVR score was created as the sum of the six categorical variables representing the six risk factors described above: hypertension risk (0/1), hypercholesterolemia (0/1), diabetes (0/1), smoking score (0/1/2), WHR (0/1), and APOE-ε status (0/1/2). The resulting composite score was on a scale $0-8$ and the higher the score, the higher the cerebrovascular burden of an individual. The UK Biobank data used to obtain the score are listed in Table 3.A.3.

3.2.3 Cognitive testing

Data from the reaction time task, thought to be a sensitive index of speed of processing, was used in the analysis (Fawns-Ritchie and Deary, 2020). We used speed of processing as a cognitive variable because it has most consistently shown a relationship with WMH load, it is normally distributed and it has considerably less missing data than some of the other cognitive test variables available in UK Biobank. The task was administered by touch screen at the same session as the MRI scan. The UK Biobank reaction time task was a variant of the 'Snap' card game, in which participants react to the presence of a pair of matching cards over 12 rounds of the game. Mean reaction time was recorded

and trials with responses below 50ms or above 2000ms were excluded.

Education is considered an important confounding variable when modelling cognitive function (Evans et al., 1993; Whalley et al., 2004). Participants responded to the question "Which of the following qualifications do you have?" as part of the questionnaire presented on touch screen tablets. A continuous variable "years of education" was defined according to the ISCED categories (Lee et al., 2018; Cheesman et al., 2020); for details see Table 3.A.3.

Missingness in the cognitive task variable resulted in 923 exclusions with further 27 exclusions due to missingness in the education variable. Those participants were excluded only when the statistical analysis included reaction time as a variable, otherwise 13,680 individuals were used.

3.2.4 MRI data

Volunteers were scanned on Siemens Skyra 3T scanners with 32 channel head coils. We used the T2-weighted fluid attenuated inversion recovery (FLAIR, $1.05 \times 1 \times 1$mm resolution) and the T1-weighted, 3D magnetization-prepared rapid gradient echo (MPRAGE, $1 \times 1 \times 1$mm resolution, T1=880ms, TR=2000ms, matrix=208x256x256) sequence as part of the longer imaging protocol[1]. The published UK Biobank pipeline (Alfaro-Almagro et al., 2018) details the spatial normalisation procedure of the T1 image to MNI 152 space. Briefly, after gradient distortion correction and reduction of the field of view (FOV) to remove non-brain space, FNIRT (Andersson et al., 2007) was used for non-linear registration to 1mm resolution MNI 152 space. FNIRT parameters were optimised for best performance on the UK Biobank's T1 image resolution and contrast.; the FNIRT configuration file used as part of the UK Biobank pipeline is available online at `https://git.fmrib.ox.ac.uk/falmagro/UK_biobank_pipeline_v_1/-/blob/master/bb_data/bb_fnirt.cnf`. All three of the above steps are combined into a single non-linear and reversible transformation. Note that participants with large ventricles were excluded from the dataset as part of the UK Biobank quality control pipeline. After gradient distortion correction, the T2 FLAIR image in native space was rigid-body transformed using FLIRT (Jenkinson et al., 2002a) to register to T1 space. We excluded individuals with non-usable or missing T1 or T2 FLAIR images, see Section 3.2.1 and

[1] `http://biobank.ctsu.ox.ac.uk/crystal/crystal/docs/brain_mri.pdf`

Figure 3.1.

The Brain Intensity Abnormality Classification Algorithm (BIANCA) (Griffanti et al., 2016) was used to segment WMHs. BIANCA is an automated method for WMH segmentation based on voxel intensity and distance to the ventricles. BIANCA's segmentation has been compared to segmentation of WMHs from two different cohorts, with different MRI sequence parameters, as well as in patient populations (a vascular and a neurodegenerative cohort). Correlations in total extracted WMH load, spatial overlap and visual rating scales has shown BIANCA to be a valid alternative to manual segmentation (Griffanti et al., 2016). When applied to the UK Biobank imaging data, it produced an output image in subject space which represented the probability per voxel of being a WMH; as part of the segmentation, it was then thresholded at 0.8 to give a binary WMH mask. The threshold of 0.8 was the optimised tuning parameter to minimise prediction error in native space when compared to manually segmented lesion masks of 12 UK Biobank individuals. We applied the estimated spatial normalisation parameters to the WMH maps; specifically, the transformation parameters for the T2-weighted FLAIR and non-linear warping were used (Andersson et al., 2007), and then we thresholded the warped WMH maps (the warping process used trilinear interpolation that introduced non-binary values) at 0.5 to get binary WMH maps in MNI space. The 0.5 threshold was used as a neutral value to preference neither enlargement or shrinkage of total lesion volume. Having a binary WMH map per participant, we estimated the WMH load as the number of WMH-affected voxels.

All preprocessing steps were performed using the FSL software[2]. Note that voxel size of $2mm^3$ was chosen for computational reasons and this implied a standard brain mask of dimension $91 \times 109 \times 91$ voxels. Binary WMH maps and WMH load (unit of measurement was $2mm^3$) were available for 13,680 participants.

3.2.5 Statistical analysis

We chose to do complete cases analysis, rather than impute missing data, and therefore excluded 5,338 individuals with missing data in one or more CVRs (see Figure 3.1 for details). All the exclusions as described in Section 3.2.1 led to a final dataset for the analysis of 13,680 participants.

[2] `https://fsl.fmrib.ox.ac.uk/fsl/fslwiki`

Univariate analysis

From each participant's WMH binary mask, where one indicates the presence of a WMH and zero the absence, respectively, we used the subject-specific summary measure log-transformed WMH load as the response variable in our first modelling step. Given the highly right-skewed distribution of WMH load across the population, we log-transformed WMH load to enable the use of standard least squares linear models.

As an exploratory step of whether and how aging related to log(WMH load) in individuals grouped by presence or absence of cerebrovascular risk factors (or by sex), we used locally estimated scatterplot smoothing (loess) (Cleveland et al., 2017) as implemented in `R` package `stats`, function `loess.smooth()`. Loess is a non-parametric local averaging method, which uses weighted regression inside windows with a fixed number of points. To determine the window, we fixed the span parameter to 20%, which means that the horizontal window surrounding a target observation contains 20% of its nearest neighbours. Then, a weighted polynomial was fitted to the data within the window and the predicted response at the target point was the fitted value. The fitted smooth curve provided a graphical overview of underlying patterns in the dataset. We used the resulting fitted curves to explore (i) the linear age effect assumption, (ii) the need for an age by sex interaction, and (iii) the effect of risk factors on log(WMH load) and whether it varied across age.

Multiple linear regression was used to formally assess the dependence of log(WMH load) on the cerebrovascular risk factors, while controlling for known confounders (age, sex, head size). We controlled for head size which is a recognised confounding variable in MRI studies generally and in UK Biobank specifically (Alfaro-Almagro et al., 2020). Because head size correlates with sex, spurious correlations can arise between sex and MRI variables if head size is not controlled for. A recent paper on deconfounding UK Biobank MRI data (Alfaro-Almagro et al., 2020), recommends the minimal set of confounding variables includes age, sex, age-sex interaction and head size scaling (Section 2.4.1). We also explored the inclusion of an age-sex interaction term to the models as one of the confounds.

With N subjects and Y_i the random variable representing the log(WMH load) for each subject $i\,(i = 1, \ldots, N)$, suppose all Y_i's are independent and Normally distributed with

means μ_i=$\mathbb{E}(Y_i)$ and variance σ^2, which is the same across subjects. Then a normal linear model can be written as

$$Y_i \sim N(\mu_i, \sigma^2) \qquad \text{(random)} \tag{3.1}$$

$$\mathbb{E}(Y_i) = \mu_i = \boldsymbol{x}_i^\top \boldsymbol{\beta} \qquad \text{(systematic)} \tag{3.2}$$

where $\boldsymbol{\beta}$ is an P-vector of parameters, and $\boldsymbol{x}_i$ denotes the P-vector of subject-specific covariates for subject i and X is the full rank design matrix with rows $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N$.
To assess the importance of each explanatory variable, maximum likelihood estimates (MLEs) of the regression coefficients were explored along with 95% confidence intervals (CIs) and p-values (significance level 0.05). Note that for the discrete explanatory variables (such as sex, hypertension risk, smoking score, etc.), level 0 of the factor variable was fitted to be the baseline, e.g. the estimated regression coefficient $\hat{\beta}$ for hypertension risk estimated the effect of hypertension risk 1 on the outcome in comparison to the effect of hypertension risk 0. If the factor variable had more than 2 levels, here smoking score and APOE-ε status, a dummy variable was created for each of the two contrasts with category 0 used as baseline - category 2 compared to category 0 and category 1 compared to category 0 - so two regression coefficients were estimated for the two contrasts. We used adjusted R^2 (R_a^2) as a measure of goodness of fit, which is interpreted as the percentage of variance explained. We also computed partial R_a^2, which measures the additional variation explained by each explanatory variable, after adjustment for the other predictors. We fitted various multiple regression models aiming to (i) outline the risk factors which were significant predictors of log(WMH load), and (ii) to determine the models used for the spatial voxel-wise analysis.

Voxel-wise analysis

Voxel-wise analysis was employed to assess how different contributors to the cerebrovascular burden related to the spatial distribution of WMHs.

Mass-univariate voxel-wise modelling of WMH masks requires a generalized linear model (GLM), e.g. logistic regression or probit regression, to account for the binary nature of the WMH masks. Since we now want to model WMH probability at each voxel, let $Y_i(s_j)$ denote a Bernoulli random variable with probability of success $p_i(s_j)$, where

$Y_i(s_j)$ represents the presence ($Y_i(s_j)$=1) or absence of a WMH for subject $i, i = 1, \ldots, N$ at voxel $s_j\,(j = 1, \ldots, M)$. Assume $Y_1(s_j), \ldots, Y_N(s_j)$ are independent random variables across subjects i and voxels s_j. In contrast to the normal linear model, every GLM has a link function g, which is a monotonic function that relates the expectation of the random outcome to the systematic component. Note that the link function is the identity link function for the normal linear model in Equations (3.1, 3.2). The GLM can be written as

$$\begin{aligned} [Y_i(s_j) \mid p_i(s_j)] &\sim \text{Bernoulli}(p_i(s_j)) && \text{(random)} \\ g(\mathbb{E}[Y_i(s_j) \mid p_i(s_j)]) &= \eta_i(s_j) && \text{(link)} \\ \eta_i(s_j) &= \boldsymbol{x}_i^\top \boldsymbol{\beta}(s_j) && \text{(systematic),} \end{aligned}$$

where $\boldsymbol{\beta}(s_j)$ is a P-vector of parameters at each voxel s_j.
For this analysis, we have chosen probit link[3] Φ^{-1}, where Φ indicates the standard normal cumulative distribution function, so the model can be written as probit regression

$$\mathbb{P}(Y_i(s_j)=1 \mid \eta_i(s_j)) = \Phi(\boldsymbol{x}_i^\top \boldsymbol{\beta}(s_j)).$$

At each voxel s_j, we obtain the MLEs $\hat{\boldsymbol{\beta}}(s_j)$ through iterative optimization, such as iteratively reweighted least squares (IWLS) (Green, 1984). The MLE is the default choice of an estimator due to its optimal asymptotic properties. However, in finite samples, the MLE may demonstrate significant bias and large variance. Furthermore, in binary response models, there is positive probability for the MLE to have at least one infinite component, which results in issues with common inference procedures, such as Wald tests and Wald-type CIs. Such infinite estimates result when the data are separated (Albert and Anderson, 1984), that is a covariate or a combination of covariates perfectly separates outcome measurements (e.g. female participants in the dataset having a WMH at a particular voxel where no male does). If the MLE is infinite, inference becomes impossible and test statistics are unstable.

Logistic or probit regression has been often used in the voxel-wise brain WMH mapping literature (Rostrup et al., 2012; Lampe et al., 2019b), but potential convergence and

[3]While logit link is often used, probit and logit links give very similar results and we used probit link for comparability with other Bayesian probit-link models we investigate.

bias problems have not been discussed to our knowledge. To address both limitations - infinite and biased MLEs - we propose the use of test statistics (standardized coefficients) based on mean bias-reduced (BR) estimates $\tilde{\boldsymbol{\beta}}$. The bias-correction approach, which we focused on in the work, was first introduced in Firth (1993) for logistic regression and then further developed for exponential family GLMs (Kosmidis and Firth, 2009; Kosmidis et al., 2020). This method guarantees finite estimates when total separation is observed and it ensures estimates with asymptotically smaller bias than what the MLE has. Bias reduction is achieved by subtracting the first-order bias in each iteration of the optimization; see Kosmidis and Firth (2009) for the exact form of the adjustments. To obtain mean BR estimates, we use the `R` package brglm2 (Kosmidis et al., 2020), which adds additional functionalities to the `R` function `glm()`. Note that as in the univariate analysis, we chose level 0 as the baseline for factor variables, i.e. one regression coefficient was estimated voxel-wise for binary explanatory variables and two for discrete variables with three levels (level 2 vs baseline (level 0) and level 1 vs baseline), respectively.

To determine the size of the effect of each explanatory variable, we explored test statistics (standardized coefficients or z-scores) based on mean BR estimates $\tilde{\boldsymbol{\beta}}$ and their associated p-values. While the mean BR estimates ensure better performance when there are few WMH at a voxel, like other authors we excluded voxels when the WMH count fell too low; for example, Rostrup et al. (2012) and Lampe et al. (2019b) required at least 5 participants. Due to the large sample size, we chose 4 as our threshold, i.e. we only considered voxels where 4 or more participants had a WMH (see Appendix 3.B for more details).

After computing the p-values across the brain, we corrected for multiple testing. Threshold-free cluster enhancement (TFCE) (Smith and Nichols, 2009) is often used in the literature, but due to the UK Biobank sample size of 13,680, TFCE would be computationally expensive to perform (c.f. 605 for Rostrup et al. (2012) and 1,825 for Lampe et al. (2019b), where authors employed the TFCE approach). Instead, we used false discovery rate (FDR) correction (Benjamini and Hochberg, 1995), i.e. we controlled the expected proportion of falsely rejected hypotheses. We favoured FDR over the most common family-wise error rate correction method, Bonferroni correction, since Bonferroni is known to be quite conservative when the comparisons are not independent, which is the case for spatially dependent WMH maps (Genovese et al., 2002; Rorden

and Karnath, 2004). The `R` function `p.adjust()` (package `stats`) was used to correct the p-values, using the $\alpha_{\text{FDR}=0.05}$ significance level. We visually inspected axial slices of the the standardized coefficients for 'significant' voxels (voxels where an explanatory variable was a significant predictor of WMH risk) to gain understanding of the localized effect of cerebrovascular risk factors and how they complemented each other. We also inspected the total number of significant coefficients across the brain per predictor across a variety of models as a measure of the spread of the effect throughout white matter.

Mediation analysis

The univariate analysis framework was also employed to explore the association between speed of processing (reaction time) and log(WMH load) (or CVR score). To better understand the underlying dependencies, we also performed mediation analysis to investigate the hypothesis that the effect of WMH load on speed of processing (cognitive task) was fully or partially explained through a given CVR (mediator); see Figure 3.A.3. The `R` package `mediation` (Tingley et al., 2014) was used for the estimation and suitable models (GLMs, LMs) were used for the mediator and outcome models (some of the mediators are discrete variables, which necessitates the use of GLMs). All models were controlled for age, sex, age by sex interaction, head size and years of education. Point estimates along with 95% percentile CIs and p-values were explored for the direct, indirect and total effects (non-parametric bootstrap, 10,000 resamples). We were mostly interested whether the indirect effect was significant, i.e. whether there was significant mediation effect, and if so, what was the proportion mediated (percentage of total effect) of WMH load on speed of processing operating (partially or fully) through the CVR factors.

3.3 Results

We analysed data from 13,680 individuals from the UK Biobank (mean age 62.9 $\pm$ 7.4 years, 7,236 female). Summary descriptive statistics of the UK Biobank sample characteristics are included in Table 3.1. The UK Biobank variables used in the current work are described in Table 3.A.3.

Table 3.1: Characteristics of UK Biobank dataset of 13,680 participants.

Characteristics	Levels (N)	Mean (SD)	Median (range)
Age (years)	—	62.9 (7.4)	63.5 (45.1; 80.7)
Sex	Men (6,444), Women (7,236)	—	—
Head size	—	1.3 (0.1)	1.3 (0.9; 1.8)
Hypertension risk	0 (7,272), 1 (6,408)	—	—
Hypercholesterolemia	0 (10,899), 1 (2,781)	—	—
Diabetes	0 (13,017), 1 (663)	—	—
Smoking (score/pack years)	0 (11,291), 1 (2,238), 2 (151)	4.9 (11.4)	0 (0; 141)
WHR (indicator/continuous)	0 (7,251), 1 (6,429)	0.9 (0.1)	0.9 (0.6; 1.2)
APOE-ε status	0 (10,226), 1 (3,150), 2 (304)	—	—
CVR score	0 (2,596), 1 (4,204), 2 (3,770), 3 (1,947), 4 (869), 5 (252), 6 (36), 7 (6), 8 (0)	1.7 (1.2)	2 (0; 7)
WMH load ($2mm^3$ voxels)	—	528.6 (667.7)	310.0 (3; 8,228)
Reaction time* (milliseconds)	—	585.4 (104.9)	569 (272; 1,559)
Years of education	7 (780), 10 (1,679), 13 (757), 15 (1,481), 19 (2,027), 20 (6,006)	16.7 (4.3) 16.7 (4.3)	19 (7; 20) 19 (7; 20)

*12,730 participants for the cognition variable reaction time and education due to missingness
SD: standard deviation; WHR: waist-to-hip ratio; APOE: apolypoprotein-E;
CVR: cerebrovascular risk; WMH: white matter hyperintensity.

3.3.1 Age by sex interactions

Figure 3.1 suggests a positive linear relationship between age and CVR score as well as between age and log(WMH load). Males had overall higher cerebrovascular burden than females with no interaction with age (loess curves nearly parallel). Log(WMH load) increased with age and the different slope of the sex-specific fitted curves suggested a potential age by sex interaction. This was further confirmed through a highly significant interaction term ($p<0.001$) in a linear model of log(WMH load), adjusted for head size and total CVR burden (R^2_a=0.26, Table 3.2). To understand the effect of sex on the log(WMH load), we used the estimated regression coefficients for sex and age:sex interaction term (Model U.1) to check how the outcome variable changed. A female participant aged 75 years would be expected to have $\exp(-0.48 + 0.01\times75) = \exp(0.27)$=1.31-fold higher WMH load than a male participant the same age. For a female participant aged 50, we get 1.02-fold difference in WMH load, respectively, which highlights the

effect of the interaction term. We therefore adjusted all subsequent univariate models of log(WMH load) and voxel-wise models of WMH masks for an age by sex interaction, something often overlooked in analyses within the existing literature.

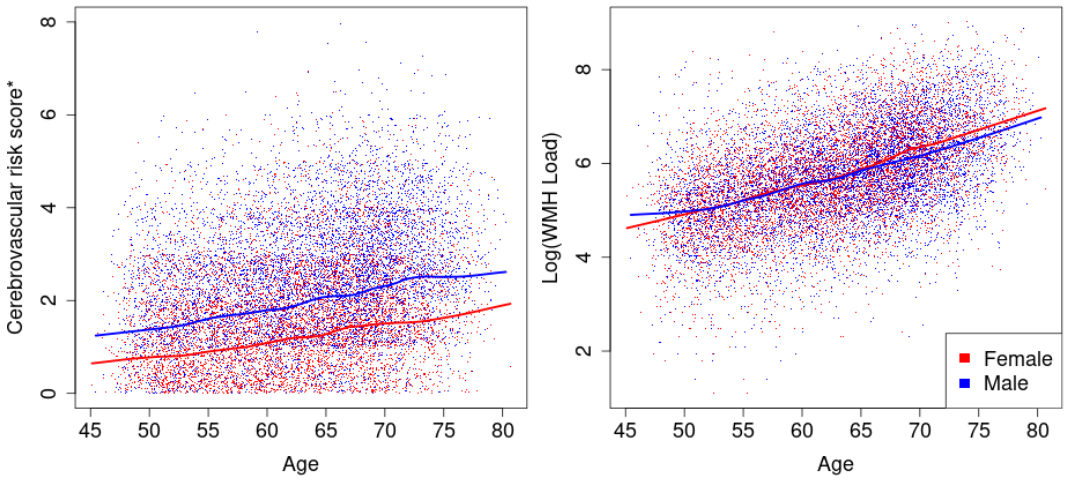


Figure 3.1: Sex-specific trends in CVR score (left) and in log(WMH load) (right) across age. Solid lines represent the loess-smoothed curve with a span of 20% and the points are the observed data points. Males have higher CVR burden than females across all ages and log(WMH load) increases across age for both sexes but potentially at different speed for males and females. *Uniform noise U(0, 1) added to the CVR scores to disperse the values in the y-axis direction (left plot).

3.3.2 White matter hyperintensity load associated with individual risk factors

Next, we explored the relationship between log(WMH load) and individual CVR factors, including hypertension risk, hypercholesterolemia, diabetes, smoking, WHR, and APOE-ε status through fitting linear regressions (Table 3.2) and risk factor specific loess-smoothed curves (Figure 3.2). Age's dominant effect was seen here: a 10-year age difference was associated with a exp(0.6)=1.82-fold difference in WMH load. All CVR factors had a marked positive effect on WMH load (Table 3.2), with hypertension risk and WHR having the highest partial R_a^2 (1.7% and 1.2%), e.g. hypertension risk explained 1.7% of the remaining variability after adjusting for confounding. The inclusion of the CVR score as a predictor led to achieving the highest R_a^2 and its partial R_a^2 of 3.1% resembled the combined effect of the risk factors.

Hypertension risk was associated with 1.27-fold higher total WMH load across all ages ($\hat{\beta}$=0.24, Table 3.2; nearly parallel loess curves, Figure 3.2). The same relationship

Table 3.2: Univariate regression summaries outlining the association between log(WMH load) and cerebrovascular risk factors (or composite score). For all CVRs, the presence of the risk has a strong positive effect on log(WMH load) when compared to its absence (discrete risk equals 0), e.g. participants who have high WHR are expected to have 0.22-fold higher log(WMH load), or exp(0.22) = 1.25-fold higher WMH load, than those who do not. Model U.1: all predictors in the model shown. Model U.2.1 - U.2.6: main effect of interest shown. All models adjusted for age, sex, head size, age-sex interaction and only one of the risk factors included in the model. $\hat{\beta}$ stands for the maximum likelihood estimate of the regression coefficient β; two regression coefficients for discrete variables with more than two levels (models U.2.4 and U.2.6): contrasting level 1 to level 0, and level 2 to level 0 (level 0 modeled as baseline).

Model / Predictor (level)	Estimate $\hat{\beta}$	95% CI	p-value	$\boldsymbol{R_a^2}$/ partial R_a^2
Model U.1				**0.260**
Intercept	2.19	(1.93; 2.46)	<0.001	
Age	0.06	(0.05; 0.06)	<0.001	0.093
Sex (Female)	-0.48	(-0.74; -0.22)	<0.001	0.001
Age:Sex (Female)	0.01	(0.01; 0.01)	<0.001	0.002
Head size	-0.29	(-0.45; -0.13)	<0.001	0.001
CVR score	0.14	(0.13; 0.15)	<0.001	0.031
Model U.2.1				**0.249**
Hypertension risk (1)	0.24	(0.21; 0.28)	<0.001	0.017
Model U.2.2				**0.240**
Hypercholesteromia (1)	0.17	(0.13; 0.21)	<0.001	0.005
Model U.2.3				**0.240**
Diabetes (1)	0.30	(0.22; 0.37)	<0.001	0.005
Model U.2.4				**0.240**
Smoking score (1)	0.15	(0.11; 0.19)	<0.001	0.006
Smoking score (2)	0.45	(0.30; 0.59)	<0.001	
Model U.2.5				**0.245**
Waist-to-hip ratio (1)	0.22	(0.15; 0.25)	<0.001	0.012
Model U.2.6				**0.237**
APOE-ε status (1)	0.05	(0.01; 0.08)	0.011	0.002
APOE-ε status (2)	0.23	(0.13; 0.34)	<0.001	

was observed for diagnosed diabetics compared to non-diabetics, with 1.35-fold greater total load across all ages for diabetic participants.

When considering the effects of hypercholesterolemia (medicated for high cholesterol) and of high WHR ratio on log(WMH load), the loess-fitted curves suggested there might be a risk factor by age interaction (Figure 3.2). For both risk factors, the log(WMH load) seemed to be the same regardless of the risk factor status over the age of 70. Multiple linear regression was used to assess the importance of hypercholesterolemia by age interaction term. A new explanatory variable (multiple of the binary hypercholesterolemia variable and age) was added to model U.2.2, but there was not enough

evidence to reject the null hypothesis of no effect (p=0.09). The WHR by age interaction was also explored (a new term added to model U.2.5) and it was marginally significant (p=0.049). However, the addition of the interaction term did not change R_a^2 to two decimal places (no higher explanatory power was achieved). Thus both interaction terms were not further explored. Both hypercholesterolemia and WHR had marked positive effect on log(WMH load) (Table 3.2).

The highest risk groups for smoking and APOE-ε status (>50 pack years and APOE $\varepsilon 4/\varepsilon 4$) were associated with a higher log(WMH load) across all ages (Figure 3.2). In linear regression analyses, the effect of smoking and APOE-ε status on log(WMH load) was found to be strong and positive (Table 3.2). For example, comparing smoking score 1 to smoking score 0 (baseline) was associated with exp(0.15)=1.16-fold increase in WMH load, and 1.57-fold increase when comparing 2 to 0, respectively.

3.3.3 Spatial distribution of white matter hyperintensities

We plotted WMH incidence, voxel-wise, across 13,680 healthy ageing individuals and reveal the expected spatial distribution of WMHs. The highest probabilities were concentrated around the periventricular areas and in deep white matter regions (Figure 3.3).

Next, we explored the effect of head size, age and sex, and their interaction on WMH probability (Figure 3.4) since they all act as confounding effects when considering the spatial distribution of WMHs associated with individual risk factors. The figure represents the standardized coefficients (z-scores based on mean BR estimates) for voxels which are significant (5% FDR correction applied) and have WMH incidence of at least 4 people (40,001 voxels, i.e. all other voxels are plotted as transparent). The age effect on WMH probability is dominant and widely spread through white matter. Also, note that the direction of the effect for those confounding variables is the same as in the univariate regression (Model U.1, Table 3.2), i.e. positive (red) for age and age by sex interaction, negative (blue) for sex and head size, but with varying effect size across the brain.

We also explored the effect of each risk factor on WMH probability (marginal models, Figure 3.5(b)) as well as the contribution of each risk factor while controlling for all other risk factors (joint model, Figure 3.5(a)). On those axial slices, the darker the colour,

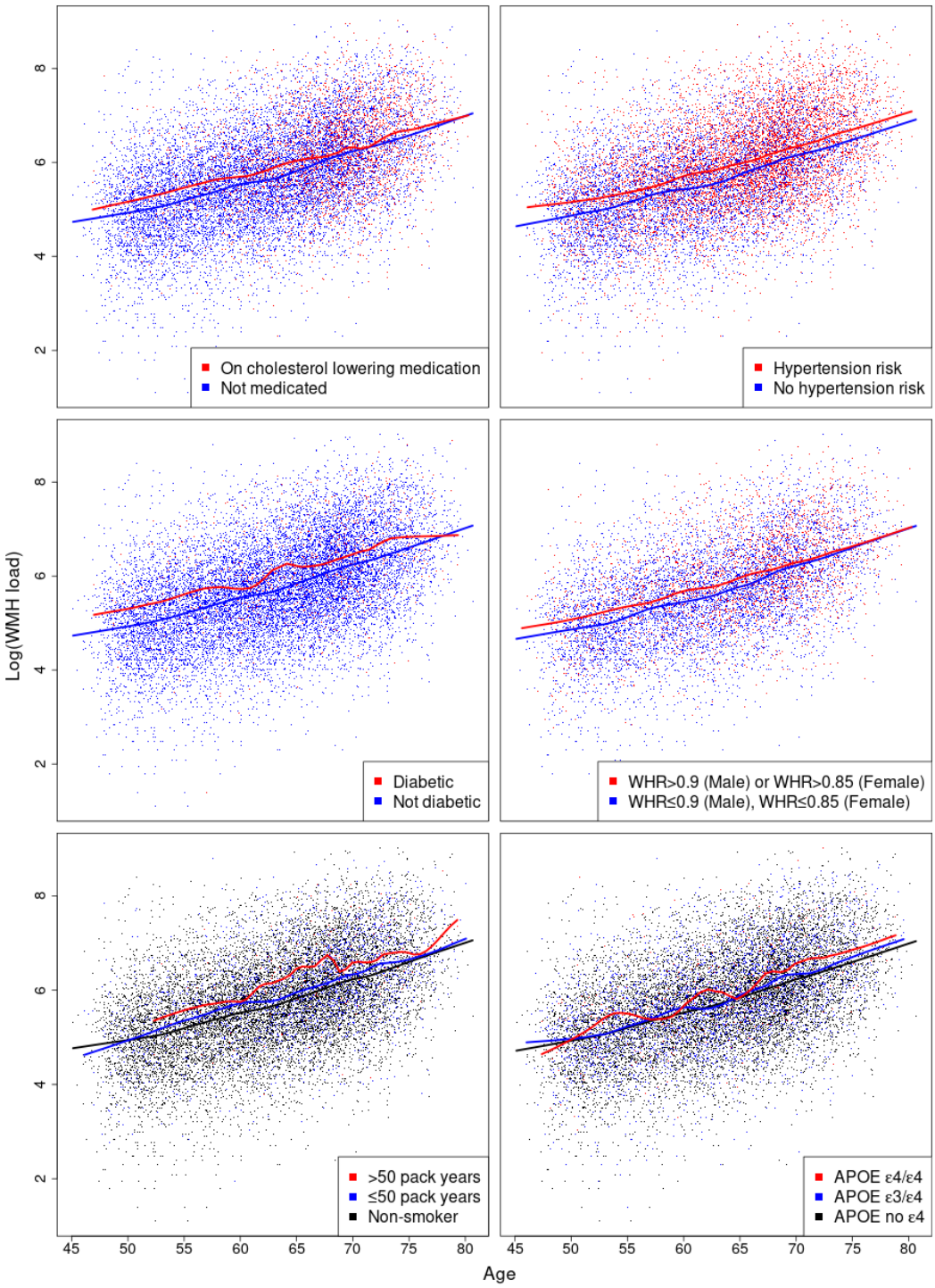


Figure 3.2: Cerebrovascular risk factor specific trends in log(WMH load) across age. Solid lines represent the loess-smoothed curve with a span of 20% and the points are the observed data points. The presence of any of the risk factors suggests higher log(WMH load). Crossing fitted curves would suggest a potential risk factor by age interaction and parallel line its absence, respectively.
WMH: white matter hyperintensity; WHR: waist-to-hip ratio; APOE: apolypoprotein-E.

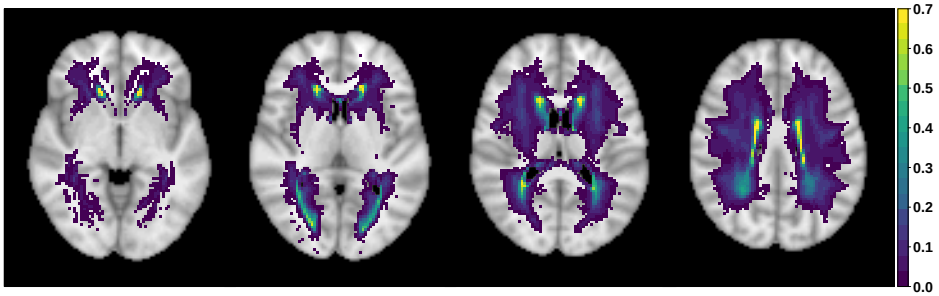


Figure 3.3: Square root transformed empirical WMH probability based on binary WMH masks of 13,680 UK Biobank individuals; axial slices z=$\{35, 40, 45, 50\}$ shown (from left to right). Square root transformation leads to more dispersed values allowing for better visualisation. Voxels with three or fewer individuals having a WMH are plotted as transparent to show a standard anatomical MRI for reference.

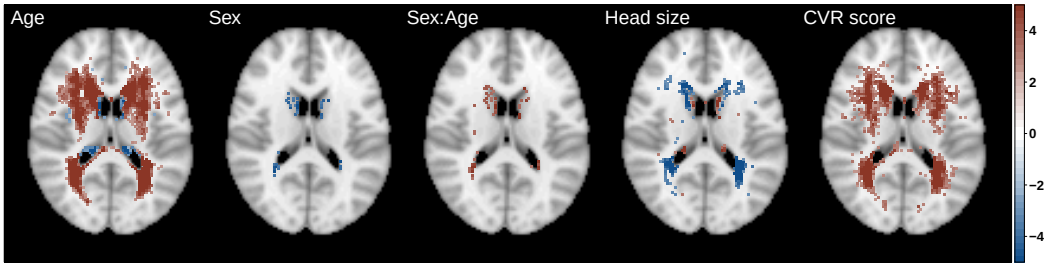


Figure 3.4: Significance maps (z-scores based on mean bias-reduced estimates) for model S.1, which includes age, sex (baseline men), age-sex interaction and head size and cerebrovascular risk (CVR) score as explanatory variables. Data on 13,680 UK Biobank individuals, and voxels with at least four individuals having a WMH explored (i.e. 0.03% WMH incidence); 5% FDR correction applied; axial slice z=45 shown.

the stronger the effect of the presence of the risk factor when compared to its absence. To quantify the WMHs associated with each risk factor, we estimated the percentage of significant voxels in a mask of 40,001 voxels for each risk factor in the marginal models and the change in the joint models, when other risk factors are controlled for (Table 3.3). This provided an additional measure of the relative importance of the different risk factors. The widest spatial distribution, in both periventricular and deep white matter, was for hypertension risk - thought to be the strongest risk factor, after age, for the presence of WMHs (Figure 3.5, Table 3.3).

In terms of spatial extent, WHR and APOE $\varepsilon 4/\varepsilon 4$ genotype stand out as the next notable risk factors with a unique spatial distribution, independent of other risk factors. The contribution of APOE $\varepsilon 4/\varepsilon 4$ genotype persists regardless of the inclusion of the other risk factors in the model (significantly affecting about 4% of the voxels analysed,

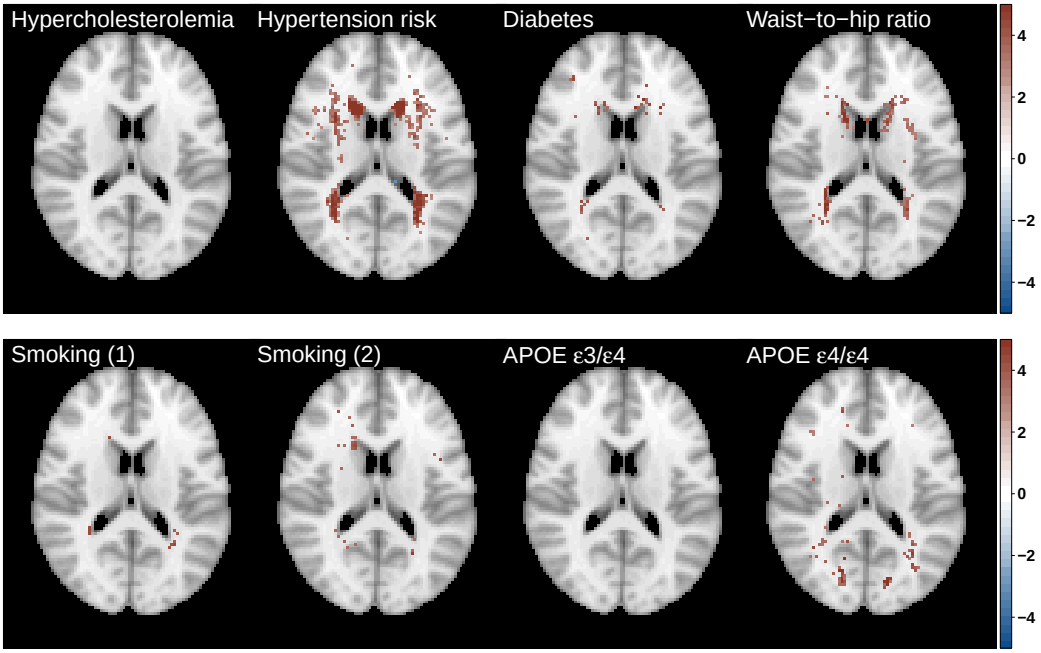
Table 3.3: Percentage and number of significant voxels across predictors for joint model S.2 (all CVR factors) and for marginal models S.3.1 - S.3.6 (one of the CVR factors). Note that for factor variables with more than two levels (smoking score and APOE-ε status), level 0 is used as a baseline and we estimate the effects of level 1 and level 2 relative to baseline. All models include the same confounding variables age, sex, age-sex interaction and head size; Voxels with at least four individuals having a WMH explored (40,001 voxels in the brain mask) and 5% FDR correction applied, i.e. % FDR-corrected voxels is out of a total of 40,001 voxels. Columns 2 and 3 complementary to Figure 3.5(a) and 3.5(b), respectively.

Predictor (level)	FDR-corrected voxels % (voxel count)	
	Joint model (S.2)	Marginal models (S.3.1-S.3.6)
Hypertension risk (1)	11.8% (4,705)	13.4% (5,366)
Hypercholesteromia (1)	0.02% (10)	1.6% (648)
Diabetes (1)	1.5% (607)	8.2% (3,270)
Smoking score (1)	0.3% (133)	1.6% (636)
Smoking score (2)	1.1% (424)	4.1% (1,633)
Waist-to-hip ratio (1)	4.6% (1,841)	10.7% (4,297)
APOE-ε status (1)	0.0% (0)	0.0% (0)
APOE-ε status (2)	4.2% (1,708)	3.9% (1,549)

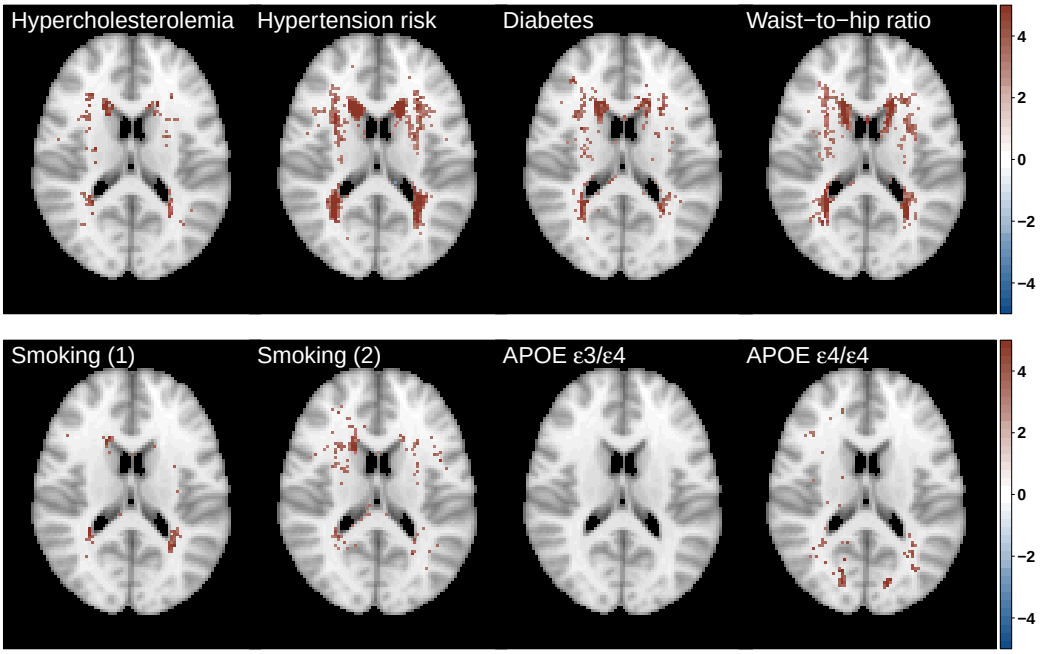
Table 3.3). We further plotted the APOE $\varepsilon 4/\varepsilon 4$ effect and found it was associated with WMH load concentrated at the boundary of the occipital and temporal lobes (Figure 3.A.1). While the spatial distribution of WMHs associated with WHR was concentrated on the periventricular areas and had a similar distribution to diabetes positive status. Notably, when examining the unique contribution of the risk factors (Figure 3.5(a)), diabetes was much reduced in spatial extent (drop from 8.2% in the marginal model to 1.5% in the joint model), suggesting its contribution typically seen in the literature may be confounded by other CVR factors. Hypertension, APOE-ε status and WHR remained important risk factors, even after controlling for other CVRs (Figure 3.5(a), Table 3.3).

The estimated effect of the composite variable CVR score on WMH probability (Figure 3.4) reflected the combined but varied effect of its constituent risk factors (Figure 3.5(a)).

Additional analyses were run to ensure our WHR findings were not driven by associated hypertension, using continuous WHR and systolic BP (as opposed to binary indicators used in the main analysis). We investigated whether the spatial distribution of WMHs is independently affected by WHR and systolic BP. Figure 3.A.2 shows the WHR effect is wider in terms of spatial extent with more spatially spread unique effects



(a) Model S.2 (confounding variables and all six CVR categorical variables).



(b) Models S.3.1-S.3.6 (confounding variables and one of the six CVR categorical variables)

Figure 3.5: Significance maps (z-scores based on mean bias-reduced estimates) for (a) model S.2 (joint) and (b) models S.3.1 - S.3.6 (marginal). All models include the same confounding variables as models S.1 (age, sex (baseline men), age-sex interaction and head size). Data on 13,680 UK Biobank individuals, and voxels with at least four individuals having a WMH explored (i.e. 0.03% WMH incidence); 5% FDR correction applied; axial slice z=45 shown.

in deep white matter (Figure 3.A.2(d)).

3.3.4 Cerebrovascular risk and speed of processing

Univariate linear regression estimated the relationship between log(WMH load) and speed of processing (reaction time in the 'Snap' task (Table 3.4)). Log(WMH load) was strongly associated with reaction time (p=0.009). The association between CVR score and reaction time did not appear to be significant after adjusting for confounding. A model including total CVR burden along with log(WMH load) showed only log(WMH load) to be a significant predictor of reaction time (p=0.011) but no explanatory power was gained (R^2_a did not change).

Table 3.4: Univariate regression summaries outlining the association between the cognition variable reaction time (as outcome) and log(WMH load) and CVR score (as explanatory variables). Model U.3.1 - U.3.3: main effects of interest shown. All models adjusted for age, sex, head size, years of education and age-sex interaction. $\hat{\beta}$ stands for the maximum likelihood estimate of the regression coefficient β.

Model / Predictor	Estimate $\hat{\beta}$	95% CI	p-value	$\boldsymbol{R^2_a}$/partial R^2_a
Model U.3.1				**0.088**
log(WMH load)	2.55	(0.63; 4.47)	0.009	<0.001
Model U.3.2				**0.088**
CVR score	0.96	(–0.57; 2.50)	0.219	<0.001
Model U.3.3				**0.088**
log(WMH load)	2.42	(0.47; 4.36)	0.015	<0.001
CVR score	0.63	(–0.93; 2.18)	0.431	<0.001

Next, mediation analysis was used to determine whether any of the CV risks could explain the relationship between speed of processing and log(WMH load). Mediation analysis (Table 3.5) found WHR to be the only CVR factor to have a significant mediation effect (for the other risk factors, see Table 3.A.4). Our results suggested 21% (9%; 83%) of the total effect of log(WMH load) on reaction time is explained through WHR.

Table 3.5: Mediation analysis. A significant proportion of the log(WMH load) (X) effect on speed of processing (Y) is explained through WHR mediation (mediator M). Models (mediator and outcome) are controlled for age, sex, age-by sex interaction, years of education and head size. See diagram 3.A.3 for an illustration of the mediation analysis. $\hat{\beta}$ stands for the maximum likelihood estimate of the regression coefficient β.

Average effect	WHR (continuous)		
	Estimate $\hat{\beta}$	95% CI	p-value
X on M	0.11	(0.10; 0.13)	<0.001
M on Y	0.05	(0.03; 0.08)	<0.001
Mediation effect	0.005	(0.003; 0.010)	$<10^{-15}$
Direct effect	0.020	(0.0003; 0.040)	0.046
Total effect	0.025	(0.006; 0.040)	0.010
Proportion mediated	0.212	(0.089; 0.830)	0.010

3.4 Discussion

White matter hyperintensity associated with individual risk factors

Population level brain imaging, lifestyle and demographic data on 13,680 healthy ageing volunteers were used to systematically investigate the association of cerebrovascular risk factors with the total burden and voxel-wise spatial distribution of WMHs. Contrary to previous reports, which typically emphasise hypertension as the main risk factor associated with WMH load, other cerebrovascular risk factors, including high waist-to-hip ratio, had similar or higher magnitude association with WMH burden, a unique spatial distribution and an independent relationship with cognition.

The contribution of known cerebrovascular risk factors to total WMH burden was examined. All CVR factors were found to be significant predictors of WMH load, motivating our exploration of the spatial distribution of the individual risk factors. Previous work in a subset (N=9,722) of the UK Biobank cohort has shown independent contributions of hypertension, diabetes, WHR and smoking pack years to the total WMH load (Cox et al., 2019). We replicate this finding, showing an additional risk of homozygous APOE ε4 status and hypercholesterolemia in a larger cohort with complete data. Hypertension is frequently found to be the most predictive risk factor for total WMH load and therefore management of blood pressure is recommended to reduce both the total burden and the progression of WMHs (Verhaaren et al., 2013). There was a higher main effect size associated with the presence of diabetes compared to hypertension with

the former associated with a 1.35-fold increase in total WMH burden compared to 1.27-fold increase for hypertension. However, risk factors including heavy smoking, APOE $\varepsilon 4/\varepsilon 4$ status and WHR had associations of similar magnitude, or greater, than hypertension. We found the presence of risk factors associated with higher total WMH load across all ages from 45−80 (Knopman et al., 2001; Debette et al., 2011). Together this points to the need for careful management of multiple risk factors across ageing for the preservation of brain vascular health.

A descriptive profile of this ageing population revealed overall higher cerebrovascular risk score (i.e. sum of risk factors) in males than females from mid through to late life. In terms of total WMH load, males and females had similar levels that increased linearly until age 65, after which total WMH load is higher in females. Several epidemiological studies corroborate this sex effect, with women appearing to consistently have a higher total WMH load than men (Sachdev et al., 2009). The age at which the difference in total WMH diverges between males and females has varied across studies, likely the result of limited sample sizes and age ranges within cohorts. One of the largest studies of WMH prevalence, the Rotterdam sample of healthy adults aged 60−90 (De Leeuw et al., 2001) found no sex differences in total WMH load, but did find sex differences across all decades in frontal periventricular WMH load (De Leeuw et al., 2001). As well as lower resolution scans (1.5 T) and a lower population size (N=1,077) compared to ours (3T and N=13,680), the study used a qualitative rating scale based on anatomical landmarks. Here, we used an objective method and voxel-wise test statistics across the whole brain to show sex effects and a sex by age interaction in 13,680 individuals concentrated in periventricular regions, with increased load in females over 65. This finding is important, given the higher risk of dementia associated with WMH burden and the higher prevalence of dementia in women. Notably, total CVR is higher in men of all ages, suggesting the increased WMH load seen in women aged over 65 may not be driven by cerebrovascular risk factors (or at least not the dominant risk factors included in our study). There is some evidence of a genetic component to WMHs, with higher heritability in women (Atwood et al., 2004; DeCarli et al., 1999). Another prominent hypothesis for higher total WMH load in women suggests the influence of sex hormones, particularly around menopause may be important. It is not yet known whether the higher incidence of AD in women is a risk or result of increased WMH load.

Longitudinal data in UK Biobank (both imaging data and health outcomes) may help to establish if the increased WMH load in older women is associated with higher incidence of dementia.

The spatial distribution of individual cerebrovascular risk factors

To date, the majority of studies have examined either total WMH load or deep versus periventricular WMHs associated with cerebrovascular risk factors. Where voxel-wise analyses have been attempted, this has either been using logistic regression (Lampe et al., 2019b), which we have discussed produces unstable test statistics or other "ad hoc" methods (Lampe et al., 2019a), but have not examined independent cerebrovascular risk factors voxel-wise.

Examination of the spatial distribution of individual cerebrovascular risk factors, led to several important findings. There were unique spatial patterns for certain risk factors and it was possible to quantify the contribution of different risk factors to total WMH load. Hypertension, WHR and APOE $\varepsilon 4$ homozygosity emerge as the dominant risk factors in terms of the spatial extent of the probability of WMHs and the number of significant voxels after controlling for head size, age, sex and their interaction. However, both hypercholesterolemia and diabetes did not reveal consistent patterns of spatial distribution in models controlling for the other risk factors, despite the latter being an important predictor of total WMH load. The finding suggests these risk factors may interact with other risk factors such as hypertension and do not present a direct path to the pathogenesis of WMHs.

The homozygous $\varepsilon 4$ genotype was revealed to be a significant predictor of WMH load. The finding is consistent with evidence of increased WMH volume in $\varepsilon 4$ carriers in the UK Biobank cohort and longitudinal evidence of WMH progression associated with the $\varepsilon 4$ genotype (Cox et al., 2017). Here we also replicate the finding from Lyall et al. (2019), in which there is no age interaction observed with APOE $\varepsilon 4$ status in terms of total WMH load, contrary to some reports that APOE $\varepsilon 4$ effects are most prominent in older age (Schiepers et al., 2012). Importantly, we extend these findings to show the spatial distribution of WMHs uniquely associated with the APOE $\varepsilon 4/\varepsilon 4$ genotype is concentrated in posterior deep white matter around the intracalcarine sulcus and extending superiorly into the temporal lobes. As evidence of the utility of our method,

these independent effects of APOE $\varepsilon 4/\varepsilon 4$ status associated with WMHs in the deep white matter of the temporal-occipital lobes would not have been obvious with an approach examining only periventricular versus deep WMHs. The proximity of these WMHs to the medial temporal lobe is notable, given the susceptibility of this region to atrophy and disruption in AD. Given the association between APOE $\varepsilon 4$ and dementia risk, it raises the question as to the potential contribution of WMHs to this risk. Longitudinal data is required to understand whether these WMHs have a role in the development or increased risk of dementia. Future studies would benefit from applying this voxel-wise method to examine how medial temporal lobe networks are impacted in the presence of WMHs in this region.

Waist-to-hip ratio above a healthy, sex-specific threshold, emerged as a critical risk factor for management in ageing. There was an independent spatial distribution of WMHs associated with WHR, that was not explained by comorbid hypertension, and was concentrated in the deep white matter and the ventricular caps. WHR was also the only risk factor to show a mediation effect on the relationship between speed of processing and total WMH burden. Waist circumference has been shown to be a reliable surrogate of visceral adiposity (Onat et al., 2004), and WHR is an especially useful measure in an ageing population because intra-abdominal fat tends to increase with age, whereas subcutaneous fat increases with degree of obesity but not age (Seidell et al., 1988). Cox et al. (2019), also noted an independent contribution of waist-to-hip ratio to WMH volume and suggest there may be metabolic and endocrine contributions to the pathogenesis of these WMHs that is distinct from arterial stiffness associated with high body mass (Cox et al., 2019). The dominance of WMHs associated with waist-to-hip ratio in deep white matter points to a possible ischaemic pathogenesis. Previous work examining deep versus periventricular WMHs associated with visceral obesity also found increased probability of WMHs in deep white matter associated with raised levels of proinflammatory cytokines (Lampe et al., 2019b). Proinflammatory cytokines are elevated in obesity and have been associated with cognitive decline and neurodegenerative processes associated with dementia (Pasha et al., 2017). This may help to explain the observed relationship to speed of processing we found with mediation analysis.

Here, we introduce a voxel-wise method that enables plotting the probability of the

presence of WMHs associated with different risk factors or variables such as cognitive scores or symptoms such as depression or anxiety. The existing literature provides a confusing picture relating different cerebrovascular risk factors to the location of WMHs, and this is largely due to different classifications of lesion locations, either total load, separated into periventricular versus deep WMHs or divided between lobes or tracts. Our method provides a quantitative approach that can be used to standardise how WMHs are spatially identified across the brain. Our method has several applications, because of the granularity of spatial localisation that can be produced at a voxel-wise level. For example, future studies might integrate voxel-wise results with other imaging modalities to examine how structural or functional networks are impacted by WMHs in specific regions.

The relationship between risk factors and cognition

Reductions in speed of processing have been most frequently associated with total WMH load, but it is not clear if all WMHs contribute to this impairment or whether particular risk factors are implicated. It is now widely accepted that cognition is reliant on distributed brain networks. The spatial distribution of WHMs may therefore directly impact particular brain networks resulting in the observed cognitive deficits. Different risk factors may increase the likelihood of WMH in certain regions and therefore differentially impact cognition. The conflicting evidence in the literature, as to whether WMHs do, or do not, correlate with cognitive performance may be in part due to conflation of WMHs associated with different CVR factors which present different spatial profiles. Beyond the spatial distribution, the contribution of individual risk factors to cognitive decline, controlling for dominant risk factors like hypertension, may inform clinical management strategies. In our analysis, WHR in particular showed a mediating effect on the relationship between speed of processing and total WMH burden. Future studies might profitably examine the longitudinal relationship between visceral adiposity and cognition to assess its importance as an early marker of cognitive decline. It is important to note that the explanatory power of the cognition models was relatively low which explains the magnitude of the regression coefficients, e.g. a change of 1 in log(WMH load) would only increase reaction time by 2.6ms, which is a significant effect but potentially not clinically meaningful.

Limitations

Our findings should be interpreted in light of some limitations in the data used and biases existing within our cohort. There are limits to the conclusions that can be drawn from cross-sectional data, but the benefit of the cohort is a very large population size with a wide age range. Longitudinal studies are often limited in the age range and number of individuals that can be feasibly sampled over time. We used outputs from the UK Biobank pipeline for the estimation of WMHs. This pipeline did not directly account for ventricular size. Ventricular size is known to increase with age and so may affect segmentation of periventricular WMHs in older individuals. Nevertheless, the UK Biobank pipeline has a number of quality assurance steps to exclude individuals with large structural deviations (such as overly enlarged ventricles) and ensure accuracy of normalisation processes (Alfaro-Almagro et al., 2018). Due to sampling biases within the UK Biobank cohort (Fry et al., 2017) and our decision to exclude non-white individuals, our sample is limited in its generalisability to other ethnicities and sociodemographic groups. The UK Biobank sample is known to be generally healthier with higher socioeconomic status than the general UK population. To a certain extent, it is interesting to see effects of cerebrovascular risk factors in a relatively healthy cohort, and the effects in the general population may be much more pronounced. We were somewhat limited in the measurements we could use to represent different risk factors, with the majority being categorical variables. Continuous measure may have shown increased sensitivity. For example, WHR is a reliable surrogate marker of visceral adiposity (Onat et al., 2004) that is sensitive to age and metabolic syndromes (Seidell et al., 1988; Shuster et al., 2011), however it lacks the precision of CT or MRI for highly specific and comprehensive assessment of intra-abdominal fat. Future studies should assess intra-abdominal directly with CT or MRI. UK Biobank has collected such data but it was unavailable at the time of analysis.

Conclusion

Our findings have some important clinical implications that may impact the management of cerebrovascular risk factors. We show the relative importance of different cerebrovascular risk factors in the contribution to total white matter hyperintensity burden

in healthy ageing and the varied spatial distribution of CVR-related WMHs across the brain. Contrary to the assumption of hypertension as the dominant risk factor associated with WMH load, we show the associations of similar magnitude with APOE $\varepsilon 4/\varepsilon 4$ status, WHR, diabetes and heavy smoking. Independent and unique spatial distributions of WMHs associated with high WHR and APOE $\varepsilon 4/\varepsilon 4$ status point to careful management and observation of these risk factors.

Waist-to-hip ratio above healthy, sex-specific thresholds emerged as a key risk factor associated with WMHs in deep white matter and ventricular caps. There was some evidence of cognitive consequences to WHR-associated WMHs suggesting visceral adiposity, as indexed by WHR, represents a risk factor for close clinical management for mitigation of cognitive decline in healthy ageing.

Data availability

The spatial distribution maps produced as part of the analysis are available at NeuroVault: `https://neurovault.org/collections/AZQTNVUF/`.

Credit authorship contribution statement

Michele Veldsman: Conceptualization, Data curation, Writing - original draft. **Petya Kindalova:** Conceptualization, Data curation, Formal analysis, Methodology, Validation, Visualization, Writing - original draft. **Masud Husain:** Funding acquisition, Supervision, Writing - review & editing. **Ioannis Kosmidis:** Conceptualization, Methodology, Supervision, Writing - review & editing. **Thomas E. Nichols:** Conceptualization, Funding acquisition, Methodology, Supervision, Writing - review & editing.

Acknowledgements

Thank you to Dr Xin-You Tai for input on the design of the study. We would also like to thank Fidel Alfaro-Almagro, Ludovica Griffanti and Mark Jenkinson for their helpful advice regarding the UK Biobank imaging pipeline and BIANCA segmentation. Thank you to the participants of the UK Biobank for their time volunteering for the study.

Appendices

3.A Complementary tables and figures

Table 3.A.1: List of codes of the UK Biobank Data-field 'Non-cancel illness' (`http://biobank.ndph.ox.ac.uk/showcase/field.cgi?id=20002`) used for exclusion of participants in the data cleaning process. 616 illnesses reported across 590 participants.

Coding	Description	Number of participants
1081	Stroke	185
1082	Transient ischaemic attack	115
1083	Subdural haematoma/haemorrhage	6
1086	Subarachnoid haemorrhage	10
1240	Neurological injury/trauma	0
1243	Psychological/psychiatric problem	0
1244	Infection of the nervous system	0
1245	Brain abscess/Intracranial abscess	0
1246	Encephalitis	2
1247	Meningitis	55
1258	Chronic neurological problem	0
1259	Motor Neuron Disease	2
1261	Multiple Sclerosis	62
1262	Parkinson's disease	36
1263	Dementia/Alzheimer's disease/Cognitive impairment	9
1264	Epilepsy	77
1266	Head Injury	14
1408	Alcoholism	10
1409	Opioid dependency	0
1410	Other dependency	0
1434	Other neurological problem	6
1491	Brain haemorrhage	10
1583	Ischaemic stroke	1
1626	Fracture skull/head	16

Table 3.A.2: Characteristics of UK Biobank dataset on 5,338 participants, who were excluded from the analysis due to missingness. There does not seem to be systematic differences between the two data sets, so the exclusions are not likely to have introduced some bias in the results of the analyses. Similar to Table 3.1 with an additional column summarising the number of missing values for each variable. Summaries calculated without missing values for each variable. The percentage values under each discrete variable in column 1 are (% of participants with risk factor present (level 1) in exclusions dataset (calculated from Table 3.A.2)) vs (% of participants with risk factor present in dataset used for the analyses (calculated from Table 3.1)), note for variables with more levels, vectors of incidences are presented.

Characteristics	Levels (N)	Mean (SD)	Median (range)	Missing
Age (years)	—	63.7 (7.5)	64.4 (46.6; 80.3)	0
Sex (53% vs 53% women%)	Men (2,496), Women (2,842)	—	—	0
Head size	—	1.3 (0.12)	1.3 (0.9; 1.8)	0
Hypertension risk (46% vs 47%)	0 (1,769), 1 (1,494)	—	—	2,075
Diabetes (5% vs 5%)	0 (4,932), 1 (241)	—	—	165
Smoking (score/pack years) (83%,16%,1%) vs (83%,16%,1%)	0 (1,827), 1 (340), 2 (25)	4.7 (11.3)	0 (0; 117.5)	3,146
WHR (indicator/continuous) (48% vs 47%)	0 (2,563), 1 (2,362)	0.9 (0.1)	0.9 (0.6; 1.2)	413
APOE-ε status (75%,23%,2%) vs (75%,23%,2%)	0 (3,622), 1 (1,092), 2 (110)	—	—	514
WMH load ($2mm^3$ voxels)	—	593 (732.1)	356 (3; 11,450)	0
Reaction time* (milliseconds)	—	594.3 (109.9)	577 (150; 1,445)	1,505
Years of education (8%,15%,5%,12%,12%,15%,45%) vs (6%,13%,6%,12%,16%,47%)	7 (468), 10 (900), 13 (330), 15 (704), 19 (946), 20 (2,752)	16.4 (4.5)	19 (7; 20)	189

*6,289 participants for the cognition variable reaction time and years of education.
SD: standard deviation; WHR: waist-to-hip ratio; APOE: apolypoprotein-E; CVR: cerebrovascular risk; WMH: white matter hyperintensity.

Table 3.A.3: List of UK Biobank variables (available at `http://biobank.ndph.ox.ac.uk/showcase/search.cgi` used in the analysis. Data on APOE-ε status and WMH masks are not part of the catalogue.

Data-Field ID	Description	Usage
31	Sex	
34	Year of birth	Age calculation
52	Month of birth	Age calculation
20002	Non-cancel illness code, self-reported	
21000	Ethnic background	
53	Date of attending assessment centre	Age calculation
25000	Volumetric scaling from T1 head image to standard space	Head size
6138	Qualifications	Years of education
	'none of the above'	7
	'CSEs or equivalent'	10
	'O levels/GCSEs or equivalent'	10
	'A levels/AS levels or equivalent'	13
	'Other professional qualification'	15
	'NVQ or HNC or equivalent'	19
	'College or University degree'	20
	'Prefer not to answer'	NA
20023	Mean time to correctly identify matches	Speed of processing
20116	Smoking status	
3436	Age started smoking in current smokers	
2867	Age started smoking in former smokers	
6194	Age stopped smoking cigarettes (current cigar/pipe or previous cigarette smoker)	Pack years calculation
2897	Age stopped smoking	
6183	Number of cigarettes previously smoked daily (current cigar/pipe smokers)	
2887	Number of cigarettes previously smoked daily	
2907	Ever stopped smoking for 6+ months	
3486	Ever tried to stop smoking	
6177	Medication for cholesterol, blood pressure or diabetes (males only)	Cholesterolemia and Hypertension risk
6153	Medication for cholesterol, blood pressure, diabetes, or take exogenous hormones (females only)	
2443	Diabetes diagnosed by doctor	Diabetes
4079	Diastolic blood pressure, automated reading	Hypertension risk
94	Diastolic blood pressure, manual reading	
4080	Systolic blood pressure, automated reading	
93	Systolic blood pressure, manual reading	
48	Waist circumference	Waist-to-hip ratio
49	Hip circumference	

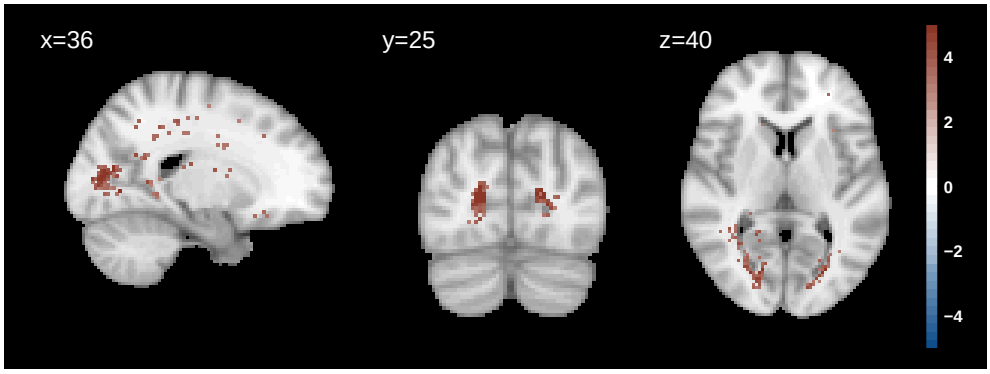
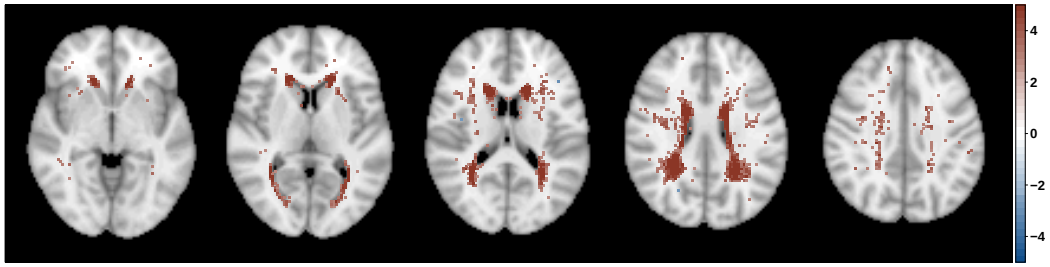
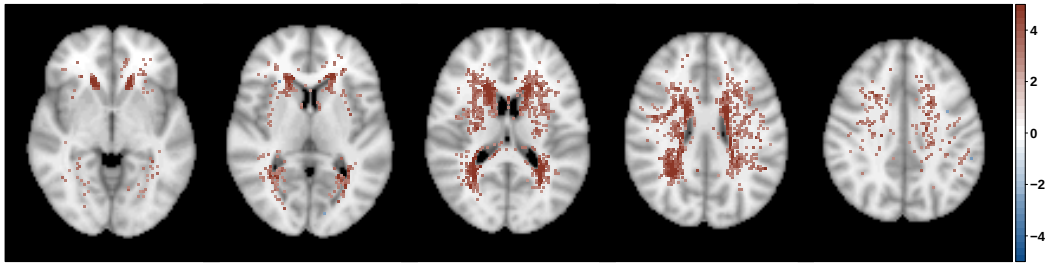


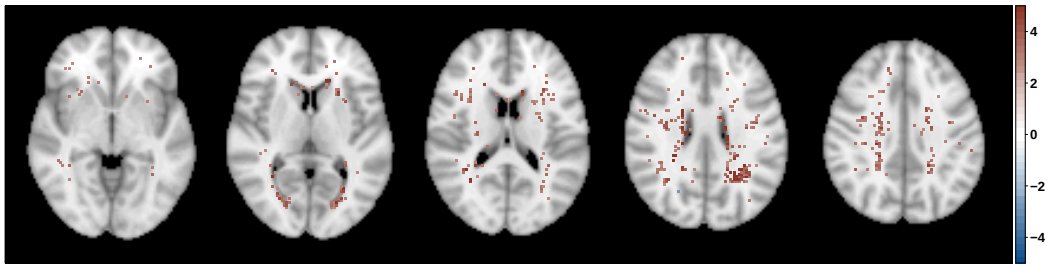
Figure 3.A.1: Significance maps (z-scores based on mean bias-reduced estimates) for APOE $\varepsilon 4/\varepsilon 4$ effect compared to no $\varepsilon 4$ alleles (marginal model). Data on 13,680 UK Biobank individuals, and voxels with at least four individuals having a WMH explored (i.e. 0.03% WMH incidence); 5% FDR correction applied; Slices x=36, y=25, z=40 shown.



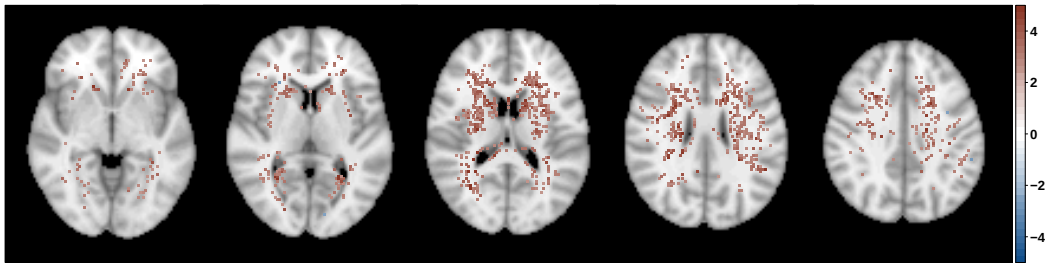
(a) Systolic BP; 4,999 voxels significant across brain.



(b) WHR; 8,574 voxels significant across the brain.



(c) Systolic BP unique effect; 2,241 voxels significant across the brain.



(d) WHR unique effect; 5,816 voxels significant across the brain.

Figure 3.A.2: Significance maps (z-scores based on mean bias-reduced estimates) across five axial slices $z = \{35, 40, 45, 50, 55\}$ in a model including age, sex, age-sex interaction, head size, systolic blood pressure and waist-to-hip ratio; data on 13,680 UK Biobank individuals and voxels with at least four individuals having a WMH explored; 5% FDR correction applied.
From top to bottom each row shows z-scores for (a) Systolic BP, (b) WHR and their 'unique' effect, respectively, (c) 'significant systolic BP/not significant WHR' and (d) 'not significant systolic BP/significant WHR'.
BP: blood pressure; WHR: waist-to-hip ratio.

Table 3.A.4: Mediation analysis exploring of the effect of log(WMH load) on speed of processing (reaction time) through cerebrovascular risk factors as mediator variable. Models (mediator and outcome) are controlled for age, sex, age-by sex interaction, years of education and head size. None of the risk factors shows a significant mediation effect. $\hat{\beta}$ stands for the maximum likelihood estimate of the regression coefficient β.

Average effect	Systolic BP (continuous)			Hypertension risk (indicator)		
	Estimate $\hat{\beta}$	95% CI	p-value	Estimate $\hat{\beta}$	95% CI	p-value
Mediation effect	–0.001	(–0.004; 0.000)	0.342	–0.002	(–0.005; 0.00)	0.222
Direct effect	0.027	(0.008; 0.050)	0.005	0.027	(0.008; 0.050)	0.005
Total effect	0.025	(0.007; 0.040)	0.006	0.025	(0.006; 0.040)	0.009
Proportion mediated	–0.054	(–0.296; 0.070)	0.346	–0.062	(–0.346; 0.050)	0.230
Average effect	Hypercholesterolemia (indicator)			Diabetes (indicator)		
	Estimate $\hat{\beta}$	95% CI	p-value	Estimate $\hat{\beta}$	95% CI	p-value
Mediation effect	0.001	(–0.000; 0.000)	0.141	0.001	(–0.001; 0.000)	0.227
Direct effect	0.024	(0.005; 0.040)	0.009	0.024	(0.006; 0.040)	0.012
Total effect	0.025	(0.006; 0.040)	0.007	0.025	(0.007; 0.040)	0.010
Proportion mediated	0.037	(–0.016; 0.170)	0.147	0.035	(–0.027; 0.160)	0.234
Average effect	Smoking[1] (continuous)			APOE-ε status (indicator[2])		
	Estimate $\hat{\beta}$	95% CI	p-value	Estimate $\hat{\beta}$	95% CI	p-value
Mediation effect	–0.001	(–0.002; 0.000)	0.045	0.000	(–0.001; 0.000)	0.885
Direct effect	0.026	(0.007; 0.050)	0.005	0.025	(0.007; 0.040)	0.009
Total effect	0.025	(0.006; 0.040)	0.008	0.025	(0.007; 0.040)	0.010
Proportion mediated	–0.043	(–0.183; 0.000)	0.053	0.002	(–0.044; 0.060)	0.886

BP: blood pressure; CI: confidence interval.
[1]log-transformed pack years used (0.05 added to 0 pack year values);
[2]APOE-ε status represented by 1 if homozygous ($\varepsilon 4/\varepsilon 4$), 0 otherwise.

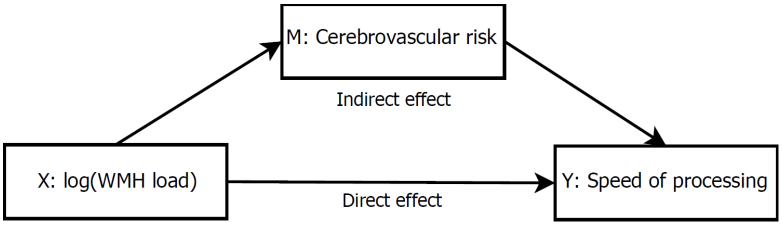


Figure 3.A.3: Mediation analysis diagram. The mediators explored are the six cerebrovascular risk factors.
All models are controlled for age, sex, age-sex interaction, and head size.
The main interest is in whether the indirect effect is significant, i.e. whether the effect of the predictor X on an outcome Y operates through a mediator variable M (fully or partially).

3.B Justification of Minimum WMH Count

We chose a lower limit $Y_{\min}$ based on the following heuristic: At any one voxel, what is the smallest WMH count that can reject a null hypothesis of zero WMH incidence, $H_0 : p = 0$? Of course, once a single WMH is observed we know H_0 must be false, but as a heuristic it seems useful to assert that, if *fewer* than $Y_{\min}$ WHM are seen, we *cannot even* reject this obviously false H_0, and this voxel should not be subject to further consideration.

This test takes the form $z = \hat{p}/\sqrt{\hat{p}(1-\hat{p})/N}$, where $\hat{p} = Y/N$. Solving $z \geq z_\alpha$ for $Y_{\min}$ shows that to obtain a minimum significance of z_α requires $Y_{\min} \geq 1/(z_\alpha^{-2} + 1/N)$. This result is virtually independent of N, giving $Y_{\min} \approx z_\alpha^2$, and for N=13,680 in particular, we found that $Y_{\min}$=4 was required to produce a z of at least 2, and hence we only considered voxels where 4 or more participants had a WMH.

Chapter 4

Penalized generalized estimating equations for relative risk regression with applications to brain lesion data

The attached preprint (Kindalova et al., 2021b) is submitted for review, and it is based on joint work with Michele Veldsman, Thomas E. Nichols and Ioannis Kosmidis. A statement of authorship form is attached at the end of the chapter.

Abstract

Motivated by a brain lesion application, we introduce penalized generalized estimating equations for relative risk regression for modelling correlated binary data. Brain lesions can have varying incidence across the brain and result in both rare and high incidence outcomes. As a result, odds ratios estimated from generalized estimating equations with logistic regression structures are not necessarily directly interpretable as relative risks. On the other hand, use of log-link regression structures with the binomial variance function may lead to estimation instabilities when event probabilities are close to 1. To circumvent such issues, we use generalized estimating equations with log-link regression structures with identity variance function and unknown dispersion parameter. Even in this setting, parameter estimates can be infinite, which we address by penalizing the generalized estimating functions with the gradient of the Jeffreys prior.

Our findings from extensive simulation studies show significant improvement over the standard log-link generalized estimating equations by providing

finite estimates and achieving convergence when boundary estimates occur. The real data application on UK Biobank brain lesion maps further reveals the instabilities of the standard log-link generalized estimating equations for a large-scale data set and demonstrates the clear interpretation of relative risk in clinical applications.

4.1 Introduction

Often data exhibit a natural clustering, such as repeated measurements taken on the same individual over time, where the individual is considered as a cluster. The fact that the measured outcomes could be correlated within each cluster brings modelling challenges which do not arise when modelling cross-sectional data, where outcomes are typically assumed to be independent.

We are motivated by a population level imaging data set from the UK Biobank (Miller et al., 2016). Data on two brain MRI scans per individual (about 2 years apart) along with demographic and lifestyle data allow us to investigate the effect of ageing and cerebrovascular risk on the spatial distribution of brain lesions. White matter lesions (Wardlaw et al., 2013) are a common finding on MRI in older populations, but their presence is not fully explained by ageing. Lesions are associated with increased risk of stroke and dementia (Wardlaw et al., 2015) and their spatial distribution is shown to vary with cerebrovascular risk factors (Veldsman et al., 2020) in a cross-sectional analysis. When modelling such binary brain lesion maps, we should account for the potential correlation between visits within each subject, but we should also ensure the desired interpretability of the estimated regression coefficients is achieved. Here, as in many medical applications, we would like to model and interpret relative risks, i.e. a ratio of probabilities. Hence, relative risk regression would be a better choice than the widely used logistic regression, which provides log-odds ratio estimates. For example, a relative risk of 1.1 for the age effect in a particular brain region suggests that lesions in that region are 10% more likely to occur if a participant is 1 year older. Lesions are rare at the population level with most voxels (volumetric pixels) having lesion incidence below 10%, which should be accounted for in the modelling to ensure stable estimates. The modelling is to be performed at the voxel level, i.e. dependence of nearby voxels is not modelled explicitly, which is known as a mass-univariate approach. Thus, the estimated relative risks could be obtained across the brain and presented in 3D spatial maps, which allows us to explore how different risk factors impact the spatial distribution of lesions. Beyond that binary brain lesion maps application motivates the current work, the methodology we develop is generally applicable for relative risk regression in longitudinal settings.

The main approaches used for modelling longitudinal data are marginal models (also known as population average models) and conditional models (such as mixed models). The former typically uses generalized estimating equations (Liang and Zeger, 1986), and the latter maximum likelihood estimation (Laird and Ware, 1982). Detailed overviews (Heagerty and Zeger, 2000; Gardiner et al., 2009; Fitzmaurice et al., 2011) and critiques (Lindsey and Lambert, 1998; Lee and Nelder, 2004; Hubbard et al., 2010) of these methods have been made, but here we focus on the two main areas where the two approaches differ; the interpretation of the parameters and the modelling assumptions.

A mixed effects model accounts for the unobserved cluster heterogeneity by the inclusion of random effects. The assumption is that the cluster-specific parameters are random variables that are distributed according to a distribution with a few unknown fixed parameters. Thus, inference is for the population, not only for the specific cohort, and we can also obtain subject-specific effects. The assumptions of the random effects model could be inadequate, for example (i) the misspecification of the random effect distribution can influence the power of the tests or can inflate Type I error rates (Litière et al., 2007), and (ii) the random effects are typically assumed to be uncorrelated with the explanatory variables, which implies that any omitted variables are uncorrelated with the explanatory variables. As this latter assumptions cannot be validated, it has led to some authors stridently arguing against the use of mixed effects models (Hubbard et al., 2010).

In contrast, generalized estimating equations (GEEs) (Liang and Zeger, 1986) model the mean unconditionally, which implies that inference can only be made about the average effect across all subjects in the given cohort. An advantage of the GEE approach is that it does not require distributional assumptions and valid inference relies solely on the correct specification of the outcome's mean and variance. The correlation within clusters is accounted for through the inclusion of a working correlation matrix which is treated as nuisance. The estimates of the regression coefficients are consistent, asymptotically unbiased, and asymptotically Normal when the number of clusters is sufficiently large even if the within subject correlation structure is misspecified (Liang and Zeger, 1986). However, GEE performance can suffer from small-sample bias (Sharples and Breslow, 1992; Sherman and Cessie, 1997), though bias corrections have been recently suggested by Paul and Zhang (2014). Another potential problem, discussed thoroughly for GLMs

for cross-sectional binary data (Mansournia et al., 2018), is called data separation (Albert and Anderson, 1984) and it is encountered when the covariates perfectly predict the outcome. A solution to this problem for the logit link GEE is proposed by Mondol and Rahman (2019), where motivated by Firth (1993), a Jeffreys-prior penalty is used to to ensure finite estimates.

When modelling binary data, either cross-sectional, or longitudinal, a logit link function is typically selected. In medical applications, the interest often lies in interpreting relative risks as opposed to odds ratios or absolute risk differences, which naturally leads to the usage of log link. Knol et al. (2012) discussed the problem of interpreting odds ratios as relative risks in cohort studies and randomized control trials (which could be acceptable in case-control studies due to the rare disease assumption (Greenland and Thomas, 1982; Greenland et al., 1986; Knol et al., 2008), typically if the outcome incidence is less than 10%). However, it is known that the odds ratio overestimates the risk ratio (when the risk ratio is higher than 1) and the higher the outcome incidence, the larger the overestimation (Zhang and Yu, 1998). Performing a simulation study to compare 8 alternative methods for obtaining adjusted risk ratios, Knol et al. (2012) recommend the use of log-Binomial regression when adjusting for multiple covariates, and Poisson regression with robust standard errors if the log-Binomial regression fails to converge, which the authors suggest happens when the outcome incidence is high. Such convergence issues happen when the success probabilities are close to one, which leads to numerical problems in iterative estimation algorithms. A fix proposed by Carter et al. (2005) is to use a quasi-likelihood with log link, identity variance function and known dispersion for Bernoulli outcomes instead of the likelihood equations. The authors show that the resulting estimates are consistent and asymptotically Normal; similar ideas are suggested by McNutt et al. (2003) and Zou (2004). Another approach to prevent the convergence problems when using log link is proposed by Fitzmaurice et al. (2014), where a Maclaurin series approximation to the Bernoulli weights in the likelihood equations is introduced.

To ensure convergence when outcome incidence goes to 1, similar developments follow for GEE, where Zou and Donner (2013) suggest the use of the so-called 'modified Poisson' GEE, i.e. GEE based on the first two moments of a Poisson model using sandwich variance-covariance and log link. Even though the simulation study undertaken by

Yelland et al. (2011a) suggests superior convergence performance of the modified Poisson GEE when compared to GEE based on the first two moments of a Binomial model, the authors state *"Surprisingly, modified Poisson regression also failed to converge on rare occasions"*. Pedroza and Truong (2017) have also compared GEE approaches to model the relative risk and they have also reported convergence problems for the modified Poisson model for small sample size settings. We believe some of the convergence issues reported for the modified Poisson GEE could be due to data separation, which is more likely to happen for rare events and thus is not reported by Carter et al. (2005) who focus on high incidence events.

In this work we propose a practical solution for modelling repeated measures binary data with a log-link GEE. To ensure stability of the estimates for high incidence events, we base our GEE on the first two moments of a Poisson model as Carter et al. (2005) do for independent data and Zou and Donner (2013) for panel data, while allowing for the estimation of the dispersion parameter. Additionally, in previous cross-sectional analysis of binary brain lesion maps (Veldsman et al., 2020), we have observed that boundary estimates occur quite often, especially for binary covariates such as sex. Thus, to avoid the problem of boundary estimates when dealing with rare outcomes, we add a Jeffreys-prior penalty as Mondol and Rahman (2019) do for logit-link GEEs. Modelling brain lesion data implies dealing with either rare events in outer white matter, or high incidence events in the periventricular areas, so our proposed modelling approach addresses both potential convergence issues and ensures the desired risk ratio interpretability through the choice of the log link function. Note that the mass-univariate approach we take, i.e. we fit a marginal model using GEE at each voxel independently, results in 19,801 regressions across the brain in the UK Biobank application considered, highlighting the need of scalable and stable estimation methods.

In Section 4.2 we start by providing an overview of the standard GEE and we then describe in detail the penalized version we propose. Those two modelling approaches for relative risk regression are then applied to simulated data sets and their performance is evaluated in terms of frequency of separation occurrence as well as estimator accuracy metrics, such as bias and mean-squared error (Section 4.3). To demonstrate the instabilities of the vanilla GEE approach and to reveal the lucid interpretation of relative risk in medical applications, we apply the methods to a subset of the UK Biobank data,

where we estimate the effect of ageing and cerebrovascular risk on lesion probability (Section 4.4).

4.2 Methods

First, we review the standard GEE approach and then describe the penalized version we propose to deal with boundary estimates in relative risk regression. We further outline the steps of the iterative estimation procedure and how we detect boundary estimates.

4.2.1 Generalized estimating equations

Suppose that there are N individuals and that each subject i ($i = 1, \ldots, N$) has a vector of correlated binary responses $\boldsymbol{y}_i = (y_{i1}, \ldots, y_{iT})^\top$ at T time points ($t = 1, \ldots, T$). Note that nothing in the present work depends on balanced data, but for notational simplicity we will assume all subjects have T observations.

The generalized estimating equations approach introduced by Liang and Zeger (1986) explicitly accounts for the correlation between repeated measures through the specification

$$\begin{aligned}
\mathrm{E}(\boldsymbol{y}_i|X_i) = \boldsymbol{\mu}_i = (\mu_{i1}, \ldots, \mu_{iT})^\top & \qquad \text{(marginal mean)} \\
V_i = W_i^{1/2} R_i(\boldsymbol{\alpha}) W_i^{1/2}/\phi & \qquad \text{(working covariance)} \\
g(\mu_{it}) = \eta_{it} & \qquad \text{(link function)} \\
\boldsymbol{\eta}_i = X_i\boldsymbol{\beta} & \qquad \text{(deterministic component)},
\end{aligned}$$

where

- y_{it} are assumed to be correlated within subject and to be independent between subjects, thus the working covariance matrix V is an $NT \times NT$ block-diagonal matrix with N blocks $V_1, \ldots, V_N$.
- $R_i(\boldsymbol{\alpha})$ is the working correlation matrix for subject i parameterised by parameters $\boldsymbol{\alpha}$, $W_i = \mathrm{diag}\{v_{it}, \ldots, v_{iT}\}$ is a $T \times T$ diagonal matrix with $v_{it} = \mathrm{Var}(\mu_{it})$ a known variance function, and ϕ is the dispersion parameter, which allows for the shrinkage or inflation of the contribution of the mean to the response variance. Here we assume that $R_i(\boldsymbol{\alpha}) = R(\boldsymbol{\alpha})$ and we subsequently suppress the index i.

- $g(.)$ denotes the link function, which is a monotonic function that relates the marginal mean for subject i at time point t to the linear predictor η_{it}.

- X_i denotes the $T \times P$ matrix of subject-specific covariates for subject i, where X_i collects time-specific covariate vectors $\boldsymbol{x}_{i1}, \ldots, \boldsymbol{x}_{iT}$ in its rows and has columns $\boldsymbol{X}_{i1}, \ldots, \boldsymbol{X}_{iP}$.

- $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_P)^\top$ is a P-vector of parameters (the regression coefficients we are interested in estimating).

The generalized estimating equations, as defined by Liang and Zeger (1986), are then

$$U(\boldsymbol{\beta}, \boldsymbol{\alpha}, \phi) = \sum_{i=1}^{N} D_i^\top V_i^{-1} (\boldsymbol{y}_i - \boldsymbol{\mu}_i) = \mathbf{0}, \tag{4.1}$$

where $D_i = \partial \boldsymbol{\mu}_i / \partial \boldsymbol{\beta}$ is a $T \times P$ matrix and U is a P-vector. The estimation steps for the regression coefficients $\boldsymbol{\beta}$ as well as the ancillary association parameters $\boldsymbol{\alpha}$ and the dispersion parameter ϕ are described in Section 4.2.3.

Choice of working correlation

The working correlation matrix $R(\boldsymbol{\alpha})$ sets the within-cluster correlation structure and it is fully characterised by the unknown parameter vector $\boldsymbol{\alpha}$. Even though its correct specification does not impact the consistency of the estimates of $\boldsymbol{\beta}$ (Liang and Zeger, 1986) and we could simply set R to the identity matrix, there is gain in efficiency in the estimation of $\boldsymbol{\beta}$ if the chosen dependence structure is close to the real one (Wang and Carey, 2003).

The most commonly used correlation matrices rely on a single association parameter α. Characteristic examples are (i) exchangeable correlation, also known as compound symmetry, where the observations within cluster share a common correlation α, and (ii) autoregressive correlation, where $\text{corr}(y_{it}, y_{it'}) = \alpha^{|t-t'|}$ $(t = 1, \ldots, T,\ t' = 1, \ldots, T)$, which implies that a natural ordering of the observations exists and the correlation decays to zero as the time separation between t and t' increases. There are other correlation structures such as Toeplitz, unstructured, etc. Crucially the choice of the correlation matrix should be considered as part of the model selection, since assuming independence can lead to quite substantial losses of efficiency (Fitzmaurice, 1995). For

informative discussions on the choice of the correlation matrix, we refer the reader to Ziegler and Vens (2010) and Westgate and Burchett (2017).

4.2.2 Penalized GEE

When dealing with rare responses or small sample size, data separation is likely to occur in logistic regression (Albert and Anderson, 1984), leading to infinite maximum likelihood estimates. To ensure finiteness of the MLE in the GLM framework, adjustments to the score equations were first introduced by Firth (1993) and the finiteness of the resulting estimates in logistic regression was later proved in Kosmidis and Firth (2021). In a similar manner, Jeffreys-prior penalty, also referred to as Firth-penalty, has been proposed in the GEE framework by Mondol and Rahman (2019) as a remedy for separation in GEEs with logit link. We refer to the Mondol and Rahman (2019) method as odds ratio penalized GEE (OR-PGEE). Adding a penalty to the standard GEE in (4.1), the PGEE for the p-th regression coefficient is of the form

$$U_p^*(\boldsymbol{\beta}, \boldsymbol{\alpha}, \phi) = U_p(\boldsymbol{\beta}, \boldsymbol{\alpha}, \phi) + A_p(\boldsymbol{\beta}, \boldsymbol{\alpha}) = 0\,.$$

A Jeffreys-prior penalty then implies adjusting the estimating equations by a vector A with p-th component

$$A_p(\boldsymbol{\beta}, \boldsymbol{\alpha}) = \frac{1}{2}\text{trace}\big[I(\boldsymbol{\beta}, \boldsymbol{\alpha}, \phi)^{-1}\frac{\partial}{\partial\beta_p}I(\boldsymbol{\beta}, \boldsymbol{\alpha}, \phi)\big]\,, \tag{4.2}$$

where $I(\boldsymbol{\beta}, \boldsymbol{\alpha}, \phi) = \text{E}(-\partial U(\boldsymbol{\beta}, \boldsymbol{\alpha}, \phi)/\partial\boldsymbol{\beta})$.

The motivation of our work is to ensure the direct interpretability of the estimated regression coefficients as relative risks through the usage of relative risk regression, so we propose the use of log-link function. To deal with the potential instability of the estimation algorithm for high incidence outcomes, we use a GEE with log-link and identity variance function ($v_{it} = \mu_{it}$) as Carter et al. (2005) do in cross-sectional settings, but with unknown dispersion. Because $D_i = W_i X_i$, the standard GEE in (4.1) simplifies to

$$U(\boldsymbol{\beta}, \boldsymbol{\alpha}, \phi) = \sum_{i=1}^{N} D_i^\top V_i^{-1}(\boldsymbol{y_i} - \boldsymbol{\mu}_i) = \phi\sum_{i=1}^{N} X_i^\top W_i^{1/2} R(\boldsymbol{\alpha})^{-1} W_i^{-1/2}(\boldsymbol{y_i} - \boldsymbol{\mu}_i) \tag{4.3}$$

where $W_i = \text{diag}\{\mu_{i1}, \ldots, \mu_{iT}\}$. We refer to the above GEE with log-link as RR-GEE.

Finally, we add a Jeffreys-prior penalty to (4.3) to avoid boundary estimates. Briefly, we first show that the expected negative Jacobian matrix of U is

$$I(\boldsymbol{\beta}, \boldsymbol{\alpha}, \phi) = \phi \sum_{i=1}^{N} X_i^\top W_i^{1/2} R(\boldsymbol{\alpha})^{-1} W_i^{1/2} X_i \,. \tag{4.4}$$

Then by further taking the derivative of I and substituting in expression (4.2), the penalty term is shown to be

$$A_p(\boldsymbol{\beta}, \boldsymbol{\alpha}) = \frac{1}{2} \sum_{i=1}^{N} \sum_{t=1}^{T} h_{it} x_{itp} \,,$$

where h_{it} is the t-th diagonal element of the i-th block of the projection matrix

$$H_i = W_i^{1/2} R(\boldsymbol{\alpha})^{-1} W_i^{1/2} X_i \Big[\sum_{i=1}^{N} X_i^\top W_i^{1/2} R(\boldsymbol{\alpha})^{-1} W_i^{1/2} X_i \Big]^{-1} X_i^\top \,.$$

So, the modified estimating equations can be written as

$$U_p^*(\boldsymbol{\beta}, \boldsymbol{\alpha}, \phi) = \phi \sum_{i=1}^{N} X_{ip}^\top W_i^{1/2} R(\boldsymbol{\alpha})^{-1} W_i^{-1/2} (\boldsymbol{y}_i - \boldsymbol{\mu}_i) + \frac{1}{2} \sum_{i=1}^{N} \sum_{t=1}^{T} h_{it} x_{itp}, \tag{4.5}$$

where X_{ip}, $\boldsymbol{y}_i$, $\boldsymbol{\mu}_i$ are T-column vectors. We refer to the above penalized estimating questions as relative risk penalized GEE (RR-PGEE). The detailed steps to obtain the form of the Jeffreys-prior penalty term $A_p(\boldsymbol{\beta}, \boldsymbol{\alpha})$ for this RR-GEE are included in Appendix 4.A.

4.2.3 Parameter estimation

In a similar manner to maximum likelihood parameter estimation in GLMs, we adopt an iterative procedure to solve the estimating equations (4.5) for RR-PGEE. The iterative process is as suggested by Liang and Zeger (1986) and it entails repeated use of a quasi Newton-Raphson iteration for the estimation of $\boldsymbol{\beta}$ and method of moments estimation of $\boldsymbol{\alpha}$ and ϕ. The resulting estimates $\tilde{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$ are the RR-PGEE estimates along with $\tilde{\phi}$ and $\tilde{\boldsymbol{\alpha}}$. The iterative procedure is as follows

Step 1 Initialization.

Set the intercept term to the log mean response incidence $\tilde{\beta}_1^{(0)} = \log\big(\sum_{i,t} y_{it}/(NT)\big)$

and $\tilde{\beta}_p^{(0)} = 0$ for $p = 2, \ldots, P$.

Step 2 For $k = 1, \ldots, K$, repeat the following two steps until a convergence criterion is satisfied, e.g. $|\tilde{\beta}_p^{(k+1)} - \tilde{\beta}_p^{(k)}| < \varepsilon$ for all parameters β_p, $p = 1, \ldots, P$ and positive convergence tolerance ε, or the maximum number of iterations K has been reached without convergence.

(i) Given $\tilde{\boldsymbol{\beta}}^{(k)}$ and assuming exchangeable working correlation, i.e. all off-diagonal elements of $R(\boldsymbol{\alpha})$ are equal to α:

$$\tilde{\phi}^{(k+1)} = \{ \sum_{i=1}^{N} \sum_{t=1}^{T} \tilde{r}_{it}^{(k)\,2} / (NT - P) \}^{-1}$$
$$\tilde{\alpha}^{(k+1)} = \frac{\tilde{\phi}^{(k+1)}}{N} \sum_{i=1}^{N} \frac{1}{T(T-1)} \sum_{t \leq T-1} \tilde{r}_{it}^{(k)} \tilde{r}_{i,t+1}^{(k)} ,$$

where the Pearson residuals at iteration k are $\tilde{r}_{it} = (y_{it} - \tilde{\mu}_{it}^{(k)}) / \sqrt{\tilde{\mu}_{it}^{(k)}}$ and $\tilde{\mu}_{it}^{(k)} = \exp(\boldsymbol{x}_{it}^{\top} \tilde{\boldsymbol{\beta}}^{(k)})$.

(ii) Using $\tilde{\alpha}^{(k+1)}$ and $\tilde{\phi}^{(k+1)}$, the quasi Newton-Raphson update is

$$\tilde{\boldsymbol{\beta}}^{(k+1)} = \tilde{\boldsymbol{\beta}}^{(k)} + [I(\tilde{\boldsymbol{\beta}}^{(k)}, \tilde{\alpha}^{(k+1)}, \tilde{\phi}^{(k+1)})]^{-1} U^*(\tilde{\boldsymbol{\beta}}^{(k)}, \tilde{\alpha}^{(k+1)}, \tilde{\phi}^{(k+1)})$$

Step 3 Set $\tilde{\boldsymbol{\beta}}$, $\tilde{\alpha}$ and $\tilde{\phi}$ to the values at the final iteration of Step 2.

Using the same steps as above, we can also obtain the RR-GEE estimates $\hat{\boldsymbol{\beta}}$, where U as in Equation (4.3) is used in the Newton-Raphson update in Step 2(ii) instead of U^*. Both sets of estimates $\tilde{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\beta}}$ are obtained using the log-link function and identity variance function with unknown dispersion.

Robust variance

To estimate the variance of the estimated regression coefficients $\tilde{\boldsymbol{\beta}}$, the sandwich variance-covariance matrix $\mathrm{Var}_S(\tilde{\boldsymbol{\beta}})$ proposed by Liang and Zeger (1986), also called robust variance, is used

$$\mathrm{Var}(\tilde{\boldsymbol{\beta}}) \approx \mathrm{Var}_S(\tilde{\boldsymbol{\beta}}) = \Big(\sum_{i=1}^{N} D_i V_i^{-1} D_i \Big)^{-1} \Big\{ \sum_{i=1}^{N} D_i V_i^{-1} \mathrm{Cov}(\boldsymbol{y}_i) V_i^{-1} D_i \Big\} \Big(\sum_{i=1}^{N} D_i V_i^{-1} D_i \Big)^{-1}, \tag{4.6}$$

where to obtain the variance estimate $\widehat{\text{Var}}_S(\tilde{\boldsymbol{\beta}})$, we replace $\text{Cov}(\boldsymbol{y}_i)$ by $(\boldsymbol{y}_i - \boldsymbol{\mu}_i)(\boldsymbol{y}_i - \boldsymbol{\mu}_i)^\top$ and $\boldsymbol{\beta}, \phi$ and α by their estimates; the same is done to obtain $\widehat{\text{Var}}_S(\hat{\boldsymbol{\beta}})$. The consistency of $\tilde{\boldsymbol{\beta}}$ and $\widehat{\text{Var}}_S(\tilde{\boldsymbol{\beta}})$ does not depend on the correct choice of the working correlation matrix R, but only on the correct specification of the outcome mean regression (Liang and Zeger, 1986).

Detection of boundary estimates

To detect boundary estimates, we define the boundary estimates criterion (BEC) for the p-th regression coefficient at iteration k

$$\widetilde{\text{BEC}}_p^{(k)} = \sqrt{\left[(I(\tilde{\boldsymbol{\beta}}^{(k)}, \tilde{\alpha}^{(k)}, \tilde{\phi}^{(k)})^{-1}\right]_{pp}} \Big/ \sqrt{\left[(I(\tilde{\boldsymbol{\beta}}^{(0)}, \tilde{\alpha}^{(0)}, \tilde{\phi}^{(0)})^{-1}\right]_{pp}}, \tag{4.7}$$

where $k = 1, \ldots, K$ and I is the expected negative Jacobian matrix as defined in Equation (4.4). Similarly, we define $\widehat{\text{BEC}}$ by replacing $\tilde{\boldsymbol{\beta}}$ with $\hat{\boldsymbol{\beta}}$ as well as the estimates of α and ϕ.

Boundary estimates occur if this ratio diverges as k grows for any of the P estimated parameters. The ratio depends on the expected negative Jacobian of the estimating equations as given in Equation (4.4), which gives us information on the curvature of the estimating function U. To give some intuition behind this criterion, in the single parameter setting, when boundary estimates occur for large values of the parameter β the Jacobian will have value almost zero, thus I^{-1} will be diverging to large values. For $k = 1, \ldots, K$, we monitor $\widetilde{\text{BEC}}_p^{(k)}$, and if it diverges as k grows for any of the P estimated parameters, it suggests we have detected boundary estimates; similarly for $\widehat{\text{BEC}}_p^{(k)}$.

4.3 Simulation study

To compare the performance of our proposed penalized GEE for relative risk regression (RR-PGEE) and the associated standard GEE (RR-GEE), we perform a simulation study in a similar manner to Mondol and Rahman (2019). The main aim of the simulation study is to investigate the frequency of boundary estimates for correlated binary data and to evaluate how the modified GEE performs in comparison to the standard GEE in those cases.

4.3.1 Simulation setup

We simulate balanced correlated binary data y_{it} for subject i at time point t ($i = 1, \ldots, N$, $t = 1, \ldots, T$). Specifying our model as per Section 4.2.1, the deterministic component of our generative model is of the form

$$\eta_{it} = \beta_1 + \beta_b x_{1it} + \beta_c x_{2it},$$

where

- β_1 is the intercept term, which determines the response incidence in the data set.
- a binary covariate X_1 is considered, which is time-invariant (e.g. sex) and its values are sampled from Bernoulli distribution with probability of success c. The inclusion of a time-invariant binary covariate could potentially lead to separation and the regression coefficient β_b and its standard error could be diverging to infinity.
- a continuous covariate X_2 is considered, which is time-varying with equally-spaced values $\{0.2, 0.4, 0.6, \ldots\}$ for each subject i.

As the link function is log-link, the marginal mean is $\mu_{it} = \exp(\eta_{it})$. By specifying the marginal mean and variance of $\boldsymbol{y}$ and the correlation matrix $R(\boldsymbol{\alpha})$, the method proposed by Qaqish (2003) is used to simulate correlated binary data. Briefly, Qaqish introduced a family of multivariate Bernoulli distributions with a conditional linear property, which provides an efficient approach to simulate correlated binary variables. The code available as part of the `binarySimCLF R` package has been adjusted to account for our choices of link function and identity variance function $\text{Var}(\mu_{it}) = \mu_{it}$.

The fixed simulation parameters are as follows (i) fixed number of time points $T = 4$, resulting in $\boldsymbol{x}_{2i} = (0.2, 0.4, 0.6, 0.8)^\top$ across all subjects i, with the regression coefficient $\beta_c = 0.2$; (ii) correlation structure is set to exchangeable with correlation parameter α. The base simulation has parameters fixed at: regression coefficients $\boldsymbol{\beta} = (\beta_1, \beta_b, \beta_c)^\top = (-4, 1.6, 0.2)^\top$, sample size $N = 50$, time-invariant binary covariate proportion of 1's set to $c = 0.2$, within-subject correlation $\alpha = 0.4$. We vary each of five simulation parameters one at a time, keeping the others fixed as above, as follows

- $\beta_1 \in \{-4, -3, -2\}$: the overall response incidence, with smallest value giving rare

events and greatest chance of boundary estimates,

- $\beta_b \in \{1.2, 1.4, 1.6, 1.8, 2.0\}$: the binary covariate effect, with smaller effects increasing the chance of boundary estimates,

- $c \in \{0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8\}$: the proportion of 1's for the binary time-invariant covariate, with higher proportions increasing the risk of boundary estimates if the response incidence is low,

- $\alpha \in \{0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8\}$: the strength of the within-subject correlation,

- $N \in \{25, 50, 75, 100, 500, 1000\}$: the sample size, with the smallest sample sizes resulting in higher chance of boundary estimates.

4.3.2 Model evaluation

We simulate $R = 1,000$ data sets for each of the simulation scenarios listed above. For each repetition r ($r = 1, \ldots, R$), RR-GEE and RR-PGEE are fitted to obtain estimates $\hat{\boldsymbol{\beta}}_r$ and $\tilde{\boldsymbol{\beta}}_r$, respectively, along with their sandwich variance and boundary estimates criterion values. Note that for the simulation study, the true value of the dispersion parameter ϕ is set to 1, but it is estimated using RR-GEE or RR-PGEE when fitting the marginal models. The tolerance for convergence is set to $\varepsilon = 10^{-4}$ and the maximum number of iterations to $K = 25$.

To judge whether boundary estimates occur, we focus on the boundary estimates criterion (BEC) in Equation (4.7) obtained from fitting RR-GEE and we keep record of those values at each iteration of the IRLS algorithm. We state that boundary estimates occurred if BEC at the final iteration is greater than 10 for any of the three regression coefficients. Note that for some of the simulated data sets, either of the modeling approaches could return missing values due to numerical instabilities (referred to as 'failed IRLS'), or it can fail converging for the maximum number of permitted iterations K (referred to as 'non-converging').

To compare the two modelling approaches in terms of estimator accuracy, the bias and mean-squared error (MSE) of the regression coefficient β_b are estimated. For RR-GEE, bias is calculated as $\mathrm{B}(\hat{\beta}_b) = \frac{1}{R'} \sum_r \hat{\beta}_{b,r} - \beta_b$ and MSE as $\mathrm{MSE}(\hat{\beta}_b) = \frac{1}{R'} \sum_r (\hat{\beta}_{b,r} - \beta_b)^2$, where $\hat{\beta}_{b,r}$ is the RR-GEE estimate for the r-th simulated data set, β_b is the true

value, and summation is over the R' data sets where estimation did not fail. Similarly, $\mathrm{B}(\tilde{\beta}_b)$ and $\mathrm{MSE}(\tilde{\beta}_b)$ are obtained using RR-PGEE estimates $\tilde{\beta}_{b,\,r}$ across repetitions. To further explore the performance of the variance estimators of the regression coefficients, we compare the average sandwich variance $\widehat{\mathrm{Var}}_S(\hat{\beta}_b)$ with the variance of the estimates $\hat{\beta}_b$ across repetitions for RR-GEE and similarly for RR-PGEE $\tilde{\beta}_b$. If the sandwich variance correctly estimates the variance of the estimators, the ratio of those two quantities should be close to 1.

4.3.3 Results

Motivating example

As an illustrative example, we show the results from fitting RR-GEE and RR-PGEE to the 1,000 simulated data sets under the base simulation setup, i.e. $\boldsymbol{\beta} = (\beta_1, \beta_b, \beta_c)^\top = (-4, 1.6, 0.2)^\top$, $N = 50$, $c = 0.2$, $\alpha = 0.4$.

On 233 out of 1,000 simulated data sets boundary estimates occurred, meaning at least one of the three regression coefficients has BEC greater than 10. Note that the threshold of 10 is chosen empirically, for a histogram of the BEC values for β_b, see Figure 4.B.1. We randomly selected one of the 233 boundary estimates data sets and the results are summarized in Table 4.1. Increasing the number of iterations highlights the rapid increase in BEC as well as the elevated regression coefficient and z-score for the binary covariate obtained using RR-GEE, where $\hat{z}_b = \hat{\beta}_b / \sqrt{\widehat{\mathrm{Var}}_S(\hat{\beta}_b)}$, while the estimates based on RR-PGEE are more stable and the iterative algorithm converges after 10 iterations.

Table 4.1: Performance of one boundary estimates data set for the effect of the binary covariate ($p = 2$) across K=$\{5, 15, 25\}$ maximum number of iterations. The simulation parameters are set to $\boldsymbol{\beta} = (\beta_0, \beta_b, \beta_c)^\top = (-4, 1.6, 0.2)^\top$, $N = 50$, $c = 0.2$, $\alpha = 0.4$.

Iterations K	$\widehat{\mathrm{BEC}}_2^{(K)}$	$\hat{\beta}_b$	$\hat{z}_b$	$\widetilde{\mathrm{BEC}}_2^{(K)}$	$\tilde{\beta}_b$	$\tilde{z}_b$
5	4.01	6.95	9.24	1.61	5.19	7.26
15*	666.89	16.70	17.76	1.42*	4.55*	5.32*
25	98,978.90	26.70	28.55	–	–	–

*RR-PGEE converges after 10 iterations. RR-GEE does not converge and stops after the maximum number of iterations K.

One warning here is that if we explore the z-scores for the data sets where boundary estimates occur, we get large absolute values of $\hat{z}_b$ for most of those data sets when

fitting RR-GEE. The reality is that users tend to extract the z-scores without realizing that software packages handle boundary estimates differently and that the user may receive no warning about boundary estimates.

Of the remaining 767 data sets, 18 had numerical underflow or non-convergence failure for either of the methods. Figure 4.1 explores the remaining 749 estimated coefficients for converging data sets with finite estimates. The highest density of points is close to the true value of $\beta_b = 1.6$ for both RR-GEE and RR-PGEE, with RR-PGEE demonstrating shrinkage to the true value. Also, RR-PGEE leads to slightly higher z-scores than RR-GEE due to the shrinkage of the standard errors towards zero.

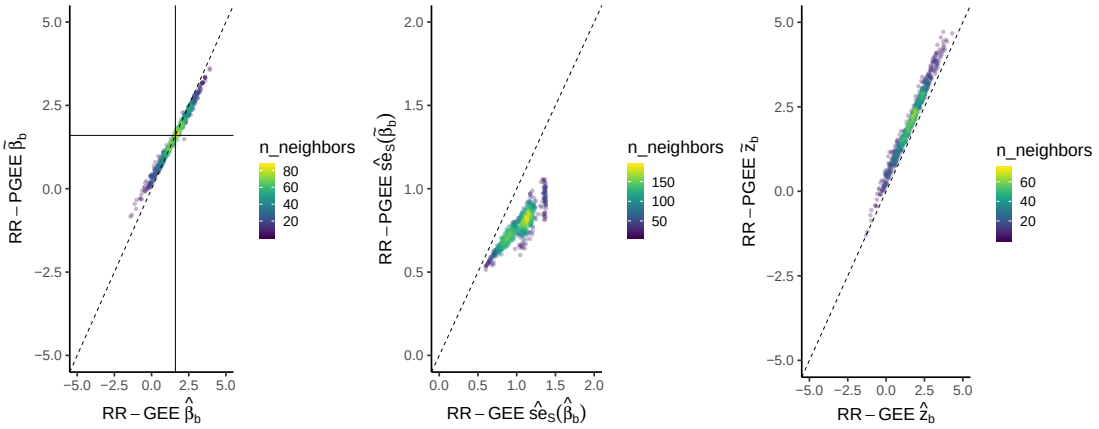


Figure 4.1: Illustration of converging data sets with finite estimates. (Left) Estimated regression coefficients for the binary covariate β_b estimated from 749 data sets, with solid lines at the true value of 1.6, (Middle) sandwich standard error, and (Right) estimated z-scores (estimated coefficient divided by sandwich standard error) for RR-PGEE vs RR-GEE. Dashed line is the identity line; those are density plots, i.e. the brighter the colour, the higher the density of the points. The plot summarises the 749 non-separated data sets out of 1,000 data sets simulated under the base simulation scenario: $\boldsymbol{\beta} = (\beta_1, \beta_b, \beta_c)^\top = (-4, 1.6, 0.2)^\top$, $N = 50$, $c = 0.2$, $\alpha = 0.4$.

Boundary estimates

The convergence issues we run into when fitting both RR-GEE and RR-PGEE are summarized in Table 4.2. The number of boundary estimates data sets out of 1,000 simulated under each scenario increases: (i) the rarer the event as controlled by the intercept term β_1, (ii) the weaker the effect of the binary covariate β_b, (iii) the stronger the association parameter α, (iv) the higher the proportion of 1's for the binary covariate, (v) the lower the sample size N. Given that the main aim of the penalized GEE for relative risk regression is to ensure finite estimates, our results reveal that the BEC

value is above the chosen threshold of 10 for a few of the data sets but IRLS always converges. It happens that IRLS fails for RR-PGEE as it does for RR-GEE, but it happens less frequently for most simulation scenarios. Overall, RR-PGEE performs well when boundary estimates are detected, i.e. when the BEC value for RR-GEE diverges. Note that for some data sets, both methods do not converge for the specified maximum number of iterations and the frequency of this happening is similar between the methods; a further check into those data sets (as in the last column of Table 4.2) shows that the outcome incidence is exactly zero for most of those data sets.

Estimator accuracy

In Table 4.3, we present the bias and mean-squared error of the binary covariate estimates $\hat{\beta}_b$ and $\tilde{\beta}_b$. We explore the data sets with finite and converging estimates to investigate how the penalty impacts the estimates, but we also look at the RR-PGEE unconditional summaries including estimates for boundary estimates data sets (in parenthesis). The results reveal that the bias is comparable between the two modelling approaches and RR-PGEE has lower MSE across the simulation scenarios, except for high proportion of 1's for the binary covariate ($c \in \{0.6, 0.7, 0.8\}$, Table 4.3). The response incidence is very low ($\exp(\beta_1) = \exp(-4) = 0.02$), so if we further explore sample sizes $N = 100$ and $N = 500$ for varying proportions of 1's in the binary covariate (see Table 4.B.1), the MSE and bias performance is encouraging, i.e. both are decreasing when sample size increases indicating consistency of the RR-PGEE estimates.

Table 4.4 reports on the variance properties, showing that RR-PGEE has smaller variance in all settings considered, i.e. $\text{Var}(\tilde{\beta}_b) < \text{Var}(\hat{\beta}_b)$. The accuracy of the estimated variance is assessed through the relative variance, i.e. mean of $\widehat{\text{Var}}(\hat{\beta}_b)$ across repetitions divided by variance of the estimated coefficients $\hat{\beta}_{b,r}$ (Table 4.4), and similarly for $\tilde{\beta}_b$. For the finite and converging data sets, the RR-GEE seems to be slightly overestimating the true variance of the binary covariate coefficient and RR-PGEE is slightly underestimating it, with ratios approaching 1 as N increases.

Note that the unconditional summaries for RR-PGEE (in parenthesis) exhibit stable performance for the data sets where boundary estimates are detected for RR-GEE. Furthermore, MSE and bias values decrease towards 0 and variance ratios go to 1 with increasing N, which we would expect from a consistent estimator.

Table 4.2: Simulation performance across simulation scenarios. The simulation parameters are set to $\boldsymbol{\beta} = (\beta_1, \beta_b, \beta_c)^\top = (-4, 1.6, 0.2)^\top$, $N = 50$, $c = 0.2$, $\alpha = 0.4$ (in bold) and each row of the table summarizes 1,000 replications with just one parameter altered, e.g. the first block of rows represents scenarios when β_1 is altered and all other parameters stay fixed.

Simulation setup		RR-GEE			RR-PGEE		
Parameter/Value		BEC>10 (non-conv.)	Failed IRLS	Non-conv. & BEC≤10	BEC>10 (non-conv.)	Failed IRLS	Non-conv. & BEC≤10
β_1	**-4**	233 (229)	7	11	16 (0)	4	11
	-3	8 (0)	0	0	0 (0)	0	0
	-2	0 (0)	0	0	0 (0)	0	0
β_b	1.2	308 (306)	11	33	23 (0)	4	33
	1.4	265 (263)	9	21	16 (0)	4	21
	1.6	233 (229)	7	11	16 (0)	4	11
	1.8	204 (197)	7	9	14 (0)	3	8
	2.0	186 (178)	9	4	13 (0)	4	6
α	0.2	137 (132)	1	3	8 (0)	4	3
	0.3	183 (177)	4	6	17 (0)	3	5
	0.4	233 (229)	7	11	16 (0)	4	11
	0.5	271 (267)	10	24	13 (0)	6	24
	0.6	317 (313)	23	39	15 (0)	7	42
	0.7	366 (362)	29	45	14 (0)	8	51
	0.8	422 (417)	39	71	16 (0)	11	75
c	**0.2**	233 (229)	7	11	16 (0)	4	11
	0.3	207 (207)	4	7	5 (0)	0	6
	0.4	244 (244)	1	3	1 (0)	0	2
	0.5	304 (304)	0	1	0 (0)	0	0
	0.6	397 (397)	1	1	0 (0)	0	0
	0.7	500 (500)	1	0	0 (0)	0	0
	0.8	643 (643)	0	0	0 (0)	0	0
N	25	476 (473)	11	122	8 (0)	29	118
	50	233 (229)	7	11	16 (0)	4	11
	75	110 (107)	5	4	13 (0)	0	6
	100	35 (33)	2	0	5 (0)	0	0

β_1 and β_b: true intercept and binary covariate regression coefficients; α: within-cluster correlation coefficient; c: proportion of 1's in binary covariate; N: number of subjects; BEC: boundary estimates criterion; IRLS: iteratively-reweighted least squares.

Table 4.3: Bias and mean-squared error of the binary covariate coefficient β_b across simulation scenarios. The simulation parameters are set to $\boldsymbol{\beta} = (\beta_1, \beta_b, \beta_c)^\top = (-4, 1.6, 0.2)^\top$, $N = 50$, $c = 0.2$, $\alpha = 0.4$ (in bold) and each row of the table summarizes 1,000 replications with just one parameter altered, e.g. the first block of rows represent scenarios when β_1 is altered and all other parameters stay fixed. Summaries included are conditional on both estimation methods converging and the estimates being finite (no boundary estimates), with unconditional summaries for RR-PGEE included in parenthesis.

Simulation setup		Finite & converging				
Parameter	Value	N.sim	B($\hat{\beta}_b$)	B($\tilde{\beta}_b$)	MSE($\hat{\beta}_b$)	MSE($\tilde{\beta}_b$)
β_1	**-4**	749 (969)	-0.04	-0.05 (0.19)	0.88	0.61 (2.18)
	-3	992 (1000)	0.03	0.03 (0.04)	0.53	0.42 (0.50)
	-2	1000 (1000)	0.03	0.02 (0.02)	0.12	0.11 (0.11)
β_b	1.2	648 (940)	0.12	0.15 (0.11)	0.89	0.62 (2.23)
	1.4	705 (959)	0.04	0.05 (0.16)	0.87	0.61 (2.26)
	1.6	749 (969)	-0.04	-0.05 (0.19)	0.88	0.61 (2.18)
	1.8	780 (975)	-0.05	-0.09 (0.23)	0.82	0.58 (2.07)
	2.0	801 (977)	-0.09	-0.15 (0.24)	0.82	0.61 (2.02)
α	0.2	859 (985)	0.02	-4×10^{-3} (0.12)	0.74	0.54 (1.51)
	0.3	807 (975)	0.01	-0.01 (0.14)	0.82	0.58 (1.77)
	0.4	749 (969)	-0.04	-0.05 (0.19)	0.88	0.61 (2.18)
	0.5	695 (957)	-0.04	-0.05 (0.18)	0.87	0.60 (2.36)
	0.6	621 (936)	-0.04	-0.05 (0.22)	0.90	0.60 (2.67)
	0.7	560 (927)	-0.04	-0.06 (0.24)	0.91	0.60 (2.90)
	0.8	468 (898)	0.04	-0.01 (0.29)	0.95	0.60 (3.13)
c	**0.2**	749 (969)	-0.04	-0.05 (0.19)	0.88	0.61 (2.18)
	0.3	782 (989)	-0.09	-0.18 (0.24)	0.85	0.64 (1.85)
	0.4	752 (997)	-0.19	-0.33 (0.20)	0.84	0.69 (1.58)
	0.5	695 (1000)	-0.27	-0.47 (0.14)	0.76	0.72 (1.31)
	0.6	601 (1000)	-0.43	-0.67 (0.02)	0.84	0.90 (1.12)
	0.7	499 (1000)	-0.64	-0.92 (-0.14)	1.03	1.27 (0.96)
	0.8	357 (1000)	-0.93	-1.26 (-0.42)	1.46	1.99 (0.89)
N	25	391 (845)	-0.02	-0.05 (0.56)	0.84	0.61 (4.04)
	50	749 (969)	-0.04	-0.05 (0.19)	0.88	0.61 (2.18)
	75	881 (981)	0.04	0.02 (0.14)	0.83	0.60 (1.40)
	100	963 (995)	0.05	0.03 (0.07)	0.75	0.56 (0.82)
	500	1000 (1000)	7×10^{-4}	-0.01 (-0.01)	0.13	0.12 (0.12)
	1000	1000 (1000)	7×10^{-3}	4×10^{-3} (4×10^{-3})	0.06	0.06 (0.06)

B: bias; MSE: mean squared error.

Table 4.4: Accuracy of variance estimates of the binary covariate coefficient β_b across simulation scenarios. We compare the average sandwich variance $\widehat{\text{Var}}_S(\hat{\beta}_b)$ with the variance of the estimates $\hat{\beta}_b$ across N.sim repetitions for RR-GEE and the same for RR-PGEE, i.e. $\tilde{\beta}_b$. The simulation parameters are set to $\boldsymbol{\beta} = (\beta_1, \beta_b, \beta_c)^\top = (-4, 1.6, 0.2)^\top$, $N = 50$, $c = 0.2$, $\alpha = 0.4$ (in bold) and each row of the table summarizes 1,000 replications with just one parameter altered, e.g. the first block of rows represent scenarios when β_1 is altered and all other parameters stay fixed. Summaries included are conditional on both estimation methods converging and the estimates being finite (no boundary estimates), with unconditional summaries for RR-PGEE included in parenthesis.

Simulation setup		Finite & converging						
Parameter	Value	N.sim	mean($\widehat{\text{Var}}_S(\hat{\beta}_b)$)	Var($\hat{\beta}_b$)	ratio	mean($\widehat{\text{Var}}_S(\hat{\beta}_b)$)	Var($\tilde{\beta}_b$)	ratio
β_1	**-4**	749 (969)	1.08	0.88	1.22	0.60 (0.57)	0.61 (2.15)	0.98 (0.26)
	-3	992 (1000)	0.45	0.53	0.84	0.32 (0.32)	0.42 (0.50)	0.76 (0.64)
	-2	1000 (1000)	0.11	0.12	0.92	0.10 (0.10)	0.11 (0.11)	0.89 (0.89)
β_b	1.2	648 (940)	1.18	0.88	1.34	0.64 (0.58)	0.60 (2.22)	1.06 (0.26)
	1.4	705 (959)	1.13	0.87	1.29	0.62 (0.58)	0.60 (2.24)	1.03 (0.26)
	1.6	749 (969)	1.08	0.88	1.22	0.60 (0.57)	0.61 (2.15)	0.98 (0.26)
	1.8	780 (975)	1.00	0.82	1.23	0.58 (0.54)	0.58 (2.02)	1.00 (0.27)
	2.0	801 (977)	0.94	0.81	1.15	0.55 (0.51)	0.58 (1.96)	0.94 (0.26)
α	0.2	859 (985)	0.89	0.74	1.21	0.53 (0.51)	0.54 (1.49)	0.98 (0.34)
	0.3	807 (975)	0.99	0.82	1.22	0.57 (0.55)	0.58 (1.75)	0.98 (0.31)
	0.4	749 (969)	1.08	0.88	1.22	0.60 (0.57)	0.61 (2.15)	0.98 (0.26)
	0.5	695 (957)	1.15	0.87	1.33	0.63 (0.58)	0.59 (2.33)	1.06 (0.25)
	0.6	621 (936)	1.23	0.90	1.37	0.65 (0.59)	0.60 (2.62)	1.09 (0.23)
	0.7	560 (927)	1.30	0.91	1.43	0.68 (0.60)	0.60 (2.85)	1.13 (0.21)
	0.8	468 (998)	1.35	0.95	1.42	0.70 (0.61)	0.61 (3.04)	1.15 (0.20)

Table continues on next page.

Table 4.4 continued.

Simulation setup		Finite & converging						
Parameter	Value	N.sim	mean($\widehat{\text{Var}}_S(\hat{\beta}_b)$)	Var($\hat{\beta}_b$)	ratio	mean($\widehat{\text{Var}}_S(\hat{\beta}_b)$)	Var($\tilde{\beta}_b$)	ratio
c	**0.2**	749 (969)	1.08	0.88	1.22	0.60 (0.57)	0.61 (2.15)	0.98 (0.26)
	0.3	782 (989)	1.01	0.85	1.19	0.57 (0.52)	0.60 (1.80)	0.95 (0.29)
	0.4	752 (997)	0.96	0.80	1.20	0.54 (0.47)	0.58 (1.55)	0.94 (0.31)
	0.5	695 (1000)	0.94	0.69	1.35	0.52 (0.43)	0.50 (1.29)	1.04 (0.33)
	0.6	601 (1000)	0.93	0.66	1.42	0.50 (0.38)	0.46 (1.12)	1.08 (0.34)
	0.7	499 (1000)	0.93	0.62	1.49	0.48 (0.34)	0.43 (0.94)	1.11 (0.36)
	0.8	357 (1000)	0.92	0.59	1.57	0.46 (0.31)	0.41 (0.72)	1.14 (0.43)
N	25	391 (845)	1.31	0.84	1.55	0.70 (0.65)	0.60 (3.72)	1.16 (0.17)
	50	749 (969)	1.08	0.88	1.22	0.60 (0.57)	0.61 (2.15)	0.98 (0.26)
	75	881 (981)	0.83	0.82	1.00	0.50 (0.49)	0.60 (1.38)	0.84 (0.35)
	100	963 (995)	0.67	0.75	0.90	0.43 (0.43)	0.55 (0.82)	0.78 (0.52)
	500	1000 (1000)	0.12	0.13	0.96	0.11 (0.11)	0.12 (0.12)	0.93 (0.93)
	1000	1000 (1000)	0.06	0.06	1.00	0.06 (0.06)	0.06 (0.06)	1.00 (1.00)

B: bias; MSE: mean squared error; ratio: mean($\widehat{\text{Var}}_S(\hat{\beta}_b)$)/Var($\hat{\beta}_b$) for $\hat{\beta}_b$ and likewise for $\tilde{\beta}_b$.

4.4 Application to brain lesion data

4.4.1 Data

To demonstrate the performance of the two methods for relative risk regression - RR-GEE and RR-PGEE - we apply them to a subset of the UK Biobank data set (Miller et al., 2016). White matter lesions are an extremely common finding on MRI in older adults, with age being the strongest predictor of lesions, followed by cerebrovascular risk (CVR) factors. CVR factors include smoking and hypertension; below we define a CVR score using 6 variables. It is known that the presence of CVR increases the risk of stroke and dementia (Wardlaw et al., 2015). Thus, we aim to investigate the effect of ageing and cerebrovascular risk on the spatial distribution of lesions, exploring both the cross-sectional and longitudinal effects.

Participants are selected according to the flow chart in Figure 4.1 resulting in a data set of 1,578 healthy ageing individuals; the same selection criterion is used as in Veldsman et al. (2020). The data available for the analysis include binary brain lesion maps at two time points along with data on cerebrovascular risk factors and other variables.

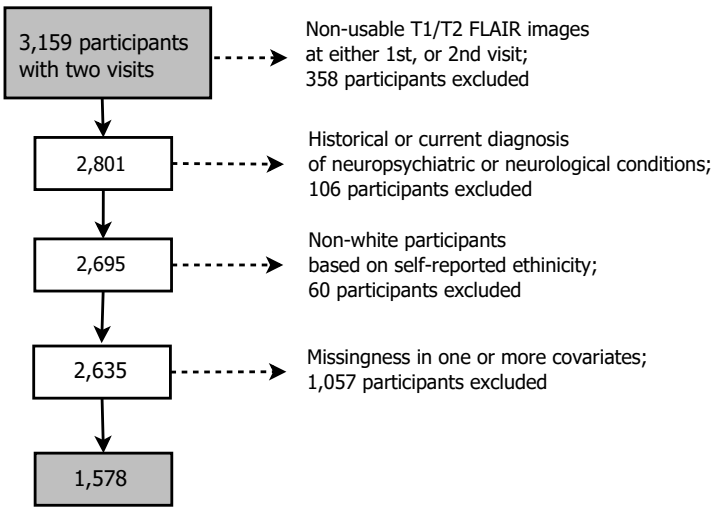


Figure 4.1: Diagram demonstrating the flow of gradually refining participants starting from all UK Biobank participants with available T1 and T2 FLAIR structural brain images at the two imaging visits. Most common neuropsychiatric/neurological conditions in decreasing number of participants: stroke (29), transient ischaemic attack (18), etc. (see Table 4.B.2 for a full list).

MRI data To generate the binary brain lesion maps, we use two MRI imaging modalities (T1-weighted and T2-weighed FLAIR) as input to the lesion segmentation

software BIANCA (Griffanti et al., 2016). The resulting lesion maps are in native space (i.e. subject-specific) with 1 indicating the presence of a lesion and 0 the absence. To facilitate the analysis, those binary maps are first registered to a common space ($2 \times 2 \times 2\text{mm}^3$ Montreal Neurological Institute (MNI) template) by applying spatial normalisation, and are then binarised with a 0.5 threshold. The image preprocessing, the lesion segmentation step as well as the derivation of the estimated spatial normalisation parameters are part of the published UKB preprocessing pipeline (Alfaro-Almagro et al., 2018). Binary lesion maps of dimension $91 \times 109 \times 91$ voxels (volumetric pixels), are available for 1,578 subjects at two time points. The lesion volume at each time point is also available as part of the UKB variables' catalogue[1].

Cerebrovascular risk score A cerebrovascular risk score is calculated as a sum of six categorical variables representing six risk factors, as described by Veldsman et al. (2020). The risk factors include hypertension, hypercholesterolemia, smoking, diabetes, waist-to-hip ratio and the APOE-ε (apolypoprotein-E) status, with categorical variables (0/1/2) for smoking and APOE-ε status, and binary variables (1 indicating presence of the risk factor) for the rest of the risk factors. The resulting score can range from 0 to 8 and is computed for both time points.

Confounding variables The minimal set of confounding variables, as suggested by Alfaro-Almagro et al. (2020), includes age, sex, age by sex interaction and head size scaling.

4.4.2 Analysis setup

We have correlated binary data on $N = 1,578$ subjects at $T = 2$ time points, where each subject i comes with a binary map $y_{it}(s) \in \{0, 1\}$, $s \in \mathcal{B} \subset \mathbb{R}^3$ at time point t. $\mathcal{B}$ is the human brain and we consider M cubic cells s_j (voxels) as a discretization of the 3D brain on a regular rectangular grid, where s_j denotes the jth voxel within the brain ($j = 1, \ldots, M$). We model the available binary lesion maps voxel-wise, i.e. we fit a model at each voxel marginally, ignoring the spatial dependence in the brain, in what is known as a mass-univariate approach. Using the same notation as in Section 4.2, the

[1] `http://biobank.ndph.ox.ac.uk/showcase/search.cgi`

marginal model is of the form

$$
\begin{aligned}
\mathrm{E}(y_{it}(s_j)|\boldsymbol{x}_{it}) = \log(\mu_{it}(s_j)) &= \boldsymbol{x}_{it}^{\top}\boldsymbol{\beta}(s_j) \\
&= \beta_1(s_j) + \beta_2(s_j)\ \text{Age (visit 1)}_{it} + \beta_3(s_j)\ \text{Time difference}_{it} \\
&\quad + \beta_4(s_j)\ \text{CVR score (visit 1)}_{it} + \beta_5(s_j)\ \text{CVR score difference}_{it} \\
&\quad + \beta_6(s_j)\ \text{Sex}_{it} + \beta_7(s_j)\ \text{Head size}_{it} \\
&\quad + \beta_8(s_j)\ \text{Age (visit 1):Sex}_{it} \\
&\quad + \beta_9(s_j)\ \text{Age (visit 1):Time difference}_{it}\,, \qquad (4.1)
\end{aligned}
$$

where $\boldsymbol{x}_{it}$ is a P-vector of time-specific covariates for subject i at time t. The proposed model above includes eight explanatory variables and $P = 9$ with an intercept. We follow the recommended practice of splitting time-varying variables into pure cross-sectional and pure longitudinal variables (Neuhaus and Kalbfleisch, 1998). This leads to five time-invariant variables, i.e. their value is the same across the $T = 2$ time points, namely Age (visit 1), CVR score (visit 1), Sex, Head size (average across two visits) and Age (visit 1) by Sex interaction. The remaining three explanatory variables are time-varying; Time difference captures the longitudinal effect of age and is constructed as a T-vector $(\text{Time difference}_{i1}, \text{Time difference}_{i2})^{\top} = (0, \text{Age (visit 2)}_i - \text{Age (visit 1)}_i)^{\top}$ for subject i; similarly for CVR score and Age (visit 1) by Time difference interaction. Note that prior to fitting the model using RR-GEE and RR-PGEE, Age (visit 1), CVR score (visit 1) and Head size are demeaned across all subjects.

At each voxel, we obtain RR-GEE estimates $\hat{\boldsymbol{\beta}}(s_j)$ and RR-PGEE estimates $\tilde{\boldsymbol{\beta}}(s_j)$ along with sandwich standard errors and the associated z-scores. We also keep track of the boundary estimates criterion (as defined in Equation (4.7)) to detect boundary estimates. To explore the interpretability of the estimated coefficients and the choice of the link function, we fit the penalized GEE for logistic regression (OR-PGEE) as introduced by Mondol and Rahman (2019) and obtain a P-vector of estimates $\boldsymbol{\beta}^*(s_j)$ across voxels s_j. The first two moments of the response are set to $\mathrm{E}(y_{it}) = \mu_{it}$ and $\mathrm{Var}(y_{it}) = \mu_{it}(1 - \mu_{it})/\phi$ and the link function used is logit. We have adjusted the Mondol and Rahman (2019) `R` code to use the same starting values as ours and to use the same convergence criterion. For rare events, we would expect the estimated relative risk and odds ratios to be highly similar. We also transform the estimated log-odds

to relative risks to allow for further comparisons. For example, the transformation to get the relative risk for Age at visit 1 from the estimated log-odds $\boldsymbol{\beta}^*(s_j)$ is of the form $\text{RR}(\beta_2^*(s_j)) = \exp(\beta_2^*(s_j))/((1-p_0(s_j)) + p_0(s_j)\exp(\beta_2^*(s_j))$, where $p_0(s_j) = \exp(\beta_1^*(s_j))/(1+\exp(\beta_1^*(s_j)))$ and $\beta_1^*(s_j)$ is the log-odds for the intercept term and $\beta_2^*(s_j)$ for age, respectively.

In contrast to the simulation study, we allow for the estimation of the dispersion parameter $\phi(s_j)$ across voxels s_j. We use exchangeable correlation matrix with parameter $\alpha(s_j)$ to measure the strength of the within-subject correlation. The convergence tolerance for all three methods is set to $\varepsilon = 0.001$ with maximum number of iterations set to $K = 25$.

Lesions are known to be disproportionately located in periventricular or in deep white matter regions, with varying lesion incidence across the brain. Thus, in a similar way to cross-sectional voxel-wise analysis (Rostrup et al., 2012; Lampe et al., 2019a; Veldsman et al., 2020), we exclude voxels when the lesion count is too low. We have chosen 6 as our threshold, i.e. the models are fitted only at voxels where 6 or more lesions are present across all subjects and both visits. To identify the voxels of interest, we create maps of the lesion incidence across the brain, where $p_1(s_j)$ denotes the lesion incidence at voxel s_j for visit 1 and $p_2(s_j)$ for visit 2, respectively, and $\boldsymbol{p}_1$ and $\boldsymbol{p}_2$ denote the corresponding images.

Given the large number of regressions performed across the brain, when exploring the z-scores we report results based on a fixed threshold of ± 1.96 at 5% significance level, or we adopt a false discovery rate (FDR) correction (Benjamini and Hochberg, 1995) at 5% significance level. To gain better understanding of the localized effect of ageing and CVR score on lesion occurrence both cross-sectionally and longitudinally, we also explore spatial maps of the unthresholded/uncorrected (or 'raw') z-scores and relative risks, i.e. exploring all voxels where the models are fitted. Note that the chosen log link function allows us to interpret the exponent of the estimated coefficients directly as the relative risk.

4.4.3 Results

The analysis to follow is based on UKB data on 1,578 individuals (mean age 62.7 ± 7.2 years, 781 men) with sample characteristics included in Table 4.1. The summaries for

CVR score and lesion volume across the two visits suggest a slight increase in both over time.

Table 4.1: Characteristics of UK Biobank dataset of 1,578 participants.

Characteristics	Mean (SD)	Median (range)
Age, visit 1 (years)	62.7 (7.2)	62.9 (50.0; 79.1)
Time between visits (years)	2.2 (0.1)	2.2 (2.0; 2.8)
Sex (baseline female)	49.5% Men	—
Head size scaling*	1.3 (0.1)	1.3 (0.9; 1.7)
CVR score, visit 1	1.7 (1.3)	2.0 (0; 6)
CVR score, visit 2	1.8 (1.2)	2.0 (0; 7)
Lesion volume, visit 1 (mm^3)	4,626 (5,327)	2,813 (94; 53,608)
Lesion volume, visit 2 (mm^3)	5,050 (5,894)	3,024 (136; 54,713)

*average of head size scaling for visits 1 and 2
N: number of participants; SD: standard deviation; CVR: cerebrovascular risk.

Spatial distribution of lesions

The lesion incidence across 1,578 participants at their first scan (Figure 4.2) highlights the periventricular and deep white matter regions as expected. The empirical relative risk $\boldsymbol{p}_2/\boldsymbol{p}_1$ suggests that the lesion incidence increases from visit 1 to visit 2 for some deep white matter areas. 19,801 voxels with six or more lesions present across all subjects and both visits are considered for the analysis, i.e. 19,801 regressions across the brain as outlined in Section 4.4.2.

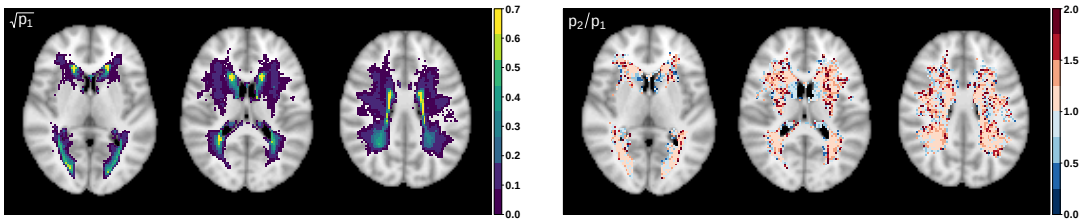


Figure 4.2: Square–root transformed empirical lesion incidence $\boldsymbol{p}_1$ for visit 1 (left) and empirical relative risk $\boldsymbol{p}_2/\boldsymbol{p}_1$ (right) based on binary lesion maps of 1,578 UKB participants across two visits; axial slices $z = \{40, 45, 50\}$ shown from left to right. 19,801 voxels with six or more individuals having a lesion across two visits are plotted with the remaining voxels plotted as transparent to show the standard anatomical MRI for reference. Square-root transformed incidence is used to better visualise the structure in the low incidence regions.

Boundary estimates

As described in Section 4.4.2, the selected marginal model is fitted using RR-GEE and RR-PGEE at 19,801 voxels across the brain. Boundary estimates are more likely to occur for binary covariates but we monitor the BEC values as defined in Equation (4.7) to detect boundary estimates for any of the covariates in the model. The threshold on BEC values is chosen empirically, see Figure 4.B.2. So, we set the threshold for BEC to 10 as we did for the simulations in Section 4.3. We get diverging standard errors for 4% of the voxels (763 voxels out of 19,801) and missing values for 1% (198 voxels) for RR-GEE and 90 voxels and 2 missing values for RR-PGEE, respectively. However, the iterative algorithm converges for 350 out of those 763 voxels for RR-GEE and for 58 out of 90 for RR-PGEE, which suggests that the threshold for the boundary estimates detection criterion may be too conservative. If we increase the threshold to 100, 337 voxels have boundary estimates (2%) for RR-GEE and none for RR-PGEE. We stick to the more conservative threshold of 10 for the remainder of the analysis. The separated voxels seem to occur at the voxels with the lowest incidence rates, as might be expected (Figure 4.B.3).

Interpretation

We inspect the total number of significant coefficients (FDR-corrected or thresholded at ± 1.96) to quantify the spread of the effect of each covariate throughout white matter (see Table 4.2). The cross-sectional effect of age and CVR score have the most widely spread association with lesion probability with their longitudinal effects almost diminishing after FDR-correction. We further explore the effect of each covariate on lesion probability by plotting the raw z-scores obtained by the RR-PGEE marginal model (Figure 4.3); for the FDR-corrected results see Figure 4.B.7, and Figures 4.B.4 and 4.B.5 for the equivalent RR-GEE results. The intra-subject correlation α and the dispersion parameter ϕ are also estimated voxel-wise (Figures 4.4 and 4.B.6) revealing (i) positive within-subject correlation, and (ii) higher dispersion parameter estimates in areas of lower lesion incidence, i.e. since ϕ is a precision parameter, its higher values imply underdispersion at the boundaries.

Clearly the cross-sectional effects of age and CVR score are the most widely spread

Table 4.2: Number of significant voxels across predictors for RR-GEE and RR-PGEE estimates. Voxels with at least six individuals having a lesion across two visits explored (19,801 voxels in the brain mask). (Left) 5% FDR correction applied, i.e. number of FDR-corrected voxels is out of a total of 19,801 voxels, and (Right) fixed threshold of ± 1.96 applied. Columns 2 and 3 complementary to Figures 4.B.5 and 4.B.7, respectively.

Predictor	FDR-corrected voxels		$\lvert z \rvert > 1.96$	
	RR-GEE	RR-PGEE	RR-GEE	RR-PGEE
Age (visit 1)	9,420	11,376	10,719	12,291
Time difference	147	604	2,597	4,293
CVR score (visit 1)	4,274	6,935	6,975	8,975
CVR score difference	274	1,199	2,355	3,896
Sex	1,022	2,367	3,621	5,148
Head size	2,099	3,675	4,682	6,293
Age (visit 1):Time difference	160	1,111	2,275	4,003
Age (visit 1):Sex	993	2,043	2,989	4,374

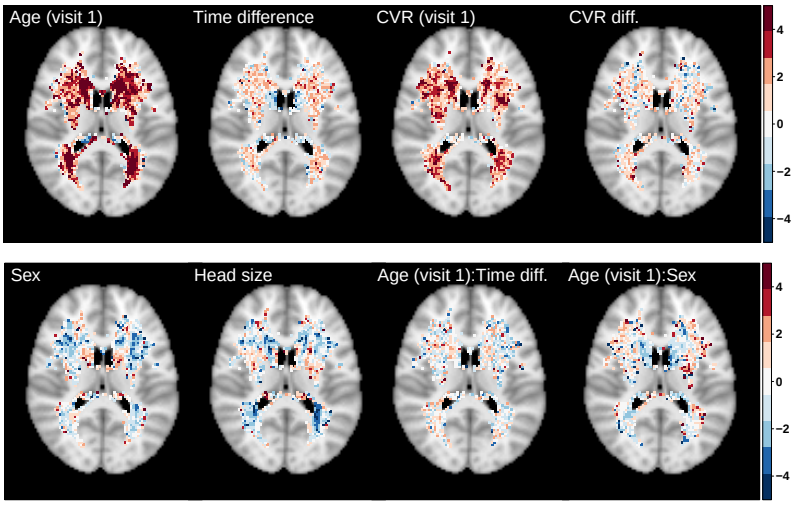


Figure 4.3: Significance maps (z-scores) based on RR-PGEE estimates $\tilde{\boldsymbol{\beta}}$. Data on 1,578 UKB participants across two visits and 19,801 voxels with six or more individuals having a lesion across two visits explored. Axial slice $z = 45$ shown.

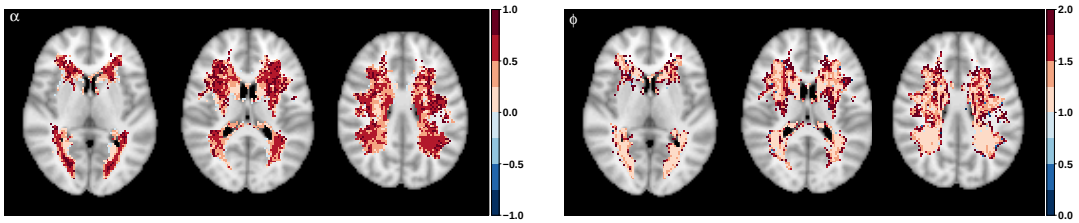


Figure 4.4: Correlation coefficient $\tilde{\alpha}$ (left) and dispersion parameter $\tilde{\phi}$ (right) based on fitting the marginal model using RR-PGEE. Data on 1,578 UKB participants across two visits and 19,801 voxels with six or more individuals having a lesion across two visits explored. Axial slices $z = \{40, 45, 50\}$ shown.

throughout white matter (Table 4.2, Figure 4.3) but z-scores suggest areas where we have detected strong associations between lesion probability and the covariates. The relative risks provide an insight into the magnitude of the effect. Without masking out any voxels, we plot the RR-PGEE relative risks for the four main covariates of interest in Figure 4.5. The baseline age relative risk suggests a remarkably uniform increase in lesion probability across the brain of about 10% (see Table 4.3 for relative risk summaries by empirical lesion incidence). In contrast, the baseline CVR relative risk suggests that an increase of 1 unit is associated with 36-39% increase in lesion probability in deep white matter where the lesion incidence is the lowest (up to 0.5%).

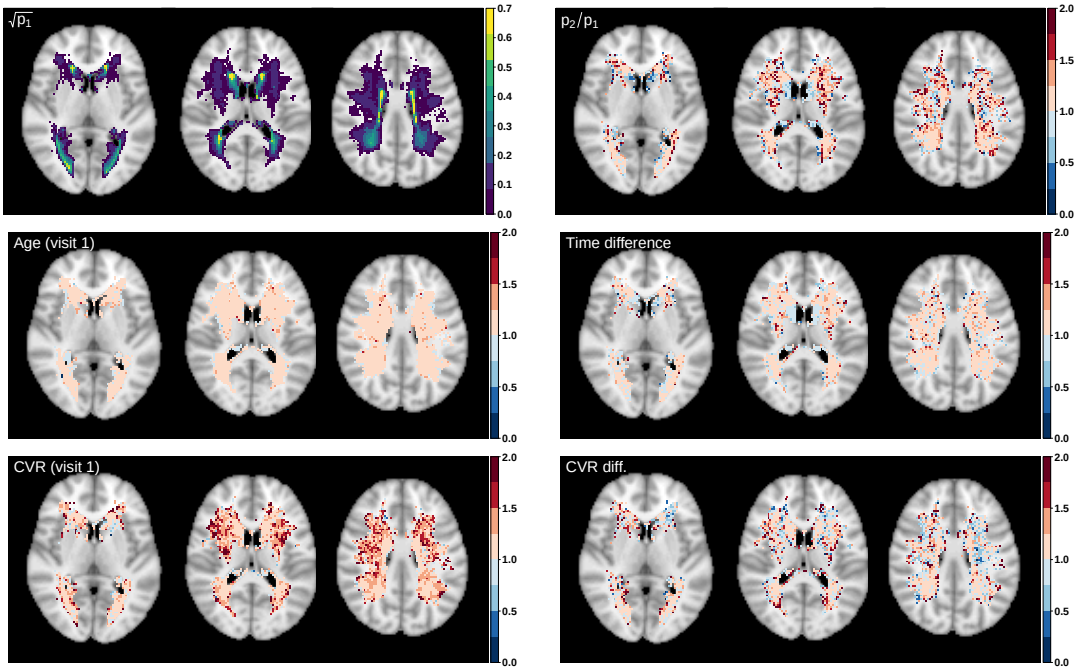


Figure 4.5: Relative risks $\exp(\tilde{\boldsymbol{\beta}})$ for four of the predictors in the RR-PGEE marginal model along with the square–root transformed empirical lesion incidence $\boldsymbol{p}_1$ for visit 1 and the empirical relative risk $\boldsymbol{p}_2/\boldsymbol{p}_1$ (same as Figure 4.2) for reference. Data on 1,578 subjects across two visits and 19,801 voxels explored. Axial slices $z = \{40, 45, 50\}$ shown.

Relative risk vs odds ratio modelling

We further explore the relative risks obtained from relative risk regression (RR-GEE and RR-PGEE) and from logit-link GEE (OR-PGEE method by Mondol and Rahman 2019). Figure 4.6 explores the obtained relative risks for age at visit 1 and it reveals that (i) both penalized GEE methods result in shrinkage of the relative risks towards 1 when compared to the RR-GEE, i.e. shrinkage of the estimated coefficients towards

0, (ii) applying a penalty to the log-odds or to the log-relative risks seems to lead to largely similar results.

Table 4.3: Relative risks across empirical lesion incidence $p_1$at visit 1 based on RR-PGEE estimates $\tilde{\boldsymbol{\beta}}$. Mean and standard deviation (SD) are taken across voxels within each bin as defined in column 1. Data on 1,578 subjects across two visits and 19,801 regressions performed with the iterative algorithm failing for 2 voxels.

p_1	Voxel count	Mean (SD)				
		$p_2 - p_1$	Relative risk $\exp(\tilde{\beta})$			
			Age (visit 1)	Time difference	CVR (visit 1)	CVR difference
$[0; 0.0025)$	1,668	0.0017 (0.0010)	1.11 (0.20)	1.52 (2.02)	1.39 (0.76)	1.18 (0.75)
$[0.0025; 0.005)$	6,232	0.0004 (0.0016)	1.11 (1.04)	1.07 (0.34)	1.36 (0.49)	1.19 (1.57)
$[0.005; 0.01)$	4,810	0.0007 (0.0024)	1.09 (0.09)	1.05 (0.21)	1.29 (0.28)	1.08 (0.45)
$[0.01; 1)$	7,089	0.0050 (0.0101)	1.08 (0.06)	1.05 (0.12)	1.20 (0.16)	1.07 (0.20)
$[0; 1)$	19,799	0.0022 (0.0066)	1.10 (0.59)	1.09 (0.64)	1.29 (0.40)	1.12 (0.94)

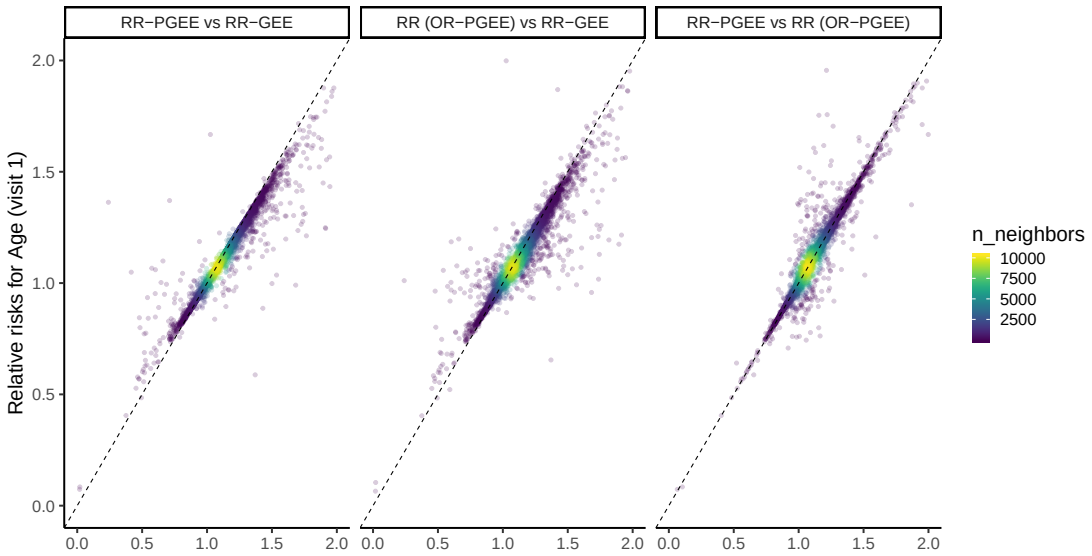


Figure 4.6: Relative risk for age at visit 1 obtained across the three methods. The title of each panel is in the form 'y vs x', i.e. relative risks for age obtained by fitting RR-PGEE plotted against relative risks obtained by fitting RR-GEE. RR (OR-PGEE) are the transformed relative risks resulting from the odds ratios (OR (OR-PGEE)) obtained by fitting OR-PGEE. Data on 1,578 subjects across two visits and 19,801 voxels explored.

Performance

We can obtain in-sample predictions voxel-wise $\hat{\eta}_{it}(s_j) = \boldsymbol{x}_{it}^\top \tilde{\boldsymbol{\beta}}(s_j)$ across subjects i and and visits t using the RR-GEE estimates $\hat{\boldsymbol{\beta}}(s_j) = (\hat{\beta}_1(s_j), \hat{\beta}_2(s_j), \ldots, \hat{\beta}_9(s_j))^\top$ for the intercept and the eight explanatory variables as in (4.1). Similarly, we obtain voxel-wise

predictions $\tilde{\eta}_{it}(s_j)$ using RR-PGEE estimates $\tilde{\boldsymbol{\beta}}(s_j)$ across subjects i and visits t. Note, we have $N \times T = 1578 \times 2 = 3,156$ predicted maps for each method with predicted linear predictor values spread across 19,801 voxels.

We then find that all 3,156 maps have at least one non-negative value ($\hat{\eta}_{it}(s_j)>0$ for at least one voxel s_j for given subject i and time point t) using the RR-GEE estimates (if we do not exclude voxels with boundary estimates) and 1,731 maps (55%) using RR-PGEE estimates, respectively. However, in the majority of predicted maps, where we observe non-negative prediction values, only up to ten voxels out of about 20,000 voxels run into this issue (71% out of 3,156 maps using RR-GEE estimates and 73% out of 1,733 maps using RR-PGEE estimates). Note that the penalty we introduce in the estimating equations does not ensure the predicted values result in valid probability values, but its use seems to alleviate the problem.

Implementation details

Performing 19,801 regressions for all three methods - RR-GEE, RR-PGEE and OR-PGEE - with data on 1,578 subjects at two time points would take long to perform serially. Our approach is to split the total number of regressions into subsets of 500 and then execute each subset of 500 regressions as a single-core job, i.e. a total of 40 jobs running in parallel for each of the methods. Our implementation of RR-GEE and RR-PGEE is entirely in `R` as is the OR-PGEE one by Mondol and Rahman (2019). Each job of 500 regressions takes about 50-60 minutes for RR-GEE, 2-2.5 hours for RR-PGEE and 3-4 hours for OR-PGEE. Note that our code is not optimised and, in particular, includes computation of boundary estimates criterion values at each iteration that would not be needed in routine usage.

4.5 Discussion

Taking a marginal approach to modelling correlated binary data, in this paper we have introduced penalized generalized estimating equations for relative risk regression (RR-PGEE). Our work was motivated by binary brain lesion data derived from MRI scans, since brain lesions have varying incidence across the brain. As a result, odds ratios estimated from GEE with logistic regression structures cannot always safely approximate

risk ratios. On the other hand, use of log-link regression structures with the binomial variance function may lead to estimation instabilities when event probabilities are close to 1. To obtain finite estimates when dealing with rare outcomes or small sample size, we introduced a Jeffreys-prior penalty to the GEE for relative risk regression in a similar manner to bias-reduction methods in GLMs, while using the identity variance function and unknown dispersion for extra stability of the estimates.

We tested the performance of the RR-PGEE estimates through extensive simulations. The penalized approach provided finite estimates and achieved convergence even for simulated data sets where RR-GEE showed signs of separation though estimates diverging to infinity and lack of convergence, or the iterative estimation procedure failed. The inclusion of the penalty also offered some reduction in the mean squared error of the estimates when boundary estimates were not present, with the bias being comparable between the two approaches.

Applying the alternative modelling approaches to a subset of the UK Biobank data, we explored the association between brain lesions, and ageing and cerebrovascular risk. The RR-PGEE approach resulted in stable estimates across the entire brain with RR-GEE resulting in about 5% missing or diverging estimates. The alternative penalized logit-link GEE (Mondol and Rahman, 2019) resulted in highly similar estimates (transforming the odds ratios to risk ratios).

The longitudinal effects of age and CVR score did not result in largely significant estimates across the brain, but this replicates what has been previously shown in the clinical literature (Sachdev et al., 2007). CVR factors and age are undoubtedly strong predictors of lesion incidence and total burden of lesions when measured cross-sectionally. In contrast, baseline total lesion load is a better predictor of lesion incidence, than CVR burden or age, when measured longitudinally. The cross-sectional effect of CVR score showed an interesting localized effect in deep white matter, which is likely the result of small vessel disease. The increase in deep white matter lesion probability associated with CVR reflects the impact of CVR factors like hypertension on small vessels. Also, age seemed to have almost homogeneous effect on lesion probability across the brain suggesting increased risk of lesion occurrence with ageing.

Limitations

Throughout the simulation scenarios, we observed that the sandwich variance slightly underestimated the true variance when using RR-PGEE. In small samples, the sandwich variance is known to underestimate the true variance and thus small-sample bias-correction to the sandwich estimator is considered in the literature (Fay and Graubard, 2001; Mancl and DeRouen, 2001; Morel et al., 2003). Studying appropriate small-sample sandwich estimator adjustments is out of the scope of this work and we believe it does not impact the big data application that motivated our own work. Non-convergence could be due to the choice of the starting values for the IRLS algorithm and more elaborate testing schemes for starting values can result in convergence for the data sets where the boundary estimates criterion was lower than the selected threshold but both methods did not converge.

When exploring the relative risks arising from the UKB data analysis, we do not restrict our attention to voxels passing FDR-correction. We emphasize on the smoothness of the relative risk maps and on the suggested patterns, but we mostly want to highlight the potential of the method and the usefulness of those maps. Surely, clinical interpretations should be approached with extra caution.
We observe that the empirical lesion risk around the ventricles is lower than 1, suggesting that some of the lesions disappear over time. We believe that this phenomenon is not biological but could rather be explained by a mixture of the following: the lesion segmentation is leading to more false positives when lesion load is low (younger subjects), the spatial normalisation to MNI is not optimized for white matter, the caudate is known to be shrinking with age and ventricles are known to enlarge with age which might lead to some registration challenges too. None of those potential segmentation and registration challenges should have affected the modelling approach we suggested, we just note that the estimates in the periventricular areas should be interpreted with caution.

One of the main disadvantages of RR-GEE and RR-PGEE is that they do not guarantee the predicted probabilities would lie in the 0-1 range. If the linear predictor is greater than 0, the resulting predicted probability would be above 1. This did happen very rarely for the simulated data sets but it happened more often for the noisier real

data set for single voxels. The main aim of the current work was to provide stable and interpretable estimates, i.e. inference, as opposed to prediction. We have demonstrated that the estimated relative risks were largely similar to the relative risks obtained from logistic regression (transforming odds ratios to risk ratios) ensuring that inference was not affected by the unsupported predicted values.

Acknowledgments

We are grateful to Ludovica Griffanti and Stephen Smith for their valuable input on the lesion segmentation and registration practices in the UK Biobank. The computational aspects of this research were supported by the Wellcome Trust Core Award Grant Number 203141/Z/16/Z and the NIHR Oxford BRC. The views expressed are those of the author(s) and not necessarily those of the NHS, the NIHR or the Department of Health.

Data availability statement

The `R` code used to fit the marginal models using RR-GEE and RR-PGEE is available at the GitHub repository `https://github.com/petyakindalova/PGEE_Illustration` along with an illustration of the code used for the simulation study. The spatial relative risk maps produced as part of the real data application are available at NeuroVault `https://neurovault.org/collections/SUDHNHAA/`.

Appendices

4.A Derivation of penalty term

To obtain the form of the Jeffreys-prior penalty term $A_p(\boldsymbol{\beta}, \boldsymbol{\alpha})$ for this RR-GEE (Equation (4.2)), we first take the derivative of U wrt β_p by using the product and chain rules:

$$[J]_p = \phi \sum_{i=1}^{N} X_i^\top \Big(\frac{\partial}{\partial \beta_p} W_i^{1/2}\Big) R(\boldsymbol{\alpha})^{-1} W_i^{-1/2} (\boldsymbol{y}_i - \boldsymbol{\mu}_i)$$
$$+ \phi \sum_{i=1}^{N} X_i^\top W_i^{1/2} R(\boldsymbol{\alpha})^{-1} \Big(\frac{\partial}{\partial \beta_p} W_i^{-1/2}\Big) (\boldsymbol{y}_i - \boldsymbol{\mu}_i)$$

$$- \phi \sum_{i=1}^{N} X_i^\top W_i^{1/2} R(\boldsymbol{\alpha})^{-1} W_i^{-1/2} \Big(\frac{\partial}{\partial \beta_p} \boldsymbol{\mu}_i \Big) \,. \tag{4.A.1}$$

Since $\mathrm{E}(\boldsymbol{y}_i) = \boldsymbol{\mu}_i$ by assumption and $\boldsymbol{y}_i - \boldsymbol{\mu}_i$ come linearly into (4.A.1), the p-th row of the expected negative Hessian matrix I is

$$[I]_p = \phi \sum_{i=1}^{N} X_i^\top W_i^{1/2} R(\boldsymbol{\alpha})^{-1} W_i^{-1/2} [D_i]^p \,,$$

where $[D_i]^p$ is the p-th column of D_i. So,

$$I(\boldsymbol{\beta}, \boldsymbol{\alpha}, \phi) = \phi \sum_{i=1}^{N} X_i^\top W_i^{1/2} R(\boldsymbol{\alpha})^{-1} W_i^{-1/2} W_i X_i = \phi \sum_{i=1}^{N} X_i^\top W_i^{1/2} R(\boldsymbol{\alpha})^{-1} W_i^{1/2} X_i \tag{4.A.2}$$

We further take the derivative of I and by using the product and chain rules we get

$$\begin{aligned}
\frac{\partial}{\partial \beta_p} I(\boldsymbol{\beta}, \boldsymbol{\alpha}, \phi) &= \phi \sum_{i=1}^{N} X_i^\top \{ \frac{\partial}{\partial \beta_p} W_i^{1/2} R(\boldsymbol{\alpha})^{-1} W_i^{1/2} + W_i^{1/2} R(\boldsymbol{\alpha})^{-1} \frac{\partial}{\partial \beta_p} W_i^{1/2} \} X_i \\
&= \phi \sum_{i=1}^{N} X_i^\top \{ \frac{1}{2} W_i^{1/2} Q_p R(\boldsymbol{\alpha})^{-1} W_i^{1/2} + W_i^{1/2} R(\boldsymbol{\alpha})^{-1} \frac{1}{2} W_i^{1/2} Q_p \} X_i \\
&= \phi \sum_{i=1}^{N} X_i^\top \{ \frac{1}{2} Q_p W_i^{1/2} R(\boldsymbol{\alpha})^{-1} W_i^{1/2} + \frac{1}{2} W_i^{1/2} R(\boldsymbol{\alpha})^{-1} W_i^{1/2} Q_p \} X_i \\
&\qquad\qquad \text{(commute diagonal matrices)} \\
&= \phi \sum_{i=1}^{N} X_i^\top \{ \frac{1}{2} Q_p B_i + \frac{1}{2} B_i Q_p \} X_i \,,
\end{aligned} \tag{4.A.3}$$

where $Q_p = \mathrm{diag}(x_{i1p}, \ldots, x_{iTp})$ and $B_i = W_i^{1/2} R(\boldsymbol{\alpha})^{-1} W_i^{1/2}$. Combining expressions (4.A.2) and (4.A.3) we construct the penalty term:

$$\begin{aligned}
A_p(\boldsymbol{\beta}, \boldsymbol{\alpha}) &= \frac{\phi}{2} \sum_{i=1}^{N} \mathrm{trace} \Big\{ I(\boldsymbol{\beta}, \boldsymbol{\alpha}, \phi)^{-1} X_i^\top \{ \frac{1}{2} Q_p B_i + \frac{1}{2} B_i Q_p \} X_i \Big\} \\
&= \frac{\phi}{2} \sum_{i=1}^{N} \mathrm{trace} \Big\{ X_i I(\boldsymbol{\beta}, \boldsymbol{\alpha}, \phi)^{-1} X_i^\top \{ \frac{1}{2} Q_p B_i + \frac{1}{2} B_i Q_p \} \Big\} \\
&= \frac{\phi}{4} \sum_{i=1}^{N} \mathrm{trace} \{ X_i I(\boldsymbol{\beta}, \boldsymbol{\alpha}, \phi)^{-1} X_i^\top Q_p B_i \} + \frac{\phi}{4} \sum_{i=1}^{N} \mathrm{trace} \{ X_i I(\boldsymbol{\beta}, \boldsymbol{\alpha}, \phi)^{-1} X_i^\top B_i Q_p \} \\
&= \frac{\phi}{4} \sum_{i=1}^{N} \mathrm{trace} \{ B_i X_i I(\boldsymbol{\beta}, \boldsymbol{\alpha}, \phi)^{-1} X_i^\top Q_p \} + \frac{\phi}{4} \sum_{i=1}^{N} \mathrm{trace} \{ Q_p X_i I(\boldsymbol{\beta}, \boldsymbol{\alpha}, \phi)^{-1} X_i^\top B_i \} \\
&\qquad / B_i \text{ is symmetric so } (B_i X_i I(\boldsymbol{\beta}, \boldsymbol{\alpha}, \phi)^{-1} X_i^\top Q_p)^\top = Q_p X_i I(\boldsymbol{\beta}, \boldsymbol{\alpha}, \phi)^{-1} X_i^\top B_i /
\end{aligned}$$

$$= \frac{\phi}{2} \sum_{i=1}^{N} \text{trace}\{B_i X_i I(\boldsymbol{\beta}, \boldsymbol{\alpha}, \phi)^{-1} X_i^\top Q_p\}$$
$$= \frac{1}{2} \sum_{i=1}^{N} \sum_{t=1}^{T} h_{it} x_{itp}\,, \tag{4.A.4}$$

where h_{it} is the t-th diagonal element of the i-th block of the projection matrix

$$H_i = W_i^{1/2} R(\boldsymbol{\alpha})^{-1} W_i^{1/2} X_i \Big[\sum_{i=1}^{N} X_i^\top W_i^{1/2} R(\boldsymbol{\alpha})^{-1} W_i^{1/2} X_i \Big]^{-1} X_i^\top .$$

4.B Complementary tables and figures

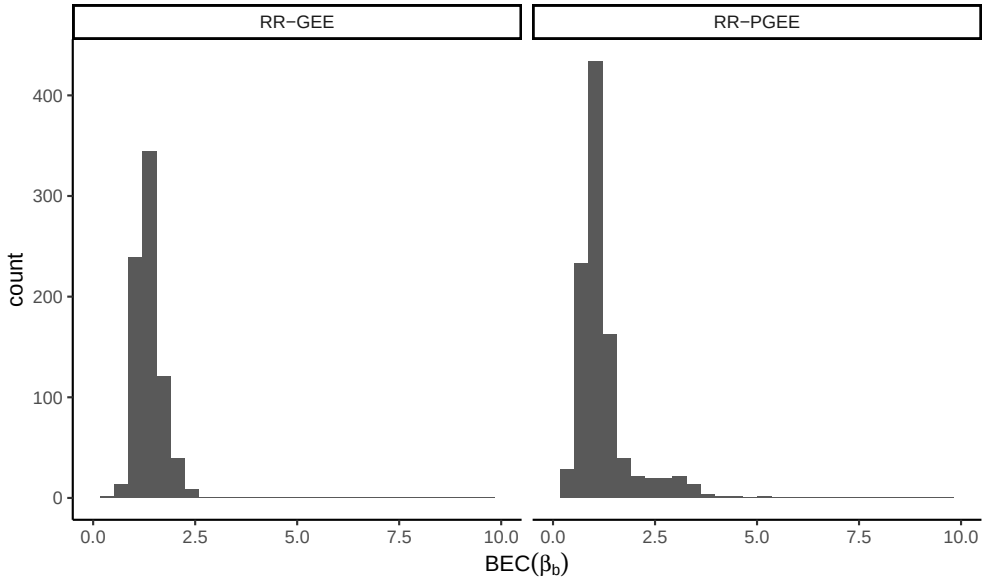


Figure 4.B.1: Boundary estimates criterion threshold is empirically selected and set to 10 for the simulation study. (Left) ($\widehat{\text{BEC}}_2$) and (Right) ($\widetilde{\text{BEC}}_2$) across 1,000 simulated data sets from the base simulation scenario; parameters are set to $\boldsymbol{\beta} = (\beta_1, \beta_b, \beta_c)^\top = (-4, 1.6, 0.2)^\top$, $N = 50$, $c = 0.2$, $\alpha = 0.4$ and all existing BEC values are shown for the fixed x-axis limits.

Table 4.B.1: Bias and mean-squared error of the binary covariate coefficient β_b across simulation scenarios. The simulation parameters are set to $\boldsymbol{\beta} = (\beta_1, \beta_b, \beta_c)^\top = (-4, 1.6, 0.2)^\top$, $\alpha = 0.4$ and each row of the table summarizes 1,000 replications with just one parameter altered, e.g. the first block of rows represent scenarios when c is altered for sample size $N = 100$ and for sample size $N = 500$, respectively. Summaries included are conditional on both estimation methods converging and the estimates being finite (no boundary estimates), with unconditional summaries for RR-PGEE included in parenthesis.

$\boldsymbol{N = 100}$		Finite & converging				
Parameter	Value	N.sim	B($\hat{\beta}_b$)	B($\tilde{\beta}_b$)	MSE($\hat{\beta}_b$)	MSE($\tilde{\beta}_b$)
c	**0.2**	963 (995)	0.05	0.03 (0.07)	0.75	0.56 (0.82)
	0.3	957 (1000)	0.11	0.02 (0.12)	0.66	0.49 (0.86)
	0.4	941 (1000)	0.10	-0.04 (0.11)	0.67	0.50 (0.85)
	0.5	906 (1000)	0.07	-0.11 (0.10)	0.60	0.44 (0.84)
	0.6	840 (1000)	0.03	-0.19 (0.11)	0.57	0.44 (0.85)
	0.7	760 (1000)	-0.10	-0.37 (0.02)	0.53	0.49 (0.79)
	0.8	615 (1000)	-0.29	-0.63 (-0.12)	0.55	0.69 (0.66)
$\boldsymbol{N = 500}$		Finite & converging				
Parameter	Value	N.sim	B($\hat{\beta}_b$)	B($\tilde{\beta}_b$)	MSE($\hat{\beta}_b$)	MSE($\tilde{\beta}_b$)
c	**0.2**	1000 (1000)	7×10^{-4}	-0.01 (-0.01)	0.13	0.12 (0.12)
	0.3	1000 (1000)	0.01	-0.01 (-0.01)	0.11	0.11 (0.11)
	0.4	1000 (1000)	0.02	-0.01 (-0.01)	0.11	0.11 (0.11)
	0.5	1000 (1000)	0.03	-0.02 (-0.02)	0.15	0.13 (0.13)
	0.6	1000 (1000)	0.08	0.02 (0.02)	0.19	0.16 (0.16)
	0.7	999 (1000)	0.11	0.03 (0.03)	0.28	0.22 (0.23)
	0.8	993 (1000)	0.18	0.03 (0.04)	0.45	0.30 (0.34)

B: bias; MSE: mean squared error.

Table 4.B.2: List of codes of the UK Biobank Data-field 'Non-cancel illness' (`http://biobank.ndph.ox.ac.uk/showcase/field.cgi?id=20002`) used for exclusion of participants in the data cleaning process. 118 illnesses reported across 106 participants.

Coding	Description	Number of participants
1081	Stroke	29
1082	Transient ischaemic attack	18
1083	Subdural haematoma/haemorrhage	2
1086	Subarachnoid haemorrhage	0
1240	Neurological injury/trauma	2
1243	Psychological/psychiatric problem	0
1244	Infection of the nervous system	0
1245	Brain abscess/Intracranial abscess	1
1246	Encephalitis	1
1247	Meningitis	14
1258	Chronic neurological problem	0
1259	Motor Neuron Disease	1
1261	Multiple Sclerosis	10
1262	Parkinson's disease	4
1263	Dementia/Alzheimer's disease/Cognitive impairment	1
1264	Epilepsy	13
1266	Head Injury	4
1408	Alcoholism	1
1409	Opioid dependency	0
1410	Other dependency	0
1434	Other neurological problem	6
1491	Brain haemorrhage	5
1583	Ischaemic stroke	1
1626	Fracture skull/head	5

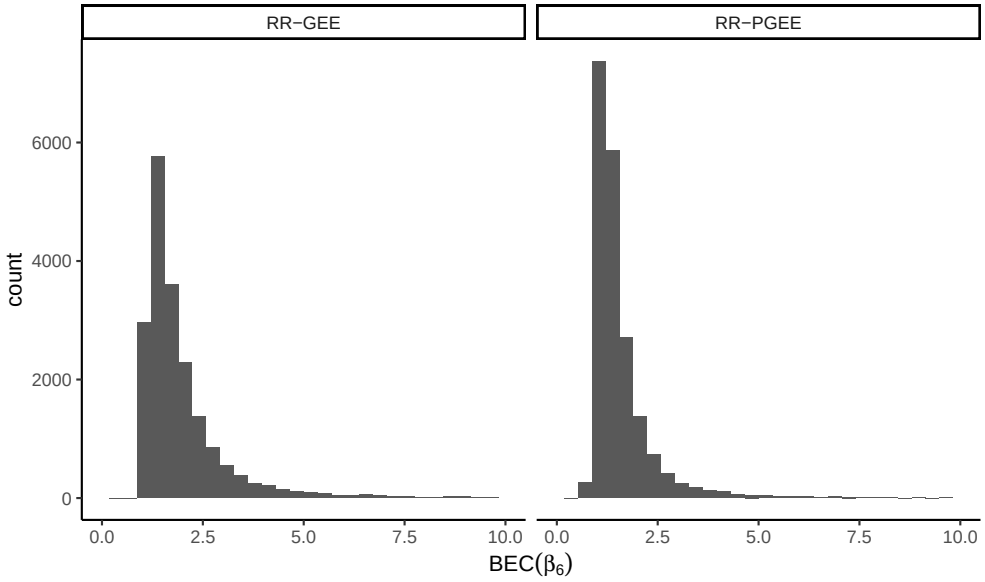


Figure 4.B.2: Boundary estimates criterion threshold is empirically selected and set to 10 for UKB analysis and the BEC values are shown for the sex covariate. (Left) ($\widehat{\text{BEC}}_6$) and (Right) ($\widetilde{\text{BEC}}_6$) across 19,801 voxels; all existing BEC values are shown for the fixed x-axis limits. Voxels not shown on the histograms include 763 voxels with BEC values greater than 10 for any of the 9 covariates for RR-GEE and 90 for RR-PGEE, respectively, as well as any missing values due to IRLS failure.

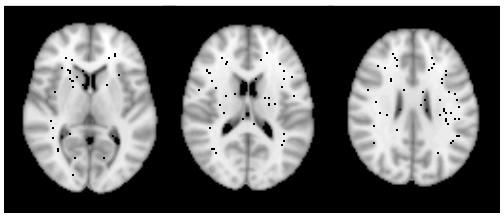


Figure 4.B.3: Voxels with boundary estimates in black, where about 5% of voxels have boundary estimates criterion values higher than 10 or missing values due to failure of the IRLS algorithm. Voxels with boundary estimates seem to occur in regions of low lesion incidence. Axial slices $z = \{40, 45, 50\}$ shown.

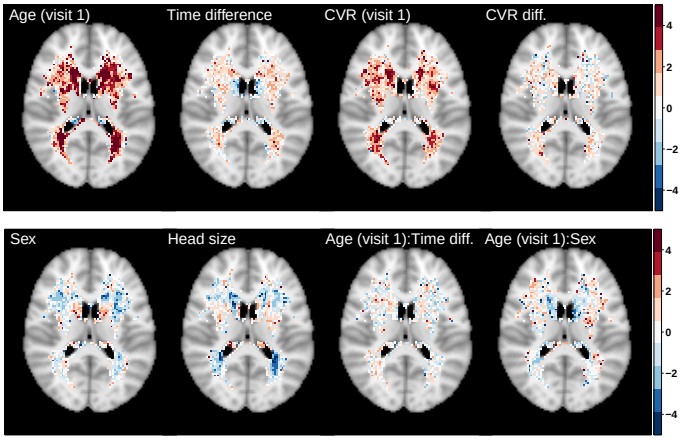


Figure 4.B.4: Significance maps (z-scores) based on RR-GEE estimates $\hat{\boldsymbol{\beta}}$. Data on 1,578 UKB participants across two visits and 19,801 voxels with six or more individuals having a lesion across two visits explored. Axial slice $z = 45$ shown.

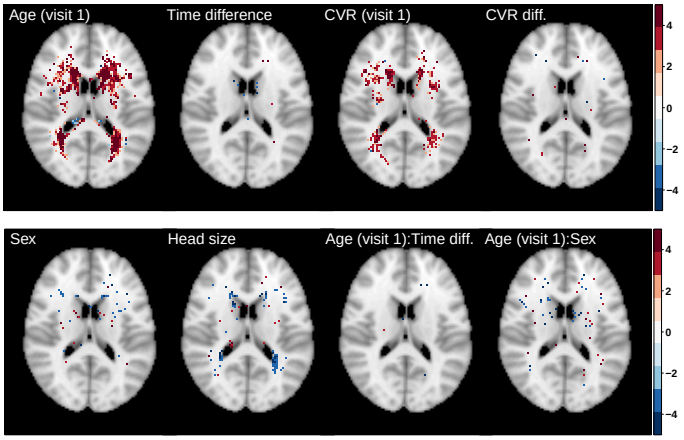


Figure 4.B.5: Significance maps (z-scores) based on RR-GEE estimates $\hat{\boldsymbol{\beta}}$. Data on 1,578 UKB participants across two visits and 19,801 voxels with six or more individuals having a lesion across two visits explored; 5%-FDR correction applied. Axial slice $z = 45$ shown.

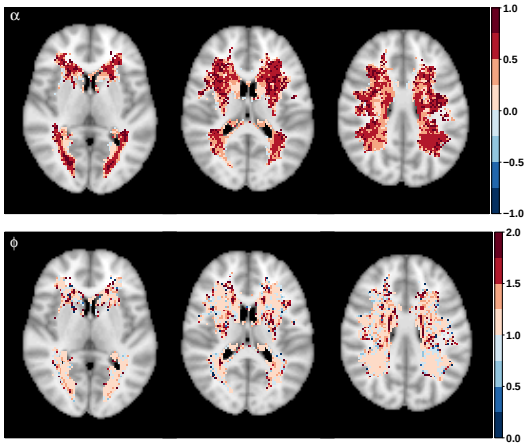


Figure 4.B.6: Correlation coefficient $\hat{\alpha}$ (top) and dispersion parameter $\hat{\phi}$ (bottom) based on fitting the marginal model using RR-GEE. Data on 1,578 UKB participants across two visits and 19,801 voxels with six or more individuals having a lesion across two visits explored. Axial slices $z = \{40, 45, 50\}$ shown.

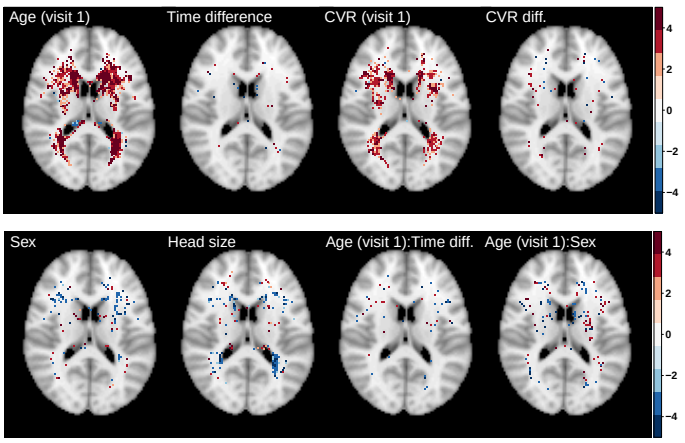


Figure 4.B.7: Significance maps (z-scores) based on RR-PGEE estimates $\tilde{\boldsymbol{\beta}}$. Data on 1,578 UKB participants across two visits and 19,801 voxels with six or more individuals having a lesion across two visits explored; 5%-FDR correction applied. Axial slice $z = 45$ shown.

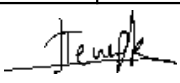
Statement of Authorship for joint/multi-authored papers for PGR thesis

To appear at the end of each thesis chapter submitted as an article/paper

The statement shall describe the candidate's and co-authors' independent research contributions in the thesis publications. For each publication there should exist a complete statement that is to be filled out and signed by the candidate and supervisor **(only required where there isn't already a statement of contribution within the paper itself).**

Title of Paper	Penalized generalized estimating equations for relative risk regression with applications to brain lesion data
Publication Status	□ Published □ Accepted for Publication ✔ **Submitted for Publication** □ Unpublished and unsubmitted work written in a manuscript style
Publication Details	Manuscript was submitted to *Statistics in Medicine* on 1st November 2021. It is available on bioRxiv https://doi.org/g4rh

Student Confirmation

Student Name:	Petya Kindalova		
Contribution to the Paper	The following section "Author contributions" was submitted to the journal. **Petya Kindalova:** Conceptualization, Data curation, Formal analysis, Methodology, Validation, Visualization, Writing - original draft. **Michele Veldsman:** Clinical interpretation of the results & contributions to writing. **Thomas E. Nichols:** Conceptualization, Funding acquisition, Methodology, Supervision, Writing - review & editing. **Ioannis Kosmidis:** Conceptualization, Funding acquisition, Methodology, Supervision, Writing - review & editing.		
Signature		Date	02/11/2021

Supervisor Confirmation

By signing the Statement of Authorship, you are certifying that the candidate made a substantial contribution to the publication, and that the description described above is accurate.

Supervisor name and title: Professor Thomas E. Nichols		
Supervisor comments Really outstanding work, reflected in these two publications, and a third one in submission.		
Signature 	Date	2 Nov 2021

Supervisor name and title: Professor Ioannis Kosmidis		
Supervisor comments Petya Kindalova's work is set to have a broad impact on: • The regression modelling of brain lesions from MRI data and inference about relevant effects and risk factors • The design of computer experiments for the assessment of competing methods for the task (through the novel simulation framework) • The estimation of relative risks in general longitudinal studies (through the penalized GEE approach that is developed) Overall, this is solid methodological and applied work, making a significant contribution to the modelling of brain lesions and making generally-applicable advances on relative risk regression with longitudinal data.		
Signature 	Date	3 Nov 2021

This completed form should be included in the thesis, at the end of the relevant chapter.

Chapter 5

Final remarks

5.1 Summary of the thesis

Brain lesions are a common finding on MRI in elderly individuals, but their presence cannot be fully explained by ageing. The literature to date suggests that they are markers of vascular burden and their presence triples the risk of stroke, for example. It has been shown that lesion size, location and growth dynamics are important for diagnosis and treatment of neurological conditions such as multiple sclerosis. Given the rich and growing size of neuroimaging data sets available, accurate and computationally efficient lesion mapping methods are necessary.

Method comparison with a novel simulation framework

In Chapter 2 the main question we aim to answer is whether the potential gains in estimator accuracy justify the use of a more computationally intensive spatial modelling approach as opposed to a mass-univariate approach to modelling voxel-wise binary lesion data.

To answer this question, we compare three alternative approaches for modelling binary lesion masks, two of which rely on voxel-wise fitting of generalized linear models using maximum likelihood and bias-reduction, and the other is a Bayesian spatial method that imposes spatial regularisation. To allow for a fair comparison, we develop a novel simulation framework of artificial lesion masks, which mimics features of real lesion masks such as lesion count. Our results show that bias-reduced estimates overcome the instabilities of maximum likelihood estimates, and scale well for large data sets due

to parallel implementation. Contrary to the assumption of spatial dependence being key in lesion mapping, our findings suggest that voxel-wise bias reduction and spatial modelling result in largely similar estimates, which gives an advantage to the computationally efficient bias-reduced estimates in biobank-scale neuroimaging data sets. We further provide a lesion mask simulator to the neuroimaging community by sharing our code online.

Cerebrovascular risk-related lesions

Hypertension is thought to be the strongest predictor of the presence of white matter lesions, but less is known about the contribution of other risk factors such as smoking or high waist to hip ratio. Taking advantage of a large data set of healthy ageing UK Biobank participants, we attempt to disentangle the contribution of individual risk factors to lesion presence, which has been difficult in past applications due to limited population sizes.

In Chapter 3, we use data on 13,680 healthy ageing individuals to examine the contribution of cerebrovascular risk factors to the total lesion load, spatial distribution of lesions, and the impact of risk factors on cognition. Contrary to an emphasis of hypertension as the main risk factor in the existing literature, we find that waist-to-hip ratio, diabetes, heavy smoking, hypercholesterolemia and homozygous APOE $\varepsilon 4$ status are important risk factors associated with total lesion burden and warrant careful control across ageing. Waist-to-hip ratio shows independent effects as well as a relationship with speed of processing that points to the management of visceral adiposity as a key target to mitigate cognitive decline in ageing.

Our contributions include introducing the bias-reduced estimates as a voxel-wise lesion mapping approach as well as presenting important clinical findings that have relevance to those researching ageing, vascular and neurodegenerative diseases. We further share all resulting spatial maps through a publicly available repository for colleagues to explore the results for themselves.

Penalized generalized estimating equations

Motivated by the repeated UK Biobank brain lesion data and as a natural extension to the cerebrovascular risk-related lesion analysis presented in Chapter 3, the work

included in Chapter 4 introduces penalized generalized estimating equations for relative risk regression for modelling correlated binary data.

Adopting a log-link GEE with identity variance function and unknown dispersion is the first step to ensuring the desired interpretability of relative risks and also addressing some of the convergence issues of log-Binomial regression. However, boundary estimates can still occur, e.g. we have observed diverging estimates for at least 2% of voxels for the neuroimaging data set considered, leading to difficulties with inference. To achieve finite estimates in the presence of boundary estimates, we add a Jeffreys-prior penalty to the estimating equations.

Our extensive simulation study demonstrates the superior convergence performance of the penalized GEE over the standard GEE and we have empirically shown the consistency behaviour of the estimator. Even though the large-scale voxel-based application does not reveal clearly localized longitudinal effects of cerebrovascular risk and age, it demonstrates the huge potential of the modelling approach as the UKB sample continues to grow, with its stable estimates and scalability when using parallel computing. We further share our code and resulting spatial maps in publicly available repositories to facilitate reproducible research.

5.2 Future directions

Method comparison with a novel simulation framework

Using data from longitudinal studies such as the UK Biobank, the simulation framework proposed in Chapter 2 could be extended to facilitate comparison of longitudinal modelling approaches. The principle challenge here is that, in addition to regression and smoothness parameters, parameters on the longitudinal correlation would also need to be estimated and incorporated into the simulation framework.

Cerebrovascular risk-related lesions

With the number of UK Biobank participants being scanned twice growing, a similar analysis to the one presented in Chapter 3 could become feasible using longitudinal data on cerebrovascular risk factors and cognitive performance, and fitting the approach discussed in Chapter 4. However, another data set with longer follow-up times between

visits might be more suitable to detect any meaningful changes over time.

Our findings suggested that waist-to-hip ratio mediates the association between total lesion burden and cognitive decline, as represented by speed of processing. It will be of interest for future analysis to use data on intra-abdominal fat assessed through MRI instead of waist-to-hip ratio. Such data should become part of the UKB catalogue and it has been collected, but were not available at the time of analysis. Additionally, the localized effect of vascular risk factors on the spatial distribution of lesions could become even more pronounced if the analysis is to be repeated for 100,000 lesion masks (the number of UKB participants to be scanned) and the parallel implementation of the method would still be feasible.

Penalized generalized estimating equations

The small-sample properties of the sandwich variance estimator have been studied in the literature since it is known to underestimate the true variance. Studying appropriate adjustments to the sandwich estimator could be pursued in future work, with some existing approaches being the work by Mancl and DeRouen (2001) or by Morel et al. (2003).

The main aim of the penalized log-link GEE was to provide stable and interpretable estimates, i.e. making inferences, not predictions. However, a potential disadvantage of the use of log link is the that predicted probabilities can go above 1, i.e. they are not bounded between 0 and 1. One suggested remedy to this issue for cross-sectional data is to maximize the likelihood over the restricted parameter space by using an adaptive barrier algorithm (Luo et al., 2014), where the authors use `constrOptim` function from the `stats` package in `R`. Recently, Schwendinger et al. (2021) compared a variety of optimization solvers and suggested modern conic programming as superior to linear programming, yet again for cross-sectional data. A separate research project could look into ways to restrict the GEE predicted values to supported predicted values.

Another consideration for future longitudinal analysis could look into lesion resolution as opposed to lesion accumulation, i.e. some of the estimated risk factor effects suggesting lower lesion incidence across time. The challenge would be to separate any potential registration and lesion segmentation issues from any true effects over time, especially given the short follow-up times.

Bibliography

Abraham, H. M. A., Wolfson, L., Moscufo, N., Guttmann, C. R., Kaplan, R. F., and White, W. B. (2016). Cardiovascular risk factors and small vessel disease of the brain: Blood pressure, white matter lesions, and functional decline in older persons. *Journal of Cerebral Blood Flow and Metabolism*, 36(1).

Albert, A. and Anderson, J. A. (1984). On the existence of maximum likelihood estimates in logistic regression models. *Biometrika*, 71(1).

Alfaro-Almagro, F., Jenkinson, M., Bangerter, N. K., Andersson, J. L., Griffanti, L., Douaud, G., ..., and Smith, S. M. (2018). Image processing and Quality Control for the first 10,000 brain imaging datasets from UK Biobank. *NeuroImage*, 166.

Alfaro-Almagro, F., McCarthy, P., Afyouni, S., Andersson, J. L., Bastiani, M., Miller, K. L., ..., and Smith, S. M. (2020). Confound modelling in UK Biobank brain imaging. *NeuroImage*, 224:117002.

Andersson, J. L. R., Jenkinson, M., and Smith, S. M. (2007). Non-linear registration aka spatial normalisation. Internal Technical Report TR07JA1, Oxford Centre for Functional Magnetic Resonance Imaging of the Brain, Department of Clinical Neurology, Oxford University, Oxford, UK.

Arellano, M. and Hahn, J. (2010). Understanding bias in nonlinear panel models: Some recent developments. In *Advances in Economics and Econometrics: Theory and Applications, Ninth World Congress*, volume 3. Cambridge University Press.

Atwood, L. D., Wolf, P. A., Heard-Costa, N. L., Massaro, J. M., Beiser, A., D'Agostino, R. B., and DeCarli, C. (2004). Genetic variation in white matter hyperintensity volume in the Framingham study. *Stroke*, 35(7).

Awad, I. A., Johnson, P. C., Spetzler, R. F., and Hodak, J. A. (1986). Incidental subcortical lesions identified on magnetic resonance imaging in the elderly. II. Postmortem pathological correlations. *Stroke*, 17(6).

Barkhof, F. (2002). The clinico-radiological paradox in multiple sclerosis revisited. *Current Opinion in Neurology*, 15(3).

Benjamin, E. J., Blaha, M. J., Chiuve, S. E., Cushman, M., Das, S. R., Deo, R., ..., and Muntner, P. (2017). Heart Disease and Stroke Statistics'2017 Update: A Report from the American Heart Association. *Circulation*, 135(10).

Benjamini, Y. and Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(1).

Boffa, R. J., Constanti, M., Floyd, C. N., and Wierzbicki, A. S. (2019). Hypertension in adults: Summary of updated NICE guidance. *The BMJ*, 367.

Bycroft, C., Freeman, C., Petkova, D., Band, G., Elliott, L. T., Sharp, K., ..., and Marchini, J. (2018). The UK Biobank resource with deep phenotyping and genomic data. *Nature*, 562.

Cannon, M. J., Warner, L., Taddei, J. A., and Kleinbaum, D. G. (2001). What can go wrong when you assume that correlated data are independent: An illustration from the evaluation of a childhood health intervention in Brazil. *Statistics in Medicine*, 20(9-10).

Carter, R. E., Lipsitz, S. R., and Tilley, B. C. (2005). Quasi-likelihood estimation for relative risk regression models. *Biostatistics*, 6(1).

Chard, D. T., Jackson, J. S., Miller, D. H., and Wheeler-Kingshott, C. A. (2010). Reducing the impact of white matter lesions on automated measures of brain gray and white matter volumes. *Journal of Magnetic Resonance Imaging*, 32(1).

Charil, A., Zijdenbos, A. P., Taylor, J., Boelman, C., Worsley, K. J., Evans, A. C., and Dagher, A. (2003). Statistical mapping analysis of lesion location and neurological disability in multiple sclerosis: Application to 452 patient data sets. *NeuroImage*, 19(3).

Cheesman, R., Coleman, J., Rayner, C., Purves, K. L., Morneau-Vaillancourt, G., Glanville, K., ..., and Eley, T. C. (2020). Familial Influences on Neuroticism and Education in the UK Biobank. *Behavior Genetics*, 50(2).

Cleveland, W. S., Grosse, E., and Shyu, W. M. (2017). Local regression models. In *Statistical Models in S*. Chapman and Hall/CRC.

Cordeiro, G. M. and McCullagh, P. (1991). Bias Correction in Generalized Linear Models. *Journal of the Royal Statistical Society: Series B (Methodological)*, 53(3).

Cox, S. R., Lyall, D. M., Ritchie, S. J., Bastin, M. E., Harris, M. A., Buchanan, C. R., ..., and Deary, I. J. (2019). Associations between vascular risk factors and brain MRI indices in UK Biobank. *European Heart Journal*, 40(28).

Cox, S. R., Ritchie, S. J., Dickie, D. A., Pattie, A., Royle, N. A., Corley, J., ..., and Deary, I. J. (2017). Interaction of APOE e4 and poor glycemic control predicts white matter hyperintensity growth from 73 to 76. *Neurobiology of Aging*, 54.

Dalton, C. M., Bodini, B., Samson, R. S., Battaglini, M., Fisniku, L. K., Thompson, A. J., ..., and Chard, D. T. (2012). Brain lesion location and clinical status 20 years after a diagnosis of clinically isolated syndrome suggestive of multiple sclerosis. *Multiple Sclerosis Journal*, 18(3).

De Bresser, J., Kuijf, H. J., Zaanen, K., Viergever, M. A., Hendrikse, J., Biessels, G. J., ..., and Zwanenburg, J. (2018). White matter hyperintensity shape and location feature analysis on brain MRI; Proof of principle study in patients with diabetes. *Scientific Reports*, 8.

De Leeuw, F. E., De Groot, J. C., Achten, E., Oudkerk, M., Ramos, L. M., Heijboer, R., ..., and Breteler, M. M. (2001). Prevalence of cerebral white matter lesions in elderly people: A population based magnetic resonance imaging study. The Rotterdam Scan Study. *Journal of Neurology Neurosurgery and Psychiatry*, 70(1):9–14.

Debette, S. and Markus, H. S. (2010). The clinical importance of white matter hyperintensities on brain magnetic resonance imaging: Systematic review and meta-analysis. *BMJ (Online)*, 341.

Debette, S., Seshadri, S., Beiser, A., Au, R., Himali, J. J., Palumbo, C., ..., and DeCarli, C. (2011). Midlife vascular risk factor exposure accelerates structural brain aging and cognitive decline. *Neurology*, 77(5):461–468.

DeCarli, C., Reed, T., Miller, B. L., Wolf, P. A., Swan, G. E., and Carmelli, D. (1999). Impact of apolipoprotein E $\epsilon 4$ and vascular disease on brain morphology in men from the NHLBI twin study. *Stroke*, 30(8).

Dice, L. R. (1945). Measures of the Amount of Ecologic Association Between Species. *Ecology*, 26(3).

Dufouil, C., De Kersaint-Gilly, A., Besançon, V., Levy, C., Auffray, E., Brunnereau, L., ..., and Tzourio, C. (2001). Longitudinal study of blood pressure and white matter hyperintensities: The EVA MRI cohort. *Neurology*, 56(7).

Enzinger, C., Smith, S., Fazekas, F., Drevin, G., Ropele, S., Nichols, T., ..., and Matthews, P. M. (2006). Lesion probability maps of white matter hyperintensities in elderly individuals: Results of the Austrian stroke prevention study. *Journal of Neurology*, 253(8).

Evans, D., Beckett, L., Albert, M., Hebert, L., Scherr, P., Funkenstein, H., and Taylor, J. (1993). Level of education and change in cognitive function in a community population of older persons. *Annals of Epidemiology*, 1(3):71–77.

Fawns-Ritchie, C. and Deary, I. J. (2020). Reliability and validity of the UK Biobank cognitive tests. *PLoS ONE*, 15(4).

Fay, M. P. and Graubard, B. I. (2001). Small-sample adjustments for Wald-type tests using sandwich estimators. *Biometrics*, 57(4).

Fazekas, F., Chawluk, J. B., Alavi, A., Hurtig, H. I., and Zimmerman, R. A. (1987). MR signal abnormalities at 1.5 T in Alzheimer's dementia and normal aging. *American Journal of Neuroradiology*, 149(2).

Fazekas, F., Kleinert, R., Offenbacher, H., Schmidt, R., Kleinert, G., Payer, F., ..., and Lechner, H. (1993). Pathologic correlates of incidental mri white matter signal hyperintensities. *Neurology*, 43(9):1683–1683.

Fazekas, F., Schmidt, R., and Scheltens, P. (1998). Pathophysiologic mechanisms in the development of age-related white matter changes of the brain. *Dementia and Geriatric Cognitive Disorders*, 9(SUPPL. 1).

Fernández-Val, I. and Weidner, M. (2016). Individual and time effects in nonlinear panel models with large N, T. *Journal of Econometrics*, 192(1).

Filli, L., Hofstetter, L., Kuster, P., Traud, S., Mueller-Lenke, N., Naegelin, Y., ..., and Bendfeldt, K. (2012). Spatiotemporal distribution of white matter lesions in relapsing-remitting and secondary progressive multiple sclerosis. *Multiple Sclerosis Journal*, 18(11).

Firth, D. (1993). Bias Reduction of Maximum Likelihood Estimates. *Biometrika*, 80(1).

Fitzmaurice, G., Laird, N., and Ware, J. (2011). *Applied longitudinal Analysis (2nd Edition)*. John Wiley & Sons, Hoboken.

Fitzmaurice, G. M. (1995). A Caveat Concerning Independence Estimating Equations with Multivariate Binary Data. *Biometrics*, 51(1).

Fitzmaurice, G. M., Lipsitz, S. R., Arriaga, A., Sinha, D., Greenberg, C., and Gawande, A. A. (2014). Almost efficient estimation of relative risk regression. *Biostatistics*, 15(4).

Friston, K. J., Worsley, K. J., Frackowiak, R. S., Mazziotta, J. C., and Evans, A. C. (1994). Assessing the significance of focal activations using their spatial extent. *Human Brain Mapping*, 1(3).

Fry, A., Littlejohns, T. J., Sudlow, C., Doherty, N., Adamska, L., Sprosen, T., ..., and Allen, N. E. (2017). Comparison of Sociodemographic and Health-Related Characteristics of UK Biobank Participants with Those of the General Population. *American Journal of Epidemiology*, 186(9).

Gardiner, J. C., Luo, Z., and Roman, L. A. (2009). Fixed effects, random effects and GEE: What are the differences? *Statistics in Medicine*, 28(2).

Ge, T., Müller-Lenke, N., Bendfeldt, K., Nichols, T. E., and Johnson, T. D. (2014). Analysis of multiple sclerosis lesions via spatially varying coefficients. *Annals of Applied Statistics*, 8(2).

Genovese, C. R., Lazar, N. A., and Nichols, T. (2002). Thresholding of statistical maps in functional neuroimaging using the false discovery rate. *NeuroImage*, 15(4).

Ghaznawi, R., Geerlings, M. I., Jaarsma-Coes, M., Hendrikse, J., and de Bresser, J. (2021). Association of White Matter Hyperintensity Markers on MRI and Long-term Risk of Mortality and Ischemic Stroke: The SMART-MR Study. *Neurology*, 96(17).

Godin, O., Tzourio, C., Rouaud, O., Zhu, Y., Maillard, P., Pasquier, F., ..., and Dufouil, C. (2010). Joint effect of white matter lesions and hippocampal volumes on severity of cognitive decline: The 3C-Dijon MRI study. *Journal of Alzheimer's Disease*, 20(2).

Green, P. J. (1984). Iteratively Reweighted Least Squares for Maximum Likelihood Estimation, and Some Robust and Resistant Alternatives. *Journal of the Royal Statistical Society: Series B (Methodological)*, 46(2).

Greenland, S. and Thomas, D. C. (1982). On the need for the rare disease assumption in case-control studies. *American Journal of Epidemiology*, 116(3).

Greenland, S., Thomas, D. C., and Morgenstern, H. (1986). The rare-disease assumption revisited: A critique of "estimators of relative risk for case-control studies". *American Journal of Epidemiology*, 124(6).

Griffanti, L., Jenkinson, M., Suri, S., Zsoldos, E., Mahmood, A., Filippini, N., ..., and Zamboni, G. (2018). Classification and characterization of periventricular and deep white matter hyperintensities on MRI: A study in older adults. *NeuroImage*, 170.

Griffanti, L., Zamboni, G., Khan, A., Li, L., Bonifacio, G., Sundaresan, V., ..., and Jenkinson, M. (2016). BIANCA (Brain Intensity AbNormality Classification Algorithm): A new tool for automated segmentation of white matter hyperintensities. *NeuroImage*, 141.

Hagens, M. H., Burggraaff, J., Kilsdonk, I. D., Ruggieri, S., Collorone, S., Cortese, R., ..., and Wattjes, M. P. (2019). Impact of 3 Tesla MRI on interobserver agreement in clinically isolated syndrome: A MAGNIMS multicentre study. *Multiple Sclerosis Journal*, 25(3).

Hall, P. and Martin, M. A. (1988). Exact convergence rate of bootstrap quantile variance estimator. *Probability Theory and Related Fields*, 80(2).

Heagerty, P. J. and Zeger, S. L. (2000). Marginalized multilevel models and likelihood inference. *Statistical Science*, 15(1).

Hemerik, J., Goeman, J. J., and Finos, L. (2020). Robust testing in generalized linear models by sign flipping score contributions. *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, 82(3).

Holland, C. M., Smith, E. E., Csapo, I., Gurol, M. E., Brylka, D. A., Killiany, R. J., ..., and Greenberg, S. M. (2008). Spatial distribution of white-matter hyperintensities in Alzheimer disease, cerebral amyloid angiopathy, and healthy aging. *Stroke*, 39(4).

Howard, V. J. (2013). Reasons underlying racial differences in stroke incidence and mortality. *Stroke*, 44(6).

Hubbard, A. E., Ahern, J., Fleischer, N. L., Laan, M. V. D., Lippman, S. A., Jewell, N., ..., and Satariano, W. A. (2010). To GEE or not to GEE: Comparing population average and mixed models for estimating the associations between neighborhood risk factors and health. *Epidemiology*, 21(4).

Jeerakathil, T., Wolf, P. A., Beiser, A., Massaro, J., Seshadri, S., D'Agostino, R. B., and DeCarli, C. (2004). Stroke risk profile predicts white matter hyperintensity volume: The Framingham study. *Stroke*, 35(8).

Jenkinson, M., Bannister, P., Brady, M., and Smith, S. (2002a). Improved optimization for the robust and accurate linear registration and motion correction of brain images. *NeuroImage*, 17(2):825–841.

Jenkinson, M., Bannister, P., Brady, M., and Smith, S. (2002b). Improved Optimization for the Robust and Accurate Linear Registration and Motion Correction of Brain Images. *NeuroImage*, 17(2).

Kim, K. W., MacFall, J. R., and Payne, M. E. (2008). Classification of White Matter Lesions on Magnetic Resonance Imaging in Elderly Persons. *Biological Psychiatry*, 64(4):273–280.

Kim, K. W., Seo, H., Kwak, M. S., and Kim, D. (2017). Visceral obesity is associated with white matter hyperintensity and lacunar infarct. *International Journal of Obesity*, 41(5).

Kincses, Z. T., Ropele, S., Jenkinson, M., Khalil, M., Petrovic, K., Loitfelder, M., ..., and Enzinger, C. (2011). Lesion probability mapping to explain clinical deficits and cognitive performance in multiple sclerosis. *Multiple Sclerosis Journal*, 17(6).

Kindalova, P., Kosmidis, I., and Nichols, T. E. (2021a). Voxel-wise and spatial modelling of binary lesion masks: Comparison of methods with a realistic simulation framework. *NeuroImage*, 236.

Kindalova, P., Veldsman, M., Nichols, T. E., and Kosmidis, I. (2021b). Penalized generalized estimating equations for relative risk regression with applications to brain lesion data. *bioRxiv*.

Kloppenborg, R. P., Nederkoorn, P. J., Grool, A. M., Vincken, K. L., Mali, W. P., Vermeulen, M., ..., and Geerlings, M. I. (2012). Cerebral small-vessel disease and progression of brain atrophy : The SMART-MR study. *Neurology*, 79(20).

Knol, M. J., Le Cessie, S., Algra, A., Vandenbroucke, J. P., and Groenwold, R. H. (2012). Overestimation of risk ratios by odds ratios in trials and cohort studies: Alternatives to logistic regression. *CMAJ*, 184(8).

Knol, M. J., Vandenbroucke, J. P., Scott, P., and Egger, M. (2008). What do case-control studies estimate? Survey of methods and assumptions in published case-control research. *American Journal of Epidemiology*, 168(9).

Knopman, D., Boland, L. L., Mosley, T., Howard, G., Liao, D., Szklo, M., ..., and Folsom, A. R. (2001). Cardiovascular risk factors and cognitive decline in middle-aged adults. *Neurology*, 56(1):42–48.

Kosmidis, I. (2014). Bias in parametric estimation: Reduction and useful side-effects. *Wiley Interdisciplinary Reviews: Computational Statistics*, 6(3).

Kosmidis, I. (2020). *brglm2: Bias Reduction in Generalized Linear Models*. R package version 0.6.2.

Kosmidis, I. and Firth, D. (2009). Bias reduction in exponential family nonlinear models. *Biometrika*, 96(4).

Kosmidis, I. and Firth, D. (2010). A generic algorithm for reducing bias in parametric estimation. *Electronic Journal of Statistics*, 4.

Kosmidis, I. and Firth, D. (2021). Jeffreys-prior penalty, finiteness and shrinkage in binomial-response generalized linear models. *Biometrika*, 108(1).

Kosmidis, I., Kenne Pagui, E. C., and Sartori, N. (2020). Mean and median bias reduction in generalized linear models. *Statistics and Computing*, 30.

Laird, N. M. and Ware, J. H. (1982). Random-Effects Models for Longitudinal Data. *Biometrics*, 38(4).

Lampe, L., Kharabian-Masouleh, S., Kynast, J., Arelin, K., Steele, C. J., Löffler, M., ..., and Bazin, P. L. (2019a). Lesion location matters: The relationships between white matter hyperintensities on cognition in the healthy elderly. *Journal of Cerebral Blood Flow and Metabolism*, 39(1):36–43.

Lampe, L., Zhang, R., Beyer, F., Huhn, S., Kharabian Masouleh, S., Preusser, S., ..., and Witte, A. V. (2019b). Visceral obesity relates to deep white matter hyperintensities via inflammation. *Annals of Neurology*, 85(2).

Lancaster, T. (2000). The incidental parameter problem since 1948. *Journal of Econometrics*, 95(2).

Lancaster, T. (2002). Orthogonal parameters and panel data. *Review of Economic Studies*, 69(3).

Lee, J. J., Wedow, R., Okbay, A., Kong, E., Maghzian, O., Zacher, M., ..., and David Cesarini (2018). Gene discovery and polygenic prediction from a genome-wide association study of educational attainment in 1.1 million individuals. *Nature Genetics*, 50(8).

Lee, Y. and Nelder, J. A. (2004). Conditional and marginal models: Another view. *Statistical Science*, 19(2).

Lesaffre, E. and Albert, A. (1989). Partial Separation in Logistic Discrimination. *Journal of the Royal Statistical Society: Series B (Methodological)*, 51(1).

Li, C., Srivastava, S., and Dunson, D. B. (2017). Simple, scalable and accurate posterior interval estimation. *Biometrika*, 104(3).

Liang, K. Y. and Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, 73(1).

Lindsey, J. K. and Lambert, P. (1998). On the appropriateness of marginal models for repeated measurements in clinical trials. *Statistics in Medicine*, 17(4).

Litière, S., Alonso, A., and Molenberghs, G. (2007). Type I and type II error under random-effects misspecification in generalized linear mixed models. *Biometrics*, 63(4).

Lloyd-Jones, D. M., Wilson, P. W., Larson, M. G., Beiser, A., Leip, E. P., D'Agostino, R. B., and Levy, D. (2004). Framingham risk score and prediction of lifetime risk for coronary heart disease. *American Journal of Cardiology*, 94(1).

Lu, M. and Tilley, B. C. (2001). Use of odds ratio or relative risk to measure a treatment effect in clinical trials with multiple correlated binary outcomes: Data from the NINDS t-PA stroke trial. *Statistics in Medicine*, 20(13).

Lubin, J. H., Couper, D., Lutsey, P. L., Woodward, M., Yatsuya, H., and Huxley, R. R. (2016). Risk of cardiovascular disease from cumulative cigarette use and the impact of smoking intensity. *Epidemiology*, 27(3).

Luo, J., Zhang, J., and Sun, H. (2014). Estimation of relative risk using a log-binomial model with constraints. *Computational Statistics*, 29(5):981–1003.

Lyall, D., Cox, S., Lyall, L., Celis-Morales, C., Cullen, B., Mackay, D., ..., and Pell, J. (2019). Association between apoe e4 and white matter hyperintensity volume, but not total brain volume or white matter integrity. *Brain Imaging and Behavior*, pages 1–9.

Mancl, L. A. and DeRouen, T. A. (2001). A covariance estimator for GEE with improved small-sample properties. *Biometrics*, 57(1).

Mansournia, M. A., Etminan, M., Danaei, G., Kaufman, J. S., and Collins, G. (2017). Handling time varying confounding in observational research. *BMJ (Online)*, 359.

Mansournia, M. A., Geroldinger, A., Greenland, S., and Heinze, G. (2018). Separation in Logistic Regression: Causes, Consequences, and Control. *American Journal of Epidemiology*, 187(4).

McCarron, M. O., Delong, D., and Alberts, M. J. (1999). APOE genotype as a risk factor for ischemic cerebrovascular disease: A meta-analysis. *Neurology*, 53(6).

McNutt, L. A., Wu, C., Xue, X., and Hafner, J. P. (2003). Estimating the relative risk in cohort studies and clinical trials of common outcomes. *American Journal of Epidemiology*, 157(10).

Miller, K. L., Alfaro-Almagro, F., Bangerter, N. K., Thomas, D. L., Yacoub, E., Xu, J., ..., and Smith, S. M. (2016). Multimodal population brain imaging in the UK Biobank prospective epidemiological study. *Nature Neuroscience*, 19(11).

Minsker, S., Srivastava, S., Lin, L., and Dunson, D. B. (2017). Robust and scalable bayes via a median of subset posterior measures. *Journal of Machine Learning Research*, 18.

Mondol, M. H. and Rahman, M. S. (2019). Bias-reduced and separation-proof GEE with small or sparse longitudinal binary data. *Statistics in Medicine*, 38(14).

Morel, J. G., Bokossa, M. C., and Neerchal, N. K. (2003). Small sample correction for the variance of GEE estimators. *Biometrical Journal*, 45(4).

Moroni, F., Ammirati, E., Rocca, M. A., Filippi, M., Magnoni, M., and Camici, P. G. (2018). Cardiovascular disease and brain health: Focus on white matter hyperintensities. *IJC Heart and Vasculature*, 19.

Mortamais, M., Artero, S., and Ritchie, K. (2013). Cerebral white matter hyperintensities in the prediction of cognitive decline and incident dementia. *International Review of Psychiatry*, 25(6).

Neuhaus, J. M. and Kalbfleisch, J. D. (1998). Between- and Within-Cluster Covariate Effects in the Analysis of Clustered Data. *Biometrics*, 54(2).

Neyman, J. and Scott, E. L. (1948). Consistent Estimates Based on Partially Consistent Observations. *Econometrica*, 16(1).

Nichols, T. E. and Holmes, A. P. (2002). Nonparametric permutation tests for functional neuroimaging: A primer with examples. *Human Brain Mapping*, 15(1).

Onat, A., Avci, G. S., Barlan, M., Uyarel, H., Uzunlar, B., and Sansoy, V. (2004). Measures of abdominal obesity assessed for visceral adiposity and relation to coronary risk. *International journal of obesity and related metabolic disorders : journal of the International Association for the Study of Obesity*, 28:1018–25.

Pasha, E. P., Birdsill, A., Parker, P., Elmenshawy, A., Tanaka, H., and Haley, A. P. (2017). Visceral adiposity predicts subclinical white matter hyperintensities in middle-aged adults. *Obesity Research and Clinical Practice*, 11(2).

Paul, S. and Zhang, X. (2014). Small sample GEE estimation of regression parameters for longitudinal data. *Statistics in Medicine*, 33(22).

Pedroza, C. and Truong, V. T. T. (2017). Estimating relative risks in multicenter studies with a small number of centers - which methods to use? A simulation study. *Trials*, 18(1).

Polman, C. H., Reingold, S. C., Banwell, B., Clanet, M., Cohen, J. A., Filippi, M., ..., and Wolinsky, J. S. (2011). Diagnostic criteria for multiple sclerosis: 2010 Revisions to the McDonald criteria. *Annals of Neurology*, 69(2):292–302.

Prins, N. D., Van Dijk, E. J., Den Heijer, T., Vermeer, S. E., Jolles, J., Koudstaal, P. J., ..., and Breteler, M. M. (2005). Cerebral small-vessel disease and decline in information processing speed, executive function and memory. *Brain*, 128(9).

Qaqish, B. F. (2003). A family of multivariate binary distributions for simulating correlated binary variables with specified marginal means and correlations. *Biometrika*, 90(2).

Quenouille, M. H. (1956). Notes on Bias in Estimation. *Biometrika*, 43(3/4).

Rorden, C. and Karnath, H. O. (2004). Using human brain lesions to infer function: A relic from a past era in the fMRI age? *Nature Reviews Neuroscience*, 5(10).

Rossi, F., Giorgio, A., Battaglini, M., Stromillo, M. L., Portaccio, E., Goretti, B., Federico, A., Hakiki, B., Amato, M. P., and de Stefano, N. (2012). Relevance of Brain Lesion Location to Cognition in Relapsing Multiple Sclerosis. *PLoS ONE*, 7(11).

Rostrup, E., Gouw, A. A., Vrenken, H., Van Straaten, E. C., Ropele, S., Pantoni, L., ..., and Waldemar, G. (2012). The spatial distribution of age-related white matter changes as a function of vascular risk factors-Results from the LADIS study. *NeuroImage*, 60(3).

Rudick, R., Altay, E., Lee, J.-C., Jones, S., Hara-Cleaver, C., and Fisher, E. (2012). Inter-Rater Reliability among Clinical Raters for New MRI Lesions in MS Patients (P03.068). *Neurology*, 78(Meeting Abstracts 1).

Ryu, W. S., Woo, S. H., Schellingerhout, D., Chung, M. K., Kim, C. K., Jang, M. U., ..., and Kim, D. E. (2014). Grading and interpretation of white matter hyperintensities using statistical maps. *Stroke*, 45(12).

Sachdev, P., Wen, W., Chen, X., and Brodaty, H. (2007). Progression of white matter hyperintensities in elderly individuals over 3 years. *Neurology*, 68(3).

Sachdev, P. S., Parslow, R., Wen, W., Anstey, K. J., and Easteal, S. (2009). Sex differences in the causes and consequences of white matter hyperintensities. *Neurobiology of Aging*, 30(6).

Salvadó, G., Brugulat-Serrat, A., Sudre, C. H., Grau-Rivera, O., Suárez-Calvet, M., Falcon, C., ..., and Gispert, J. D. (2019). Spatial patterns of white matter hyperintensities associated with Alzheimer's disease risk factors in a cognitively healthy middle-aged cohort. *Alzheimer's Research and Therapy*, 11(1).

Schiepers, O. J., Harris, S. E., Gow, A. J., Pattie, A., Brett, C. E., Starr, J. M., and Deary, I. J. (2012). APOE E4 status predicts age-related cognitive decline in the ninth decade: Longitudinal follow-up of the Lothian Birth Cohort 1921. *Molecular Psychiatry*, 17(3):315–324.

Schlather, M., Malinowski, A., Oesting, M., Boecker, D., Strokorb, K., Engelke, S., ..., and Pfaff, B. (2020). *RandomFields: Simulation and Analysis of Random Fields*. R package version 3.3.8.

Schmidt, R., Fazekas, F., Kapeller, P., Schmidt, H., and Hartung, H. P. (1999). MRI white matter hyperintensities: Three-year follow-up of the Austrian Stroke Prevention Study. *Neurology*, 53(1).

Schmidt, R., Ropele, S., Enzinger, C., Petrovic, K., Smith, S., Schmidt, H., ..., and Fazekas, F. (2005). White matter lesion progression, brain atrophy, and cognitive decline: The Austrian stroke prevention study. *Annals of Neurology*, 58(4).

Schwendinger, F., Grün, B., and Hornik, K. (2021). A comparison of optimization solvers for log binomial regression including conic programming. *Computational Statistics*, 36(3).

Seidell, J. C., Oosterlee, A., Deurenberg, P., Hautvast, J. G., and Ruijs, J. H. (1988). Abdominal fat depots measured with computed tomography: Effects of degree of obesity, sex, and age. *European Journal of Clinical Nutrition*, 42(9):805–815.

Sharples, K. and Breslow, N. (1992). Regression analysis of correlated binary data: Some small sample results for the estimating equation approach. *Journal of Statistical Computation and Simulation*, 42(1-2).

Sherman, M. and Cessie, S. l. (1997). A comparison between bootstrap methods and generalized estimating equations for correlated outcomes in generalized linear models. *Communications in Statistics Part B: Simulation and Computation*, 26(3).

Shuster, A., Patlas, M., and Pinthus, J. (2011). The clinical importance of visceral adiposity: A critical review of methods for visceral adipose tissue analysis. *The British journal of radiology*, 85:1–10.

Smith, S. M. and Nichols, T. E. (2009). Threshold-free cluster enhancement: Addressing problems of smoothing, threshold dependence and localisation in cluster inference. *NeuroImage*, 44(1).

Srivastava, S., Li, C., and Dunson, D. B. (2018). Scalable Bayes via barycenter in Wasserstein space. *Journal of Machine Learning Research*, 19.

Strassburger, T. L., Lee, H. C., Daly, E. M., Szczepanik, J., Krasuski, J. S., Mentis, M. J., ..., and Alexander, G. E. (1997). Interactive effects of age and hypertension on volumes of brain structures. *Stroke*, 28(7).

Sudlow, C., Gallacher, J., Allen, N., Beral, V., Burton, P., Danesh, J., ..., and Collins, R. (2015). UK Biobank: An Open Access Resource for Identifying the Causes of a Wide Range of Complex Diseases of Middle and Old Age. *PLoS Medicine*, 12(3).

Sudre, C. H., Cardoso, M. J., Frost, C., Barnes, J., Barkhof, F., Fox, N., and Ourselin, S. (2017). APOE $\epsilon 4$ status is associated with white matter hyperintensities volume accumulation rate independent of AD diagnosis. *Neurobiology of Aging*, 53.

Sundaresan, V., Griffanti, L., Kindalova, P., Alfaro-Almagro, F., Zamboni, G., Rothwell, P., ..., and Jenkinson, M. (2019). Modelling the distribution of white matter hyperintensities due to ageing on MRI images using Bayesian inference. *NeuroImage*, 185.

Tingley, D., Yamamoto, T., Hirose, K., Keele, L., and Imai, K. (2014). Mediation: R package for causal mediation analysis. *Journal of Statistical Software*, 59(5):1–38.

Van Dijk, E. J., Breteler, M. M., Schmidt, R., Berger, K., Nilsson, L. G., Oudkerk, M., ..., and Hofman, A. (2004). The association between blood pressure, hypertension, and cerebral white matter lesions: Cardiovascular determinants of dementia study. *Hypertension*, 44(5).

Van Dijk, E. J., Prins, N. D., Vrooman, H. A., Hofman, A., Koudstaal, P. J., and Breteler, M. M. (2008). Progression of cerebral small vessel disease in relation to risk factors and cognitive consequences: Rotterdam scan study. *Stroke*, 39(10).

Veldsman, M., Kindalova, P., Husain, M., Kosmidis, I., and Nichols, T. E. (2020). Spatial distribution and cognitive impact of cerebrovascular risk-related white matter hyperintensities. *NeuroImage: Clinical*, 28.

Verhaaren, B. F., Vernooij, M. W., De Boer, R., Hofman, A., Niessen, W. J., Van Der Lugt, A., and Ikram, M. A. (2013). High blood pressure and cerebral white matter lesion progression in the general population. *Hypertension*, 61(6).

Wacholder, S. (1986). Binomial regression in glim: Estimating risk ratios and risk differences. *American Journal of Epidemiology*, 123(1).

Wang, Y. G. and Carey, V. (2003). Working correlation structure misspecification, estimation and covariate design: Implications for generalised estimating equations performance. *Biometrika*, 90(1).

Wardlaw, J. M., Smith, E. E., Biessels, G. J., Cordonnier, C., Fazekas, F., Frayne, R., ..., and Dichgans, M. (2013). Neuroimaging standards for research into small vessel disease and its contribution to ageing and neurodegeneration. *Lancet Neurology*, 12(8).

Wardlaw, J. M., Valdés Hernández, M. C., and Muñoz-Maniega, S. (2015). What are white matter hyperintensities made of? Relevance to vascular cognitive impairment. *Journal of the American Heart Association*, 4(6).

Westgate, P. M. and Burchett, W. W. (2017). A Comparison of Correlation Structure Selection Penalties for Generalized Estimating Equations. *American Statistician*, 71(4).

Whalley, L. J., Deary, I. J., Appleton, C. L., and Starr, J. M. (2004). Cognitive reserve and the neurobiology of cognitive aging. *Ageing Research Reviews*, 3(4).

Wiseman, R. M., Saxby, B. K., Burton, E. J., Barber, R., Ford, G. A., and O'Brien, J. T. (2004). Hippocampal atrophy, whole brain volume, and white matter lesions in older hypertensive subjects. *Neurology*, 63(10).

World Health Organization (2008). Waist Circumference and Waist-Hip Ratio. *Report of a WHO Expert Consultation*.

Yelland, L. N., Salter, A. B., and Ryan, P. (2011a). Performance of the modified poisson regression approach for estimating relative risks from clustered prospective data. *American Journal of Epidemiology*, 174(8).

Yelland, L. N., Salter, A. B., and Ryan, P. (2011b). Relative risk estimation in cluster randomized trials: A comparison of generalized estimating equation methods. *International Journal of Biostatistics*, 7(1).

Zeger, S. L. and Liang, K. Y. (1992). An overview of methods for the analysis of longitudinal data. *Statistics in Medicine*, 11(14-15).

Zhang, J. and Yu, K. F. (1998). What's the relative risk? A method of correcting the odds ratio in cohort studies of common outcomes. *Journal of the American Medical Association*, 280(19).

Ziegler, A. and Vens, M. (2010). Generalized estimating equations: Notes on the choice of the working correlation matrix. *Methods of Information in Medicine*, 49(5).

Zijdenbos, A. P., Dawant, B. M., Margolin, R. A., and Palmer, A. C. (1994). Morphometric Analysis of White Matter Lesions in MR Images: Method and Validation. *IEEE Transactions on Medical Imaging*, 13(4).

Zou, G. (2004). A Modified Poisson Regression Approach to Prospective Studies with Binary Data. *American Journal of Epidemiology*, 159(7).

Zou, G. Y. and Donner, A. (2013). Extension of the modified Poisson regression model to prospective studies with correlated binary data. *Statistical Methods in Medical Research*, 22(6):661–670.

Zou, K. H., Warfield, S. K., Bharatha, A., Tempany, C. M., Kaus, M. R., Haker, S. J., Wells, W. M., Jolesz, F. A., and Kikinis, R. (2004). Statistical Validation of Image Segmentation Quality Based on a Spatial Overlap Index. *Academic Radiology*, 11(2).