Scalable and Interpretable Spatial Models for Neuroimaging Applications

Yifan Yu

Keble College

University of Oxford

A thesis submitted for the degree of

Doctor of Philosophy

Hilary 2025

Acknowledgements

First and foremost, I would like to express my deepest gratitude to my supervisor, Professor Thomas E. Nichols. His unwavering support, insightful guidance, and constant inspiration have been invaluable throughout my PhD journey. It has been a true privilege to be mentored by someone I regard as a research role model. His depth and breadth of knowledge have been both humbling and enlightening. A special acknowledgement goes to Tom, whose thoughtfulness and readiness to offer support during challenging times, both academically and personally, have meant so much to me.

I am also deeply grateful to Professor Angela Laird and Dr. Alejandro De La Vega for their insightful feedback on my work. My heartfelt thanks go to Dr. James Kent for guiding me in contributing to the NiMARE package, and for the countless hours he dedicated to teaching me formalised coding practices and improving the efficiency and readability of my codes. I would also like to express my sincere gratitude to my internship manager, Martina De Stefani, and all my colleagues at Amazon. This opportunity to experience the difference between academic and industrial research, has been invaluable in helping me make decisions about my future career path.

I would also like to thank Professor Ludovica Griffanti and Professor Simon Eickhoff for their insightful feedbacks and stimulating discussions during my PhD viva. Their comments have profoundly deepened my understanding of data interpretation and will undoubtedly shape my future research endeavours. I am also grateful for the considerable time they dedicated to thoroughly reviewing my thesis. My appreciation also goes to thank Professor Saad Jbabdi and Dr Natalie Staplin for the invaluable feedback they provided during my transfer and confirmation of status assessments. My sincere thanks also go to the wonderful labmates and colleagues of Tom's research group, past and present. In particular, I would like to express my gratitude to Anna Menacher, Kan Keeratimahat, Lav Radosavljevic, Saba Ishrat, Angeline Lee, Thomas Maullin-Sapey, Alex Bowring, Sam Davenport, George Hustchings, Emma Prevot,

Yang Sun, Anya Topiwala, Bernd Taschler, Konstantin Shestopaloff, Habib Ganjgahi, for their support, collaboration and inspiring discussions throughout my PhD journey.

Moreover, I feel incredibly fortunate to be surrounded by so many amazing and caring friends, far too many to name individually. I would like to especially thank Natalia Hong, Kevin Wang, Hang Yuan, Chenyang Wang from Keble College; Xi Lin, Linying Yang, Sahra Ghalebikesabi, Yixuan He, Hanwen Xing, Yanzhao Yang, Ning Miao, Chao Zhang, Yutong Lu, Zhixiao Zhu, Guneet Singh Dhillon from the Department of Statistics, as well as Sijia Yao, Kangning Zhang, Ziyun Liang and many others from other departments at Oxford, I am also grateful for friends from earlier chapters of my life, including those from high school or undergraduate studies – Zhiqi Wang, Kaiyue Zhang, Tianxiao Wang, Yongtong Chen, Wenxuan Dong, Luyang Cui, Yuru Bai, Zhimeng Shi and Xue Lin.

Finally, above all, my deepest thanks go to Jin Xu for his love and companionship, for the countless train tickets between Cambridge and Oxford, between Tübingen and Luxembourg. I couldn't imagine my PhD journey without him. He has brought immense joy, unwavering support and continual inspirations to both my life and research. Additionally, I am eternally grateful to my parents Zhisan Yu and Xiao Xia. Their unconditional support and constant encouragement gave me the freedom to pursue my passions, and the thought of them has always brought me comfort during my most difficult times.

Abstract

Neuroimaging techniques, such as functional magnetic resonance imaging (fMRI) and structural MRI, have become essential tools for understanding function and pathology of the human brain. fMRI allows researchers to identify brain regions associated with specific cognitive and behavioural processes by measuring blood-oxygen-level-dependent (BOLD) signals. Structural MRI provides high-resolution mapping of brain anatomy, allowing for the identification of morphological alternations, such as white matter lesions, which improves our understanding of neurodegenerative and cerebrovascular diseases. Despite recent developments in data sharing and availability of large-scale neuroimaging cohorts, several common analytical challenges remain. Individual neuroimaging studies often rely on relatively small sample sizes, which limits statistical power and reduces the generalisability of findings. Many existing analytical approaches struggle with balancing model complexity, interpretability and scalability. Another critical limitation is the lack of methods that explicitly model spatial dependence in neuroimaging data while maintaining computationally efficiency. Addressing these challenges is crucial for improving the reliability, reproducibility, and clinical relevance of neuroimaging-based research.

This thesis addresses these challenges by developing statistical methods that explicitly capture spatial dependence in neuroimaging data, while ensuring computational efficiency and strong scalability. The proposed methodological frameworks are built upon generalised linear models (GLMs) that incorporate spatial components through spline parametrisations or Gaussian kernels. These models provide flexibility to accommodate either globally constant or spatially varying covariates, support probability or intensity estimation within the regression framework, and enable flexible voxel-wise statistical testing for spatial homogeneity or group comparisons. To address computational challenges posed by large-scale neuroimaging datasets, this thesis introduces several dimension reduction and efficiency-enhancing strategies, including model factorisation, parallel computing, memory-efficient algorithms. In addition, it investigates robust and accurate inference methods, such as standard error estimation using

sandwich estimator, and a bootstrap-based approach for non-parametric inference.

Our proposed methods are applied to two important neuroimaging applications, demonstrating their interpretability, scalability, and robustness: (1) coordinate-based meta-analysis (CBMA) of task- or stimulus-based fMRI studies, and (2) brain lesion probability estimation from T2-weighted MRI data, as well as inference on the spatially varying effects of risk factors. In the context of CBMA, the developed models offer a more accurate and spatially smooth approximation of activation intensity functions from CBMA data. They also enable a systematic investigation of potential sources of variability. This framework is further extended to accommodate multiple-group comparisons, allowing for inference on differences in activation patterns across groups. For brain lesion mapping, similar principles are employed to estimate the voxel-wise effects of clinical risk factors (such as age and cardiovascular conditions) on lesion distribution across the brain.

Declaration

I declare that this thesis is entirely my own work, except where specific reference is made to the work of others. The content of this thesis is original and has not been submitted, in whole or in part, for consideration for any other degree or qualification at this or any other university or institution. Any material that is the outcome of work done in collaboration is indicated as such within the text.

- The work presented in Chapter 3 has been published in *Biostatistics* journal under the title "Neuroimaging Meta Regression for Coordinate Based Meta Analysis Data with a Spatial Model" (Yu et al., 2024). This work was also presented as a poster at the *Organization for Human Brain Mapping* (OHBM) Annual Meetings in 2021 and 2022.
- The work presented in Chapter 4 shall be submitted for publication shortly. This work was previously presented as a poster at the *Organization for Human Brain Mapping* (OHBM) Annual Meetings in 2023.
- The work presented in Chapter 5 shall be submitted for publication shortly.

Yifan Yu

April 2025

Contents

Chapter 1

Introduction

1.1 Motivation

Over the past few decades, neuroimaging has become essential to advancing our understanding of brain function and pathology. Functional neuroimaging modalities such as functional magnetic resonance imaging (fMRI) and positron emission tomography (PET) infer neural activity by measuring changes in cerebral blood flow, oxygenation levels, and metabolic processes [Logothetis, 2008, Raichle, 1998]. These techniques have significantly contributed to cognitive neuroscience by allowing researchers to identify brain regions associated with a wide range of cognitive processes, such as perception, memory, attention, and decision-making [Cabeza and Nyberg, 2000, Petersen and Posner, 2012]. Complementing functional methods, structural magnetic resonance imaging (MRI) provides high-resolution anatomical images, which are essential for identifying and characterising structural brain abnormalities. For example, white matter lesions, which are frequently observed in ageing populations and are associated with neurological conditions such as small vessel disease and dementia, can be readily visualised using structural MRI [Wardlaw et al., 2013b, Debette and Markus, 2010].

As the field of neuroimaging has rapidly expanded, so too has the need to synthesise findings across individual studies. Neuroimaging research often faces limitations such as small sample sizes, heterogeneous methodologies, and diverse data analysis pipelines. These issues reduce statistical power and increase the likelihood of false-positive results, previous studies estimate that $10 - 20\%$ of reported activation coordinates in fMRI research may reflect spurious rather than true effects [Eickhoff et al., 2012,

Button et al., 2013]. To address these limitations, researchers increasingly employ meta-analytic approaches that aggregate evidence from multiple studies to improve statistical robustness and reveal consistent activation patterns across diverse experimental designs [Turkeltaub et al., 2002, Fox et al., 2005].

Coordinate-based meta-analysis (CBMA) remains one of the most widely adopted methods in neuroimaging meta-analysis. In contrast to image-based meta-analysis (IBMA), which requires access to full statistical parametric maps, CBMA relies solely on reported peak activation coordinates extracted from published studies. Its prominence is largely attributable to historical limitations in data sharing, when entire statistical maps were rarely available. These constraints led to the development of large-scale coordinate databases such as BrainMap [Laird et al., 2005] and Neurosynth [Yarkoni et al., 2011], which aggregate activation foci across thousands of studies. Despite its widespread application, CBMA is constrained by the inherent sparsity and incompleteness of coordinate-based data. Specifically, it lacks critical information regarding the spatial extent and statistical magnitude of activation clusters, which restricts the precision of effect size estimation and increases uncertainty in the localisation of true effects [Salimi-Khorshidi et al., 2009, Eickhoff et al., 2009]. Additionally, heterogeneity in statistical thresholding, reporting practices and anatomical labelling across studies further contributes to variability in the meta-analytic results [Samartsidis et al., 2019]. These limitations can affect the validity of statistical inferences, particularly when estimating the consistency of activation across studies or comparing effects between experiment conditions.

In parallel, structural neuroimaging research has focused on brain abnormalities such as white matter hyperintensities (WMHs), which are frequently observed in ageing populations. WMHs appear as regions of increased signal intensity on T2-weighted and fluid-attenuated inversion recovery (FLAIR) MRI sequences and are widely recognised markers of cerebral small vessel disease [Wardlaw et al., 2013b, Debette and Markus, 2010]. Their presence and volume are strongly associated with vascular risk factors, including hypertension, diabetes, smoking and hypercholesterolemia [DeCarli et al., 2005b, Schmidt et al., 1999]. The spatial distribution of WMHs, particularly in periventricular and deep white matter regions, has been associated with cognitive impairment, reduced processing speed, executive dysfunction and an increased risk of developing dementia [Gunning-Dixon and Raz, 2000, Debette and Markus, 2010]. Quantifying and mapping WMHs is essential for understanding their clinical relevance. Modern lesion mapping methods often involve estimating voxel-wise lesion probabilities

across the brain and assessing their associations with demographic and clinical risk factors [Rostrup et al., 2012]. These analyses provide valuable insights into the underlying mechanisms of disease, support early risk stratification, and inform the development of personalised prevention and intervention strategies.

Despite differences in data types and research objectives, CBMA and lesion mapping share common methodological challenges, most notably, the need to accurately model spatial dependence across brain regions. Conventional voxel-wise statistical approaches typically treat each voxel independently, ignoring the inherent spatial correlations present in neuroimaging data. This assumption of independence can result in inefficient parameter estimates, inflated false positive rates and reduced statistical power, ultimately leading to results that are difficult to interpret or potentially misleading [Worsley et al., 2002, Nichols and Holmes, 2002]. Accounting for spatial structure is particularly critical in large-scale neuroimaging datasets such as the UK Biobank [Miller et al., 2016] and the Adolescent Brain Cognitive Development (ABCD) study [Casey et al., 2018]. These datasets also highlight the growing demand for statistical methods that are both spatially informed and scalable, while remaining computationally efficient. As neuroimaging studies continue to increase in size and complexity, there is an growing need for analytical approaches that balance computational feasibility with statistical rigour and interpretability [Bowring et al., 2019].

Although recent advances in Bayesian spatial models have enabled the explicit incorporation of spatial dependence in neuroimaging analyses, offering improved accuracy and more interpretable results [Bowman, 2007, Montagna et al., 2018], these models typically rely on computationally intensive algorithms, such as Markov Chain Monte Carlo (MCMC), and often require parallel computing or approximation techniques for practical implementation [Rue et al., 2009]. While these methods are well-suited for capturing complex spatial patterns, their high computational demands present significant challenges when scaling to large neuroimaging datasets. In contrast, classical mass-univariate approaches, although simple and computationally efficient, fail to account for spatial dependence and may produce unstable or suboptimal results, particularly in sparse or noisy data scenarios. This trade-off between model complexity and computationally feasibility remains a central challenge in developing robust and scalable frameworks for neuroimaging analysis.

Given these limitations, there remains a clear and ongoing need for methodological frameworks that explicitly model spatial dependencies, flexibly incorporate publication-

level covariates or clinical risk factors, and scale efficiently to large neuroimaging datasets. The objective of this thesis is to develop practical statistical approaches that strike a balance between interpretability, statistical validity and computational efficiency, thereby facilitating robust and reproducible analyses in neuroimaging research.

1.2 Overview

This thesis presents a unified methodological framework to address key analytical challenges in neuroimaging. Through three interconnected projects, it advances statistical methods that explicitly model spatial dependencies, incorporate covariates that are either globally consistent or spatially varying, and ensure scalability for large and complex datasets.

Chapter 3 addresses foundational methodological challenges in coordinate-based meta-analysis (CBMA), a widely used technique for synthesising activation coordinates reported across multiple fMRI studies. Existing CBMA approaches include kernel-based methods, which offer limited statistical interpretability and flexibility, and Bayesian methods, which provide greater interpretability but are computationally intensive. To overcome these limitations, this project introduces a frequentist spatial generalised linear model (GLM) that employs spline-based smoothing to explicitly capture spatial dependence. This model also supports meta-regression with publication-level covariates, enabling the investigation of heterogeneity across studies.

Chapter 4 extends the framework introduced in Chapter 3 to accommodate multi-group CBMA, addressing the practical need to compare activation patterns across different experimental or clinical groups, which is a common challenge in cognitive neuroscience. The proposed multi-group model estimates group-specific spatial intensity functions while sharing globally consistent publication-level covariates. To mitigate issues related to data sparsity (i.e., insufficient numbers of foci per group), the model incorporates regularisation techniques alongside robust inference procedures based on bootstrapping and resampling. These improvements ensure numerical stability and valid statistical inference even in challenging scenarios. The methodological developments from this project have been implemented as a module within NiMARE, an accessible, open-source software library.

Finally, Chapter 5 demonstrates the broader applicability of the spatial modelling framework by extending its statistical principles to a different neuroimaging task: voxel-wise lesion probability mapping using structural MRI data. Unlike CBMA, which aggregates sparse data across multiple studies, lesion mapping utilises individual-level imaging data to estimate how clinical risk factors (such as age, hypertension, or genetic markers) influence lesion incidence across the brain. Conventional mass-univariate methods for this task ignore spatial dependencies, while advanced Bayesian spatial models are computationally intensive for large-scale datasets. To bridge this gap, we propose a spatial GLM approach specifically adapted for lesion data, again leveraging spline-based smoothing, scalable computational strategies through model factorisation, and robust inference techniques. These methodological innovations ensure an efficient and reliable estimation of lesion probability maps, even in large datasets such as the UK Biobank [Miller et al., 2016].

Taken together, these three projects demonstrate how a cohesive spatial modelling framework can be effectively applied across a broad range of neuroimaging analyses. Progressing from single-group meta-regression to multi-group comparisons, and ultimately to individual-level lesion mapping, the thesis highlights the flexibility, generalisability, and practical utility of spatially informed GLMs. Collectively, this work provides the neuroimaging community with robust, interpretable, and computationally efficient analytical tools that enhance the reliability and reproducibility of scientific discovery.

Chapter 2

Background

In this chapter, we provide background material that forms the foundation of our research. We begin with a broad overview of neuroimaging modalities and their applications in neuroscience in Section 2.1. Section 2.2 then introduces statistical modelling techniques for neuroimaging data, with a focus on the structure of generalised linear models (GLMs) and their estimation and optimisation methods. In Section 2.3, we outline key statistics preliminaries that are employed throughout this thesis. Finally, Sections 2.4 and 2.5 describe the two primary neuroimaging applications to which our proposed methods are applied: meta-analysis in neuroimaging studies, and the clinical relevance, pathophysiology and segmentation of white matter hyperintensities (WMHs).

This chapter provides a concise overview of key neuroimaging statistics concepts referenced throughout this thesis. It is not intended as a comprehensive introduction and therefore omits detailed explanations of concepts that are not central to the core contributions of this work. Additional information related to Section 2.1, Section 2.4 and Section 2.5 can be found in Poldrack et al. [2011], Jenkinson and Chappell [2018] and Brown and Semelka [2011]. Broader neuroimaging statistical methods discussed in Section 2.2 and 2.3 are covered in more detail by Lazar [2008], Fox [2015] and Dobson and Barnett [2018].

2.1 Neuroimaging

In this section, we first provide a brief introduction to neuroimaging applications in neuroscience (Section 2.1.1). We then narrow the focus to task- or stimulus-based functional magnetic resonance imaging (fMRI) and T2-weighted structural MRI, discussed in Sections 2.1.2 and 2.1.4, respectively. Following this, Section 2.1.3 outlines the standard preprocessing pipeline used for fMRI data prior to statistical analysis. Finally, Section 2.1.5 presents several examples of publicly available large-scale neuroimaging datasets, reflecting the growing emphasis on data sharing and open access practices within the neuroscience research community.

2.1.1 Neuroimaging in Neuroscience

Historically, neuroscience as a distinct and integrative scientific discipline has experienced rapid growth over the past few decades, although its foundations date back much further. Early investigations of the nervous system can be traced to ancient civilizations, while systematic neuroanatomical studies began in the 19th century [Finger, 2001, Shepherd, 2015]. Modern neuroscience began to take shape in the latter half of the 20th century, driven by major breakthroughs in molecular and cellular biology, such as the discovery and characterisation of ion channels and neurotransmitters [Kandel, 2001], as well as advances in electrophysiological techniques, notably the development of patch-clamp recording for measuring neuronal activity at the single-cell level [Neher and Sakmann, 1976]. At the same time, the emergence of non-invasive neuroimaging techniques, including magnetic resonance imaging (MRI), functional MRI (fMRI), positron emission tomography (PET), and electroencephalography (EEG), provided unprecedented insights into the structure and function of the living human brain [Raichle, 2009]. Over the past thirty years, neuroscience has evolved into a highly interdisciplinary and data-intensive field, integrating perspectives and methodologies from biology, psychology, physics, computer science, mathematics, and biomedical engineering. This convergence has allowed researchers to investigate the brain across multiple spatial and temporal scales, from molecular signalling and synaptic transmission to complex neural circuits and large-scale brain networks underlying perception, cognition and behaviour [Sejnowski et al., 2014].

Neuroimaging methods have become central to neuroscience research, providing

powerful non-invasive tools for investigating how the brain gives rise to cognitive functions, emotional states, and behavioural responses. Among these techniques, fMRI has emerged as one of the most widely used modalities. By measuring blood-oxygen-level-dependent (BOLD) signals as an indicator of neural activity, fMRI enables researchers to identify and characterise brain regions and large-scale functional networks involved in various cognitive processes [Ogawa et al., 1990, Logothetis, 2008]. This approach has profoundly advanced our understanding of the neural correlates of attention, memory, language, decision-making, and affective processing [Petersen and Sporns, 2015]. Complementary to fMRI, EEG and MEG offer millisecond-level temporal resolution, making them indispensable for investigating the timing of neural events. EEG, in particular, is valuable for studying event-related potentials (ERPs), which reveal the sequential stages of sensory processing and decision-making [Luck, 2014]. Similarly, MEG has been used to capture the fast dynamics of neural oscillations during language comprehension, perceptual integration, and other cognitive processes [Baillet, 2017]. These electrophysiological techniques can detect rapid neural communication that fMRI (with slower temporal resolution) may overlook. To achieve a more comprehensive understanding of both the timing and localisation of brain activity, EEG or MEG is often combined with fMRI in multimodal neuroimaging studies [Debener et al., 2006]. Meanwhile, PET enables non-invasive visualisation of biochemical processes and molecular targets in the living brain. By using radiolabelled tracers, PET can measure neurotransmitter system activity, receptor binding potential, and metabolic processes during specific cognitive or emotional states. However, due to its limited temporal resolution, PET is better suited for examining sustained mental states rather than capturing rapid, event-related neural responses. Additionally, diffusion tensor imaging (DTI) and related diffusion-based techniques are widely used to map structural brain connectivity by measuring the anisotropic diffusion of water molecules along white matter tracts [Basser et al., 1994]. In cognitive neuroscience, DTI has proven invaluable for exploring how the microstructural integrity of specific white matter tracts (often quantified using metrics such as factional anisotropy (FA)) relates to individual differences in cognitive functions, including executive control and working memory [Madden et al., 2008].

Neuroimaging methods are also essential tools in clinical neuroscience, supporting the diagnosis of neurological disorders, guiding therapeutic interventions, and monitoring disease progression. MRI, with its high anatomical resolution, is routinely used to detect brain tumours, strokes and neurodegenerative atrophy [Wardlaw et al.,

2013b]. In neurosurgical planning, particularly for brain tumours, task-based fMRI is increasingly employed pre-surgically to localise critical functional areas such as the language and motor cortices, relative to the lesion [Petrella et al., 2006, Bizzi et al., 2008]. DTI is another valuable technique: when tumours are located near key white matter pathways, DTI can reveal displacement or infiltration of these fibres, allowing surgeons to plan safer surgical routes that minimise damage to essential tracts [Wedeen et al., 2008, Essayed et al., 2017]. Together, these imaging modalities provide crucial information to guide decisions about whether to operate and how to maximise tumour resection while preserving neurological function.

In clinical psychiatry, neuroimaging is not yet routinely used for diagnosis, but research findings are increasingly being translated into potential biomarkers. For example, PET has been employed to visualise neurotransmitter abnormalities, in psychiatric conditions such as depression or schizophrenia [Howes and Kapur, 2009]. fMRI and EEG are also being explored as tools for identifying biomarkers, including altered functional connectivity in major depressive disorder and distinctive EEG patterns in autism spectrum disorder [Drysdale et al., 2017, Dinstein et al., 2012]. One emerging area is pain assessment, where fMRI patterns have been investigated as objective correlates of pain perception [Woo et al., 2017]. Although such applications remain largely experimental, they demonstrate the growing potential of neuroimaging methods to impact clinical practice and even extend into commercial domains.

Neuroimaging technology continues to advance rapidly, offering higher spatial and temporal resolution, faster acquisition time, and novel capabilities are expanding the boundaries of brain research. High-resolution imaging, in particular, has seen dramatic progress with the development of ultra-high-field MRI scanners (e.g. 7 Tesla and beyond), which provide finer anatomical detail and greater sensitivity to microstructural features [Uğurbil, 2014]. These advancements have enabled more precise mapping of brain morphology, including detailed measurements of cortical thickness, surface area, and gyrification [Fischl, 2012]. In functional imaging, the implementation of multiband echo-planar imaging (EPI) has substantially increased temporal resolution, allowing researchers to capture rapid fluctuations in brain activity and enhance functional connectivity analyses [Smith et al., 2013].

Another major frontier in neuroimaging is the integration of machine learning and AI, which is rapidly transforming data analysis, interpretation and clinical translation. In structural and functional MRI, a wide range of machine learning algorithms, ranging

from classical models such as support vector machines (SVMs) and random forests to more advanced deep learning architectures like convolutional neural networks (CNNs) and graph neural networks (GNNs), have been applied to diverse tasks [Wen et al., 2020, Ktena et al., 2018]. These tasks include brain tissue segmentation, lesion detection, disease classification and the prediction of behavioural and cognitive phenotypes [Lundervold and Lundervold, 2019, Litjens et al., 2017]. By capturing complex spatial and temporal patterns in high-dimensional neuroimaging data, these models can uncover subtle brain-behaviour relationships and facilitate data-driven discovery of imaging biomarkers, and identify early indicators of neurological and psychiatric disorders [Vieira et al., 2017].

2.1.2 Functional Magnetic Resonance Imaging (fMRI)

Functional magnetic resonance imaging (fMRI) is a non-invasive neuroimaging technique that enables the investigation of brain activity by detecting changes in blood oxygenation and cerebral blood flow associated with neural activation. These changes are typically captured through the blood-oxygen-level-dependent (BOLD) signal, which reflects the ratio of oxygenated to deoxygenated hemoglobin, as an indirect indicator of local neuronal activity. fMRI can be applied in both resting-state paradigms, where participants are not engaged in any specific task, and task-based paradigms, in which participants perform cognitive, sensory, or motor tasks designed to elicit specific brain responses [Ogawa et al., 1990, Smith et al., 2009]. In both approaches, fMRI involves the acquisition of a time series of T2*-weighted images, typically collected at repetition times (TRs) ranging from 1 to 3 seconds. This results in a four-dimensional dataset (three spatial dimensions over time) that captures temporal fluctuations in BOLD signal across the brain. By analysing these signal changes, researchers can identify brain regions or large-scale functional networks whose activity is temporally correlated with either externally presented stimuli (in task-based fMRI studies) or spontaneous intrinsic neural activity (in resting-state studies). Such analyses allows the investigation of functional connectivity, revealing coherent patterns of activity across different brain regions. This temporal information is critical for understanding the dynamic organisation of functional brain networks [Poldrack et al., 2011, Power et al., 2011].

fMRI offers relatively high spatial resolution (typically in the range of $2 - 3$ mm isotropic), which allows for precise localisation of brain activity at the level of cortical

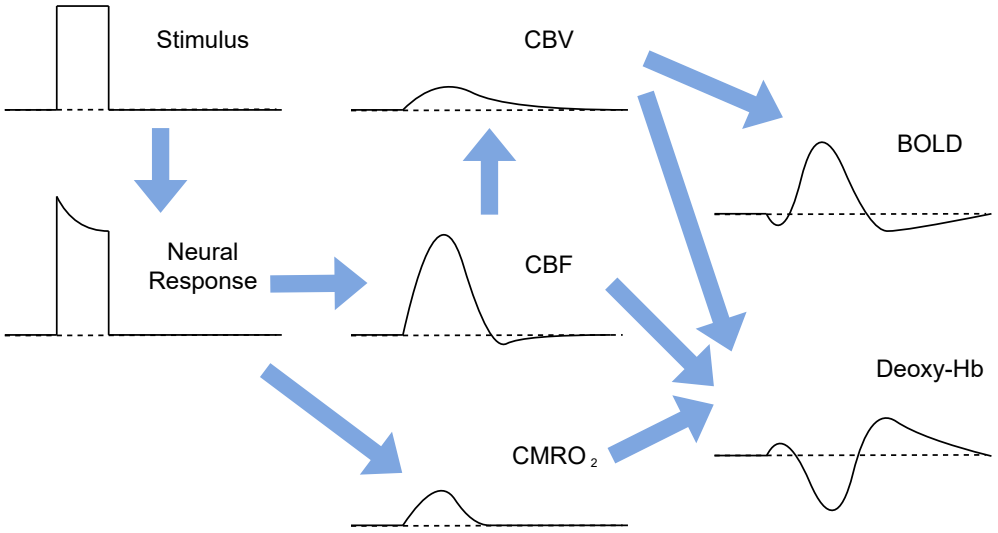


Figure 2.1: Physiological factors influencing the BOLD response. In response to a stimulus, the BOLD signal is modulated by a combination of neuronal activity, cerebral blood flow (CBF), cerebral blood volume (CBV) and the cerebral metabolic rate of oxygen consumption ($CMRO_2$). This illustration is adapted from the models presented in Buxton [2012].

and subcortical structures [Logothetis, 2008]. However, its temporal resolution is inherently constrained by the characteristics of the haemodynamic response function (HRF), as illustrated in Figure 2.1. The BOLD signal reflects vascular responses to underlying neuronal activity, typically peaking around $4-6$ seconds after stimulus onset and returning to baseline within $12-20$ seconds [Glover, 1999]. As a result, fMRI is less suited for capturing rapid or transient neural events compared to electrophysiological methods like EEG or MEG. Despite this limitation, fMRI remains one of the most powerful and widely used neuroimaging tools in cognitive, affective, and clinical neuroscience. It enables non-invasive mapping of brain function and supports investigations into the neural correlates of psychological processes, including perception, attention, memory, language, emotion, and decision-making [Poldrack et al., 2011].

Recent advances in acquisition methods, such as multiband imaging, have significantly improved temporal resolution, enabling whole-brain coverage with shorter repetition times (TRs) [Feinberg and Yacoub, 2012]. In parallel, novel analytical approaches such as functional connectivity analysis, independent component analysis (ICA), and graph-theoretical modelling, have significantly expanded the utility of fMRI

beyond the investigation of localised brain activation. These methods allow for the characterisation of large-scale functional brain networks and facilitate research into how patterns of inter-regional communication are altered in various neurological and psychiatric disorders [Smith et al., 2011, van den Heuvel and Sporns, 2019].

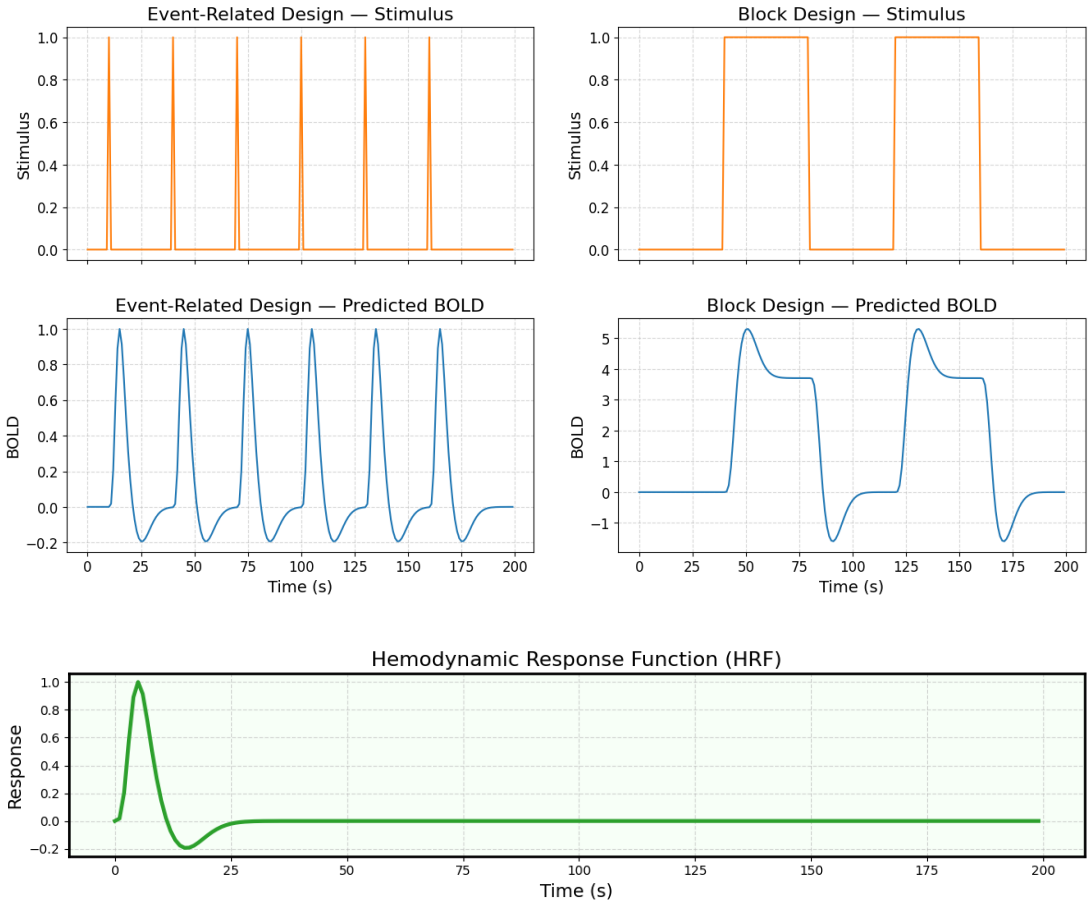


Figure 2.2: Event-related (left) and block (right) designs in task-based fMRI. For each experimental design, the predicted BOLD signal is obtained by convolving the stimulus onset timing (top) with an approximation of the haemodynamic response function (HRF, bottom).

Among the two principal paradigms in fMRI: resting-state fMRI and task-based fMRI, this thesis primarily focuses on task-based fMRI. In task-based fMRI, participants engage in specific cognitive, sensory or motor tasks during image acquisition, enabling the identification of brain regions whose activity is modulated by experimentally controlled conditions [Poldrack et al., 2011]. Experimental designs in task-based fMRI typically follow one of three formats: block designs, even-related designs, or hybrid combinations. As illustrated in Figure 2.2, block designs group similar stimuli or tasks into discrete time intervals (blocks), which improves statistical power by eliciting strong

and sustained BOLD signal changes [Friston et al., 1999]. In contrast, event-related designs present stimuli in a randomised and temporally separated format, which allows for finer temporal resolution and the modelling of transient neural responses [Josephs et al., 1997]. The choice of design depends on the cognitive precesses under investigation, as well as the trade-offs between signal strength, temporal resolution, and flexibility.

Experimental conditions in task-based fMRI are typically modelled using the general linear model (GLM), which estimates the relationship between the observed BOLD signal and a set of predicted haemodynamic responses based on the timing and nature of experimental stimuli [Friston et al., 1994a]. This involves convolving the onset times of each condition with a canonical haemodynamic response function (HRF), which approximates the delayed and dispersed nature of the BOLD response following neural activation. The resulting design matrix allows for voxel-wise statistical inference to identify brain regions whose activity is significantly associated with specific task conditions [Poldrack et al., 2011].

2.1.3 Preprocessing in fMRI

fMRI data require extensive preprocessing prior to statistical analysis to correct for various sources of noise and artefacts, and to standardise the data across subjects and scanning sessions. These preprocessing steps are essential for enhancing data quality, improving the signal-to-noise ratio (SNR) and ensuring accurate spatial alignment of anatomical structures across participates [Poldrack et al., 2011, Esteban et al., 2019]. The most commonly used preprocessing procedures are summarised below:

- **Distortion Correction** is a critical step that compensates for spatial deformations in acquired fMRI images, primarily caused by magnetic field inhomogeneities. These distortions are most pronounced in echo-planar imaging (EPI), and are particularly evident in regions near air-tissue interfaces due to susceptibility-induced magnetic field gradients. Accurate distortion correction improves co-registration between functional and anatomical images, enhances anatomical fidelity and increases the validity of group-level analyses in functional imaging studies [Jezzard and Balaban, 1995, Poldrack et al., 2011].

- **Motion Correction** addresses artefacts introduced by head movement during image acquisition, which can significantly degrade the quality and reliability of fMRI data. Even subtle movements can introduce substantial errors in the BOLD signal. These motion-related artefacts might result in spurious activations, temporal misalignment of voxels, reduced sensitivity in detecting true neural effects, ultimately compromising statistical power and increasing the risk of false positives and false negatives in group-level analyses. Accurate motion correction is therefore essential for valid inference in both task-based and resting-state fMRI studies [Friston et al., 1996, Power et al., 2012, Satterthwaite et al., 2012].

- **Bain Extraction** involves isolating the intracranial brain from non-brain tissues in MRI images. These non-brain structures are generally not relevant for most neuroimaging analyses and, if not removed, can adversely affect downstream preprocessing steps, including spatial normalisation, inter-subject registration, tissue segmentation and functional alignment [Smith, 2002]. Inaccurate brain extraction might lead to distorted estimates of brain morphology, increased inter-subject variability, and compromised statistical inference. To address these challenges, various automated tolls have been developed, including the Brain Extraction Tool (BET) in FSL [Smith, 2002], 3dSkullStrip in AFNI [Cox, 1996], as well as more recent hybrid or deep learning-based approaches such as ROBEX [Iglesias et al., 2011] and DeepBrain [Kleesiek et al., 2016].

- **Spatial Smoothing** involves convolving the imaging data with a spatial filter, most commonly a Gaussian kernel, which replaces the signal intensity at each voxel with a weighted average of its neighbouring voxels. The degree of smoothing is determined by the full-width at half-maximum (FWHM) of the Gaussian kernel, which typically ranges from 4 to 12 mm, depending on the spatial resolution of the data and the size of expected size of activation clusters [Hopfinger et al., 2000, Mikl et al., 2008]. While this process effectively blurs the image, it effectively suppresses high-frequency spatial noise and enhances the sensitivity for detecting spatially extended patterns of brain activation, making statistical inference more robust and reliable.

- **Co-registration** aims to align images acquired from different modalities or sessions into a common anatomical space. In most fMRI studies, this process specifically refers to the alignment of low-resolution, T2*-weighted functional images to high-resolution T1-weighted structural images from the same participant [Ashburner and Friston, 1997].

- **Registration** typically refers inter-subject registration or spatial normalisation, in which each subject's MRI scan is transformed into a common anatomical space or standard brain template (e.g., the MNI152 or Talairach atlas). Registration usually involves estimating a combination of affine and non-linear deformations that more accurately match the individual's brain morphology to the geometry of the reference template [Ashburner and Friston, 1999].

- **Bias Field Correction** addresses low-frequency spatial intensity variations in MRI scans that are unrelated to actual tissue characteristics. These artifacts are especially pronounced at higher magnetic field strengths (e.g., 3T and above). If uncorrected, bias fields can distort tissue contrast, impair tissue segmentation, and reduce the accuracy of subsequent steps such as image registration and spatial normalisation. Bias field correction algorithms aim to estimate and remove this smoothly varying multiplicative field, thereby restoring consistent intensity values for the same tissue type across the image [Sled et al., 1998, Tustison et al., 2010].

- **Slice Timing Correction** compensates for temporal differences in slice acquisition within a single volume. In standard echo-planar imaging (EPI) acquisition, slices are acquired sequentially rather than simultaneously, resulting in temporal offsets across the volume. Depending on the repetition time (TR) and the total number of slices, these offsets can reach several hundred milliseconds. Such time differences can introduce temporal mismatches between the recorded signal and the actual neural events, potentially affecting the accuracy of BOLD responses, particularly in event-related designs where precise temporal alignment is critical [Henson et al., 1999, Sladky et al., 2011].

2.1.4 Structural Magnetic Resonance Imaging (Structural MRI)

Structural Magnetic Resonance Imaging (Structural MRI) is a non-invasive neuroimaging technique that provides high-resolution, three-dimensional representations of brain's anatomical structures. In contrast to fMRI, which captures temporal fluctuations in the BOLD signal associated with neuronal activity, structural MRI provides static images that accurately reflect the brain's morphology with high spatial precision. Structural MRI has been essential in identifying morphological changes related to age-

ing, neurodevelopmental processes, and a wide range of neuropathological conditions, including Alzheimer's disease, multiple sclerosis, and schizophrenia.

Structural MRI includes a variety of imaging modalities designed to visualise different tissue properties within the brain at high spatial resolution. These modalities primarily differ in their sensitivity to tissue-specific magnetic properties, such as proton density and relaxation times (T1, T2 and T2*), which reflect the underlying microstructural composition of neural tissues [McRobbie et al., 2017]. T1-weighted imaging provides excellent contrast between gray and white matter, and is widely employed for anatomical segmentation, volumetric quantification, and surface-based morphometric analyses. In contrast, T2-weighted and FLAIR (Fluid-Attenuated Inversion Recovery) sequences are particular sensitive to fluid content and pathological changes such as inflammation, demyelination and edema, making them especially valuable for detecting white matter lesions and cerebrovascular abnormalities [Filippi et al., 2016, Wardlaw et al., 2013a]. The choice of imaging modality and sequence parameters thus depends on the specific anatomical structures or pathological processes of interest, enabling a comprehensive assessment of brain structures in both clinical diagnostics and neuroscience research.

While structural MRI is a powerful and widely used tool for visualising brain anatomy, it also has several notable limitations. One major constraint is its limited sensitivity to subtle or early-stage pathological changes, particularly those do not produce overt morphological alternations. Microstructural abnormalities, early neurodegenerative processes, or mild inflammatory responses may remain undetected on conventional T1- or T2-weighted scans until significant tissue degeneration has occurred [Jack et al., 2010]. Additionally, structural MRI is susceptible to various artifacts that can compromise image quality and the interpretability of morphometric data. One of the most prevalent source of artifact is head motion during acquisition, which can introduce blurring, ghosting, and spatial distortion, particularly during longer scan sessions. [Reuter et al., 2015]. Even subtle motion can systematically bias estimates of cortical thickness, gray matter volume and other structural metrics. Moreover, magnetic field inhomogeneities, especially near air–tissue interfaces can lead to signal dropouts and geometric distortions, further limiting the reliability of structural measurements in these regions [Jezzard and Balaban, 1995]. These challenges highlight the importance of optimised acquisition protocols, effective motion correction techniques, and rigorous quality control procedures in structural MRI research.

2.1.5 Large-scale Neuroimaging Datasets

Over the past few decades, advances in brain imaging have significantly deepened our understanding of human brain structure and function. However, many early neuroimaging findings were constrained by small sample sizes, limited statistical power, and a lack of population diversity, which are factors that known to undermine the reliability and reproducibility of findings [Button et al., 2013, Marek et al., 2022]. These methodological limitations have raised critical concerns about the generalisability and robustness of neuroimaging-based inferences.

In response, the field has increasingly moved toward more rigorous and reproducible scientific practices, large-scale and demographically diverse studies are increasingly recognised as essential for generating reliable and generalisable insights. Notable initiatives include Neurosynth, which provides large-scale, automated meta-analytic synthesis of published fMRI studies, providing a data-driven framework for synthesising activation patterns across the literature [Yarkoni et al., 2011]; The UK Biobank, a population-based imaging cohort with over $100,000$ participants, associating brain imaging with extensive genetic, behavioural and health data [Miller et al., 2016]; The Human Connectome Project (HCP), which offers high-resolution, multi-modal imaging and behavioural assessments from a well-characterised sample of healthy adults [Van Essen et al., 2013]. These resources have become foundational to the emerging field of population-level neuroscience, enabling more robust and replicable inference and discovery.

2.1.5.1 Neurosynth

Neurosynth is a large-scale, publicly available meta-analytic dataset that systematically links psychological concepts to patterns of brain activation by automatically extracting and aggregating data from the fMRI literature. Developed by Yarkoni et al. [2011], Neurosynth dataset consists of activation coordinates and associated cognitive terms parsed from thousands of published fMRI studies using text mining and natural language processing techniques. It enables the generation of statistical maps through both forward inference (estimating the likelihood of activation given a cognitive term) and reverse inference (estimating the likelihood of a cognitive process given observed brain activation), supporting robust identification of consistent structure–function associations across the literature [Yarkoni et al., 2011]. By making

large-scale activation data openly accessible in a standardised format, Neurosynth facilitates transparency, reproducibility, and scalability in cognitive neuroscience. It also addresses methodological challenges such as analytic flexibility, publication bias and limited statistical power in individual studies [Poldrack et al., 2013].

2.1.5.2 UK Biobank

The UK Biobank is a large-scale, prospective epidemiological cohort study designed to investigate the genetic, environmental and lifestyle factors that influence human health and disease. It includes approximately $500,000$ participants, aged 40 and 69 at the time of recruitment ($2006 - 2010$), the majority of whom are of white British ancestry [Bycroft et al., 2018]. Although the cohort is not fully representative of the general UK population (healthier, older, and less ethnically diverse), it provides an unprecedented resource for population-level research, due to its extensive breadth and depth of phenotypic, genotypic, and imaging data [Fry et al., 2017].

An imaging extension of the UK Biobank, launched in 2014, aims to collect multimodal brain imaging data, including structural MRI, diffusion MRI and resting-state fMRI, from up to $100,000$ participants. This unprecedented large-scale initiative facilitates in-depth investigations into the neural correlates of ageing, disease susceptibility, and brain–behaviour relationships across a broad population [Miller et al., 2016]. The richness, standardisation and open accessibility of the UK Biobank dataset have made it a cornerstone of modern neuroimaging research, particularly for studies focused on improving reproducibility, statistical power and generalisability in brain-wide association and population neuroscience.

In the real data application presented in Chapter 5, we utilise the brain lesion masks derived from the T2-weighted FLAIR MRI scans from the UK Biobank to investigate associations between lesion incidence and potential risk factors, such as ageing and CVR factors.

2.2 Statistical Modelling in Neuroimaging

In this section, we present generalised linear models (GLMs), which constitute one of the most widely used statistical modelling frameworks in neuroimaging. We begin

by introducing the fundamental definition and structure of GLMs in Section 2.2.2. This is followed by a discussion of key estimation techniques, including maximum likelihood estimation and score equations, in Section 2.2.3. Finally, Section 2.3 addresses optimisation and computational consideration relevant to the application of GLMs.

2.2.1 Generalised Linear models (GLMs)

We first start with standard linear models and then extend to generalised linear models (GLMs). In its classical form, the linear model assumes that a continuous response variable $Y_i \in \mathbb{R}$ where i indexes observations (samples), and is linearly related to a set of corresponding explanatory variables $X_i = [x_{i1}, x_{i2}, \ldots, x_{ip}] \in \mathbb{R}^p$ through a linear predictor $\eta_i = X_i^\top \beta$, where $\beta = [\beta_1, \ldots, \beta_p] \in \mathbb{R}^p$ denotes a vector of unknown coefficients. The model is expressed as

$$Y_i = X_i^\top \beta + \varepsilon_i, \quad \varepsilon_i \sim \mathcal{N}(0, \sigma^2), \tag{2.1}$$

where the errors ε_i are assumed to be independent and identically distributed (i.i.d.) Gaussian random variables with constant variance σ^2. This formulation implies that the conditional expectation of Y_i given X_i is $\mu_i = \mathbb{E}[Y_i \mid X_i] = X_i^\top \beta$, and the variance is constant across all observations [Montgomery et al., 2021, Seber and Lee, 2003]. While the linear model is computationally efficient and analytically tractable, admitting closed-form solutions via ordinary least squares (OLS), its applicability is limited to scenarios where the response is continuous, and approximately normally distributed with constant variance.

However, in many real-world problems, responses are binary, count-valued, or otherwise non-normal. These limitations motivate the extension to generalised linear models (GLMs), which preserve the linear structure of the predictor but relax the assumptions on the distribution of Y_i and the relationship between the expected response μ_i and the linear predictor η_i, thereby accommodating a broader class of responses through appropriate link functions and maximum likelihood estimation. Unlike linear models, which assume that the response variable is continuous and normally distributed with constant variance, GLMs extend this framework by allowing the response variable to follow any distribution from the exponential family (to be described). This flexibility makes GLMs suitable for modelling a wide range of outcome types, including binary outcomes, count data, and positively skewed continuous outcomes, among others.

Additionally, GLMs relax the assumption of a direct (identity) relationship between the expected response and the linear predictor by introducing a link function, which transforms the expected response to ensure that predictions remain within appropriate ranges (e.g., probabilities between 0 and 1 for binary outcomes), thereby enabling non-linear relationships to be modelled within a linear modelling framework [McCullagh, 2019, Dobson and Barnett, 2018].

GLMs have three main components:

1. **Stochastic Component (Exponential Family Distribution):**
   Rather than assuming the observed response Y_i is always Gaussian, we extend it to a probability distribution from the exponential family, a broad class of distributions that include many commonly encountered in practice. A univariate distribution belongs to the exponential family if it can be written in the canonical form:
   $$f(y \mid \theta) = h(y) \exp\left(\eta(\theta) \cdot T(y) - A(\theta)\right),$$
   where θ is the canonical (natural) parameter, and $h(y), \eta(\theta), T(y), A(\theta)$ are known functions that define the distribution, where:

   - $\eta(\theta)$ is the **natural parameter**,
   - $T(y)$ is the **sufficient statistic**,
   - $A(\theta)$ is the **log-partition function**,
   - $h(y)$ is the **base measure** (always non-negative).

   The exponential family is in its canonical form if $\eta(\theta) = \theta$ which is always possible by letting the transformed parameters $\eta(\theta)$ to be the parameters. Many well-known distributions can be seen as special cases of the exponential family and we give a few examples below:

   Bernoulli Distribution

   Original form:
   $$f(y \mid p) = p^y (1-p)^{1-y}, \quad y \in \{0, 1\}$$

   Exponential family form:
   $$f(y \mid \theta) = \exp\left(y\theta - \log(1 + e^\theta)\right), \quad \theta = \log\left(\frac{p}{1-p}\right)$$

   Components:

- $\eta(\theta) = \theta$
- $T(y) = y$
- $A(\theta) = \log(1 + e^{\theta})$
- $h(y) = 1$

Normal Distribution (Known Variance σ^2)

Original form:

$$f(y \mid \mu) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y-\mu)^2}{2\sigma^2}\right)$$

Exponential family form:

$$f(y \mid \mu) = \exp\left(-\frac{y^2}{2\sigma^2}\right) / \sqrt{2\pi\sigma^2} \exp\left(\frac{\mu}{\sigma^2} y - \frac{\mu^2}{2\sigma^2}\right)$$

Components:

- $\eta(\mu) = \frac{\mu}{\sigma^2}$
- $T(y) = y$
- $A(\mu) = \frac{\mu^2}{2\sigma^2}$
- $h(y) = \exp\left(-\frac{y^2}{2\sigma^2}\right) / \sqrt{2\pi\sigma^2}$

Poisson Distribution

Original form:

$$f(y \mid \lambda) = \frac{e^{-\lambda}\lambda^y}{y!}, \quad y \in \mathbb{N}$$

Exponential family form:

$$f(y \mid \theta) = \frac{1}{y!} \exp\left(y\theta - e^{\theta}\right), \quad \theta = \log \lambda$$

Components:

- $\eta(\theta) = \theta$
- $T(y) = y$
- $A(\theta) = e^{\theta}$
- $h(y) = \frac{1}{y!}$

2. **Deterministic Component (Linear Predictor):**
   As in standard linear regression, the effect of covariates is modelled through a linear combination. Let $X_i = [X_{i1}, X_{i2}, \ldots, X_{ip}]$ be the vector of covariates for the i-th observation. The linear predictor η_i is defined as:

$$\eta_i = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_p X_{ip},$$

   where $\beta = [\beta_0, \beta_1, \ldots, \beta_p]$ is the vector of regression coefficients. Note that it is equivalent to insert 1 as the first dimension of X_i and let $X_i = [1, X_{i1}, X_{i2}, \ldots, X_{ip}]$ so we can write:

$$\eta_i = X_i \beta^\top$$

   This linear form provides interpretability and tractability, and it remains central to the GLM structure, even when the response is not linearly related to the predictors.

3. **Link Function:**
   To connect the linear predictor η_i with the mean of the response variable $\mu_i = \mathbb{E}[Y_i]$, a link function $g(\cdot)$ is introduced:

$$g(\mu_i) = \eta_i = X_i^\top \beta.$$

   The link function must be monotonic and differentiable, and it ensures that the domain of μ_i matches the range of the linear predictor η_i. The inverse link function, $g^{-1}(\eta)$, maps the linear predictor back to the mean of the distribution.

   A special case is the canonical link function, defined such that $\eta_i = \theta_i$, i.e., the linear predictor equals the canonical parameter of the exponential family distribution. Using the canonical link often simplifies the mathematical form of the likelihood and the estimation process.

   However, in neuroimaging applications, the use of the canonical link function is not always required, and can sometimes be suboptimal. Researchers often choose alternative link functions that provide more interpretable model outputs, especially when the objective involves communicating effect sizes or latent processes. For example, while the canonical logit link is standard in logistic regression with a Bernoulli response, the probit link is sometimes preferred when the interpretation aligns with an underlying latent Gaussian process, as is common in hierarchical or Bayesian models [Albert and Chib, 1993, McCullagh, 2019]. Moreover, canonical links do not always provide the best empirical fit for

the relationship between predictors and brain responses. In functional MRI or lesion mapping studies, activation intensity or lesion presence may exhibit non-linear associations with variables such as age, clinical risk factors, or cognitive performance. In such scenarios, alternative link functions (e.g., identity, log or complementary log-log links) can provide better model fit and improved predictive performance [Kim et al., 2021, Datta et al., 2020]. Non-canonical link functions can also improve numerical stability and convergence behaviour in iterative optimisation procedures. This advantage becomes particularly relevant when using regularisation techniques or approximate Bayesian inference, where canonical link assumptions can introduce estimation instability or overly aggressive shrinkage [Woolrich et al., 2009, Carpenter et al., 2017]. Furthermore, many advanced neuroimaging models, especially those incorporating spatial priors, hierarchical structures or latent factor representations, derive from the classical GLM framework. In these contexts, the flexibility to choose a non-canonical link function enables better alignment with the model structure and improved compatibility with inference procedures such as Markov chain Monte Carlo (MCMC) or variational methods [Penny et al., 2005, Montagna et al., 2018].

2.2.2 Estimation Methods

GLMs are estimated using methods that extend the classical least squares framework of linear models to accommodate non-Gaussian responses and non-linear link functions. Parameters are typically estimated via maximum likelihood estimation (MLE), which generally lacks closed-form solutions due to the non-linearity introduced by the link function and the structure of the likelihood function. As a result, estimation relies on iterative numerical techniques, most notably the iteratively re-weighted least squares (IRLS) algorithm. IRLS can be interpreted as a special case of the Fisher scoring method, an optimisation approach that substitutes the expected information matrix for the observed one, which improves numerical stability and convergence efficiency [McCullagh, 2019]. These iterative procedures are fundamental to the practical implementation of GLMs in statistical software. In this section, we illustrate each of these estimation methods in detail.

1. **Maximum likelihood estimation (MLE):**

Assume we have n independent observations (X_i, Y_i), for $i = 1, \ldots, n$, modelled using a generalised linear model (GLM). Under this model, each response Y_i follows a distribution from the exponential family. The parameter vector β of the GLM is commonly estimated through maximum likelihood estimation (MLE). The joint likelihood function for all observations is expressed as the product of their individual probability density (or mass) functions:

$$L(\beta) = \prod_{i=1}^{n} f(y_i \mid \beta) = \prod_{i=1}^{n} \left[h(y_i) \exp\left(\eta_i T(y_i) - A(\eta_i)\right) \right], \tag{2.2}$$

where $\eta_i = X_i^\top \beta$ is the linear predictor, and the functions $h(y_i), T(y_i), A(\eta_i)$ are defined by the chosen exponential family distribution.

For numerical stability, we typically maximise the log-likelihood function, which is defined as:

$$l(\beta) = \log L(\beta) = \sum_{i=1}^{n} \left[\eta_i T(y_i) - A(\eta_i) + \log h(y_i) \right]. \tag{2.3}$$

2. **Fisher Scoring and Iteratively Re-weighted Least Squares (IRLS):**
We begin by introducing the score equations. To obtain the maximum likelihood estimator $\hat{\beta}$, we differentiate the log-likelihood function with respect to β and set the resulting expressions equal to zero. These equations, known as the *score equations*, are given by:

$$\frac{\partial l(\beta)}{\partial \beta} = \sum_{i=1}^{n} \left[T(y_i) - \frac{\partial A(\eta_i)}{\partial \eta_i} \right] \frac{\partial \eta_i}{\partial \beta} = \sum_{i=1}^{n} \left[T(y_i) - \mathbb{E}[T(Y_i)] \right] \frac{\partial \eta_i}{\partial \beta} = 0.$$

Since $\eta_i = X_i^\top \beta$, we have:

$$\frac{\partial \eta_i}{\partial \beta} = X_i,$$

simplifying the score equations to:

$$\sum_{i=1}^{n} \left[T(y_i) - \mathbb{E}[T(Y_i)] \right] X_i = 0.$$

For most GLMs, the sufficient statistic is simply $T(y_i) = y_i$, and thus the score equations further simplify to:

$$\sum_{i=1}^{n}(y_i - \mu_i)X_i = 0,$$

where $\mu_i = \mathbb{E}[Y_i]$ is linked to β via the inverse of the link function $g^{-1}(\eta_i)$.

Typically, the score equations do not have closed-form solutions. Thus, iterative numerical methods such as Newton-Raphson method or Fisher scoring are employed to approximate solutions. Under the terminology of the Newton-Raphson method, the score vector $U(\beta) = \nabla l(\beta)$ is the gradient of the log-likelihood, and the Hessian matrix $H(\beta) = \nabla^2 l(\beta)$ is the matrix of second derivatives. Fisher scoring replaces the observed Hessian with its expectation, known as the Fisher information matrix. This leads to the update rule:

$$\beta^{(m+1)} = \beta^{(m)} + \left[I(\beta^{(m)})\right]^{-1} U(\beta^{(m)}),$$

While the iteratively re-weighted least squares (IRLS), a practical method for estimating parameters in GLMs, solves a series of weighted least squares problem of the form

$$\beta^{(m+1)} = (X^\top W^{(m)} X)^{-1} X^\top W^{(m)} z^{(m)},$$

where $W^{(m)}$ is a diagonal weight matrix at iteration m, depending on the current estimates. The diagonal entries of the weight matrix are given by:

$$w_{ii} = \frac{1}{[g'(\mu_i)]^2 \operatorname{Var}(Y_i)}.$$

and the working response $z^{(m)}$ is a first-order Taylor approximation of the inverse link function around the current estimate. It transforms the non-linear GLM problem into a locally linear regression:

$$z^{(m)} = \eta^{(m)} + \frac{y - \mu^{(m)}}{g'(\mu^{(m)})}$$

This procedure is implemented algorithmically through the Iteratively Re-weighted Least Squares (IRLS) method:

Algorithm 1 Iteratively Re-weighted Least Squares (IRLS)

1: **Initialize** $\beta^{(0)}$
2: **for** $m = 0, 1, 2, \ldots$ *until convergence* **do**
3: Compute linear predictor: $\eta_i^{(m)} = X_i^\top \beta^{(m)}$
4: Compute mean: $\mu_i^{(m)} = g^{-1}(\eta_i^{(m)})$
5: Compute weights:

$$w_{ii}^{(m)} = \frac{1}{\left[g'(\mu_i^{(m)})\right]^2 \cdot \mathrm{Var}(Y_i)}$$

6: Compute adjusted response:

$$z_i^{(m)} = \eta_i^{(m)} + \frac{y_i - \mu_i^{(m)}}{g'(\mu_i^{(m)})}$$

7: Update parameters:

$$\beta^{(m+1)} = \left(X^\top W^{(m)} X\right)^{-1} X^\top W^{(m)} z^{(m)}$$

The IRLS algorithm is equivalent to Fisher scoring when the link function is canonical (e.g., the logit link for Bernoulli responses or the log link for Poisson responses), and the weights used in IRLS match the Fisher information structure.

2.2.3 Optimisation and Computation

In addition to IRLS, the limited-memory Briyden-Fletcher-Goldfarb-Shanno (L-BFGS) algorithm is a widely used quasi-Newton optimisation method for fitting GLMs, particularly in high-dimensional or regularised settings. Unlike IRLS, which relies on repeated matrix inversions, L-BFGS approximates the inverse Hessian matrix using a limited amount of memory and gradient information from previous iterations [Byrd et al., 1995]. This approach significantly reduces computational and memory burdens, making it well-suited for large-scale problems where explicit computation of the full Hessian is infeasible or numerically unstable. This method is especially effective in penalised GLMs, such as those involving L_1 (lasso) or L_2 (ridge) regularisation, where the optimisation landscape may be non-smooth or requires efficient gradient-based optimisation [Ng, 2004]. Due to its scalability and robustness, L-BFGS has been widely adopted in modern statistical and machine learning toolkits, including scikit-learn, PyTorch, TensorFlow and R.

As neuroimaging datasets continue to expand in both size and complexity, driven by advances in high-resolution imaging, large-scale cohort studies and multi-modal data integration, the computational demands of fitting GLMs have increased substantially. This challenges is particularly evident in mass-univariate analysis, where a separate GLM is fitted to each voxel or region, resulting in thousands to millions of individual model fits. Therefore, parallelisation has emerged as a key strategy for scaling GLM estimation to large datasets. Since voxelwise or subject-level GLMs are independent, they can be evaluated in parallel across multiple CPUs or GPUs, thereby optimising the efficiency of computational resources. This approach significantly reduces runtime and also facilitates the adoption of more advanced statistical models, such as regularised GLMs and Bayesian hierarchical models, which introduce additional computational complexity.

2.3 Neuroimaging Statistics

In this section, we describe how statistical inference methods are applied to analyse the regression coefficients obtained in Section 2.2. First, in Section 2.3.1, we introduce global tests used to assess overall model fitness. Following this, in Section 2.3.2, we outline the localised inference approaches for evaluating the significance of spatial homogeneity or group differences at the voxel level, as well as using the inverse of the Fisher information matrix to estimate standard errors. Next, in Section 2.3.3, we describe an alternative method, the robust variance estimator (sandwich estimator), which provides a more accurate and reliable standard error estimates, particularly under model misspecification. Afterwards, Section 2.3.4 addresses the issue of multiple comparison correction, especially in the context of voxel-level statistical inference in neuroimaging data. Finally, we discuss the use of resampling and bootstrapping techniques in situations where parametric statistical tests might not provide reliable inference.

2.3.1 Global Tests of Model Fitness

In neuroimaging statistics, global tests of model fitness are used to evaluate whether a statistical model sufficiently captures the overall structure of the observed data across the entire brain or within a predefined region of interest (ROI). These tests evaluate

whether the residuals or unexplained variance are consistent with the assumptions of the model, thereby providing a measure of model adequacy. Unlike voxel-wise tests, which evaluate local effects at individual voxels, global tests are sensitive to widespread deviations from model assumptions, such as spatial correlation misspecification or incorrect error distributions. Common approaches include likelihood ratio tests, deviance statistics, and information criteria such as the Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) [Akaike, 1998, Schwarz, 1978]. In neuroimaging, the General Linear Model (GLM) often employs global tests to validate the overall adequacy of the design matrix and the specification of the noise model [Friston et al., 1994a]. Moreover, permutation-based global inference methods have been developed to address the complex spatial dependencies in neuroimaging data [Nichols and Holmes, 2002]. These approaches do not depend on parametric assumptions regarding the distribution of test statistics, making them especially useful when standard assumptions are violated [Winkler et al., 2014].

The likelihood ratio test (LRT) is a classical method for comparing the goodness-of-fit of two nested model: a simpler (reduced) model that is a special case of a more complex (full) model. Let $\mathcal{L}_0$ and $\mathcal{L}_1$ denote the maximum likelihoods under the reduced and full models, respectively. The LRT statistics is defined as:

$$\Lambda = -2\log\left(\frac{\mathcal{L}_0}{\mathcal{L}_1}\right) = -2[\log\mathcal{L}_0 - \log\mathcal{L}_1] \tag{2.4}$$

which, under standard regularity conditions and under the null hypothesis H_0 that the reduced model is correct, asymptotically follows a chi-square distribution:

$$\Lambda \sim \chi^2_k \tag{2.5}$$

where k is the difference in the number of parameters between the full and reduced models [Wilks, 1938].

In Generalised Linear Models (GLMs), the LRT is widely used to evaluate whether the inclusion of additional predictors significantly improves model fit. For example, in neuroimaging applications, the LRT can be employed to test whether the addition of a clinical covariate, such as diagnosis, age or interaction term, significantly improves the explanation of voxel-wise data variation [Friston et al., 1994a]. Beyond evaluating predictor inclusion, the LRT is also useful for assessing distributional assumptions within the GLM framework. For example, when modelling count data in neuroimaging (e.g., lesion count or event-related spike data), the LRT can be used to compare a

Poisson model to a more flexible Negative Binomial model, which accommodates overdispersion. A significant LRT result in this context indicates that the Negative Binomial distribution provides a better fit, highlighting the importance of accurately capturing variance structure in the data [Cameron and Trivedi, 2013].

While comparing non-nested models or models with different levels of complexity, information criteria, such as Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) are commonly employed as quantitative tools for model selection. These criteria balance model fit and complexity by introducing a penalty term for the number of parameters, thereby reducing the risk of overfitting. In neuroimaging, careful model selection is particularly critical for accurately characterising underlying brain activity or structural features while avoiding overfitting.

Given a statistical model with likelihood $\mathcal{L}$, number of estimated parameters k, and sample size n, AIC and BIC are defined as:

$$\begin{aligned} \text{AIC} &= -2\log\mathcal{L} + 2k, \\ \text{BIC} &= -2\log\mathcal{L} + k\log n. \end{aligned} \tag{2.6}$$

The AIC, derived from information theory, aims to identify the model that minimises the expected Kullback-Leibler (KL) divergence between the true data-generating process and the fitted model [Akaike, 1974]. In contrast, the BIC, grounded in Bayesian principles, approximates the Bayes factor under certain regularity conditions. As the sample size n increases, BIC generally favours more parsimonious models, reflecting its stronger penalty for model complexity [Schwarz, 1978].

In neuroimaging, these information criteria are applied across a wide range of applications, including the selection of explanatory variables in generalised linear models (GLMs) [Friston et al., 1994a], the comparison of alternative spatial priors in Bayesian hierarchical models [Penny et al., 2005], and the evaluation of distributional assumptions in count data models (e.g., comparing Poisson vs. Negative Binomial models) [Cameron and Trivedi, 2013, Yu et al., 2024]. While these criteria are often computed at the voxel level, caution is required when interpreting results in the presence of spatial correlation, as standard assumptions of independent observations may be violated.

2.3.2 Localised Inference Using Chi-square Tests

In neuroimaging, localised inference refers to statistical testing conducted at the voxel level, aiming to identify brain regions where the observed signal significantly deviates from that expected under a specified null hypothesis. This voxel-wise analytic framework is fundamental to both functional and structural neuroimaging studies, allowing the detection of spatially specific patterns of brain activation or morphological variation at high resolution [Friston et al., 1994a, Worsley et al., 1996]. Typically, the null hypothesis assumes no effect or no difference in signal intensity across experimental conditions or subject groups, and voxel-level test statistics are used to evaluate deviations from this assumption. Given that neuroimaging datasets often consist of hundreds of thousands of voxels, localised inference must rigorously address the multiple comparisons problem to ensure valid statistical inference. Accordingly, correction procedures are applied to control for either the family-wise error rate (FWER) or the false discovery rate (FDR), as discussed in more details in Section 2.3.4.

We propose that localised inference can be formulated in a general and flexible statistical framework, capable of accommodating a wide range of group-level comparisons and model structures. Specifically, let $C \in \mathbb{R}^{m \times S}$ be a contrast matrix designed to test linear hypotheses involving S groups. At each voxel j, the null hypothesis can be expressed as,

$$H_0 : C\hat{\theta}_j = \mathbf{0}_{m \times 1}$$

where $\hat{\theta}_j$ denotes the estimated model parameters at voxel j, typically derived from a generalised linear model (GLM) fitted independently at each spatial location [Friston et al., 1994a]. The contrast matrix C defines the specific hypothesis being tested, such as group differences, condition effects or interaction terms. To test the null hypothesis, a voxel-wise test statistic is given by

$$(C\hat{\theta}_j)^\top (CV_jC^\top)^{-1}(C\hat{\theta}_j) \xrightarrow{D} \chi^2_m \tag{2.7}$$

where V_j denotes the estimated covariance matrix of $\hat{\theta}_j$, and m is the rank of the contrast matrix C. Under standard regularity conditions and assuming that $\hat{\theta}_j$ is asymptotically normally distributed, this test statistic follows a chi-square distribution with m degrees of freedom [Rao et al., 1973]. Such quadratic forms are widely used in neuroimaging for multi-group or multivariate inference, especially in analyses involving complex experimental designs or hierarchical modelling frameworks [Friston, 2002].

In the special case where only a single model parameter is tested against a null hypothesis of homogeneity, Equation 2.7 simplifies to the classical Wald test:

$$Z = \frac{\hat{\theta}_j - \theta_0}{SE(\hat{\theta}_j)} \xrightarrow{D} \mathcal{N}(0,1) \tag{2.8}$$

where $SE(\hat{\theta}_j)$ denotes the standard error of the estimator $\hat{\theta}_j$, and θ_0 is the value of the parameter under the null hypothesis. This univariate form of localised inference is widely employed in voxel-wise analyses to assess the significance of individual effects, typically by comparing estimated activation or contrast values against zero (i.e., no effect) or an alternative baseline. It is particularly useful in mass-univariate framework due to its computational efficiency and straightforward interpretation.

In the GLM commonly used in neuroimaging analyses, a standard approach for estimating the covariance matrix of the model parameters involved in Equation 2.7 is to invert the observed Fisher information matrix. The Fisher information quantifies the amount of information that the observed data contain about an unknown parameter on which the likelihood function depends. Under standard regularity conditions, the inverse of the Fisher information provides an asymptotically consistent estimator of the covariance matrix of the maximum likelihood estimates (MLEs) [Casella and Berger, 2002, Pawitan, 2001]. Specifically, for a parameter vector θ the covariance matrix $Cov(\hat{\theta}_j)$ is typically approximated by $I(\hat{\theta})^{-1}$, where $I(\hat{\theta})$ denotes the observed Fisher information evaluated at the MLE. This approach is particularly advantageous in large-sample settings, where the asymptotic normality of the MLE ensures valid inference. In neuroimaging studies, where thousands of voxel-wise hypotheses are tested simultaneously, accurate estimator of the parameter covariance structure is essential for valid statistical inference. It directly affects the type I error rates and the reliability of test statistics [Efron, 2010].

2.3.3 Robust Variance Estimation (Sandwich Estimator)

However, under model misspecification, such as when the assumed likelihood does not accurately represent the data-generating process, using the inverse of the Fisher information matrix to estimate the covariance of parameter estimates can result in biased and inconsistent inference. This is because the Fisher information relies on correct model specification, including assumptions of homoscedasticity, independence, and the correct functional form of the model. When these assumptions are violated, standard

(model-based) variance estimators typically underestimate the true variability of the parameter estimates, which can result in inflated type I error rates and overconfident inferences, thereby compromising the validity the reliability of statistical conclusions [White, 1982].

To address the limitations of model-based variance estimators under misspecification, robust variance estimation, commonly known as the sandwich estimator, provides a consistent and asymptotically valid alternative. Unlike traditional methods that rely on correct model misspecification of the full likelihood, the sandwich estimator requires only that the estimating equations are unbiased. It accommodates violations of standard assumptions such as heteroscedasticity and within-cluster correlation by using the empirical variability of the residuals to adjust the estimated covariance matrix of the parameter estimates [White, 1980, Huber, 1967]. This robustness makes the sandwich estimator particularly useful in settings involving clustered, longitudinal, or otherwise dependent data structures, where conventional variance estimators often underestimate uncertainty. It is widely applied in generalised estimating equations (GEEs), generalised linear models (GLMs), and semi-parametric regression frameworks, providing a principled approach to ensure valid statistical inference even when certain model assumptions are violated [Liang and Zeger, 1986, Zhou et al., 2020].

2.3.4 Multiple Comparison Correction

In a typical fMRI analysis, statistical inference is performed independently at each voxel, often involving approximately 2.2×10^5 voxels across the entire brain. Under the mass-univariate approach, this corresponds to conducting 2.2×10^5 hypothesis tests simultaneously (one per voxel). If the global null hypothesis were true, and the statistical maps were thresholded using a conventional uncorrected significance level of $\alpha = 0.05$, then by chance alone, one would expect approximately

$$0.05 \times 2.2 \times 10^5 = 1.1 \times 10^4$$

voxels to be incorrectly identified as significant. These false positives arise purely due to random noise, in the absence of any true underlying signal.

This issue is referred as the multiple comparison problem, where the likelihood of observing false positives increases with the number of statistical tests performed simultaneously. In neuroimaging, this issue is particularly pronounced due to the

high spatial resolution of fMRI data. Without appropriate correction for multiple comparisons, statistical maps may misleadingly suggest widespread brain activation, even in the absence of true effects.

To control this inflation of false positives, a wide variety of multiple comparison correction methods have been developed. These methods aim to maintain statistical validity across the large number of voxel-wise tests typically performed in fMRI and other neuroimaging modalities. One class of methods focuses on controlling the Family-Wise Error Rate (FWER), which is defined as the probability of making at least one false positive across all conducted tests. A classical example is the Bonferroni correction, which adjusts the significance threshold by dividing the desired α level by the number of tests (i.e., $\alpha_{adj} = \alpha/m$). While simple and stringent, Bonferroni correction is often overly conservative in neuroimaging due to the spatial correlation among voxels. More advanced approaches based on Random Field Theory (RFT) account for the smoothness and spatial structure of brain images. RFT estimates the expected Euler characteristic (EC) of the excursion set to approximate the distribution of the maximum statistic under the null hypothesis [Friston et al., 1994b, Worsley et al., 1996, Nichols and Hayasaka, 2003]. These methods allow for cluster-level and peak-level inference while preserving strong control over FWER, offering improved sensitivity over traditional methods in the analysis of spatially correlated neuroimaging data.

In contrast, False Discovery Rate (FDR)-based methods aim to control the expected proportion of false positives among the set of rejected hypotheses. Introduced by Benjamini and Hochberg [1995], FDR procedures provide a less stringent but more powerful alternative to FWER control, especially in scenarios where many true effects are expected. FDR is particularly useful in exploratory neuroimaging analyses, where the emphasis is on identifying a broad set of potentially meaningful voxels while tolerating a controlled proportion of false positives. However, standard FDR methods can be sensitive to the dependency structure among tests and may be less robust in settings with strong spatial correlation, unless appropriately adjusted to account for such dependencies [Genovese et al., 2002].

In summary, FWER control is generally preferred in clinical applications, where minimising false positives is essential to ensure reliability and reproducibility of findings. Such stringent control is particular important when statistical results may inform clinical decisions or guide follow-up research. In contrast, FDR control is more

appropriate for exploratory analyses, where the primary goal is to detect as many true effects as possible, even at the cost of accepting a limited proportion of false positives. FDR-based methods allow researchers to balance sensitivity and specificity in high-dimensional settings, making them especially useful for generating hypotheses and for identifying spatially distributed activation patterns across the brain [Benjamini and Hochberg, 1995, Genovese et al., 2002].

Additionally, non-parametric permutation tests have become increasingly popular in neuroimaging due to their flexibility, robustness and minimal reliance on distributional assumptions. Unlike traditional parametric methods, which often rely on assumptions such as multivariate normality and spatial smoothness, permutation-based inference constructs the empirical null distribution of the test statistics by randomly shuffling data labels or experimental conditions. This resampling approach provides exact control of the family-wise error rate (FWER) under the null hypothesis, regardless of the underlying data distribution [Nichols and Holmes, 2002]. Permutation tests are particularly well-suited to the complex or often non-standard data structures encountered in neuroimaging, as well as high-dimensional and spatially correlated nature of fMRI data. Because they do not rely on parametric assumptions, they inherently account for spatial dependencies within the data. Moreover, permutation methods are highly adaptable, supporting a wide range of experimental designs, including one-sample and two-sample tests, regression models and repeated measures designs. Methods such as Threshold-Free Cluster Enhancement (TFCE) [Smith and Nichols, 2009] and maximal statistic correction across space further increase their statistical sensitivity while maintaining strict control over type I error. These approaches enhance signal detection without relying on arbitrary cluster-forming thresholds, thereby offering a more data-driven and interpretable framework for identifying significant effects. As a result, permutation-based inference becomes not only more powerful but also more principled and robust in neuroimaging analysis. Despite their traditionally high computational cost, recent algorithmic advancements and the availability of high-performance computing resources have made permutation testing computationally feasible even for large-scale neuroimaging datasets.

2.3.5 Resampling and Bootstrapping Techniques

Bootstrapping is a statistical resampling technique in which multiple samples are drawn with replacement from an observed dataset to empirically estimate the sampling

distribution of a statistic. Introduced by Efron [1992], the bootstrap provides a robust, non-parametric framework to quantify the variability (e.g., standard error, confidence interval) of estimators such as mean, regression coefficients or more complex estimators by simulating the process of repeated sampling from the population. Unlike traditional parametric methods, which rely on assumptions about the underlying distribution (e.g., normality), the non-parametric bootstrapping leverages the observed data as an empirical approximation to the true population. This approach is particularly advantageous in situations where theoretical distributions are analytically intractable or when sample sizes are small and asymptotic approximations may not hold [Efron and Tibshirani, 1994].

In contrast, parametric bootstrapping assumes a known parametric model for the data-generating process. After fitting this model to the observed data (typically using maximum likelihood estimation (MLE) or Bayesian inference), the estimated parameters are used to simulate synthetic datasets from the assumed distribution [Davison and Hinkley, 1997]. For each simulated dataset, the statistic of interest is recalculated, and the empirical distribution of these bootstrapped statistics is used to approximate the sampling distribution. This method can yield more precise estimates than the non-parametric bootstrap when the underlying model is correctly specified, as the incorporation of structural assumptions often leads to reduced variance in the estimates [Efron and Tibshirani, 1994]. However, parametric bootstrapping is inherently more sensitive to model misspecification: if the assumed distribution poorly reflects the true data-generating process, the resulting inferences might be biased or misleading [Shao and Tu, 2012]. Thus, it presents a trade-off between model-based efficiency and robustness to distributional assumptions, underscoring the importance of rigorous model validation prior to its application.

Unlike analytical methods that derive variance or confidence intervals using closed-form expressions or asymptotic theory, bootstrapping relies on recalculating the statistic of interest across many resampled dataset to empirically approximate its sampling [Efron and Tibshirani, 1994, Davison and Hinkley, 1997]. To obtain stable and reliable estimates, a large number of bootstrap resamples (often ranging from hundreds to several thousands) is typically required. Each replicate involving resampling the data, re-estimating the model and computing the statistic, making the procedure computationally intensive, particularly for complex models such as generalised linear models (GLMs) or mixed-effects models. This computational burden is further amplified in high-dimensional applications like voxel-wise analyses in neuroimaging,

where model estimation must be repeated independently at each voxel, often across hundreds of thousands of voxels. Because the total workload scales linearly (or worse) with both the number of voxels and the number of bootstrap replicates, runtime and total computation demands increase rapidly. In addition, each bootstrap replicate may require saving the full-resolution intermediate outputs (e.g., test statistic maps), which can result in substantial memory and storage requirements, particularly when working with full-brain 3D or 4D datasets from modalities such as fMRI or PET.

Despite these computational challenges, bootstrapping is inherently parallelisable: each bootstrap replicate is independent and can be distributed efficiently across multiple CPU cores or computing nodes within a high-performance computing (HPC) environment. In contrast, many traditional parametric inference procedures, such as the expectation-maximisation (EM) algorithm or Markov chain Monte Carlo (MCMC) methods, involve iterative, sequential updates and are often more difficult to parallelise effectively [Gilks et al., 1995, McLachlan and Krishnan, 2008].

2.4 Meta-Analysis in Neuroimaging

In this section, we provide a brief overview of meta-analysis in neuroimaging. First, Section 2.4.1 introduces the general concept of meta-analysis. Following this, Section 2.4.2 and Section 2.4.3 then describe the two main categories of neuroimaging meta-analysis: image-based meta-analysis (IBMA) and coordinate-based meta-analysis (CBMA). Within CBMA, both kernel-based and model-based methods are discussed. Finally, Section 2.4.4 presents large-scale CBMA datasets and platforms, with a focus on Neurosynth and its associated tools.

2.4.1 Meta-Analysis

Meta-analysis is a statistical technique that systematically synthesises findings from multiple independent studies addressing a common research question. By aggregating effect sizes or measures of association, meta-analysis increases statistical power, improves the precision of effect estimation, and resolves inconsistencies that may arise from conflicting individual findings. This approach is particularly valuable in disciplines such as psychology and neuroscience, where individual studies often suffer from small sample sizes, low statistical power and methodological heterogeneity (e.g.,

differences in experimental design, measurement instruments, analysis pipelines and participant populations) [Hunter and Schmidt, 2004, Ioannidis, 2005]. In addition to estimating overall effects, meta-analysis allows for the investigation of moderators and potential sources of heterogeneity across studies through techniques such as subgroup analysis meta-regression [Thompson and Higgins, 2002]. These capabilities make meta-analysis an essential tool in evidence-based research, allowing for the synthesis of broader scientific conclusions and the formulation of more generalisable inferences from the literature [Borenstein et al., 2021].

In neuroimaging, two main methodological approaches have been developed for conducting meta-analysis of brain activation data: image-based meta-analysis (IBMA) and coordinate-based meta-analysis (CBMA). Both approaches have contributed significantly to cumulative neuroimaging research and offer complementary advantages depending on the availability and resolution of underlying data.

2.4.2 Image-Based Meta-Analysis (IBMA)

Image-based meta-analysis (IBMA) is a quantitative approach in neuroimaging that integrates unthresholded statistical maps (such as t-statistic or z-statistic images) across multiple independent studies to derive population-level inferences about brain function. Unlike coordinate-based meta-analysis (CBMA), which relies solely on reported peak activation coordinates and is thus limited to sparse, thresholded data, IBMA leverages the full spatial resolution of three-dimensional statistical images [Salimi-Khorshidi et al., 2009, Nichols and Hayasaka, 2003]. This approach allows for voxel-wise modelling of both within-study effect sizes and between-study variability, typically within a random-effects or mixed-effects framework [Wager et al., 2007]. As a result, IBMA provides more accurate estimates of activation magnitude and improved control of false positives by leveraging full statistical maps, while also preserving sub-threshold signals that may be omitted in CBMA approaches.

Several specific IBMA methods have been developed to combine statistical maps from multiple neuroimaging studies while preserving voxel-wise information. Among p-value combination methods, one widely used approach is Fisher's method, a classic meta-analytic method that aggregates independent p-values by summing their logarithmic transformations. Specifically, Fisher's method computes the test statistic $X = -2\sum_{i=1}^{k} \ln(p_i)$ where p_i are the p-values from k independent studies. Under the

null hypothesis, this statistic follows a chi-squared distribution with $2k$ degrees of freedom [Fisher, 1970]. An alternative approach is Stouffer's method, which combines standardised test statistics (Z-score) from each study, optionally incorporating study-specific weights. The combined Z-score is given by

$$Z = \frac{\sum_{i=1}^{k} w_i z_i}{\sqrt{\sum_{i=1}^{k} w_i^2}}$$

where z_i is the Z-score for the $i-$th study and w_i is a corresponding weight, often based on the sample size or inverse variance. This method allows for greater flexibility by accommodating heterogeneity in study designs and has been shown to provide increased statistical power under certain conditions [Stouffer et al., 1949]. Despite their simplicity and analytical convenience, both Fisher's and Stouffer's methods assume statistical independence across studies and do not account for spatial correlations inherent in neuroimaging data. When applied voxel-wise without adjustment, these limitations can lead to inflated false-positive rates and reduce the robustness and reliability of meta-analytic inferences.

Effect-size-based IBMA models quantify and synthesise standardised effect across studies. Fixed effects models assume that all included studies share a single true effect size, so observed variability reflects only within-study sampling error. Pooled effects (Hedges g), are typically aggregated using inverse-variance weighting, where studies with more precise estimates (i.e., smaller standard errors) contribute more heavily to the pooled effect size [Borenstein et al., 2021]. Alternatively, random-effects models introduce a between-study variance component, often denoted as τ^2. A widely adopted method for estimating this variance is the DerSimonian–Laird method [DerSimonian and Laird, 1986], which provides more conservative and generalisable inferences by allowing the true effect sizes to vary across studies and modelling them as normally distributed around an overall mean. Additionally, Hedges' g includes a correction for small-sample bias, improving the accuracy of effect size estimates when sample sizes are limited [Hedges, 1981]. This adjustment is particularly critical in neuroimaging research, where limited participant numbers are common and uncorrected estimates may be systematically inflated and biased.

Likelihood-based approaches in IBMA construct a voxel-wise likelihood from each study's effect size map and its associated variances. These models flexibly accommodate

studies of varying quality by incorporating weights based on either sample size (sample-size-based likelihood) or within-study variance (variance-based likelihood), thereby adjusting for the relative precision of individual studies [Wager et al., 2007, Salimi-Khorshidi et al., 2009]. By accounting for heteroscedasticity and imbalanced study designs, likelihood-based methods often yield more stable and statistically efficient inferences compared to unweighted alternatives. Another technique is weighted least squares (WLS), which estimates voxel-wise meta-analytic effects by minimising the weighted sum of squared deviations, where weights inversely proportional to the variance of each study's effect size estimate [Becker, 1992]. This method is particularly advantageous when effect size precision varies substantially across studies, a common scenario in neuroimaging due to differences in sample size, scanner strength and preprocessing pipelines. Additionally, permuted ordinary least square (permuted OLS) provides a non-parametric inference framework by generating empirical null distributions through permutation [Nichols and Holmes, 2002]. This resampling approach offers robust control over family-wise Type I error rate and is especially useful when assumptions of normality and homoscedasticity are violated, as is often the case in voxel-wise neuroimaging data.

Despite these methodological advantages, the broader adoption of IBMA has historically been constrained by limited data availability. Most published neuroimaging studies have traditionally reported only peak activation coordinates, while full unthresholded maps are rarely shared. However, recent initiatives such as NeuroVault [Gorgolewski et al., 2015] have significantly improved data accessibility by providing a centralised, open-access repository for unthresholded statistical maps, thereby facilitating the wider implementation of IBMA methods and more reproducible neuroimaging research.

2.4.3 Coordinate-Based Meta-Analysis (CBMA)

Unlike IBMA, which requires access to full statistical maps, CBMA relies on reported peak activation coordinates (foci) from individual studies. These coordinates are typically provided in standardised stereotaxic spaces such as Talairach or Montreal Neurological Institute (MNI) space. Because full unthresholded statistical maps are often unavailable, due to data-sharing restrictions or legacy reporting practices, CBMA provides a practical and scalable alternative for synthesising findings across large volumes of published neuroimaging literature.

In CBMA, activation coordinates from multiple studies are aggregated to estimate the likelihood of activation in each brain region, based on the assumption that true effects are more likely to recur across independent studies. To account for spatial uncertainty in the reported peaks, each coordinate is modelled using a probabilistic kernel, for example, a 3D Gaussian centred at the reported location. These kernels are then combined across studies to construct a meta-analytic map of brain activity.

CBMA methods are broadly categorised into kernel-based methods and model-based approaches. Among the kernel-based methods, activation Likelihood Estimation (ALE) is one of the most widely used. ALE estimates the probability of activation at each voxel by combining modelled activation maps from individual studies, and assesses statistical significance through permutation testing to generate a null distribution of random spatial convergence [Eickhoff et al., 2009]. Another commonly used method, Multilevel Kernel Density Analysis (MKDA), places spherical kernels around reported foci and incorporates study-level weights (typically based on sample size) to compute voxel-wise statistics. Inference is then conducted by comparing the observed activation pattern to a null distribution generated from randomly located foci [Wager et al., 2007]. Seed-based d Mapping (SDM) and its variants offer a hybrid framework that integrates both peak coordinates and available effect size estimates (when available), thereby combining features of CBMA and IBMA [Radua et al., 2012]. While kernel-based methods are computationally efficient and conceptually straightforward, they have several limitations. These include limited statistical interpretability, inability to explicitly model spatial dependence or uncertainty, and constraints on performing group-level comparisons.

Model-based CBMA is an emerging and increasingly influential framework for synthesising results from neuroimaging studies that incorporates a formal probabilistic framework. Unlike kernel-base CBMA methods which estimate activation maps by convolving reported foci with fixed spatial kernels, model-based approaches employ structured probabilistic models that allow for more rigorous statistical inference and greater flexibility. In particular, Bayesian model-based CBMA treats reported peak coordinates as stochastic realisations from latent spatial processes, thereby enabling explicit modelling of spatial uncertainty, between-study heterogeneity and inter-study variability. A prominent class of these models is grounded in spatial point process theory, where the distribution of activation foci is governed by latent spatial intensity functions. These intensity functions are typically modelled using Gaussian processes or latent factor structures to capture spatial dependencies and reduce dimensionality

[Kang et al., 2011, Montagna et al., 2018, Samartsidis et al., 2019]. Such models support full Bayesian inference, offering posterior estimates of activation probabilities, while also facilitating meta-regression and subgroup analyses through the inclusion of publication-level covariates. While Bayesian model-based methods offer improved accuracy and interpretability compared to kernel-based approaches, they are also more computationally intensive, often requiring parallel computation on GPUs [Samartsidis et al., 2019].

Despite their reliance on sparse and thresholded peak coordinates, CBMA methods remain widely used due to the accessibility of published activation coordinates and their ability to summarise findings across extensive literatures. Recent methodological developments, including improved spatial modelling, the integration of effect size information, and hybrid CBMA–IBMA frameworks, continue to improve the interpretability and statistical power of coordinate-based meta-analytic findings in cognitive and clinical neuroscience.

2.4.4 CBMA Datasets and Repositories

Several curated repositories have been developed to support CBMA by aggregating these activation coordinates along with comprehensive metadata describing experimental conditions. One prominent example is BrainMap [Laird et al., 2005], a manually curated database containing thousands of peer-reviewed functional neuroimaging experiments. Each entry in BrainMap is richly annotated with metadata, including behavioural domain (e.g., cognitive, affective, sensorimotor), paradigm class, subject population characteristics, and specific experimental contrasts, enabling structured and hypothesis-driven meta-analyses across a wide range of psychological and neuroscientific topics.

In contrast to manually curated meta-analytic datasets, Neurosynth [Yarkoni et al., 2011] adopts an automated, data-driven approach that leverages text mining and natural language processing (NLP) techniques to extract activation foci and associated cognitive terms directly from the neuroimaging literature. Neurosynth automatically parses thousands of published articles indexed in PubMed to identify reported peak activation coordinates (in MNI and Talairach space) and links them to terms derived from article abstract. By statistically modelling the co-occurrence between terms and

brain locations using a naive Bayes classifier, Neurosynth enables large-scale meta-analyses that reveal associations between psychological concepts and neural activity. This framework facilitates both forward inference (estimating the probability of brain activation given the presence of a specific term) and reverse inference (estimating the probability of a psychological process given activation at a specific brain region), supporting exploratory analyses of the brain's functional architecture. Over time, the Neurosynth framework has evolved into a broader ecosystem of tools. For example, NeuroVault [Gorgolewski et al., 2015] is a closely integrated repository designed for sharing unthresholded statistical maps, allowing for greater spatial precision and reproducibility in meta-analytic studies. Another extension, Neuroynth Compose [Kent et al., 2024], builds upon the Neurosynth database by using language models to generate natural language descriptions of brain activation patterns. This tool represents a novel step toward automated cognitive decoding, enabling researchers to translate activation maps into interpretable psychological summaries.

Building on the foundation of Neurosynth, NeuroQuery [Dockès et al., 2020] introduces a more advanced predictive modelling framework. It employs regularised regression techniques to estimate continuous voxel-wise activation patterns from arbitrary natural language queries. Unlike Neurosynth's discrete term-based approach, NeuroQuery captures the multivariate relationship between linguistic features and brain activation, allowing for more precise and interpretable mapping from text to neuroimaging data. Additionally, NeuroQuery integrates a larger and more diverse corpus of studies and leverages improved coordinate extraction and spatial normalisation procedures, enhancing both coverage and anatomical specificity. As such, it provides a powerful tool for hypothesis generation, reverse inference, and large-scale cognitive ontology mapping in neuroimaging research.

2.5 White Matter Hyperintensities (WMHs)

In this section, we provide background information on white matter hyperintensities (WMHs). We begin by discussing the clinical relevance and underlying pathophysiology of WMHs in Section 2.5.1. Next, we describe the MRI characteristics and imaging modalities commonly used to detect WMHs in Section 2.5.2. Finally, we present an overview of current brain lesion segmentation approaches in Section 2.5.3.

2.5.1 Clinical Relevance and Pathophysiology

White matter hyperintensities (WMHs) are among the most frequently observed radiological abnormalities in ageing brains and are widely recognised as key imaging biomarkers of cerebral small vessel disease (SVD) [Longstreth et al., 1996, de Leeuw et al., 2001]. These lesions appear as regions of increased signal intensity on T2-weighted and fluid-attenuated inversion recovery (FLAIR) MRI sequence, as it suppresses the signal from cerebrospinal fluid (CSF), thereby enhancing the contrast between normal and pathological tissue. WMHs are most commonly observed in the periventricular and the deep subcortical white matter, and are typically bilateral, with a relatively symmetrical distribution.

Extensive research has consistently demonstrated that WMHs are highly prevalent in older adults, with both the incidence and volume increasing significantly with age [de Leeuw et al., 2001]. In some population-based cohorts, the prevalence of WMHs in individuals over the age of 65 has been estimated to exceed 90%. Beyond ageing, WMHs are closely associated with a range of modifiable cerebral vascular risk factors, including hypertension, type 2 diabetes, cigarette smoking, and hyperlipidaemia, as well as non-modifiable factors such as genetic predisposition and sex [Longstreth et al., 1996, Debette and Markus, 2010]. Among these, hypertension has emerged as a particularly influential risk factor, likely due to its chronic adverse effects on the structural and functional integrity of small cerebral vessels.

From a pathophysiological perspective, WMHs are believed to arise from a combination of chronic cerebral hypoperfusion and blood–brain barrier (BBB) dysfunction, both of which are consequences of progressive small vessel damage. Histopathological analyses have revealed that WMHs correspond to a spectrum of tissue abnormalities, including myelin loss, axonal degeneration, gliosis, and rarefaction of the white matter [Pantoni, 2010]. These changes are primarily driven by small vessel pathologies such as arteriosclerosis, lipohyalinosis and fibrinoid necrosis affecting the small perforating arterioles that supply the deep white matter [Wardlaw et al., 2013b]. Compromised autoregulatory capacity of these vessels, along with endothelial dysfunction, leads to impaired regulation of cerebral blood flow, reduced clearance of interstitial fluid, and increased permeability of the BBB, all of which contribute to tissue injury over time.

Clinically, the presence and severity of WMHs have been associated with a range of adverse outcomes. They are a well-established risk factor for cognitive impairment

and dementia, particularly vascular cognitive impairment and mixed Alzheimer's disease pathology. WMHs are also associated with reduced processing speed, executive dysfunction, gait disturbances, and an increased risk of falls and disability in older adults [Debette and Markus, 2010]. Furthermore, accumulating evidence suggests that the progression of WMHs over time can predict future stroke and worsen functional outcomes after cerebrovascular events. As such, WMHs are increasingly recognised not only as markers of existing small vessel pathology but also as potential early indicators of subclinical cerebrovascular disease, making them a valuable target for early intervention and longitudinal monitoring.

In summary, WMHs are important neuroimaging features with substantial clinical and pathophysiological relevance. Their presence reflects underlying small vessel pathology that contributes to a wide range of neurological and cognitive impairments. Given their high prevalence and strong association with modifiable vascular risk factors, WMHs provide critical insight into the relationship between vascular health and brain ageing. Consequently, WMHs serve not only as diagnostic markers but also as potential therapeutic targets for the prevention and management of age-related neurovascular and neurodegenerative disorders.

2.5.2 MRI Characteristics and Imaging Modalities

On MRI, WMHs are characterised by distinct signal characteristic across different imaging sequences. They typically appear as hyperintense regions on T2-weighted and fluid-attenuated recovery (FLAIR) images. FLAIR imaging is particularly effective in detecting WMHs, as it suppresses the signal from cerebrospinal fluid (CSF), thereby enhancing the visibility of lesions against surrounding brain tissue. Conversely, WMHs often appear isointense or hypointensity on T1-weighted images, especially when the lesions are more severe and confluent. T1 hypointensity has been associated with more severe underlying structural damage, including demyelination, gliosis, and axonal degeneration [DeCarli et al., 2005a, Gouw et al., 2011].

Anatomically, WMHs are commonly categorised based on their spatial distribution into two primary subtypes: periventricular WMHs (PVWMHs), which are located adjacent to the lateral ventricles, and deep WMHs (DWMHs), which occur in the subcortical or deep white matter. This classification is clinically relevant, as these subtypes may reflect distinct underlying vascular pathologies and differ in their clinical

implications. PVWMHs are often associated with age-related changes and alternations in CSF dynamics, as well as chronic periventricular ischemia, whereas DWMHs are more closely linked to hypertensive arteriopathy and chronic hypoperfusion of the deep perforating arteries [Fazekas et al., 1987, Wardlaw et al., 2013b]. Some studies suggest that DWMHs are more predictive of cognitive decline, particularly in domains such as processing speed and executive function [Gouw et al., 2011, Debette and Markus, 2010].

While conventional structural MRI, particularly FLAIR and T2-weighted imaging, remains the standard approach for detecting and visually quantifying WMHs, advances in neuroimaging have enabled more refined and quantitative assessments of WMH burden and the integrity of surrounding white matter tissue. Automated and semi-automated volumetric segmentation algorithms now allow for precise measurement of WMH load, reducing inter-rater variability and improving the reliability of longitudinal tracking in both clinical and research settings [Caligiuri et al., 2015]. In addition, diffusion tensor imaging (DTI) provides valuable microstructural insights into white matter by measuring parameters such as fractional anisotropy (FA) and mean diffusivity (MD), which can detect subtle alterations in tissue integrity even in regions that appear normal on conventional MRI [Bastin et al., 2009, O'Sullivan et al., 2001]. These findings have revealed that microstructural abnormalities often extend beyond the visible borders of WMHs, suggesting a more diffuse pattern of white matter injury typically associated with cerebral small vessel disease [Maillard et al., 2014].

2.5.3 WMH Segmentation

Accurate segmentation of WMHs is essential for quantifying lesion burden, monitoring disease progression, assessing their impact on cognitive and functional outcomes, and evaluating the efficacy of interventions targeting vascular risk. Segmentation methods can be broadly categorised into manual, semi-automated and fully automated approaches,

1. **Manual segmentation**, often regarded as the gold standard due to its anatomical precision, involves voxel-wise delineation by trained experts. While it provides high accuracy, this approach is labour-intensive, time-consuming and subject to inter- and intra-rater variability [Gouw et al., 2008].

2. **Semi-automated methods** combine user input with algorithmic processing (such as intensity thresholding and edge detection) to improve efficiency and reproducibility. However, manual correction is often still required, especially in cases with complex lesion morphology or suboptimal image equality [Caligiuri et al., 2015].

3. **Fully automated algorithm**, including tools such as BIANCA (Brain Intensity Abnormality Classification Algorithm), LST (Lesion Segmentation Tool), and deep learning-based models (e.g., U-Net architectures), offer high-throughput and reproducible WMH segmentation across large-scale and multi-site datasets [Griffanti et al., 2016, Schmidt et al., 2012, Liu et al., 2024]. These methods typically integrate multimodal MRI inputs (such as T1-weighted, T2-weighted, and FLAIR sequences), and often utilise combinations of intensity, spatial and anatomical priors to improve segmentation accuracy and robustness.

In this thesis, we focus on large-scale lesion mapping studies from the UK Biobank [Miller et al., 2016], where binary WMH lesion masks have been generated using BIANCA [Griffanti et al., 2016]. BIANCA is a fully automated, supervised machine learning tool developed as part of the FMRIB Software Library (FSL), specifically proposed for the segmentation of WMHs and other brain lesions on MRI. It employs a k-nearest neighbours (k-NN) classifier to differentiate lesion from non-lesion voxels based on local intensity features and spatial coordinates. During the training phase, manually annotated examples are used to construct a model that generates probabilistic lesion maps for new subjects. Key features of BIANCA include its ability to incorporate anatomical priors, such as distance from the lateral ventricles, and the flexibility to adjust probability thresholds and feature sets to optimise performance for specific datasets. This algorithm has demonstrated reliable and reproducible WMH segmentations across various clinical and population-based studies, particularly those focused on ageing and cerebral small vessel disease. However, its accuracy can be sensitive to the quality, size and representativeness of the training dataset, requiring careful consideration when applied to heterogeneous imaging protocols or across different imaging sites.

Recent advances in machine learning, particularly in deep learning, have significantly improved the performance of automated WMHs segmentation. Convolutional neural networks (CNN), especially architectures based on U-Net and its variants, have demonstrated high segmentation accuracy comparable to, and in some cases exceeding

that of human experts in well-controlled settings [Kuijf et al., 2019, Liu et al., 2024]. These models are capable of automatically learning hierarchical and spatially invariant features from multi-contrast MRI inputs (e.g., FLAIR and T1-weighted images). This capabilities allows them to robustly detect WMHs with substantial variability in size, shape and signal intensity, thereby addressing many limitations inherent in conventional machine learning approaches. Moreover, deep learning-based methods often generalise more effectively across diverse imaging datasets, particularly when trained on sufficiently large and heterogeneous cohorts.

Despite these advances, several challenges remain. One major limitation is the generalisability of deep learning models across different MRI scanners, acquisition protocols, population cohorts and disease conditions. Such variability can lead to inconsistent segmentation outputs, reducing the applicability of trained models in real-world studies. Difference in image resolution, signal-to-noise ratio and tissue contrast can introduce systematic biases that compromise the accuracy and comparability of WMH metric across studies. These challenges highlight the critical need for harmonization of image preprocessing pipelines, implementation of robust cross-site calibration strategies, and the adoption of publicly available, standardised benchmark datasets to support rigorous and reproducible evaluation of segmentation performance [Wardlaw et al., 2013b, Kuijf et al., 2019].

2.6 Contributions of Thesis

The main contribution of this thesis is the development of efficient and scalable generalised linear model (GLM) frameworks with an integrated spatial component, which can accommodate a wide range of stochastic distributions for modelling the complex and diverse characteristics of neuroimaging data. Scalability to the large sample sizes of population-level studies is achieved through robust model factorisation techniques and accurate approximation strategies. In addition, a flexible statistical inference framework is proposed, allowing for a variety of hypothesis tests related to spatial homogeneity and group-level comparisons. The proposed frameworks support both parametric and non-parametric inference approaches, including bootstrapping and resampling, which are particularly well-suited to the sparse and high-dimensional nature of neuroimaging data. Spatial dependence between neighbouring brain locations is captured through a spatial model with spline-based parametrisation, allowing for

smooth and interpretable spatial effects. This thesis is structured into three main chapters: the first two focus on meta regression and inference framework, applied to 20 coordinate-based meta-analytic datasets; the final chapter presents a spatial lesion estimation model and its application to lesion mapping in the UK Biobank dataset.

In Chapter 3, we develop a meta-regression framework for coordinate-based meta-analysis (CBMA) data, referred to as CBMR (Coordinate-Based Meta-Regression). This framework incorporates a spline-based parametrised spatial model as a general and flexible approach for analysing CBMA datasets. We apply CBMR to a collection of 20 CBMA datasets, demonstrating its effectiveness in synthesising findings across studies. In this work, we also explore and compare multiple stochastic models and extend the framework to incorporate publication-level covariates, allowing for the investigation of potential sources of heterogeneity at both the voxel and publication levels.

In Chapter 4, we extend CBMR to support multiple groups and introduce a roughness penalty to regularise the smoothness of the spatial functions parametrised by splines. We also develop a parametric bootstrap method as an alternative for more accurate inference in datasets with an insufficient number of foci, aiming to improve both the accuracy and robustness of the inference. This extended CBMR module is implemented in the open-source Python package NiMare, which is designed for meta-regression and meta-inference on CBMA fMRI datasets.

In Chapter 5, we develop a scalable voxel-wise GLM framework that incorporates an efficient approximate factorisation method for large-scale regression. The model supports flexible inference using either Chi-square tests or, alternatively, a sandwich estimator to improve accuracy and robustness under model misspecification. In our real data application, we apply the model to UK Biobank data to identify association between risk factors (e.g., age and cardiovascular risk factors) and lesion incidence probability.

Chapter 3

Neuroimaging Meta Regression for Coordinate Based Meta Analysis Data with a Spatial Model

This chapter focuses on meta-regression and inference methodology for coordinate-based meta-analysis (CBMA) of fMRI data. We propose a generative coordinate-based meta-regression (CBMR) framework to approximate smooth activation intensity functions and investigate the effects of publication-level covariates (e.g., year of publication, sample size). To capture the spatial structure of brain activation, we employ spline parameterization, and we consider four stochastic models to account for random variation in reported foci. To evaluate the validity of the proposed CBMR framework, we estimate brain activation across 20 meta-analytic datasets, conduct voxel-level spatial homogeneity tests, and compare our results with those obtained using existing kernel-based approaches.

3.1 Introduction

Functional neuroimaging includes a number of techniques to image brain activity, including positron emission tomography (PET) and functional magnetic resonance imaging (fMRI). Starting three decades ago, PET studies were used to compare brain activity between rest and experimental conditions, producing maps of “activation", images of statistics measuring the strength of the experimental effect. Especially in the last two decades, the literature on fMRI activations has grown rapidly, which motivates

a need to integrate findings, establish consistency and explore heterogeneity across independent but related studies. However in both PET and fMRI studies, validity is challenged by common drawbacks such as small sample sizes, a high prevalence of false positives (approximately $10-20\%$ of reported foci in publications are false positives [Wager et al., 2007]), significant heterogeneity among studies and unreliable inference due to their diversity in measurements and types of analysis [Samartsidis et al., 2017]. Meta-analysis is an essential tool to address these limitations and improve statistical power by pooling evidence from multiple studies and providing insight into consistent results. While there are also applications of neuroimaging meta-analysis to resting-state fMRI and structural analysis using voxel-based morphometry, in this work we focus on fMRI but note that our work applies to data from other types of studies.

Meta-analysis in neuroimaging research is classified into two categories: image-based meta-analysis (IBMA) which uses the 3D statistical maps of original studies and coordinate-based meta-analysis (CBMA) which uses the reported spatial coordinates of activation foci in standard MNI or Talairach space. Ideally, only IBMA would be used, as there is substantial information loss by only using activation foci as compared to full statistics maps, and further accuracy loss occurs when deactivation foci are ignored [Salimi-Khorshidi et al., 2009]. However, while it is now more common to share entire statistical maps in published studies, historically, researchers typically reported only the x, y, z coordinates of peak activation (local maxima) within each activation region. While this data is sparse, with an average of fewer than 10 foci reported per publication, and there are large-scale coordinate databases (e.g., BrainMap [Laird et al., 2005], Neurosynth [Yarkoni et al., 2011]) that index thousands of studies. Hence, CBMA remains the predominant approach for neuroimaging meta-analysis.

To identify brain regions with consistent activation across studies, researchers have developed a variety of CBMA methods, which are either kernel-based or model-based. Kernel-based CBMA methods utilise spatial kernel functions to model the uncertainty around each reported focus. In contrast, model-based CBMA methods employ nuanced statistical models with assumptions about the underlying brain function. Among those kernel-based methods, activation likelihood estimation (ALE, with a Gaussian kernel), multilevel kernel density analysis (MKDA, with a uniform sphere) and signed differential mapping (SDM, with a Gaussian kernel scaled by effect size) are commonly used [Turkeltaub et al., 2002, Eickhoff et al., 2012, Wager et al., 2007, Radua et al., 2012]. None of the three methods is based on a formal statistical model, however, all

are able to obtain statistical inferences by referencing to a null hypothesis of total random arrangement of the foci [Samartsidis et al., 2017]. Voxels with significant p-values are considered regions of consistent activation. Multiple testing corrected inferences are made by controlling the family-wise error rate using the null maximum distribution [Westfall and Young, 1993] or the false discovery rate (FDR) (Benjamini-Hochberg (BH) procedure, [Benjamini and Hochberg, 1995]). However, kernel-based methods lack interpretability, generally do not allow group comparison, do not model the spatial dependence of activation foci, nor can accommodate publication-level covariates to conduct a meta-regression [Samartsidis et al., 2019].

Bayesian model-based methods address these limitations, and are categorised into parametric spatial point process models [Kang et al., 2011, Montagna et al., 2018, Samartsidis et al., 2019] and non-parametric Bayesian models [Yue et al., 2012, Kang et al., 2014]. They use explicit generative models for the data with testable assumptions. Although they generally provide advances in interpretability and accuracy over kernel-based methods, they are computationally intensive approaches and generally require parallel computing on GPUs [Samartsidis et al., 2019], and only some approaches can conduct meta-regression to estimate the effect of publication-level covariates. Further, it can be more challenging for practitioners to interpret the spatial posterior intensity functions and utilise spatial Bayesian models in practice.

In this work, we propose classical frequentist models that explicitly account for the spatial structure of the distribution of activation foci. Specifically, we develop a spatial model that takes the form of a generalised linear model (GLM), where we make use of a spline parameterization to induce a smooth response and model the entire image jointly; we allow for image-wise publication-level regressors and consider different stochastic models to find the most accurate but parsimonious fit. Although Poisson is the classic distribution for describing independent foci counts, we have previously found evidence of over-dispersion [Samartsidis et al., 2020a], and thus we further explore a Negative Binomial model, a Clustered Negative Binomial model and a Quasi-Poisson model to allow excess variation in counts data.

Our work builds on the existing methods for CBMA, while introducing key innovations. From the Bayesian work, we adopt the concept of explicit spatial models; from the kernel methods, we incorporate the idea of fixing the degree of spatial smoothness. The contribution of this meta-regression model is both methodological and practical – it provides a generative regression model that estimates a smooth intensity function

and can incorporate publication-level regressors. Meanwhile, using a crucial memory-saving model factorisation, it also offers a computationally efficient alternative to existing Bayesian spatial regression models and provides an accurate estimation of the intensity function. While our method is suitable for any CBMA data, we are particularly motivated by studies of cognition. Cognition encompasses various mental processes, including perception, intelligence, problem solving, social interactions, and can be affected by substance use. We demonstrate this meta-regression framework on previously published meta-analyses of 20 cognitive and psychological tasks, allowing generalised linear hypothesis testing on spatial effect, as well as inference on the effect of publication-level covariates.

In the remainder of this work, we present our proposed meta-regression framework, introduce the model factorisation and optimisation procedures, as well as inferences on meta-regression outcomes via statistical tests in Section 3.2. Then we explain the experiment settings in Section 3.3 and explore different variants of stochastic models on the 20 meta-analytic datasets. We describe multiple goodness-of-fit statistics to identify the most accurate model, establish valid FPR control via Monte Carlo simulation under the null hypothesis of spatial homogeneity, followed by a comparison of homogeneity tests with kernel methods in Section 4.3. Finally, Section 3.5 summarises our findings and discusses potential extension of this meta-regression framework in the future.

3.2 Methods

GLMs are described in terms of their stochastic and deterministic components. Our deterministic model features a regression structure with a spatial component utilising spline parameterization and a publication-level covariate component. For the stochastic model, we consider multiple models motivated by CBMA data characteristics. We then propose a model factorisation approach to make our methods scalable, before outlining a general inference framework.

3.2.1 Deterministic model

3.2.1.1 Generic regression structure

Assume there are N voxels in each of M publications, and then our CBMA data at voxel j for publication i is the voxelwise count of foci Y_{ij}, written as a N-vector $Y_i = [Y_{i1}, Y_{i2}, \cdots, Y_{iN}]^\top$ for publication i. We generate a spatial design matrix X $(N \times P)$ with P cubic B-spline bases (more details to follow in Section 3.2.1.2) and construct a publication-level covariates matrix Z $(M \times R)$ using R publication-level covariates from each of M publications. For the CBMA framework, the central object of interest is the voxelwise intensity function for publication i, which considers both effects of smooth spatial bases and publication-level covariates. In this setting, we concisely write the model for publication i as

$$\log(\mu_i) = \log\left[\mathbb{E}(Y_i)\right] = X\beta + (Z_i\gamma)\mathbf{1}_N \tag{3.1}$$

where β $(P \times 1)$ and γ $(R \times 1)$ are regression coefficients for spatial bases X and publication-level covariates Z, respectively, Z_i is the i^{th} row of publication-level regressors Z, and $\mathbf{1}_N$ is a N-vector of 1's; the estimated intensity is μ_{ij} for publications $i = 1, ..., M$ and voxels $j = 1, ..., N$, written as the N-vector $\mu_i = [\mu_{i1}, \mu_{i2}, \cdots \mu_{iN}]^\top$ for publication i. This model is identifiable as long as we ensure each covariate variable is mean zero, letting X capture the overall mean. The GLM for all voxels in all M publications is then

$$\log\left[\mathbb{E}(Y)\right] = (\mathbf{1}_M \otimes X)\beta + (Z \otimes \mathbf{1}_N)\gamma \tag{3.2}$$

where $Y = [Y_1, Y_2, \cdots, Y_M]^\top$ is a $(M \times N)$-vector, containing voxelwise foci count for all of M publications, and $\otimes$ is the Kronecker product. Note that our GLM has millions of rows (MN) and the spatial design matrix has billions of entries $(MN \times P)$. In consideration of implementation complexity and memory requirement, we will propose a simplified reformulation of this GLM in Section 3.2.3.

3.2.1.2 Spline parameterization

Previous work on spatial point process modelling of CBMA data has treated each publication's foci as a realisation of a doubly-stochastic Poisson process, also known as a Cox process. In some of that work, the log intensity function is parametrised by

superimposed Gaussian kernel basis functions [Montagna et al., 2018], while in others, the log intensity is a Gaussian process [Samartsidis et al., 2019]. Both the tensor product of cubic B-spline bases and the Gaussian kernel basis functions are suitable for modelling spatial intensity. Their smoothness, stability and ability to provide local support make them ideal spatial bases for CBMA applications. We choose tensor product splines for this work but, in a small evaluation, found that these two approaches have comparable performance; see Appendix A.3.1 of the Supplementary material.

A 1-dimensional cubic B-spline is a piece-wise polynomial of order 3, where pre-specified knots determine the parameterization of basis functions where the polynomial sections join. For our 3D lattice, assume there are v_x voxels along the x direction, the coefficients of v_x voxels evaluated at each of n_x B-spline bases construct a coefficient matrix C_x (size $v_x \times n_x$). Similarly, there exist another two coefficient matrices C_y and C_z (size $v_y \times n_y$ and $v_z \times n_z$) along y and z direction. The whole coefficient matrix C of 3-dimensional B-spline bases is constructed by taking the tensor product of the three coefficient matrices (see Figure 3.1 for a 2D illustration),

$$C = C_x \otimes C_y \otimes C_z \tag{3.3}$$

The matrix of C is $(v_x v_y v_z) \times (n_x n_y n_z)$, and is the basis for the entire 3D volume, while the analysis is based on a brain mask of N voxels. The design matrix X is obtained from C after a three-step process: First, rows corresponding to voxels outside the brain mask are removed; then, columns are removed if they correspond to weakly supported B-spline bases (a B-spline basis is regarded as "weakly supported" if its maximum value of coefficients evaluated at each voxel is below 0.1). Finally, the rows are re-normalised (sum to 1) to preserve the "partition of unity" property of B-spline bases.

We define our cubic B-spline bases with equally spaced knots in x, y and z dimensions, and thus we parametrise the level of spatial smoothness by the knot spacing. Larger knots spacing, smaller basis, and greater smoothness; conversely, closer knots, larger basis, and greater ability to represent fine details. Conceptually, more flexible parameterizations would allow arbitrary knots locations, but with the consideration of minimising computational complexity, we fix the design matrix X based on pre-specified knots spacing and locations. We also provide practical recommendation on parameter selection for knot configuration in section A.3.2 in the Supplementary material. While other spline applications use a dense array of knots and then control

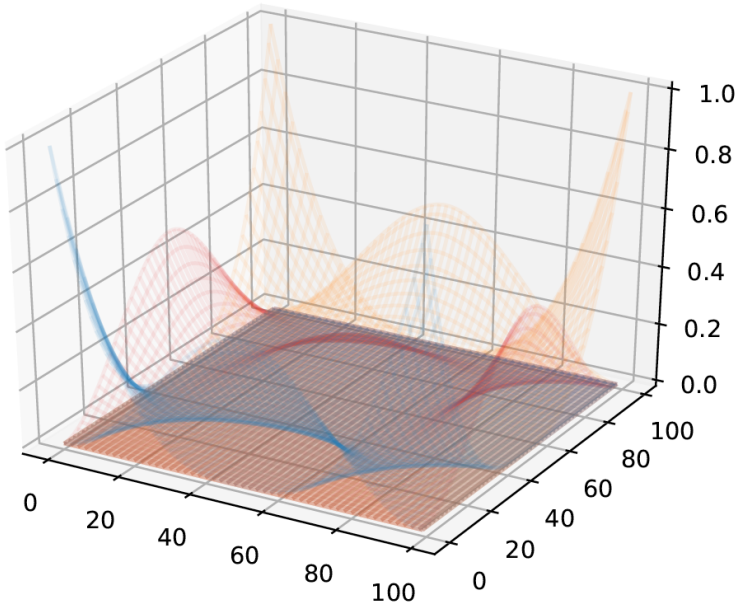


Figure 3.1: Illustration of a 2D tensor product spline basis.

smoothness with a roughness penalty, the computational and memory requirements of our spatial model demand that we judiciously select the coarsest spline spacing consistent with our application.

3.2.2 Stochastic model

Different stochastic assumptions on the CBMA foci data determine the form of statistical likelihood we use. We consider a set of four stochastic models for the distribution of foci counts at the voxel level. All of our models take the form of GLMs, where inhomogeneous intensity at each voxel is captured by the spline bases and any publication-level covariate (as per Equation 3.2). We fit our model either by maximising log-likelihood function iteratively via L-BFGS (Limited-memory Broyden–Fletcher–Goldfarb–Shanno) algorithm [Shanno, 1970] or iteratively re-weighted least squares (IRLS) for Quasi-likelihood models. We now elaborate each of these models in turn and discuss their strengths and limitations.

3.2.2.1 Poisson model

In practice, the count of foci Y_{ij} (for publications $i = 1, \cdots, M$, voxels $j = 1, \cdots, N$) is only ever 0 or 1, which strictly indicates a Binomial model. However, inspired by previous success with the Poisson point process, and the accuracy of the Poisson approximation for low-rate Binomial data [Eisenberg et al., 1966], we consider a Poisson model.

If foci arise from a realisation of a (continuous) inhomogeneous Poisson process, the (discrete) voxel-wise counts will be independently distributed as Poisson random variables, with a rate equal to the integral of the (true, unobserved, continuous) intensity function over each voxel. As the sum of multiple independent Poisson random variables is also Poisson, this also gives rise to a practical consequence: it is equivalent to either model the set of M publication-level counts or the summed counts at each voxel. Following the deterministic structure outlined in Equation (3.1), the intensity for voxel j in publication i is

$$\begin{aligned} \mathbb{E}[Y_{ij}] &= \mu_{ij} \\ \log(\mu_{ij}) &= \eta_{ij} = x_j^\top \beta + Z_i \gamma \end{aligned} \tag{3.4}$$

where $Y_{ij} \sim \text{Poisson}(\mu_{ij})$, $x_j^\top$ is the j^{th} row of spatial design matrix $X(N \times P)$, and β is the regression coefficient of spline bases. The data vector Y has a length-(MN), which is impractical to represent explicitly. Under the assumption of independence of counts across publications, the likelihood function is exactly same if we model the voxel-wise total foci count over publications instead (more details to follow in A.1.1 in the Supplementary material), which gives rise to the modified Poisson model on summed data at voxel j over all publications, $Y_{\bullet,j} = \sum_{i=1}^{M} Y_{ij}$,

$$\begin{aligned} &\mathbb{E}[Y_{\bullet,j}] = \mu_{\bullet,j}, \\ &\mu_{\bullet,j} = \sum_{i=1}^{M} \mu_{ij} = \sum_{i=1}^{M} \exp\left(x_j^\top \beta + Z_i \gamma\right) = \exp(x_j^\top \beta) \left(\sum_{i=1}^{M} \exp(Z_i \gamma) \right) \end{aligned} \tag{3.5}$$

where $\mu_{\bullet,j} = \sum_{i=1}^{M} \mu_{ij}$ is the expected sum of intensity at voxel j over publications. Under this formulation, the likelihood to be optimised is,

$$l(\theta) = l(\beta, \gamma) = \sum_{j=1}^{N} \left[Y_{\bullet,j} \log(\mu_{\bullet,j}) - \mu_{\bullet,j} - \log(Y_{\bullet,j}!) \right] \tag{3.6}$$

3.2.2.2 Negative Binomial model

While Poisson model is widely used in the regression of count data, it is recognised that counts often display over-dispersion (the variance of the response variable substantially exceeds the mean). Imposing a Poisson model based on the unrealistic assumption (variance equals mean) may underestimate the standard error, and lead to biased estimation of the regression coefficients. While Barndorff-Nielsen and Yeo [1969] proposed a formal definition of spatial Negative Binomial model, it involves Gaussian processes and complexities we sought to avoid. Hence, here we do not propose a formal point process model, but rather simply assert that the count data at each voxel follows a Negative binomial (NB) distribution independently, thus allowing for anticipated excess variance relative to Poisson [Lawless, 1987].

Our NB model uses a single parameter α shared over all publications and all voxels to index variance in excess of the Poisson model. For each publication i and voxel j, let λ_{ij} follow a Gamma distribution with mean μ_{ij} and variance $\alpha\mu_{ij}^2$; then conditioned on λ_{ij}, let Y_{ij} be Poisson with mean λ_{ij}. It can be shown that the marginal distribution of Y_{ij} follows a NB distribution with probability mass function,

$$\mathbb{P}(Y_{ij} = y_{ij}) = \frac{\Gamma(y_{ij} + \alpha^{-1})}{\Gamma(y_{ij} + 1)\Gamma(\alpha^{-1})} \left(\frac{1}{1 + \alpha\mu_{ij}}\right)^{\alpha^{-1}} \left(\frac{\alpha\mu_{ij}}{1 + \alpha\mu_{ij}}\right)^{y_{ij}}. \tag{3.7}$$

In terms of the success count and probability parameterization, $\text{NB}(r, p)$, we have $Y_{ij} \sim \text{NB}(\alpha^{-1}, \frac{\mu_{ij}}{\alpha^{-1}+\mu_{ij}})$, with mean $\mathbb{E}(Y_{ij}) = \mu_{ij}$ and variance $\mathbb{V}(Y_{ij}) = \mu_{ij} + \alpha\mu_{ij}^2$. Details on the derivation of the probability density function of the NB model can be found in A.1.2 of the Supplementary material. When $\alpha > 0$, we observe Poisson-excess variance of $\alpha\mu_{ij}^2$; or analogous to the coefficient of variation, the coefficient of excess variation is $\sqrt{\alpha\mu_{ij}^2}/\mu_{ij} = \sqrt{\alpha}$, which can be interpreted roughly as the relative excess standard deviation relative to a Poisson model.

Again, the data vector is impractical to represent explicitly, but unlike Poisson, the sum of multiple independent NB random variables doesn't follow an NB distribution. Thus, we propose a moment matching approach to approximate the mean (first moment) and variance (second moment) of this convolution of NB distributions, which significantly facilitates the simplification of the log-likelihood function. Matching the first two moments, the approximate NB distribution of the total count of foci over all

publications at voxel j is given by $Y_{\bullet,j} = \sum_{i=1}^{M} Y_{ij} \sim \text{NB}(r'_j, p'_j)$, where

$$r'_j = \frac{\mu^2_{\bullet,j}}{\alpha \sum_{i=1}^{M} \mu^2_{ij}}, \quad p'_j = \frac{\sum_{i=1}^{M} \mu^2_{ij}}{\alpha^{-1}\mu_{\bullet,j} + \sum_{i=1}^{M} \mu^2_{ij}}$$

with corresponding excess variance

$$\alpha' = \alpha \frac{\sum_{i=1}^{M} \mu^2_{ij}}{\mu^2_{\bullet,j}},$$

which gives rise to the simplified NB log-likelihood function,

$$l(\theta) \approx l(\beta, \alpha') = \sum_{j=1}^{N} \left[\log \Gamma(Y_{\bullet,j} + r'_j) - \log \Gamma(Y_{\bullet,j} + 1) - \log \Gamma(r'_j) + r'_j \log(1 - p'_j) + Y_{\bullet,j} \log p'_j\right] \tag{3.8}$$

Details on the derivations of the moment matching approach can be found in A.1.3 in the Supplementary material. We have also included a simulation in section A.1.4 in the Supplementary material, which demonstrates the accuracy of this method in approximating the sum of NB distributed variates. Our findings indicate a negligible bias (0.3098%) in the standard error estimates for the mean estimate using the moment matching approach. Furthermore, the maximum likelihood estimates (MLEs) for both methods are remarkably close to their true values.

3.2.2.3 Clustered Negative Binomial model

While the NB model can be regarded as a kind of "random effects" Poisson model, as developed above, the latent Gamma random variable introduces independent variation at each voxel. Instead, we could assert that the random (Gamma-distributed) effects are not independent voxel-wise effects, but rather latent characteristics of each publication, representing a shared effect over the entire brain for a given publication. This is, in fact, the approach used by a Bayesian CBMA method [Samartsidis et al., 2019], and in a non-imaging setting, a Poisson-Gamma model for two-stage cluster sampling [Geoffroy and Weerakkody, 2001]. Therefore, we now consider a third GLM, where at the first stage, we assume each individual publication i is sampled with a global latent value λ_i from a Gamma distribution with mean 1 and variance α, which accommodates excess

variance by the dispersion parameter α ($\lambda_i \sim Gamma(\alpha^{-1}, \alpha^{-1})$). At the second stage, conditioned on the global variable λ_i, Y_{ij} are drawn from a Poisson distribution with mean $\lambda_i \mu_{ij}$ ($Y_{ij}|\lambda_i \sim \text{Poisson}(\lambda_i \mu_{ij})$), where μ_{ij} is the expected intensity parametrised by spatial regression parameter β and covariates regression parameter γ. The marginal distribution of Y_{ij} also follows an NB distribution,

$$\mathbb{P}(Y_{ij} = y_{ij}) = \frac{\Gamma(y_{ij} + \alpha^{-1})}{\Gamma(y_{ij} + 1)\Gamma(\alpha^{-1})} \left(\frac{\alpha^{-1}}{\mu_{ij} + \alpha^{-1}}\right)^{\alpha^{-1}} \left(\frac{\mu_{ij}}{\mu_{ij} + \alpha^{-1}}\right)^{y_{ij}} \tag{3.9}$$

where $Y_{ij} \sim NB(\alpha^{-1}, \frac{\mu_{ij}}{\alpha^{-1}+\mu_{ij}})$ with mean $\mathbb{E}(Y_{ij}) = \mu_{ij}$ and variance $\mathbb{V}(Y_{ij}) = \mu_{ij} + \alpha\mu_{ij}^2$. Details on the derivation of the probability density function of the clustered NB model can be found in A.1.5 in the Supplementary material. This two-stage hierarchical Clustered NB model also introduces a covariance structure between foci within a publication, which is determined by the expected intensity of the observations as well as the dispersion parameter α (see A.1.6 in the Supplementary material). The covariance for publications i and i', and distinct voxel j and j' is,

$$\begin{cases} \mathbb{C}(Y_{ij}, Y_{i',j'}) = \alpha\mu_{ij}\mu_{ij'}, \text{ if } i = i' \\ \mathbb{C}(Y_{ij}, Y_{i',j'}) = 0, \text{ if } i \neq i' \end{cases} \tag{3.10}$$

The log-likelihood is the sum of terms over independent publications,

$$\begin{aligned} l(\beta, \alpha, \gamma) &= \sum_{i=1}^{M} \log[f(Y_{i1}, Y_{i2}, \cdots, Y_{iN})] \\ &= M\alpha^{-1}\log(\alpha^{-1}) - M\log\Gamma(\alpha^{-1}) + \sum_{i=1}^{M} \log\Gamma(Y_{i,\cdot} + \alpha^{-1}) \\ &\quad - \sum_{i=1}^{M}\sum_{j=1}^{N} \log Y_{ij}! - \sum_{i=1}^{M}(Y_{i,\cdot} + \alpha^{-1})\log(\mu_{i,\cdot} + \alpha^{-1}) + \sum_{i=1}^{M}\sum_{j=1}^{N} Y_{ij}\log\mu_{ij} \end{aligned} \tag{3.11}$$

where $Y_{i,\cdot} = \sum_{j=1}^{N} Y_{ij}$ is the sum of foci within publication i. One limitation of this model, though, is that it doesn't admit a factorisation and depends on the length-(MN) data vector (see A.1.6 in Supplementary material).

While the intra-publication dependence is well-motivated, the Clustered NB model depends on the strong assumption that excess variance is captured by the global dispersion λ_i. If there is voxel-wise independent excess variance, the previous NB model will be preferred; we assess this issue below with real data evaluations.

3.2.2.4 Quasi-Poisson model

As an alternative to the NB model, Quasi-Poisson model also allows for over-dispersed count data, and is a straightforward elaboration of the GLM. Instead of specifying a specific probability distribution for count data, the Quasi-Poisson model only requires a mean model and a variance function, $\mathbb{V}(Y_{ij}) = \theta\mu_{ij}$ (with $\theta \geq 1$). While the variance-mean relationship is linear for the Quasi-Poisson model, it is quadratic in the NB model. This results in small foci counts being weighted more and can have greater adjustment effect in the Quasi-Poisson model, which theoretically might be ideal for our scenario where most brain regions have zero or low foci counts [Ver Hoef and Boveng, 2007].

Quasi-Poisson model can be framed as a GLM, with the mean and variance for voxel j in publication i given by,

$$\begin{aligned} E[Y_{ij}] &= \mu_{ij} \\ \text{Var}(Y_{ij}) &= \theta\mu_{ij}. \end{aligned} \tag{3.12}$$

Without a likelihood function, we instead use ILRS algorithm, with the $(k+1)^{th}$ iteration given by,

$$\begin{aligned} \hat{\beta}^{[k+1]} &= \beta^{[k]} + (X^{*\top}W^{[k]}X^{*})^{-1}X^{*\top}(Y - \mu^{[k]}) \\ \hat{\gamma}^{[k+1]} &= \hat{\gamma}^{[k]} + (Z^{*\top}W^{[k]}Z^{*})^{-1}Z^{*\top}(Y - \mu^{[k]}) \end{aligned} \tag{3.13}$$

where $W = \text{diag}(\frac{\mu_{11}}{\theta}, \cdots, \frac{\mu_{1N}}{\theta}, \cdots, \frac{\mu_{M1}}{\theta}, \cdots, \frac{\mu_{MN}}{\theta})$, and $X^{*} = \mathbf{1}_M \otimes X$, $Z^{*} = \mathbf{1}_N \otimes Z$. This model can be simplified as well, though we again defer that to the next Section 3.2.3.

3.2.3 Model factorisation

Having derived the explicit log-likelihood functions for meta-regression with three stochastic likelihood-based models, as well as the updating equation for a quasi-likelihood based model, we now consider model factorisation to replace the full (MN)-vector of foci counts by sufficient statistics. Following the generic formulation of GLM proposed in Section 3.2.1.1,

$$\eta_{ij} = \log(\mu_{ij}) = \sum_{k=1}^{P} X_{jk}\beta_k + \sum_{s=1}^{R} Z_{is}\gamma_s. \tag{3.14}$$

η_{ij} is the estimated linear response from GLM, specific to each voxel j in each individual publication i. In our neuroimaging application, there are always at least $220,000$ voxels (N), hundreds or thousands of publications M, and ≈500 or more basis elements ($P = 456$ for $20mm$ knots spacing), giving rise to millions of rows (MN) and billions of entries ($MN \times (P + R)$) in a GLM formulation. Thus, we propose a reformulation of this model into a series of sufficient statistics that are never larger than M or N in dimension. First, note that the localised spatial effect μ^X and global effect of publication-level covariates μ_i^Z for publication i factorise μ_{ij} as

$$\mu_{ij} = \exp\left(\sum_{k=1}^{P} X_{jk}\beta_k + \sum_{s=1}^{R} Z_{is}\gamma_s\right) = \exp\left(\sum_{k=1}^{P} X_{jk}\beta_k\right) \exp\left(\sum_{s=1}^{R} Z_{is}\gamma_s\right) = \mu_j^X \, \mu_i^Z \tag{3.15}$$

To further simplify the log-likelihood function, we also use the fact that $Y_{ij} \leq 1$ (either 0 or 1), as there will never be more than one foci at the same location in a given publication. Define the following notation:

- let N-vector $\mu^X = \exp(X\beta)$ be the vector of spatial effects;
- let M-vector $\mu^Z = \exp(Z\gamma)$ be the vector of global publication-level covariates effects;
- as already defined, $Y_{\bullet,j} = \sum_{i=1}^{M} Y_{ij}$ is sum of foci counts at voxel j across all publications, and define the N-vector $Y_{\bullet,} = [Y_{\bullet,1}, \cdots, Y_{\bullet,N}]^\top$;
- and let $Y_{i,\bullet} = \sum_{j=1}^{N} Y_{ij}$ be the sum of foci counts for publication i across all voxels, and define the M-vector $Y_{,\bullet} = [Y_{1,\bullet}, \cdots, Y_{M,\bullet}]^\top$.

The simplified factorisation of total log-likelihood functions or IRLS updating equation are specific to each stochastic model. Full details are provided in A.2 in the Supplementary material; in summary:

- Poisson model:

$$l(\beta, \alpha) = Y_{\bullet,}^\top \log(\mu^X) + Y_{,\bullet}^\top \log(\mu^Z) - \left[\mathbf{1}^\top \mu^X\right]\left[\mathbf{1}^\top \mu^Z\right], \tag{3.16}$$

- NB model: As described in Section 3.2.2.2, we approximate a sum of independent NB variables again as a NB:

$$Y_{\cdot,j} = \sum_{i=1}^{M} Y_{ij} \sim \text{NB}(r'_j, p'_j) = \text{NB}\left(\frac{(\mu_j^X)^2 [\mathbf{1}^\top \mu^Z]^2}{\alpha' \sum_{i=1}^{M} (\mu_j^X \mu_i^Z)^2}, \frac{\sum_{i=1}^{M} (\mu_j^X \mu_i^Z)^2}{(\alpha')^{-1} \mu_j^X [\mathbf{1}^\top \mu^Z]^\top + \sum_{i=1}^{M} (\mu_j^X \mu_i^Z)^2} \right) \tag{3.17}$$

with dispersion parameter $\alpha' = \frac{\alpha \sum_{i=1}^{M} (\mu_j^X \mu_i^Z)^2}{(\mu_j^X)^2 [\mathbf{1}^\top \mu^Z]^2}$. The log-likelihood function is given by,

$$l(\alpha', \beta, \gamma) = \sum_{j=1}^{N} \left[\log \Gamma(Y_{\cdot,j} + r'_j) - \log \Gamma(Y_{\cdot,j} + 1) - \log \Gamma(r'_j) + r'_j \log(1 - p'_j) + Y_{\cdot,j} \log p'_j \right], \tag{3.18}$$

- Clustered NB model:

$$\begin{aligned} l(\alpha, \beta, \gamma) = {} & M\alpha^{-1} \log(\alpha^{-1}) - M \log \Gamma(\alpha^{-1}) + \sum_{i=1}^{M} \log \Gamma(Y_{i,\cdot} + \alpha^{-1}) \\ & - \sum_{i=1}^{M} (Y_{i,\cdot} + \alpha^{-1}) \log(\alpha^{-1} + \mu_{i,\cdot}) + Y_{\cdot,\cdot}^\top \log(\mu^X) + Y_{\cdot,\cdot}^\top \log(\mu^Z) \end{aligned} \tag{3.19}$$

where dispersion parameter α measures the excess variance across all publications and all voxels,

- Quasi-Poisson model:

$$\begin{aligned} \hat{\beta}^{[j+1]} &= \beta^{[j]} + (X^\top W^{[j]} X)^{-1} X^\top (Y_{\cdot,} - (\mu^X)^{[j]}) \\ \hat{\gamma}^{[j+1]} &= \hat{\gamma}^{[j]} + (Z^\top V^{[j]} Z)^{-1} Z^\top (Y_{\cdot,\cdot} - (\mu^Z)^{[j]}) \end{aligned} \tag{3.20}$$

where $W = \text{diag}(\frac{\mu_1^X}{\theta}, \cdots, \frac{\mu_N^X}{\theta})$ and $V = \text{diag}(\frac{\mu_1^Z}{\theta}, \frac{\mu_2^Z}{\theta}, \cdots, \frac{\mu_M^Z}{\theta})$.

3.2.4 Model optimisation

For likelihood-based models (Poisson, NB and clustered NB model; Section 3.2.2.1 - Section 3.2.2.3) without publication-level covariates, we employ Fisher scoring for the iterative optimisation of parameters in GLMs. Fisher scoring replaces the gradient and Hessian of Newton's method with the score and observed Fisher's information,

respectively [Longford, 1987]. Writing θ for all parameters, the updating equation at the $(k+1)^{th}$ iteration is,

$$\theta^{[k+1]} = \theta^{[k]} + I(\theta^{[k]})^{-1} \frac{\partial}{\partial \theta^{[k]}} l(\theta^{[k]}) \tag{3.21}$$

where the observed Fisher information is $I(\theta^{[k]}) = \mathbb{E}\left[-\frac{\partial^2 l(\theta)}{\partial\theta\partial\theta^\top}\right]_{\theta=\theta^{[k]}}$.

For the Poisson model, $\theta = [\beta, \gamma]$, the Fisher information is given by,

$$I(\theta) = I(\beta, \gamma) = \begin{bmatrix} -\frac{\partial^2 l}{\partial\beta\partial\beta^\top} & -\frac{\partial^2 l}{\partial\beta\partial\gamma^\top} \\ -\frac{\partial^2 l}{\partial\gamma\partial\beta^\top} & -\frac{\partial^2 l}{\partial\gamma\partial\gamma^\top} \end{bmatrix} \tag{3.22}$$

with negative Hessian matrix of β, $\left(-\frac{\partial^2 l}{\partial\beta\partial\beta^\top}\right)_{P\times P} = X^\top \text{diag}(\mu^X) X$; the negative cross term $\left(-\frac{\partial^2 l}{\partial\beta\partial\gamma^\top}\right)_{P\times R} = \left(-\frac{\partial^2 l}{\partial\gamma\partial\beta^\top}\right)^\top_{R\times P} = [X^\top \mu^X][(\mu^Z)^\top Z]$; and negative Hessian matrix of γ, $\left(-\frac{\partial^2 l}{\partial\gamma\partial\gamma^\top}\right) = Z^\top \text{diag}(\mu^Z) Z$.

Likelihood-based models with publication-level covariates lead to more complicated derivations of updating equations via Fisher scoring. Instead, we use a more efficient quasi-Newton algorithm (the L-BFGS algorithm, [Shanno, 1970]), which minimises smooth, nonlinear functions without directly computing the Hessian matrix. It estimates the observed Fisher Information with gradient evaluations, significantly reducing both memory requirements and computational complexity. It is particularly well-suited for optimisation problems characterised by a large number of variables, where computing the full Hessian matrix would be computationally expensive or even infeasible due to memory constraints.

Lastly, for Quasi-likelihood models (e.g. Quasi-Poisson model, see Section 3.2.2.4) when exact likelihood functions are computationally infeasible, optimising the regression coefficients using Fisher scoring becomes impractical. Instead, we employ the Iteratively Reweighted Least Squares (IRLS) method to iteratively find the optimal regression coefficients. For the updating equations, please refer to Section A.2.4 in the Supplementary Material.

3.2.5 Statistical inference

3.2.5.1 Global test of model fitness

Among the proposed stochastic models in Section 3.2.2, the Poisson, NB and clustered NB model are likelihood-based, while Quasi-Poisson model is Quasi-likelihood based (its exact likelihood is computationally infeasible). To compare the goodness of fit from a global perspective, we will utilise likelihood-based comparison criteria (e.g., LRT and Akaike information criterion (AIC)) with likelihood-based models, as well as other global model fitness criteria across all stochastic models within this meta-regression framework.

Likelihood-based model selection criteria LRT uses the difference in log-likelihoods to test the null hypothesis that the true model is the smaller nested model. Since the Poisson model is nested in both the NB model and the clustered NB model with a dispersion parameter $\alpha = 0$, for the null hypothesis H_0: dispersion parameter $\alpha = 0$, the likelihood-ratio test statistic is given by,

$$\lambda_{LR} = -2\left[l(\hat{\theta}_0) - l(\hat{\theta})\right]$$

where $l(\hat{\theta}) = l(\hat{\alpha}, \hat{\beta}, \hat{\gamma})$ is the maximum log-likelihood of the NB model or clustered NB model without any constraint on parameters, and $l(\hat{\theta}_0) = l(\hat{\alpha} = 0, \hat{\beta}, \hat{\gamma})$ is the maximum log-likelihood of the NB model or clustered NB model with the dispersion parameter α constrained at 0 (i.e. Poisson model). The test statistic is Chi-square distributed with 1 degree of freedom.

AIC is an alternatives to LRT which also address the trade-off between the goodness of fit and the simplicity of the model, and they address the overfitting problem by penalising the number of parameters in the model. To measure the goodness of fit of a model M on dataset D,

$$AIC = 2k - 2l(\hat{\theta}) \tag{3.23}$$

where $l(\hat{\theta})$ is the maximised log-likelihood function of the model M, k is the number of parameters in model M and n is the number of data points in the dataset D. The model with the smaller AIC is believed to be a better fit to the dataset.

Bias and variance of estimation For the purpose of selecting the best model in terms of goodness of fit across a variety of datasets, we extend the model comparisons to include all stochastic models proposed in Section 3.2.2, including the Quasi-Poisson

model. As the central outcome of this meta-regression framework is voxel-wise intensity estimation for each publication, with the effect of publication-level covariates being considered, it's natural to utilise bias and variance of intensity estimation as new criteria stated below,

- Relative bias of the estimated total sum of intensity (per publication), compared with the averaged sum of foci count (per publication) across multiple datasets;
- Relative bias of standard deviation (SD) in each of x, y, z dimension, compared with the actual Standard deviation in foci count (per publication) across multiple datasets;
- Relative bias of voxel-wise variance between the actual foci count (per publication) and the intensity estimation (per publication).

Here, relative bias is evaluated instead of absolute bias, especially when applied to a variety of datasets with diverse foci counts.

3.2.5.2 Localised inference with Wald tests on μ_{ij}^X and η_{ij}^X

While our model is parameterised by P basis elements, users want to make inference at each of the N voxels. Hence, we provide localised inference on estimated spatial intensity μ_{ij}^X (or $\eta_{ij}^X = \log(\mu_{ij}^X)$) and the regression coefficient of publication-level covariates (γ) via Wald tests.

Test of spatial homogeneity: In the CBMA context, the most basic inference is a test of homogeneity to identify regions where more foci arise than would be expected if there were no spatial structure. Precisely, we use the null hypothesis on voxelwise intensity estimation or estimated linear response, $H_0 : \mu_{ij}^X = \mu_0 = \sum_{i=1}^{M}\sum_{j=1}^{N} Y_{ij}/(MN)$ or $\eta_{ij}^X = \eta_0 = \log(\mu_0)$ at voxel j, for publication i. The standard error for β can be asymptotically estimated from the inverse of the observed Fisher Information matrix, which gives rise to the standard error for the linear response η_{ij}^X, and thus the standard error for μ_{ij}^X is obtained via the delta method (see Section A.2.5 in the Supplementary Material for details). It allows inference via Wald tests by examining voxelwise intensity estimation against the null hypothesis of homogeneity over space.

The signed Wald statistic for μ_{ij}^X or η_{ij}^X takes the form:

$$Z_{\mu^X} = \frac{\mu_{ij}^X - \mu_0}{\text{SE}(\mu_{ij}^X)}, \quad Z_{\eta^X} = \frac{\eta_{ij}^X - \eta_0}{\text{SE}(\eta_{ij}^X)} \tag{3.24}$$

where $SE(\mu_{ij}^X)$ is the standard error of the estimated spatial intensity μ_{ij}^X, and $SE(\eta_{ij}^X)$ is the standard error of the estimated linear response η_{ij}^X, and the statistics are asymptotically Gaussian. Finally, we can create p-value maps that are thresholded to control the false discovery rate (FDR) at 5% [Benjamini and Hochberg, 1995].

3.2.5.3 Inference on publication-level covariates

For the regression coefficient γ $(s \times 1)$ of publication-level covariates, we consider general linear hypothesis (GLH) tests through a contrast matrix C_γ $(m \times s)$. Under the null hypothesis,

$$H_0 : C_\gamma \gamma = \mathbf{0}_{m \times 1} \tag{3.25}$$

The test statistic follows a χ^2 distribution with m degree of freedom asymptotically,

$$(C_\gamma \hat{\gamma})^T (C_\gamma \text{Cov}(\hat{\gamma}) C_\gamma^T)^{-1} (C_\gamma \hat{\gamma}) \xrightarrow{D} \chi_m^2 \tag{3.26}$$

and in the case of a single contrast $(m = 1)$, a signed Z test can be computed. Details of GLH on publication-level covariates can be found in A.4.1 in the Supplementary material.

3.3 Experiments

3.3.1 Simulation settings

The statistical analyses of model estimation with CBMA data are conducted at the voxel level: voxelwise test statistics are evaluated to examine the significance of the experimental effect. Therefore, before investigating model fitness, we evaluate our models' false positive rates (FPR) under null settings. Due to the computationally intensive nature of these evaluations, we only evaluated the two models that showed promise in other evaluations, Poisson and NB. Under the null hypothesis of spatial homogeneity, we use Monte Carlo (MC) simulation to establish the validity of FPR

control for the test of spatial intensity (μ^X). Specifically, we will explore meta-regression with either the Poisson or NB model, with or without publication-level covariates. To ensure the validity of FPR control is applicable to all CBMA data, the sampling mechanism is either model-based or empirical, with simulated foci count always analogous to the foci count within a real dataset. Specifically, in model-based sampling, the data generating mechanism matches the regression model, with the number of publications and average foci per publication identical to the original dataset; while in empirical sampling, real data foci locations are randomly shuffled to guarantee the spatial homogeneity of the foci distribution.

3.3.2 Applications to 20 meta-analytic datasets

Cognition concerns psychological and cognitive processes that focus on learning people's perception, interpretation and response to information and stimuli. It refers to both conscious procedure and unconscious, automatic mechanisms in the brain that occur as a response to stimuli, and is highly variable across individuals [Gallagher et al., 2019]. Cognition has been studied intensively to identify brain regions involved in cognition tasks, conducted in an MRI scanner. Here we use 20 previously published meta-analytic datasets for the purpose of evaluating the accuracy and sensitivity of this meta-regression framework, as well as analysing the goodness of fit of stochastic models with respect to different CBMA datasets. These datasets involve multiple aspects of cognition research, as listed in Table 3.1.

The preprocessing steps are summarised in Figure 3.2. The discrete sampling space of our analysis is the $2mm^3$ MNI (Montreal Neurological Institute) atlas [Collins et al., 1994], with dimensions $91 \times 109 \times 91$, and $N = 228,483$ brain voxels. We first apply this brain mask to remove foci outside the brain and remove any multiple-foci (while original data peaks are always distinct, a foci count in excess of 1 can occur when Talairach coordinates are rounded to the MNI 2mm grid). We then extract all the sufficient statistics after model factorisation in Section 3.2.3, including the spatial design matrix X $(N \times P)$ generated from B-spline bases, total foci count per voxel $Y_{\cdot,}(N \times 1)$ and total foci count per publication $Y_{,\cdot}$ $(M \times 1)$ and publication-level covariates $Z(M \times R)$ if considered.

Table 3.1: Number of experiments and foci counts of 20 meta-analytic datasets.

Dataset	number of experiments	total count of foci	max foci count	average foci count
1. Social Processing	599	4934	47	8.24
2. PTSD	22	154	26	7.00
3. Substance Use	89	657	110	7.38
4. Dementia	28	1194	548	42.64
5. Cue Reactivity	275	3197	58	11.63
6. Emotion Regulation	338	3543	87	10.48
7. Decision Making	145	1225	49	8.45
8. Reward	850	6791	59	7.99
9. Sleep Deprivation	44	454	59	10.32
10. Naturalistic	122	1220	59	10.00
11. Problem Solving	282	3043	44	10.79
12. Emotion	1738	22038	203	12.68
13. Cannabis Use	81	314	16	3.88
14. Nicotine Use	13	77	23	5.92
15. Frontal Pole CBP	795	9525	57	11.98
16. Face Perception	385	2920	50	7.58
17. Nicotine Administration	75	349	24	4.65
18. Executive Function	243	2629	54	10.82
19. Finger Tapping	76	696	27	9.16
20. n-Back	29	640	69	22.07

3.4 Results

3.4.1 Simulation results

For each of the 20 meta-analytic datasets, we simulate foci distribution under a null hypothesis of spatial homogeneity, estimate spatial intensity and investigate the distribution of voxel-wise p-values for the eight different scenarios: fitting Poisson or NB model, using a model-based or empirical (random shuffling) data sampling mechanism, and including or omitting publication-level covariates. For all settings, we use a B-spline knot spacing of $20mm$ in x, y, z direction, producing $P = 456$ basis elements. The computation of test statistics depends on the covariance of regression coefficients, which is approximated by the inverse of the Fisher Information matrix of optimised parameters at maximised log-likelihood (see Section 3.2.4). Empirically, we sometimes found the p-values are underestimated, particularly below the threshold of 10^{-3}, which we believe has two causes. Firstly, the inference based on the inverse Fisher Information (FI) matrix is only asymptotic, and hence under- or over-coverage could be obtained for any finite number of publications N. Secondly, small meta-analysis with some regions having essentially no foci drive some of the β coefficients

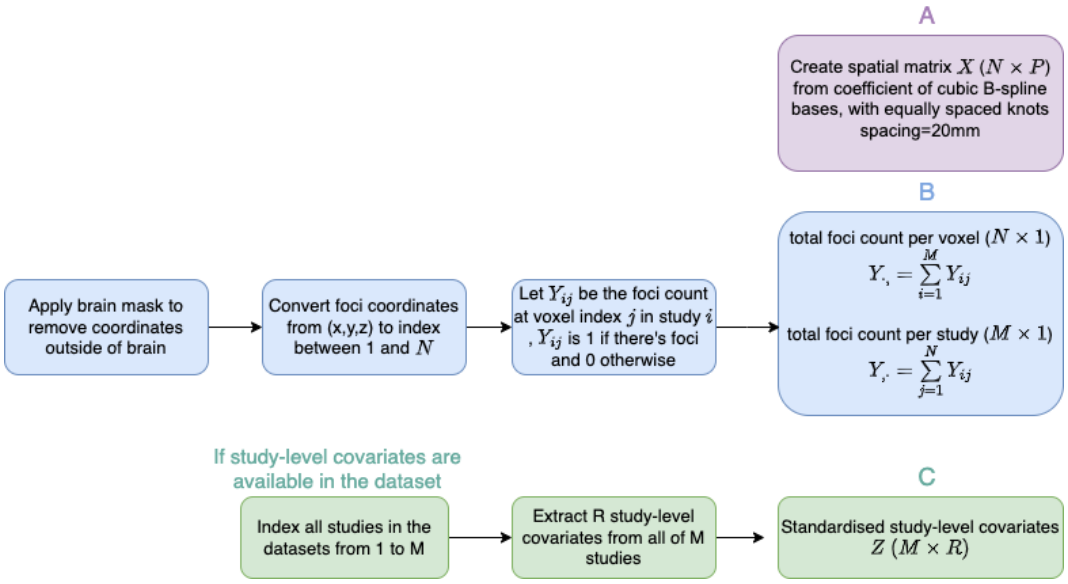


Figure 3.2: Preprocessing pipeline of meta-analytic datasets before fitting CBMR framework. Note that panel A and B are applicable to all datasets, which generate a spatial design matrix X, total foci count per voxel $Y_{\cdot,}$ $(N \times 1)$ and total foci count per publication $Y_{,\cdot}$ $(M \times 1)$. Panel C is only needed if the effect of publication-level covariates is considered, as covariates matrix Z $(M \times R)$.

to negative infinity, producing an estimated rate of zero, which in turn produces an ill-conditioned and singular FI matrix. In our experiments, we observed that datasets with a total foci count of at least 1000 generally avoided these singularity problems and produced accurate standard errors for NB model, however, this criterion also depends on the chosen spline knot spacings (we also provide a practical guideline of choosing appropriate knot spacings based on the total foci counts in Section A.4.1 in the Supplementary material). We tried various different approaches to regularise and make the FI matrix invertible but these often deflated the computed sample variances, inflating significance, and hence are not part of the proposed method.

To establish the validity of spatial homogeneity tests ($\mu_j^X = \mu_0$, $\forall j = 1, \cdots, N$) for each of the 20 meta-analytic datasets, we compute p-values and create P-P plots. We compute 100 null realisations, each producing N p-values (one for each voxel), with the null expected $-\log_{10}$ p-values ranging from $-\log_{10}(N/(N+1)) \approx 0$ to $-\log_{10}(1/(N+1)) = 5.359$. To avoid the overplotting of 100 curves on the $-\log_{10}$ P-P plots, for each ordered p-value index on the abscissa we compute the average and standard deviation (SD) of the 100 corresponding $-\log_{10}$ p-values, plotting the mean and confidence bounds at ± 1.96 SD. We rejected the null hypothesis of spatial

homogeneity at a 5% significance level, and calculated the percentage of rejected voxels out of the 228, 483 voxels located within the brain. Since the P-P plots are very similar for each of the eight scenarios, we only display the results for the setting of CBMR with an NB model without publication-level covariates, sampled with a model-based approach. Figure 3.3 shows the four representative $-\log_{10}$ P-P plots (results for all 20 datasets shown in Figure A.5 in the Supplementary material), with identity (dashed diagonal line), 5% significance (dashed horizontal line) and the FDR 5% boundary (solid diagonal line); gray shaded areas plot the point-wise 95% prediction intervals. It shows that p-values $< 0.05 \approx 10^{-1.3}$ are valid, and extreme p-values can skew liberal; the worst affected cases are datasets with very few foci (e.g. analysis 14). In general, datasets with total foci counts less than 1000 show poor behaviour.

Since multiple testing correction requires valid p-values far smaller than 0.05, we focus on controlling the FDR in these null simulations. None of the 20 datasets have valid FDR control (PP-plots or prediction intervals fall above the 5% Benjamini-Hochberg threshold). However, the PP plots generally show valid p-values $< 10^{-3}$, and if we truncate p-values by replacing any p-value smaller than 10^{-3} with that value, we obtain valid (if conservative) FDR control (Table 3.2). This pragmatic approach could impact power, but empirical results (Section 3.4.2) suggest that the inferences based on truncated p-values remain sensitive.

Table 3.2: The percentage of invalid FDR control (before/after p-value truncated at 10^{-3}) in 20 meta-analytic datasets over 100 realisations.

Dataset	Before	After	Dataset	Before	After
1. Social Processing	44%	0%	2. PTSD	100%	0%
3. Substance Use	26%	0%	4. Dementia	16%	0%
5. Cue Reactivity	28%	0%	6. Emotion Regulation	23%	0%
7. Decision Making	18%	0%	8. Reward	43%	0%
9. Sleep Deprivation	30%	0%	10. Naturalistic	22%	0%
11. Problem Solving	26%	0%	12. Emotion	100%	0%
13. Cannabis Use	63%	0%	14. Nicotine Use	94%	0%
15. Frontal Pole CBP	90%	0%	16. Face Perception	19%	0%
17. Nicotine Administration	54%	0%	18. Executive Function	22%	0%
19. Finger Tapping	22%	0%	20. n-Back	27%	0%

3.4.2 Results from 20 meta-analytic datasets

We first evaluate the goodness of fit among likelihood-based stochastic models (Poisson, NB and clustered NB model) via comparisons of maximised log-likelihood and AIC. As

shown in Figure A.6 and Figure A.7 in A.4.3 of the Supplementary material, CBMR with the NB model outperforms the other two likelihood-based stochastic models in every dataset. This is not surprising as the NB model is the only likelihood-based model that allows for the anticipated excess variance relative to Poisson at the voxel level; clustered NB is better than Poisson for the majority of these 20 meta-analytic datasets, but only by a small margin. It is conceivable that although a publication-wise global dispersion parameter exists in the clustered NB models, CBMA data is just as well specified by a Poisson model at the voxel level. LRT comparison of nested models rejects the null Poisson model vs. NB for all datasets, with p-value less than 10^{-8}; the Poisson null is rejected in favour of the clustered NB model for the majority of the 20 meta-analytic datasets (with p-value less than 10^{-8}) (see Table A.9 in Appendix A.4.3 of the Supplementary material).

For all methods we also use three metrics to assess model fit:

- **Relative Absolute Bias of Intensity Sum:** This metric compares the sum of CBMR estimated intensity over the space to the total number of observed foci counts within the dataset.
- **Relative Absolute Bias of Intensity Standard Deviation (SD) in the x, y, z directions:** This metric evaluates the SD of CBMR intensity estimation compared to the empirical distribution of foci counts in the dataset in all three directions.
- **Relative Absolute Bias of Variance at the voxel-wise level:** This metric measures the discrepancy between the asymptotic variance of the fitted CBMR model empirical variance of foci counts in the dataset, calculated at each voxel and averaged over voxels with at least one foci.

These metrics were calculated for each of the 20 datasets and are presented using box plots to illustrate the variability and distribution of the results.

Plots in Figure 3.4a suggest that the four evaluated stochastic models (Poisson, NB, clustered NB and Quasi-Poisson model) produce consistently estimate the intensity accurately, with the median relative bias of estimated publication-wise total foci count less than 1.0%, among which the Poisson model has the lowest median relative absolute bias (0.05%), across 20 meta-analytic datasets. However, all four stochastic

models tend to slightly overestimate the publication-wise total foci count in these datasets. The Quasi-Poisson, in particular, shows a more variable relative absolute bias across the 20 datasets. The results in Figure 3.4b suggest that the CBMR framework also provides an accurate estimation of standard variation (SD) of intensity across the x, y, z dimensions. The relative bias is controlled below 0.25% for all stochastic models in 20 meta-analytic datasets, and estimated intensity along the x axis are the most accurate (with the smallest SD bias). As shown in Figure 3.4c, CBMR with the Poisson model and clustered NB model display a negative bias in variance, which suggests that excess variance cannot be explained by the Poisson assumption. The publication-specific over-dispersion modelled by the clustered NB is insufficient, as this model also has negative bias. Small relative bias is found in both NB and Quasi-Poisson model (with median 0.78% and 1.25%), with less variation in relative bias across multiple datasets with the NB model, which suggests both models are capable of dealing with excess variance in CBMA data.

Overall, we regard these evaluations as evidence that NB model is preferred. While it has slightly more bias in the publication-wise total foci count (Figure 3.4a), its variance estimation is considerably more precise compared to the Poisson model (Figure 3.4c).

3.4.3 Comparison with ALE

Activation likelihood estimation (ALE) is one of the widely used kernel-based CBMA methods. Under the null hypothesis of ALE, any observed clustering of activation foci is purely due to chance rather than reflecting true convergence or systematic effects. For each focus, ALE generates a map with a Gaussian kernel centred at the location, and then combines pairs of maps using the probability of a union of events rule ($P(A \cup B) = P(A) + P(B) - P(A \cap B)$) at each voxel [Turkeltaub et al., 2002]. It appears to model the probability that one or more foci arise at a given voxel, conditional on the total number of foci over all publications.

While the null hypothesis of CBMR assumes spatial homogeneity of activation intensity across the brain. Specifically, it assumes the estimated activation intensity at each voxel does not exceed the average intensity. Despite these differences in the null hypotheses, we compared our CBMR results to ALE, conducting tests for spatial homogeneity across space with both approaches. For simplicity, we only demonstrate the comparison of detected activation regions on the Cue Reactivity dataset (total

foci count of 3197) [Hill-Bowen et al., 2021], and only the z-value map generated by the CBMR with the NB model is presented here as a representative example. For comparison purposes, we show z-statistic values at all voxels significant at $\alpha = 0.05$ uncorrected in Figure 3.5. Here, we choose FWHM=14 to obtain comparative spatial resolution between ALE and CBMR. Evidence for consistent activation is found in the left cerebral cortex, frontal orbital cortex, insular cortex, left and right accumbens, with exact activation regions differing slightly between ALE and CBMR, and ALE detecting more voxels. This disparity reflects their different sensitivities to spatial clustering. ALE's null hypothesis is highly responsive to even modest clusters of reported foci, meaning mild spatial clusters can readily produce significant results. CBMR, in contrast, requires a consistent elevation of voxel intensity above the global mean, thereby imposing a stricter threshold that typically yields fewer significant voxels. For additional comparisons of ALE and CBMR activation regions across other meta-analytic datasets (among the 20 datasets), please refer to Appendix A.5.

Another criterion of consistency is the Dice Similarity Coefficient (DSC), the intersection of ALE and CMBR significant voxels divided by the average number of significant voxels. As shown in Table 3.3, ALE appears generally more sensitive than CBMR, regardless of foci counts in the datasets, though DSC varies from 71.89% to 80.33% on the datasets with more than 1200 foci counts, which demonstrates good similarity between the methods. While for datasets with less than 1200 foci counts, the DSC increases as the number of foci grows, this is potentially caused by numerical instability in CBMR's standard error estimation, which rely on the inverse Fisher information matrix. A promising avenue for future work is to develop a more robust, reliable standard error estimation for CBMR when the number of foci is limited.

ALE evaluates the experimental effect by testing probabilistic maps (generated by a Gaussian kernel) against the null hypothesis, CBMR estimates activation intensity and conducts hypothesis testing at the voxel level. Both methods ultimately produce statistical maps representing either probabilities or intensities, allowing for similar meta-analytic interpretations. Specifically, in analysing a Social processing dataset with $4,934$ activation foci (Lobo et al. [2021]), both ALE and CBMR identify the activation in key social processing regions, including the medial prefrontal cortex, temporoparietal junction, lateral occipital cortex, and precuneus. These findings align well with established neural processing pathways: lateral occipital regions initially detect visually-defined social cues, subsequently feeding information to the temporoparietal junction and precuneus for constructing perspectives and contextual

Table 3.3: Number of voxels in activation regions of ALE (FWHM=14) and CBMR (with the NB model), denoted as $|AR_{ALE}|$ and $|AR_{CBMR}|$, based on uncorrected p-values with 5% significance level, as well as Dice similarity coefficient in 20 meta-analytic datasets (Datasets are listed in an ascending order according to total number of foci).

Dataset	n_foci	$\lvert AR_{CBMR}\rvert$	$\lvert AR_{ALE}\rvert$	$\lvert AR_{CBMR} \cap AR_{ALE}\rvert$	DSC
14. Nicotine Use	77	1312	12431	1154	17.79%
2. PTSD	154	6306	15866	5067	45.71%
13. Cannabis Use	314	11841	18390	8235	54.48%
17. Nicotine Administration	349	11546	18916	8028	52.71%
9. Sleep Deprivation	454	10250	15461	5732	44.59%
20. n-Back	640	19404	31512	17627	69.24%
3. Substance Use	657	19024	26477	13602	59.79%
19. Finger Tapping	696	19067	33914	17939	67.72%
4. Dementia	1194	16244	30437	12464	53.41%
10. Naturalistic	1220	22328	29442	15344	59.28%
7. Decision Making	1225	28284	36735	23372	71.89%
12. Emotion	2038	57698	67699	48847	77.91%
18. Executive Function	2629	33848	46679	31698	78.73%
16. Face Perception	2920	41682	53109	36710	77.45%
11. Problem Solving	3043	38466	51315	34757	77.43%
5. Cue Reactivity	3197	41242	52371	37301	79.69%
6. Emotion Regulation	3543	36602	48157	31176	73.56%
1. Social Processing	4934	48376	61136	40740	74.40%
15. Frontal Pole CBP	9525	53165	65339	47595	80.33%
8. Reward	6791	43048	51721	37711	79.59%

understanding, and ultimately converging in the medial prefrotal cortex for higher-order inferential processes and valuation [Frith and Frith, 2007, Van Overwalle, 2009].

Some researchers have proposed a stringent threshold ($\alpha = 0.0001$) on uncorrected p-values to reduce type I error [Turkeltaub et al., 2002], while a more principled approach is to control the false discovery rate (FDR) via Benjamini-Hochberg (BH) procedure. Figure 3.6 shows a comparison of results using a 5% FDR threshold, where CMBR (NB) p-values use a 10^{-3} truncation, and Table 3.4 shows a comparison of the number of detected voxels.

It is seen that CBMR generally detects fewer voxels than ALE Table 3.4, however these two approaches are not directly comparable. In previous work ([Samartsidis et al., 2017]) we demonstrated that ALE behaves like a fixed-effect model, where significance can be driven by a tiny fraction of "real" publications mixed with purely noise publications. In contrast, with our model, heterogeneity can be captured by the NB excess variance term and make inferences more sceptical. As a result, direct

comparisons of sensitivity seem akin to comparing the power of a fixed effects model (that neglects an important source of variation) to a mixed effects model, where the fixed effects model will always be more powerful by design. We also add that CBMR is grounded in a generative statistical model, accommodates publication-level covariates and produces standard errors on interpretable parameters.

Instead of relative power, our goal here is to demonstrate that the detected activation regions produced by both methods are roughly consistent. For this purpose, we found that the DSC varies between 70.55% and 79.76% for datasets with more than 1225 foci counts, indicating consistency of activation regions between ALE and CBMR approach after FDR correction.

Table 3.4: Number of voxels in activation regions of ALE (FWHM=14) and CBMR (with the NB model), denoted as $|AR_{ALE}|$ and $|AR_{CBMR}|$, based on FDR corrected p-values (using BH procedure) with 5% significance level, as well as Dice similarity coefficient in 20 meta-analytic datasets (Datasets are listed in an ascending order according to total number of foci). Roughly, datasets with at least 1000 foci show reasonable similarity between ALE and CBMR.

Dataset	n_foci	$\lvert AR_{CBMR}\rvert$	$\lvert AR_{ALE}\rvert$	$\lvert AR_{CBMR} \cap AR_{ALE}\rvert$	DSC
14. Nicotine Use	77	209	0	0	0.00%
2. PTSD	154	0	1201	0	0.00%
13. Cannabis Use	314	313	152	17	7.31%
17. Nicotine Administration	349	1338	943	522	45.77%
9. Sleep Deprivation	454	176	0	0	0.00%
20. n-Back	640	11456	17725	10212	69.99%
3. Substance Use	657	3145	2082	1225	46.87%
19. Finger Tapping	696	12410	23837	11590	63.95%
4. Dementia	1194	5126	7931	3142	48.13%
10. Naturalistic	1220	4192	3241	1861	50.07%
7. Decision Making	1225	15331	20468	12628	70.55%
18. Executive Function	2629	26039	37797	24690	77.67%
16. Face Perception	2920	28893	38193	25533	76.12%
11. Problem Solving	3043	28221	39091	25675	76.29%
5. Cue Reactivity	3197	30382	38847	27375	78.57%
6. Emotion Regulation	3543	23388	31620	20056	72.92%
1. Social Processing	4943	34317	45263	28555	71.76%
8. Reward	6791	33021	39743	28728	78.96%
15. Frontal Pole CBP	9525	44030	55251	39594	79.76%
12. Emotion	22038	50480	57321	41918	77.77%

3.4.4 Effect of publication-level covariates

Here we demonstrate how CBMR, unlike ALE, can estimate the effect of publication-level covariates. We integrate two publication-level covariates, publication-wise (square root) sample size and year of publication (after centring and standardisation) into the CBMR framework on each of the 20 meta-analytic datasets. We find, for example, on Cue Reactivity dataset, the year of publication is not significant ($Z = -0.6880, p = 0.4915$), while sample size is significant ($Z = 6.1454, p < 10^{-8}$); interpreting the γ parameter for sample size finds that a doubling of sample size results in an expected 26.15% increase in the publication-wise spatial intensity (See Table A.10 for $p-$values and $Z-$scores of publication-level covariates on each of the 20 meta-analytic datasets).

3.5 Discussion

In this work we have presented a meta-regression framework with a spatial model as a general approach for CBMA data, where we have considered multiple stochastic models and allowed for publication-level factors (e.g., sample size and year of publication). Our approach uses spline parameterization to model the smooth spatial distribution of activation foci, and fits a generalised linear model with different variants of voxelwise (Poisson model, NB model and Quasi-Poisson model) or publication-wise (Clustered NB model) statistical distributions. Our approach is a computationally efficient alternative to previous Bayesian spatial regression models, providing the flexibility and interpretability of a regression model while jointly modelling all of space. For comparison, using the Cue Reactivity dataset as an example, the implementation of Bayesian log-Gaussian Cox process regression required approximately 30 hours on an NVIDIA Tesla K20c GPU card [Samartsidis et al., 2019], in contrast to approximately 537.52 seconds (about 9 minutes) required for CBMR with the NB model on an Intel Xeon Gold 6340R CPU. Furthermore, grounded in a generalised linear model, we believe that our meta-regression framework is more comprehensible to practitioners, relative to inference on the spatial posterior intensity function.

Through simulations on synthetic data (with simulated foci counts analogous to those in each of 20 meta-analytic datasets), we demonstrated valid FDR control for the spatial homogeneity null hypothesis after a truncation of $p-$values below 10^{-3}. According to 20 meta-analytic datasets, we found that the NB model is the most accurate stochastic

model in model comparisons via LRT and AIC, as well as having the smallest relative bias in both mean and variance of intensity estimation (per publication), while the Poisson and clustered NB model cannot explain the over-dispersion observed in foci count. Meanwhile, we also compared the findings of activation regions from both the ALE and CBMR approach, and justified the validity and robustness of CBMR, especially on the datasets with relatively high foci count, e.g., datasets with at least 1000 total foci.

There are a few limitations in our work. Here we have only considered a single group of publications. In future work, we plan to extend our method to estimate the spatial intensity function of multiple groups (e.g., multiple types of stimuli within a cognitive task), so that we can investigate the consistency and difference in activation regions through group comparison. Additionally, our current analysis is limited to the global effects of publication-level covariates, a pragmatic decision given common application with 10's-100's of publications. We recognise, however, that this approach might not be appropriate in cases where there are significant spatial variations in covariate effects. Ideally we would add a basis function to express each covariate effect, though this would likely be infeasible without many 1000's of publications. Alternatively we could use a coarse parcellations (e.g. 3-6 regions) and allow parcel-specific regression coefficients for each region.

We are currently not using regularisation term on spatial regression coefficients of CBMR. Initially we considered a Firth-type penalty which indeed guarantees convergent estimates (especially in brain regions without any foci) and removes the first-order asymptotic bias term of maximum likelihood estimates, but we found it also causes significant overestimation of intensity at the edge of brain mask. The edge effect induced by Firth-type penalty relates to the structure of the Jeffreys prior and higher variance associated with edge and corner basis elements. However, it's plausible to consider regularising likelihood functions with alternative penalty terms (e.g., L_1 or L_2 norm) in the future, though requiring hyper-parameter tuning. We estimate the variance of voxel-wise spatial intensity using the covariance of spatial regression coefficients found by inverting the Fisher Information matrix. This can be numerically unstable because the dimension of Fisher Information matrix is large (there are hundreds or even thousands elements in spline bases), and it might even be numerically singular for datasets with low foci count since most voxels have near-zero intensity estimation. We have tried many approaches to improve numerical stability, including adding an extremely small epsilon (10^{-6}) or 1% of the largest diagonal

element on the diagonal of the Fisher Information matrix, or computing the Fisher Information assuming the null hypothesis of homogeneity is true. However, all of these efforts produced underestimation of the variance of voxel-wise spatial intensity and led to invalid p-values. In future work, we might consider non-parametric methods to estimate the covariance of spatial regression coefficient instead of the inverse of Fisher Information, or add a regularisation term on B-spline roughness to avoid very negative spatial regression coefficients.

Another important direction is a combined IBMA-CBMA analysis, were we extend our model to include continuous effect size maps. One possible approach is to combine separate coordinate and intensity models using Markov melding in a fully Bayesian framework for joining probabilistic sub-models. In this approach, evidence from each different source is specified in each sub-model, and the sub-models are joined while preserving all information and uncertainty [Goudie et al., 2019]. Such an approach might enrich the inference obtained from CBMR by integrating the magnitude of CBMA activation or even image-based meta-analysis data.

Another direction of interest is investigating the variability caused by different meta-analysis pipelines. This consideration is important, as we have observed significant variation in activation regions due to the different sensitivity in analysis pipelines in each publication. In fact, the CBMR's publication-level covariates already allow it to accommodate variations in analysis pipelines by including the specific pipeline used as a publication-level covariates and understanding its impact at a global level.

Finally, another direction to consider is a zero-inflated stochastic model (e.g., Poisson or NB model) as the current datasets only consist of publications with at least one focus, there might be inflated zero foci count than observed. Excess zeros are separated and modelled independently in zero-inflated models, which might provide a more accurate approximation for low-rate Binomial data, as was found useful when modelling image-wise total foci count [Samartsidis et al., 2020a].

3.6 Software

Implementation in the form of Python and Pytorch code can be found in Github repository. CBMR framework has also been implemented and integrated into NiMARE python package.

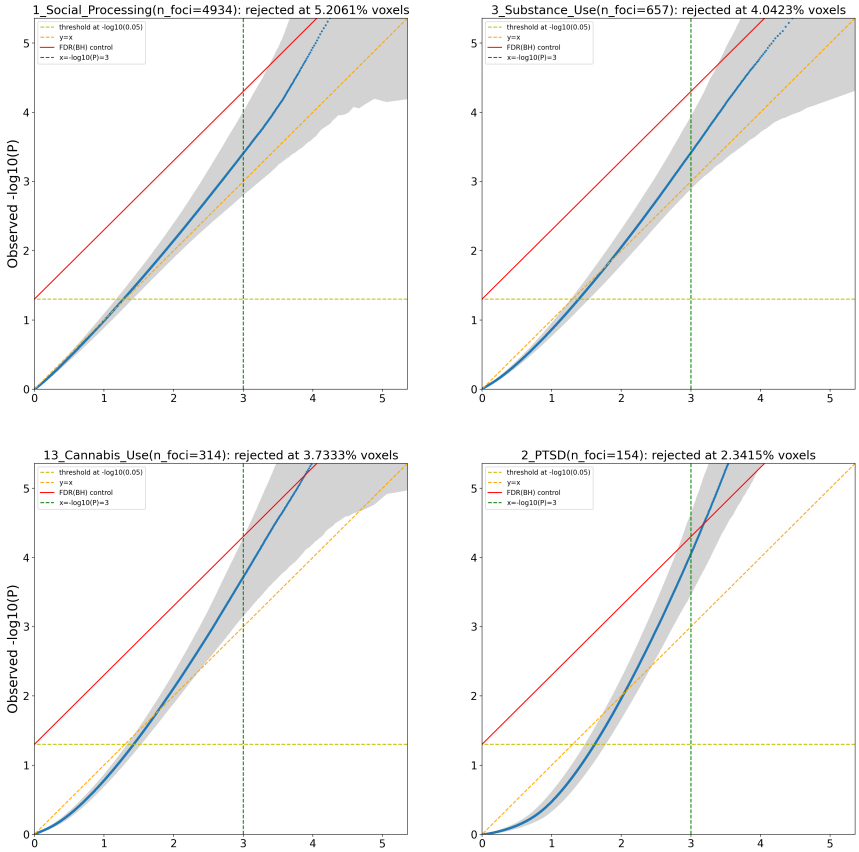
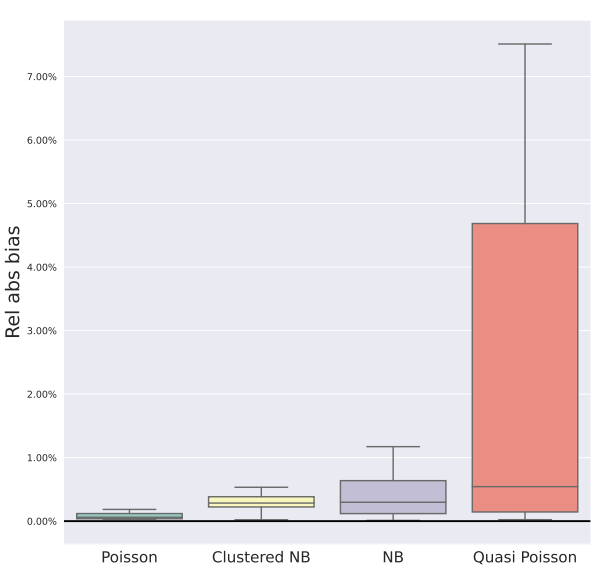
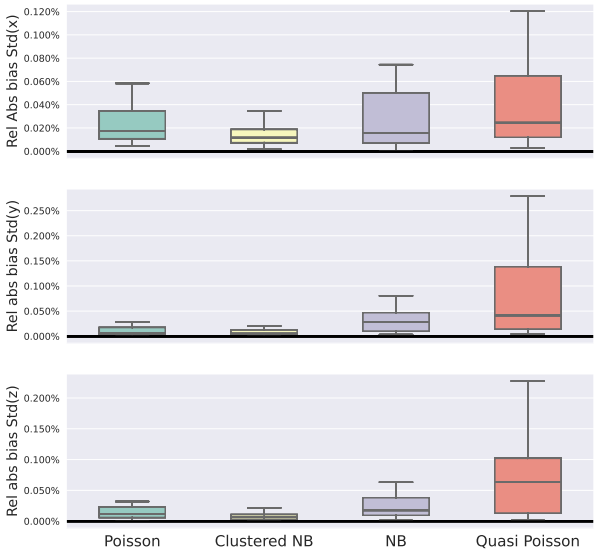


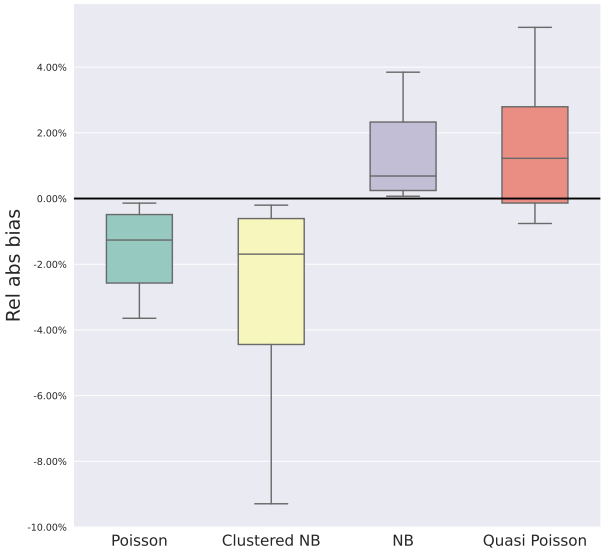
Figure 3.3: P-P plot of null p-values, $-\log_{10}$ scale, showing four representative meta-analytic datasets (Social Processing, Substance Use, Cannabis Use and PTSD datasets), estimated by CBMR with NB model without publication-level covariates, with null data generated with a model-based approach. CBMR's p-values are generally valid for $p < 0.001$, especially for publications with 1000's of foci.



(a) Relative absolute bias of intensity sum

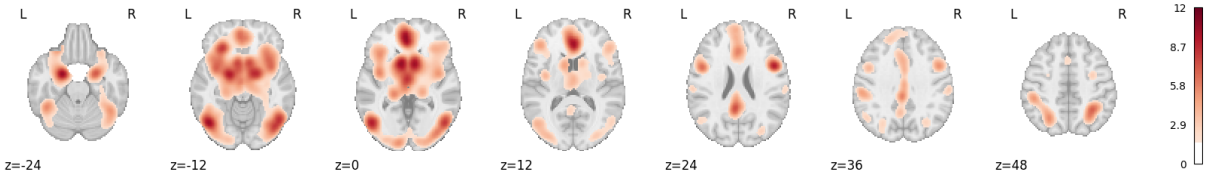


(b) Relative absolute bias of intensity SD in x,y,z directions

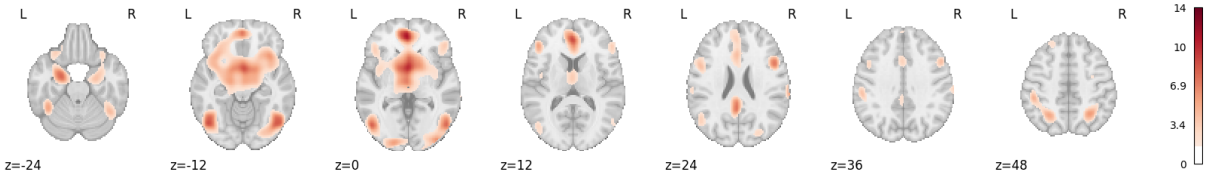


(c) Relative absolute bias of voxelwise intensity variance

Figure 3.4: Results from bias-related model comparison criteria, fitted with four stochastic models on each of 20 meta-analytic datasets: (a) Box plot of relative absolute bias of estimated intensity sum (per publication); (b) Box plot of relative absolute bias of estimated intensity SD in x,y,z directions (per publication); (c) Box plot of relative absolute bias of voxelwise estimated intensity variance (per publication).

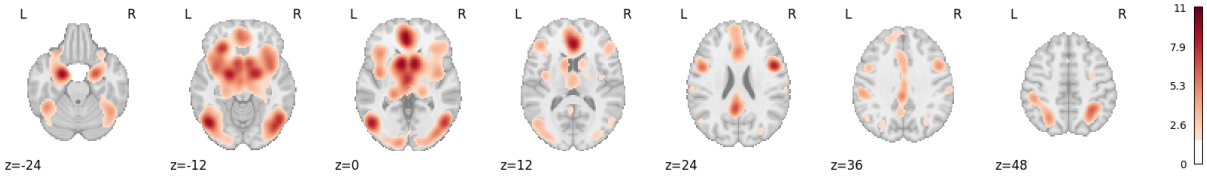


(a) Z-score map generated by ALE

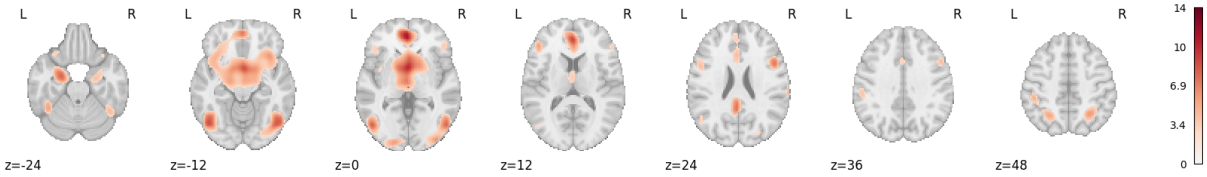


(b) Z-score map generated by CBMR (with NB model)

Figure 3.5: Activation maps (for significant uncorrected p-values, $p \leq 5\%$, displayed as Z-scores) generated by ALE (with FWHM=14) and CBMR (with NB model) on the Cue Reactivity dataset with axial slices at $z = -24, -12, 0, 12, 24, 36, 48$. Both methods identify similar regions for significant evidence against the null of spatial homogeneity. While ALE finds more significant voxels, it represents a fixed-effects type of analysis not directly comparable with CBMR inferences (see text).



(a) Z-score map generated by ALE



(b) Z-score generated by CBMR (with NB model)

Figure 3.6: Activation maps (for significant FDR corrected p-values under 5% significance level, presented in Z-scores) generated by ALE and CBMR with FDR correction (by BH procedure) with truncated p-values of Cue Reactivity dataset. The figure is shown with axial slices at $z = -24, -12, 0, 12, 24, 36, 48$. Under the null hypothesis of spatial homogeneity, activation regions with z-scores corresponding to corrected p-values below the significance level 0.05 are highlighted.

Chapter 4

CBMR: Meta Regression and Inference for Coordinate Based Meta Analysis Data Across Multiple Groups

The previous chapter introduced a classical frequentist meta-regression framework (CBMR) that explicitly captures the spatial structure of activation foci distributions. Formulated as a generalised linear model (GLM), the CBMR framework incorporates a spline-based spatial component and publication-level covariates, while addressing over-dispersion in foci counts through the use of Negative Binomial (NB) and Clustered NB models. In this chapter, we extend the framework by introducing a roughness penalty on the spatial spline bases to enhance numerical stability, improve the accuracy of statistical inference, and enable reliable group-wise intensity estimation and comparison, even in scenarios with limited foci per publication. Additionally, we replace traditional Wald-based inference with parametric bootstrapping to address covariance underestimation in small datasets, allowing for more robust testing of spatial homogeneity and group equality. The proposed enhancements are validated through simulation studies and application to a Cue Reactivity dataset, demonstrating its advantages over existing kernel-based methods.

4.1 Introduction

4.1.1 Background

To address the limitations of both kernel-based and Bayesian model-based CBMA approaches, we have proposed a classical frequentist meta-regression framework that explicitly captures the spatial structure of the activation foci distribution [Yu et al., 2024]. This model is formulated as a generalised linear model (GLM) and consists of two key components: a spatial effect, which incorporates a spline parameterization to generate a smooth response across the entire brain image; and a global effect to account for publication-level covariates specific to each publication. Four different stochastic models within the GLM framework have been considered: While Poisson is the classic distribution for approximating foci distribution (as a low-rate Binomial distribution) at the voxel level, Samartsidis et al. [2020b] has found evidence of over-dispersion in CBMA data. To address this, we further explore a Negative Binomial model to account for the excess variation in foci data. We omit comparisons with Clustered NB model and Quasi-Poisson models, as they have been shown to be incapable of accommodating voxel-wise independent excess variance within a publication and generated a poorer model fit [Yu et al., 2024]. Meanwhile, there are practical challenges in the implementation and optimisation of this meta-regression approach for real fMRI datasets. With fewer than 10 reported foci per publication on average, the values of spatial regressors become highly negative during optimisation. This leads to difficulties in convergence and poses challenges for statistical inference, particularly when estimating the covariance structure between different voxels. In this work, we demonstrate that applying a roughness penalty to the spatial spline parameterization improves the numerical stability of the meta-regression and enhance the precision of statistical inference. This modification also enables the estimation of group-wise intensity functions for multiple groups, and facilitates group comparisons for spatial activation intensity. By introducing this penalty, we overcome the strict limitation imposed by the minimum number of foci per group for meta-regression, making comparisons across multiple groups possible.

A further limitation of our previous meta-regression framework was that the covariance structure of spatial intensity across different voxel locations, estimated from the inverse Fisher Information, could encounter numerical issues in small datasets, leading to underestimation of covariance. To address this, we replaced the parametric inference

based on the Wald test with parametric and non-parametric bootstrapping approaches, allowing for tests of spatial homogeneity within each group and group equality in multi-group datasets. Additionally, we explored parallelizing code execution to accelerate the bootstrapping process, making it a computationally feasible alternative to the traditional parametric Wald test. We then demonstrate the validity of bootstrap-based statistical inference on both simulated and real datasets, comparing its activation maps with that of traditional kernel-based methods.

In this paper, we present a coordinate-based meta regression and inference (CBMR) framework for multiple groups, a Python-based tool that allows for the estimation of both group-specific spatial regressors and regressors for publication-level covariates, as well as statistical inference for spatial homogeneity and equality of group-specific intensity functions. The CBMR tool is integrated into the Python package NiMARE [Salo et al., 2022], and will be accessible through a web-based platform, Neurosynth Compose. This platform allows customised neuroimaging meta-analyses using either self-uploaded data or data imported directly from the Neurosynth database, providing a wide range of CBMA methods with no programming experience required. Our current implementation of the CBMR framework consists of meta-regression and meta-inference modules. The meta-regression module can be executed independently to estimate group-specific intensity functions, while the meta-inference module uses the optimised regressors from the meta-regression module as input and supports flexible (single or multiple, independent or simultaneous) hypothesis testing on either spatial homogeneity or group equality, which can be easily specified with a contrast matrix.

In the following sections, we first provide background on spline parameterization for modelling spatial dependence, as well as the stochastic models, parameter estimation and inferences in CBMR. Following this, we give preliminary statistical information describing the single-group CBMR and its extension to multi-group settings. In the method section, we outline the computational pipeline of CBMR, starting with input specification and dimension reduction, followed by parallelised execution of optimisation, parameter estimation and finally, inference using either the parametric Wald test or a bootstrapping approach. Next, we evaluate the validity and performance of CBMR through simulations and comparisons with existing kernel-based and model-based approach on real datasets. Finally, we conclude with a real dataset example of cue reactivity task.

4.1.1.1 Spatial model: spline parameterization

Gaussian and uniform kernels are commonly used in kernel-based CBMA methods to model the spatial distribution of reported foci, smoothing and estimating the probability of activation around each focus to capture spatial uncertainty in neuroimaging data effectively. In contrast, model-based CBMA methods have previously treated each publication's foci as realisations of a doubly-stochastic Poisson process, also known as a Cox process, in spatial point process modelling of CBMA data. In some of these model-based approaches, the log intensity function is parametrised either by superimposed Gaussian kernel basis functions or as a Gaussian process [Montagna et al., 2018, Samartsidis et al., 2019]. These previous studies highlight the importance of applying spatial models to explain spatial uncertainty in neuroimaging data.

Here, we propose a spatial model parametrised by a tensor product of cubic B-spline basis functions. This spatial basis is chosen for its smoothness, stability, and flexibility, as the level of spatial smoothness is parametrised by knots spacing: larger knots spacing generates fewer basis functions and thus greater smoothness, while closer knots produce more basis functions and enhance the model's ability to capture fine details. After setting the knot spacing uniformly across the x, y and z directions, we construct a B-spline curve as a linear combination of the B-spline basis functions in each direction. We then evaluate the coefficients at each voxel corresponding to the B-spline bases to construct a coefficient matrix for each direction. The three-dimensional coefficient matrix of B-spline bases is then constructed by taking the tensor product of the three coefficient matrices along each of the x, y, z directions, further details are outlined in [Yu et al., 2024].

We assert that the spatial model parametrised by spline bases is capable of efficiently capturing spatial uncertainty. This is supported by both its demonstrated effectiveness in previous experiments within single-group CBMR settings and comparison with alternative spatial models, such as Gaussian kernels. Minimal differences were observed between these two spatial models in both simulated and real datasets, as detailed in [Yu et al., 2024]. Accordingly, we believe it is reasonable to adopt this spatial model in the current work.

4.1.1.2 CBMR parameter estimation

A vast amount of literature exists on the development of tools and methodologies for generalised linear models (GLM). Since the formalization of GLMs in 1972 [Nelder and Wedderburn, 1972], iterative re-weighted least squares (IRLS) has been recognised as a reliable and efficient computational approach for parameter estimation, effectively addressing the complexity introduced by the non-linear relationships. IRLS became the standard method for parameter estimation in GLMs. Later, the Newton-Raphson method and its variation, Fisher scoring, were proposed and widely adopted due to their faster convergence and improved efficiency and numerical stability, particularly on highly non-linear optimisation surfaces [Jennrich and Sampson, 1976]. Since the 1990s, regularized estimation methods that add a penalty term (e.g., Lasso, Ridge, and Elastic Net) to the likelihood function have also been developed, encouraging sparsity and stability in parameter estimation [Tibshirani, 1996, Hoerl and Kennard, 1970, Zou and Hastie, 2005]. More recently, several tools and software built upon these foundational methods have been developed for GLMs parameter estimation. Among the most popular are R packages such as *glmnet* [Friedman et al., 2010], *MASS* [Ripley et al., 2013] and *lme4* [Bates, 2014], as well as Python packages *statsmodels* [Seabold and Perktold, 2010], which have made GLM parameter estimation accessible and scalable, supporting MLE, IRLS, and Bayesian methods. These tools, along with advancements in computing, enable efficient parameter estimation for GLMs, even with large datasets and complex models.

However, in meta-regression of fMRI data, parameter estimation is performed for a model with two components: the spatial effect which includes hundreds of thousands of different voxels within the brain mask for each publication, and the global effect of publication-level covariates which moderates the intensity function of a specific publication by a constant. For a large-scale, voxel-wise GLM analysis to fully optimise the computational efficiency, it is essential to vectorize computation across voxels. Many existing GLM tools and software are developed with operations that are not fully vectorised, especially when dealing with complex or large-scale data structures. Handling high-dimensional data across iterative computations without careful memory management can limit vectorization, and GLMs applied to sparse or irregular data further complicate vectorization due to the challenges brought by sparse matrices. While for likelihood functions with regularization terms (e.g., Lasso, Ridge, Elastic Net), additional iterative processes like coordinate descent are required [Friedman

et al., 2010], making it even more difficult to fully vectorize and parallelize these computations. Operations that are not amenable to vectorisation create bottlenecks in large-scale GLM optimisation, as they must be executed separately for each voxel in each publication, significantly slowing down computation. As a result, many existing software for GLMs analysis is not suitable for large-scale or complex CBMA data.

Additionally, we believe that efforts should focus on reducing the dimensionality of the variables rather than the combined product of the number of publications and voxels used as dimensions. We provide rigorous proofs demonstrating that the GLM with various stochastic models can be simplified to equivalent forms with sufficient statistics, with dimensions no grater than either the number of voxels or the number of publications [Yu et al., 2024]. We will continue to follow this approach in the current work.

As an efficient and fundamental approach for parameter estimation in GLMs, Maximum Likelihood Estimation (MLE) is widely used to optimize the model by maximizing the probability of the observed data given a set of parameters, under the assumptions of the GLM. One effective optimization algorithm for this task is the Limited-memory Broyden-Fletcher-Goldfarb-Shanno (L-BFGS), a quasi-Newton method known for its memory and computational efficiency in handling high-dimensional data and parameters, as it approximates the Hessian matrix rather than computing it explicitly. L-BFGS is also chosen for its faster convergence compared to simpler gradient descent methods, especially in scenarios with complex or irregular likelihood curvatures [Liu and Nocedal, 1989]. We have observed its effectiveness in optimizing single-group CBMR scenarios [Yu et al., 2024], and we will extend it to the current work, a more complex multi-group CBMR setting.

4.1.1.3 Inference

For CBMA, the central object of interest is to identify the brain activation regions associated with a specific cognitive task, or to find differences in activation regions in respond to similar but distinct stimuli. This requires fMRI GLM analyses to conclude with significance-based hypothesis tests, conducting either homogeneity tests or group comparison tests at the voxel-wise level using Wald test statistics. The Wald-based hypothesis testing procedures in the neuroimaging applications include tests on both single and multiple parameters. The single-parameter test assesses whether the

estimated intensity at each voxel is greater than the average intensity expected under a random distribution of foci, whilst the multiple-parameter test evaluates whether a specific linear combination of group-wise estimated intensities is distinguishable from zero at each voxel. In some circumstances, it may also be of practical interest to assess multiple flexible group comparison hypotheses simultaneously. As the effect of publication-level covariates is an additional component of CBMR, single- or multiple-parameter hypothesis testing is also applicable to these covariates, This allows for assessment of whether a specific publication-level covariate has a significant effect or whether multiple publication-level covariates have equivalent effects in CBMR, following the same Wald test procedure.

In existing GLM tools and software, Wald tests are commonly implemented to assess the significance of individual coefficients or groups of coefficients, typically for testing whether they are significantly different from zero. However, they are not generally implemented for more flexible group comparisons, such as testing if two or more groups of coefficients are equivalent, for example, the summary.glm function in R packages *stats* and Python package *statsmodel* do not support such comparisons. Additionally, as the hypothesis testing of localised spatial intensity or log-transformed spatial intensity at each voxel, a key focus in CBMA application, is not supported by most popular GLM tools.

When a GLM involves multiple hypotheses, such as testing estimated group-specific intensity against homogeneity or comparing groups voxel-by-voxel, multiple testing corrections are applied to control for false positives. Without correction, the probability of encountering at least one false positives increases with the number of tests. For example, in localised tests across $228,483$ voxels (within a MNI152 2mm brain mask), even a 5% false positive rate could result in a substantial number of false positives. In neuroimaging data, multiple testing corrections are applied by controlling either the family wise error rate, using the null maximum distribution [Westfall and Young, 1993] or the false discovery rate (FDR) using the Benjamini-Hochberg procedure [Benjamini and Hochberg, 1995]. FWER correction is a more stringent approach, as it minimises the chance of any false positives across the entire set of hypothesis tests. However, it often reduces statistical power and leads to fewer significant results, for instance, Bonferroni corrections can be overly conservative and may excessively penalise neuroimaging datasets. In contrast, FDR correction is more powerful in large-scale testing scenarios, where hundreds of thousands of tests are conducted simultaneously,

FDR correction improves statistical power and allows for more significant findings while still controlling the overall rate of false discoveries among detected results.

In our previous work on single-group CBMR, both single- and multiple-hypothesis testing of estimated spatial intensity involve the standard error of the estimated intensity or log-transformed estimated intensity. This is derived from the standard error of spatial regression coefficients using the inverse of the Fisher Information matrix, with additional transformations such as the delta method applied. However, in practice, we observed numerical singularity in the Fisher Information matrix, particularly for smaller datasets where the total number of foci is below 200 [Yu et al., 2024]. This motivates us to explore parametric bootstrapping as an alternative to parametric inference based on the Fisher Information matrix. By obtaining p-values from the tail of the null bootstrap distribution, we avoid the numerical instability caused by extremely small estimated intensity values close to zero. Although this approach increases computational complexity by requiring thousands of bootstrap samples, it provides a more numerical stable solution, See Section 4.2.1.4 for more details.

4.1.2 Preliminaries

In this section, we provide a brief overiew and description of the multi-group CBMR. To simplify notation, we begin with the definition of the single-group CBMR in Section 4.1.2.1. Following this, we explain how the definition and notation from Section 4.1.2.1 are extended to the multi-group CBMR setting in Section 4.1.2.2.

4.1.2.1 The single-group CBMR

In the simplest single-group settings, a CBMR with M publications (each containing N voxels) is assumed to take the following form:

$$\log(\mu_i) = \log[\mathbb{E}(Y_i)] = X\beta_i + (Z_i\gamma)\mathbf{1}_N \tag{4.1}$$

where Y_{ij} is the voxelwise count of foci at voxel j for publication i (either 0 or 1 in practice), and $N-$vector $Y_i = [Y_{i1}, Y_{i2}, \cdots, Y_{iN}]^\top$ represents CBMA data for publication i. The spatial design matrix $X(N \times P)$ is generated with spline parameterization with P cubic B-spline bases as detailed in Section 4.1.1.1 (See also [Yu et al., 2024]),

and a publication-level covariates matrix $Z(M \times R)$ is created with R publication-level covariates from M publications followed by standardisation as pre-processing procedure. The estimated intensity is μ_{ij} for publications $i = 1, \cdots, M$ and voxels $j = 1, \cdots, N$, written as the $N-$vector $\mu_i = [\mu_{i1}, \mu_{i2}, \cdots, \mu_{iN}]^\top$ for publication i. This model is identifiable as long as each covariate variable has a mean of zero, allowing X to capture the overall mean. The GLM for all voxels in all M publications is then

$$\log[\mathbb{E}(Y)] = (\mathbf{1}_M \otimes X)\beta + (Z \otimes \mathbf{1}_N)\gamma \tag{4.2}$$

where $Y = [Y_1, Y_2, \cdots, Y_M]^\top$ is an $(M \times N)-$vector, containing voxelwise foci count for all M publications, and $\otimes$ denotes the Kronecker product. Given that our GLM has millions of rows (MN) and the spatial design matrix has billions of entries ($MN \times P$), we proposed a simplified reformulation of this GLM to reduce complexity and memory requirement. A comprehensive discussion of this reformulation, along with a more detailed introduction to the four stochastic models and the notations used in this section, is provided in our previous work, [Yu et al., 2024].

4.1.2.2 The multi-group CBMR

In the multi-group CBMR setting, a dataset is categorised into multiple groups, we fit group-wise activation intensity functions and generate group-specific statistical maps. Adapting the notation of the previous section, this can be represented as:

$$\log(\mu_{g(i)}) = \log[\mathbb{E}(Y_i)] = X\beta_{g(i)} + (Z_i\gamma)\mathbf{1}_N \tag{4.3}$$

where the subscript $g(i)$ represents the group that includes publication i. In equation 4.3, the spatial design matrix, parametrised by spline bases with pre-defined knot spacing, remains fixed across all groups, and the regression coefficient for publication-level covariates, γ, is shared among all groups, while the regression coefficient $\beta_{g(i)}$ is specific to each group. By incorporating group-specific spatial effects while retaining shared global publication-level covariates, Equation 4.3 generalises the conventional form of the single-group CBMR model to the multi-group CBMR framework.

Given a total of M publications divided into G groups, we reorder the publication indices according to their respective groups and assume that group g contains M_g

publications ($M = \sum_{i=1}^{G} M_g$). The GLM for all voxels across all M_g publications for group g can be represented as:

$$\log[\mathbb{E}(Y_g)] = (\mathbf{1}_{M_g} \otimes X)\beta_g + (Z_g \otimes \mathbf{1}_N)\gamma \tag{4.4}$$

where $Y_g = [Y_1, Y_2, \cdots, Y_{M_g}]^\top$ and $Z_g = [Z_1, Z_2, \cdots, Z_{M_g}]^\top$ represent the voxelwise foci counts and publication-level covariates for all M_g publications within group g. Accordingly, the GLM for all voxels across the M publications is formulated by vertically concatenating equation 4.4 for each group. To address the substantial memory and computational demands, a similar reformulation procedure is applied to the multi-group CBMR.

4.2 Methods

This section outlines the computational pipeline employed by CBMR to conduct multi-group meta-regression and meta-inference on CBMA data, as well as the simulations and real-data examples, with results presented in Section 4.3. To begin, Section 4.2.1 provides a detailed overview of the stages involved in the CBMR computational pipeline. Next, Section 4.2.2 describes the simulations designed to evaluate the accuracy and performance of CBMR. Finally, Section 4.2.3 presents a real-world application using the Cue Reactivity dataset, demonstrating the practical implementation of CBMR.

4.2.1 The CBMR pipeline

Figure 4.1 presents a visual overview of the CBMR pipeline as an activity diagram. The pipeline is divided into four stages: meta-regression, parameter estimation and inference and output. Each stage is described in detail in Sections 4.2.1.1 through 4.2.1.4. The implementation of CBMR algorithm in the Python package NiMARE, adheres to these same four stages, as illustrated in Figure 4.1 [Salo et al., 2022].

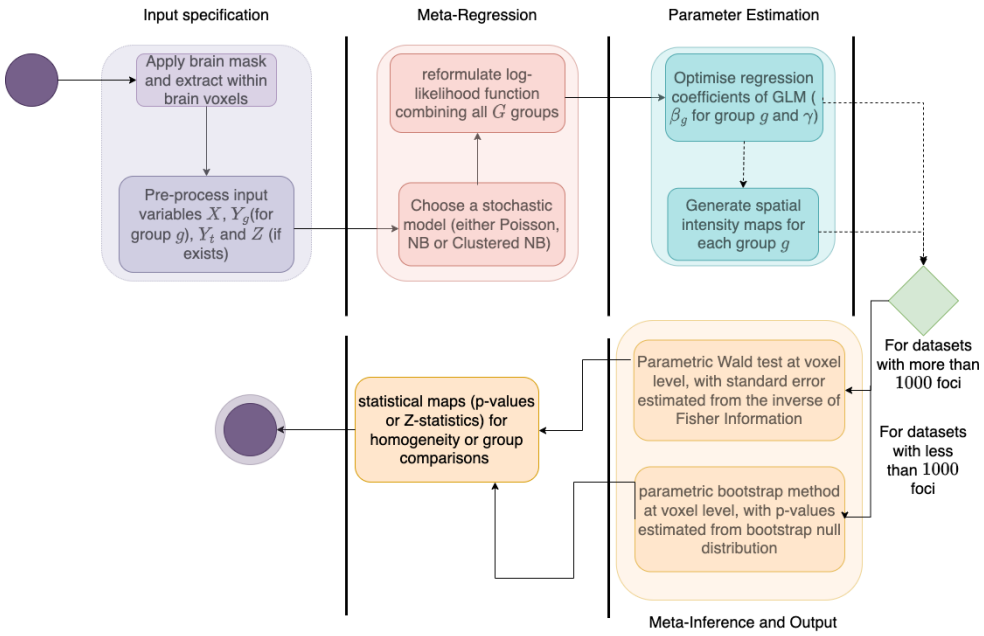


Figure 4.1: The activity diagram provides a detailed view of the CBMR pipeline. The pipeline begins and ends with nodes represented by a dark purple circle and nested dark purple and light purple circles, respectively. Decision nodes are depicted as diamonds, where decisions are made based on whether the number of foci exceeds a specified threshold, while computational stages are represented by vertical bars. Panels separated by these vertical bars correspond to distinct stages within the CBMR pipeline. The entire pipeline is divided into four main stages: input specification, meta-regression, parameter estimation, and meta-inference and output.

4.2.1.1 Input specification

Figure 4.2 illustrates the preprocessing steps required to generate all the necessary input variables for the CBMR pipeline. The preprocessing begins by applying a brain mask to exclude all voxels outside the brain. The default brain mask is the MNI152 2mm template in the code implementation of CBMR. The voxel space has dimensions has dimensions of $91 \times 109 \times 91$ in the x, y and z directions, resulting in a total of $902,629$ voxels. However, most of these voxels fall outside of the brain mask. Applying the brain mask is therefore a crucial step to eliminate redundant voxels and avoid unnecessary computations involving non-brain regions in subsequent processing steps. Next, we select equally spaced knots (with a default spacing of $10mm$) to construct cubic B-spline bases along the x, y and z directions, assuming the numbers of B-spline bases are n_x, n_y and n_z, respectively. The coefficients of these basis functions over v_x, v_y and v_z voxels yield design matrices with shape $n_x \times v_x$, $n_y \times v_y$ and $n_z \times v_z$ for

each dimension. These dimension-specific design matrices are then combined using the tensor product to construct a comprehensive design matrix for further analysis. For more details on the spatial model parametrised by spline bases, refer to Section 4.1.1.1 and [Yu et al., 2024].

In fMRI publications, activation foci are typically reported as their x, y and z coordinates. A CBMA dataset often contains hundreds or thousands of such foci from numerous publications. It is common and straightforward to compute voxelwise foci count across the entire brain for each publication. Building on the previous single-group CBMR model, our current objective is to investigate group-specific activation intensity functions and perform subsequent CBMR inference analyses. To achieve this, we define multiple groups with clear selection criteria, categorize all publications into these groups, and store voxel-wise foci counts separately for each group to support the analysis. The importance of simplified model factorisation has been highlighted in our previous work, aiming to reduce dimensionality and alleviate computational complexity [Yu et al., 2024]. Specifically, we adopt the following two approaches for different stochastic models:

- **Poisson model**: The total voxel-wise foci counts across all publications are assumed to follow a Poisson distribution with the mean equal to the sum of the estimated mean from each publication. This leverages the additive property of the Poisson process for computational simplicity and interpretability.
- **NB model**: By matching the first and second moments (mean and variance), we approximate the likelihood function under the assumption that the convoluted voxel-wise foci counts follow a Negative Binomial (NB) distribution.

In this work, we adopt the same convention of applying above model factorisation methods to simplify the log-likelihood functions. However, unlike previous approaches, we first categorise all publications into multiple groups and then apply these factorisation methods at the group level. Following group-level model factorisation, the sufficient statistics are reduced to dimensions no greater than either the number of publications within each group or the number of voxels within the brain mask, as detailed below,

- Let $Y_{gj} = \sum_{i=1}^{M_g} Y_{ij}$ be the sum of foci counts at voxel j across all M_g publications within in the group g, and the $N-$vector $Y_g = [Y_{g1}, Y_{g2}, \cdots, Y_{gN}]^\top$;

- Let $Y_{ti} = \sum_{j=1}^{N} Y_{ij}$ be the sum of foci counts for publication i across all voxels, and the $M-$vector $Y_t = [Y_{t1}, Y_{t2}, \cdots, Y_{tM}]$;
- Let N-vector $\mu_g^X = \exp(X\beta_g)$ be the vector of localised spatial effects of publications in group g;
- Let M-vector $\mu^Z = \exp(Z\gamma)$ be the vector of global publication-level covariate effects.

In the CBMR pipeline that incorporates the effects of publication-level covariates, an additional input variable, Z_g (with dimension $M \times R$) is introduced. This variable is constructed by extracting R publication-level covariates from M publications. Common examples of publication-level covariates include sample size, year of publication and participant age. It is important to standardise the publication-level covariates to have a mean of 0 and a standard error of 1 during preprocessing. This standardisation ensures that X captures the overall mean, enabling more straightforward and comparable analyses of spatial intensity functions in subsequent steps.

Another potential input variable introduced during CBMR preprocessing is the roughness penalty matrix J (with dimensions $P \times P$) of spline bases. During optimisation using L-BFGS, we observed challenges with datasets that have insufficient foci counts, where some elements of the group-wise spatial regression coefficients β_g are driven to highly negative values. This results in an overly flexible and detailed representation of the foci distribution. While such flexibility can model complex patterns, it often leads to overfitting, unnecessarily intricate functions, and causes numerical instability. To address these issues, we incorporate a roughness penalty that penalises overly flexible spatial functions (combinations of spline bases) by discouraging overly complex or "rough" spatial functions. This ensures smoother and more stable solutions. Details on constructing the roughness penalty matrix J are provided in Appendix B.1.

4.2.1.2 Meta-regression

Following the preprocessing of CBMR input variables, the next step is to evaluate different stochastic models to identify the most accurate but parsimonious fit. We will present the statistical formulations, as well as advantages and drawbacks of all the stochastic models proposed in our previous work, except for the Quasi-Poisson

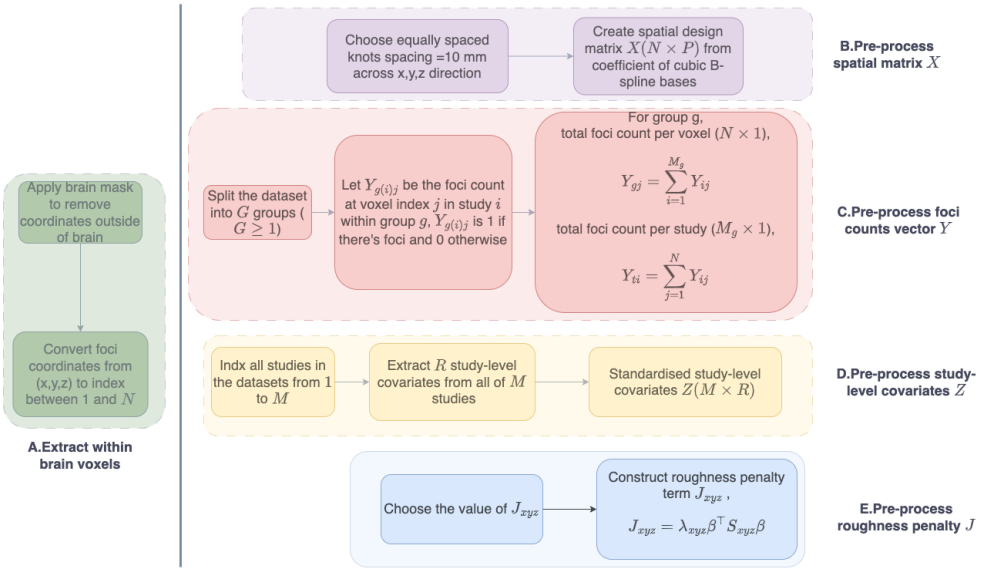


Figure 4.2: This preprocessing pipeline for multi-group meta-analytic datasets is applied before fitting coordinate-based meta-regression (CBMR) framework. Panel A, B and C are applicable to all datasets and used to generate the spatial design matrix X, total foci count per voxel $Y_g(N \times 1)$ for group g and total foci count per publication $Y_t(M \times 1)$. Panel D is required only when considering the effects of publication-level covariates, in which case the covariates matrix $Z(M \times R)$ is included. Panel E is recommended for datasets with insufficient foci counts, as it discourages overly complex spatial functions and improves numerical stability.

model. The exclusion of the Quasi-Poisson model is due to its characteristics as a Quasi-likelihood-based model that requires optimisation using the IRLS algorithm. This approach complicates the implementation of optimisation process, as it does not support the use of L-BFGS algorithm for maximising likelihood functions. Furthermore, as demonstrated in our previous work, the Quasi-Poisson model is similar to the NB model in explainability of excess variation in foci count data but exhibits inferior performance [Yu et al., 2024].

The Poisson model is the simplest stochastic model option in the CBMR pipeline, both in terms of statistical formulation and computational complexity. In practice, the count of foci Y_{ij} (for publication $i = 1, \cdots, M$ and voxel $j = 1, \cdots, N$) is always 0 or 1, which strictly indicates a Binomial model. Therefore, we adopt the Poisson model, inspired by its previous success with Poisson point process and the accuracy of Poisson approximations for low-rate Binomial data. One appealing property of the Poisson process is that the sum of multiple Poisson random variables is also Poisson. This allows for practical flexibility: it is equivalent to model either the set

of M_g publication-level counts or the summed counts at each voxel for each group g. Following the model structure outlined in Equation 4.3, the intensity for voxel j in publication i for group g is,

$$\begin{aligned} E[Y_{g(i)j}] = \mu_{g(i)j} &= \mu^X_{g(i)j} \cdot \mu^Z_i \\ \log[\mu_{g(i)j}] = \eta_{g(i)j} &= x_j^\top \beta_{g(i)} + Z_i\gamma \end{aligned} \tag{4.5}$$

where publication i belongs to group $g(i)$, $Y_{g(i)j} \sim Poisson(\mu_{g(i)j})$, $x_j^\top$ is the j^{th} row of spatial design matrix $X(N \times P)$, and $\beta_{g(i)}$ is spatial regression coefficients of group $g(i)$. Under the assumption of independence of counts across publications, the total likelihood function is exactly same if we model the voxelwise total foci count over publications for each group instead, and the likelihood to be optimised is,

$$\begin{aligned} l(\theta) = l(\beta_1, \cdots, \beta_G, \gamma) &= \sum_{g=1}^{G} l(\beta_g, \gamma) \\ &= \sum_{g=1}^{G}\sum_{j=1}^{N}[Y_{gj}\log(\mu_{gj}) - \mu_{gj} - \log(Y_{gj}!)] \\ &= \sum_{g=1}^{G} Y_g^\top \log(\mu_g^X) + Y_t^\top \log(\mu^Z) - \sum_{g=1}^{G}[\mathbf{1}^\top \mu_g^X][\mathbf{1}^\top \mu_g^Z] \end{aligned} \tag{4.6}$$

For more detailed derivations, refer to Appendix B.2.1.

While Poisson model is widely used for the regression of count data, foci counts often exhibit over-dispersion in practice, where the variance of the response variable substantially exceeds the mean. In such case, imposing a Poisson model may underestimate the standard error and lead to biased estimates of the regression coefficients. To address this, we propose modelling the count data at each voxel as independently following a group-specific Negative Binomial (NB) distribution, which accounts for excess variance relative the Poisson model [Lawless, 1987]. The NB model employs a group-specific single parameter, α_g, shared across all publications within group g and all voxels, to index the variance in excess of Poisson model. Specifically, for group g, publication i, and voxel j, let $\lambda_{g(i)j}$ follows a Gamma distribution with mean $\mu_{g(i)j}$ and variance $\alpha_g\mu^2_{g(i)j}$. Conditioned on $\lambda_{g(i)j}$, let Y_{ij} follow a Poisson distribution with mean $\lambda_{g(i)j}$. Unlike Poisson, the sum of multiple independent NB random variables doesn't follow an NB distribution. To address this, we propose moment matching approach to approximate the first two moments (mean and variance) of the convolution of NB distributions. This significantly simplifies the log-likelihood function. By matching the

first two moments, the approximate NB distribution of the total count of foci across all publications within group g at voxel j is given by $Y_{gj} = \sum_{i=1}^{M_g} Y_{ij} \sim NB(r'_{gj}, p'_{gj})$ where

$$r'_{gj} = \frac{\mu_{gj}^2}{\alpha_g \sum_{i=1}^{M_g} \mu_{ij}^2}, p'_{gj} = \frac{\sum_{i=1}^{M_g} \mu_{ij}^2}{\alpha_g^{-1}\mu_{gj} + \sum_{i=1}^{M_g} + \sum_{i=1}^{M_g} \mu_{ij}^2} \tag{4.7}$$

with corresponding excess variance for each group g,

$$\alpha'_g = \alpha_g \frac{\sum_{i=1}^{M_g} \mu_{ij}^2}{\mu_{gj}^2} \tag{4.8}$$

which gives rise to the simplified NB log-likelihood function,

$$\begin{aligned} l(\theta) &= l(\beta_1, \cdots, \beta_G, \alpha'_1, \cdots, \alpha'_G, \gamma) = \sum_{g=1}^{G} l(\beta_g, \alpha'_g, \gamma) \\ &= \sum_{g=1}^{G} \sum_{j=1}^{N} [\log \Gamma(Y_{gj} + r'_{gj}) - \log \Gamma(Y_{gj} + 1) - \log \Gamma(r'_{gj}) \\ &+ r'_{gj} \log(1 - p'_{gj}) + Y_{gj} \log(p'_{gj})] \end{aligned} \tag{4.9}$$

For more detailed derivations, refer to Appendix B.2.2

4.2.1.3 Parameter estimation

The most computationally intensive stage of CBMR pipeline is the estimation of the unknown model parameters $(\beta_g, \alpha_g, \gamma)$ for each group g. A common approach for estimating these group-specific parameters is Maximum Likelihood Estimation (MLE), based on reformulated log-likelihood functions tailored to each stochastic model described in Section 4.2.1.2. To efficiently optimise these parameters, the CBMR pipeline employs the L-BFGS algorithm, a quasi-Newton method well-suited for problems involving large-scale datasets and high-dimensional parameter spaces. By approximating the Hessian matrix rather than computing and storing it directly, the L-BFGS algorithm achieves significant computational efficiency, making it ideal for

the CBMR pipeline [Liu and Nocedal, 1989]. Considering the log-likelihood function is non-convex for both the NB model, a more cautious optimisation strategy is adopted. Specifically, a smaller learning rate is used during L-BFGS optimisation to reduce the risk of the algorithm becoming trapped in a local optimum rather than converging to the global optimum.

For the Poisson model in CBMR, the group-specific spatial regression coefficient β_g is initialise either with random values uniformly distributed within the range $[-0.01, 0.01]$ or with values assuming spatial homogeneity of foci locations. Both initialisation strategies allow the L-BFGS algorithm to converge effectively. To address the non-convexity of log-likelihood functions for NB model, we propose using the optimised spatial regression coefficient β_g from the Poisson model as the initialisation for these two models, improving the stability and robustness of the optimisation process. During optimisation, we iteratively optimise the group-wise dispersion parameter α_g while keeping the group-specific spatial regression coefficient β_g and if applicable, the coefficient of publication-level covariates γ fixed. Subsequently, α_g is fixed, and other variables are optimised in alternating iterations until convergence. Pseudocode for this alternating iterations is provided by 3.

Algorithm 2 Alternating iterations for CBMR with NB model

Assign initial estimates to group-specific parameters β_g, α_g and group-shared γ
while *Current* $l(\theta)$ *and previous* $l_{prev}(\theta)$ *differ by more than a predefined tolerance* **do**
    Evaluate the previous log-likelihood using $l_{prev}(\theta) = \sum_{g=1}^{G} l_g(\beta_g, \alpha_g, \gamma)$
    while *Current* $l(\alpha_g)$ *and previous* $l_{prev}(\alpha_g)$ *differ by more than predefined tolerance (*$1e^{-9}$ *by default)* **do**
        Update the group-wise dispersion parameter α_g for each group g using L-BFGS algorithm, while keeping β_g and γ (if applicable) fixed.
    while *Current* $l(\beta_g, \gamma)$ *and previous* $l_{prev}(\beta_g, \gamma)$ *differ by more than predefined tolerance (*$1e^{-9}$ *by default)* **do**
        Update the group-specific parameters β_g and the group-shared parameter γ (if applicable), while keeping α_g fixed.
    Recompute the current log-likelihood values $l(\theta)$ and calculate the difference from the previous log-likelihood values.

In the implementation of the CBMR parameter estimation stage, we use the built-in function *scipy.optimize. minimize(method='L-BFGS-B')* from Scipy to minimise the objective function (negative log-likelihood function). This L-BFGS function was chosen for its well-documented and user-friendly interface, which simplifies integration

into the pipeline. To address the increased computational demands of the parametric bootstrap method (see Section 4.2.1.4 for details), we also implemented parallelisation to accelerate computation. Furthermore, we implemented the code in JAX to take advantage of its automatic differentiation capabilities. This allows for efficient approximation of the observed Fisher information matrix using the optimised regression coefficients for inference based on the Wald test (see Section 4.2.1.4 for details), without the need to explicitly derive the Hessian matrix of the log-likelihood function.

Using the optimised CBMR regression coefficients, we can construct group-specific estimated intensity maps to intuitively visualise the brain activation patterns. For more rigorous inference, allowing the identification of brain regions with significant p-values from statistical maps, we will further implement meta-inference pipelines, with further details provided in Section 4.2.1.4.

4.2.1.4 Meta-inference and output

The final stage of the CBMR pipeline involves performing inference (for both the homogeneity test and group comparison test) on the group-specific estimated intensity maps and outputting the analysis results as statistical maps in NIfTI format. To conduct voxel-wise hypothesis testing for both types of test, CBMR adopts an approach similar to that used in the popular GLM python package *statsmodels* and the R function *glm()*. In this approach, the group-specific estimated spatial intensity μ_g^X or its log-transformed counterpart η_g^X are used to construct test statistics at voxel-wise level, as well as their standard errors.

Assuming a contrast matrix $C(m \times S)$ is provided for S involved groups, a voxel-wise null hypothesis $H_0 : C\hat{\theta}_j = \mathbf{0}_{m\times 1}$ for voxel j can be specified. For simplicity, we assume that any redundant columns containing only zero elements (corresponding to groups not involved in the contrast) are removed before proceeding with the analysis. CBMR computes the corresponding test statistics as:

$$(C\hat{\theta}_j)^\top (CV_jC^\top)^{-1}(C\hat{\theta}_j) \xrightarrow{D} \chi^2_m \tag{4.10}$$

where $\hat{\theta}_j$ represents either the estimated intensity $[\hat{\mu}^X_{1j}, \cdots, \hat{\mu}^X_{Sj}]^\top$ or its log-transformed value $[\hat{\eta}^X_{1j}, \cdots, \hat{\eta}^X_{Sj}]^\top$, in practice, we recommend using the log-transformed values $\hat{\eta}^X_{gj}$, as they corresponds to the linear response of GLMs, avoiding the additional

approximation required to transform $\hat{\eta}_{gj}^{X}$ to $\hat{\mu}_{gj}^{X}$ for estimating standard errors. As the inverse of the Fisher information gives the asymptotic variance of the estimates of spatial regression coefficients $\beta_1, \cdots, \beta_S$ for the S involved groups. Based on the deterministic structure of GLMs $\hat{\eta}_{g}^{X} = X\beta_g$, we approximate the variance of $\hat{\eta}_{g}^{X}$ for group g as $X^{\top} Var(\beta_g) X$, where X is the spatial design matrix. Additionally, $V_j(S \times S)$ represents the covariance matrix constructed from the estimated variance of $\hat{\eta}_{g}^{X}$ at the j^{th} voxel across all S groups. The degrees of freedom for the statistical test are determined by the number of rows in the contrast matrix C. To calculate the corresponding p-values, the test statistics are approximated using a Chi-square distribution.

In scenarios where only one group is involved ($m = S = 1$), the statistical test simplifies to a Wald test, with the null hypothesis formulated as $C(\hat{\theta}_j - \theta_0) = 0$. This can be further simplified to the following form,

$$W_j = \frac{\hat{\theta}_j - \theta_0}{SE(\hat{\theta}_j)} \tag{4.11}$$

where $\hat{\theta}_j$ represents either the estimated intensity $\hat{\mu}_{gj}^{X}$ for the involved group g or its log-transformed value $\hat{\eta}_{gj}^{X}$. The corresponding voxel-wise p-value p_j is calculated under the assumption that the Wald test statistics W_j follows a standard normal distribution.

In practice, approximating group-wise spatial regression coefficients by inverting the Fisher Information matrix often leads to numerical instability. This issue is particularly pronounced in datasets with an insufficient number of foci, especially during group comparisons involving multiple CBMR groups, where standard error estimations rely independently on the inversion of group-specific Fisher information matrix. A practical threshold for ensuring reliable inference is at least 200 foci per group. This numerical instability arises primarily due to the high dimensionality of the Fisher Information matrix, often consisting of hundreds or even thousands of spline bases elements. With fewer foci, the Fisher Information matrix can become numerically singular, as most voxels have near-zero intensity estimates.

Despite efforts to improve numerical stability during the optimisation process, such as adding a roughness penalty to prevent coefficients from being driven to highly negative values, we observed that approximating group-wise spatial regression coefficient by inverting Fisher Information matrix often results in numerical instability. This is

particularly prevalent in datasets with an insufficient number of foci, with a practical threshold of at least 200 foci per group required for reliable inference. The instability arises due to the high dimensionality of the Fisher Information matrix, which can have hundreds or even thousands of elements corresponding to the spline basis functions. For datasets with a low foci count, the Fisher Information matrix can become numerically singular because most voxels have near-zero intensity estimates. We have experimented with several approaches to improve this instability, including adding a small epsilon (10^{-6}) or 1% of the largest diagonal element to the diagonal of the Fisher Information matrix, and computing the Fisher Information under the assumption that the null hypothesis of homogeneity is True. However, these methods consistently resulted in underestimation of the variance of voxel-wise spatial intensity, resulting in invalid p-values.

Given these challenges, we are now exploring parametric bootstrap methods as an alternative for meta-inference, rather than relying on statistical tests based on the inverse of the Fisher Information matrix. The parametric bootstrap is a resampling-based statistical technique that estimates the sampling distribution of a statistic without requiring strong parametric assumptions about the underlying data distribution. Specifically, for group-wise homogeneity test, the bootstrap process involves the following steps: for each bootstrap sample, foci are randomised under the assumption of spatial homogeneity, following a Binomial process. The CBMR regression is refitted to obtain group-wise estimated intensity values or their log-transformed values at voxel level. This procedure is repeated at least 1000 times to generate the bootstrap null distribution. Under the null distribution $H_0 : \eta_{gj}^X = \eta_{g0}$ or $\mu_{gj}^X = \mu_{g0}$, the observed values of η_{gj}^X or μ_{gj}^X for group g are compared to the bootstrap null distribution. The p-values is then calculated as the probability of observed test results as extreme as the actual results, assuming the null hypothesis is true. While for group comparison tests, a similar bootstrap procedure is applied with a slight modification: under the null hypothesis $\eta_{Aj} = \eta_{Bj}$ between group A and B, we first combine all foci counts from both groups to estimate a shared activation intensity function. Data are then regenerated from the chosen stochastic model associated with the CBMR regression, ensuring the total number of publications remain the same as before for both groups. The model is refitted for each bootstrap sample. Repeating this procedure generates the bootstrap null distribution, and p-values are calculated by comparing the actual results to the bootstrap null distribution. We assert that this method avoids the numerical issues encountered during the inference stage, although at the cost of increased

computational requirements. Its validity and effectiveness will be demonstrated in the Sections 4.2.2.

Additionally, we are also interested in investigating the global effects of publication-level covariates on group-wise spatial activation functions. For example, we aim to assess whether there is a global effect of the (square root of) sample size on spatial activation functions, or whether the influence of (square root of) sample size is stronger than that of publication year. To address these questions, we perform hypothesis testing on one or more elements of the regression coefficient vector γ, which captures the effects of the publication-level covariates. Similarly to the voxelwise hypothesis testing of spatial intensity in equation 4.10, this is achieved using a contrast matrix $C_\gamma(m \times s)$, where s denotes the number of relevant publication-level covariates after excluding irrelevant ones. The contrast matrix C_γ allows for the specification of flexible hypotheses. Under the null hypothesis $H_0 : C_\gamma\gamma = \mathbf{0}_{m\times 1}$, the test statistic is given by

$$(C_\gamma\hat{\gamma})^\top(C_\gamma Cov(\hat{\gamma})C_\gamma^\top)^{-1}(C_\gamma\hat{\gamma}) \xrightarrow{D} \chi^2_m \tag{4.12}$$

where $Cov(\hat{\gamma})$ represents the covariance structure of elements in $\hat{\gamma}$, and the p-values can be approximated using a chi-square distribution with m degrees of freedom. Note that the issue of inverting a numerically singular Fisher Information matrix is unlikely to arise when performing inference on the regression coefficients of publication-level covariates. This is because γ typically contains only a few elements (fewer than 5), resulting in a Fisher Information matrix of low dimensionality. Furthermore, since most of elements in γ are unlikely to be simultaneously close to zero, ensuring the Fisher Information matrix is not numerically singular. Therefore, we believe it is unnecessary to use bootstrap methods for inference on publication-level covariates.

4.2.2 Simulation methods

In order to quantitively evaluate and demonstrate the computational accuracy and efficiency of CBMR, extensive simulations were conducted across twelve settings. Simulated data were generated for three spatial configurations: a two dimensional grid consisting of 100×100 voxels, a three dimensional grid consisting of $100 \times 100 \times 100$ voxels, a three dimensional grid within a MNI152 2mm brain mask, containing $228,483$ voxels. Each configuration was analysed under four data generation designs. These data generation designs combined two key factors: the underlying intensity function and the spatial patterns for data generation. The underlying intensity function for

generating CBMR data is either high (an average total foci count of approximately $1,000$ per publication) or low (an average total foci count of approximately 10). Data generation followed either a homogeneous spatial intensity assumption or a scenario with two Gaussian bump signals overlaid on a background constant intensity function. The high-intensity ($1,000$ foci per publication on average) and spatial homogeneity setting is primarily used as a sanity check, in contrast, the low-intensity (10 foci per publication on average) and two bump signals setting is designed to evaluate model performance under more realistic conditions that closely reflect real-world datasets. For each simulation, the data includes three groups with identical underlying intensity functions but differing numbers of publications: 100, 100 and 500, respectively. Following data generation, CBMR regression is performed using either the Poisson or NB model. Statistical tests are conducted using either standard error estimates derived from the inverse of the Fisher Information matrix or a bootstrap approach with $1,000$ bootstrap samples.

In each simulation setting, the spatial design matrix $X(P = 2624)$ is constructed using cubic B-spline bases with knot spacing of 10mm. This design matrix is fixed and applied consistently across all groups in every simulation settings. The effect of publication-level covariates is assumed to exist in all settings, with their values generated uniformly within the range $[-1, 1]$ and standardised to have a mean of 0 and a standard deviation of 1, allowing X to capture the overall mean. During the optimisation process in each simulation, the group-specific spatial regression β_g and the shared regression coefficient γ across all groups are estimated, and then used to construct the group-specific intensity maps, as defined by equation 4.4.

In order to evaluate the accuracy and performance of parameter estimation in each simulation setting, we conduct meta-inference for both homogeneity test within each group and the group comparison test between any two groups. These tests were performed at the voxel level using two inference approaches: (i) parametric statistical tests, as detailed in Equation 4.10, and (ii) the parametric bootstrap method. After obtaining voxel-wise p-values, they were sorted in an ascending order and visualised using a PP-plot to compare the probability distribution of the observed and theoretical p-values. The x-axis represents the theoretical distribution (a uniform distribution between 0 and 1), while the y-axis represents the observed distribution. If the two distribution are similar, the sorted pairs of observed and theoretical p-values are expected to align closely along the $45-$degree diagonal line ($y = x$). Deviations from this diagonal indicate discrepancies between the observed data and the theoretical

distribution. Additionally, the group-specific estimated intensity maps produced by CBMR were compared using the mean absolute difference across the whole brain image to assess the accuracy of the CBMR regression stage. Finally, the computational time for the two inference approaches were recorded for comparison.

In summary, the simulations we have described evaluate CBMR in three key aspects: (i) the accuracy of parameter estimation during the regression stage, (ii) the performance of parametric statistical tests and parametric bootstrap method under various data generation settings, and (iii) computational time required to run the CBMR pipeline. All reported results were obtained using an HPC cluster with Intel(R) Xeon(R) Gold 6126 2.60HZ processors each with 16 GB RAM.

4.2.3 Real data methods

As a demonstration of the large-scale capabilities of CBMR, here we present an example involving a more complex data than those considered in the simulation discussed in Section 4.2.2. In this example, we use a meta-analytic cue reactivity dataset, as the cue-reactivity paradigm is a widely employed in neuroimaging studies to elicits brain activity associated with attentional, affective, and reward processes in response to appetitive stimuli. We conducted literature search for visual cue-reactivity fMRI studies focused on drugs of abuse or natural rewards published up to Aug 2020. Cue types include nicotine, alcohol, cannabis, cocaine, heroin, food or sexual stimuli. This dataset includes 546 experiments assessing visual stimuli categorised as drug-neutral ("drug", $n = 163$), natural-neutral ("natural", $n = 110$) and reward-neutral ("reward", $n = 273$). Relevant publication-level information was recorded, including participate age, sex, cue type, MRI scanner field strength and processing software [Hill-Bowen et al., 2021].

Here, we address two primary research questions using either voxel-level group-wise spatial homogeneity tests or group comparison tests between multiple groups:

- Where are the regions of activation associated with a specific group of cue-reactivity stimuli (e.g. drug-related stimuli) that show stronger estimated intensity than average, under the assumption of spatial homogeneity?

- Where do differences exist in activation regions between two stimulus types within the cue reactivity dataset (e.g., differences between drug and natural stimulus groups)?

At the pre-processing stage, all 546 experiments are categorised into three groups based on their respective visual stimulus types. Foci located outside of the MNI152 2mm brain mask are removed, and a spatial design matrix $X(P = 2624)$ is constructed using cubic B-splines with a knot spacing of 10mm. In this experiment, we include the square root of the sample size and publication year as publication-level covariates. These covariates are standardised to have a mean of 0 and a standard deviation of 1 before being integrated into the CBMR pipeline. At the CBMR regression stage, either Poisson or Negative Binomial (NB) model is employed for parameter estimation. This involves optimising the group-specific regression coefficients β_g for each group g, the group-shared regression coefficients for the effects of the publication-level covariates, and if the NB model is employed, the group-specific overdispersion parameter α_g. Using these estimates, group-specific intensity maps are then constructed for each group according to Equation 4.4. At the CBMR inference stage, both group-wise homogeneity tests and group comparison tests between any two groups are performed. Voxel-wise p-values are obtained using either statistical tests, as described in Equation 4.10, or by comparing the observed data to null distributions generated via parametric bootstrap methods with $1,000$ bootstrap samples. Activation maps (for significant uncorrected p-values under the 5% significance level) generated by these two inference methods are compared against those generated by ALE. For the group-wise homogeneity test, ALE activation maps are computed using the default full-width half maximum (FWHM) settings based on sample size, as described in [Eickhoff et al., 2012]. For group comparison, ALE subtraction analysis is employed. Both methods are implemented using the built-in functions of the Python package NiMARE [Salo et al., 2022]. Additionally, we investigate the global effects of sample size and publication year on the group-wise intensity functions, analysing if these effects are significant, as well as comparing the strength for each group g.

The primary goal of the analyses described above is to demonstrate the practical application of CBMR and to highlight its efficiency and scalability through a real-world example. To evaluate computational efficiency, the time required for parameter estimation during the regression stage was recorded for both inference methods: parametric statistical tests and parametric bootstrap. In Section 4.3.2, results are

reported for Likelihood Ratio tests, along with alternative model selection criteria such as AIC and BIC, which takes the model complexity into consideration. All analyses were conducted on an HPC cluster with Intel(R) Xeon(R) Gold 6126 2.60HZ processors each with 16 GB RAM.

4.3 Results

4.3.1 Simulation results

4.3.1.1 Parameter optimisation

Across the twelve simulation settings outlined in Section 4.2.2 (three spatial configurations combined with four data generation schemes), all parameter estimates produced by CBMR regression closely matched the ground truth. For consistency and clarity, we focus on showcasing results from the experimental design involving CBMR regression with the NB model applied to a three-dimensional brain image, as all designs demonstrated similar patterns. The observed absolute bias for intensity function estimation, averaged across all $1,000$ bootstrap samples and voxel locations, is presented in Table 4.1. We noticed that, settings with low underlying intensity and intensity functions with two bump signals posed greater challenges for CBMR regression, as reflected in larger absolute bias values. However, all results remain within the magnitude of 10^{-4}, demonstrating the validity and accuracy of CBMR regression for multiple groups under various experimental conditions.

	Spatial homogeneous intensity	With two bump signals
High intensity	1.1501×10^{-4}	2.4335×10^{-4}
Low intensity	1.6793×10^{-4}	6.4429×10^{-4}

Table 4.1: Bias for CBMR intensity function estimation under different conditions: high or low underlying intensity levels, combined with either spatially homogeneous intensity functions or intensity functions with two bump signals.

4.3.1.2 Computation time

We emphasise that, after model re-factorisation, neither the number of foci nor the number of publications affects the computation time. This is because the sufficient

statistics are reduced to the group-wise vector of voxel-wise total foci counts across all publications within group g (y_g, $N \times 1$) and the vector of total foci counts across all voxel locations within a publication (y_t, $M \times 1$). Only the number of groups influences the computation of log-likelihood function during the each iteration. Therefore, we assert that our CBMR regression stages scales efficiently with the number of publications or foci. Moreover, a larger number of publications or foci improves numerical stability and accelerates convergence during the optimisation process.

As a computationally efficient alternative to Bayesian model-based meta-regression methods, one of the key advantages of our CBMR pipeline is its simple, intuitive statistical structure and scalability. Our experiments demonstrate that the optimisation in meta-regression with multiple groups takes approximately 30 minutes on an NVIDIA GTX 1080 Graphics Card – a significant improvement compared to some Bayesian model-based methods, which require roughly 30 hours on an NIVDIA Tesla K20c GPU card [Samartsidis et al., 2019]. However, in experimental settings with an insufficient number of foci (e.g., low underlying intensity functions with either spatial homogeneity or two bump signals in our setting), we consider using parametric bootstrap methods as an alternative. This is due to the occurrence of numerical singularities in the Fisher Information matrix, which prevents the subsequent meta-inference stage. Nonetheless, the parametric bootstrap method is computationally intensive, as it requires repeated data simulations (e.g., $1,000$ bootstrap samples in our experiment) and model refitting for each sample to obtain the bootstrap null distribution. In practice, we implemented parallelisation on HPC clusters to accelerate model re-fitting. Running model-refitting on five bootstrap samples in parallel on a single HPC cluster nodes with Intel(R) Xeon(R) Gold 6126 2.60HZ processors takes approximately 40 minutes. Experiments based on parametric bootstrap methods are feasible only with parallelisation and the availability of hundreds of HPC cluster nodes. However, this approach is as computationally intensive as, or even more than the Bayesian model-based methods and is not easily accessible to users without HPC cluster resources. As a result, the CBMR regression loses one of its key advantages–computational efficiency, when applied to small meta-analytic datasets with less than 200 foci per group.

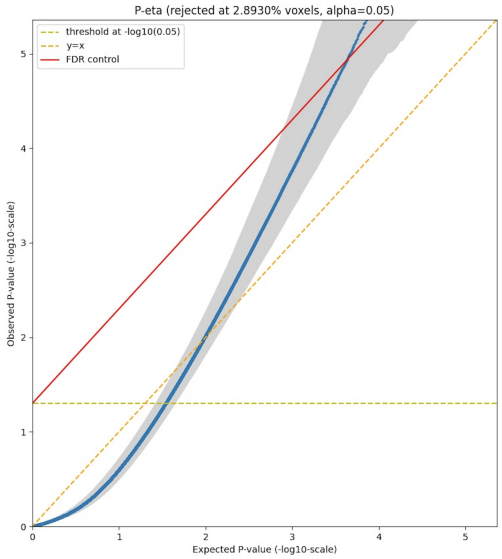
4.3.1.3 Validation of the Meta-inference stage

Following the simulation settings described in Section 4.2.2, we validate the accuracy of the meta-inference pipeline by evaluating PP-plots of voxel-wise p-values under each

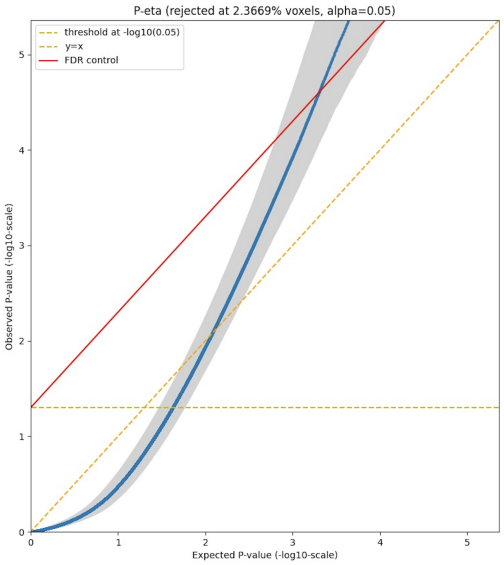
simulation scenario. These P-values are computed either using parametric statistical tests described in Equation 4.10 or through the parametric bootstrap method. A perfect alignment with the $y = x$ line would indicate that the meta-inference stage produces valid outcomes, thereby providing confidence to apply the same inference procedure to real datasets.

Since the PP-plots are very similar across the twelve scenarios, we only present results for a representative setting: CBMR inference using the parametric statistical test described in Equation 4.10. This setting compares estimated intensity functions between two groups with identical underlying intensity functions. These functions are simulated at different overall intensities ($1,000$ vs. $5,000$ foci, or $100,000$ vs. $500,000$ foci) and exhibit either spatial homogeneity or two Gaussian bump signals. Figure 4.3 displays four $-\log_{10}$ PP-plot corresponding to different underlying intensity functions in this simulation setting. The plots include the $y = x$ line (dashed diagonal line), the 5% significance (dashed horizontal line) and the FDR 5% boundary (solid diagonal line); and gray shaded areas indicating the point-wise 95% prediction intervals. The results show that for scenarios with low underlying intensity functions (both spatial homogeneous and with two bump signals), p-values $> 0.05 \approx 10^{-1.3}$ can skew conservative, while extreme p-values can skew liberal. This poor behaviour is observed in both spatially homogeneous and bump signal cases, particularly when the number of foci is insufficient. Conversely, for scenarios with high underlying intensity functions, the PP-plot lines are only slightly skewed, and the $y = x$ line falls within the point-wise 95% prediction intervals. These results support the observation that PP-plots exhibit poor behaviour when the number of foci falls below a certain threshold (1000 foci in our previous one-group CBMR experiment [Yu et al., 2024]). In such cases, inference results based on parametric statistical tests become unreliable due to numerical singularity in the Fisher information matrix encountered during practical implementation.

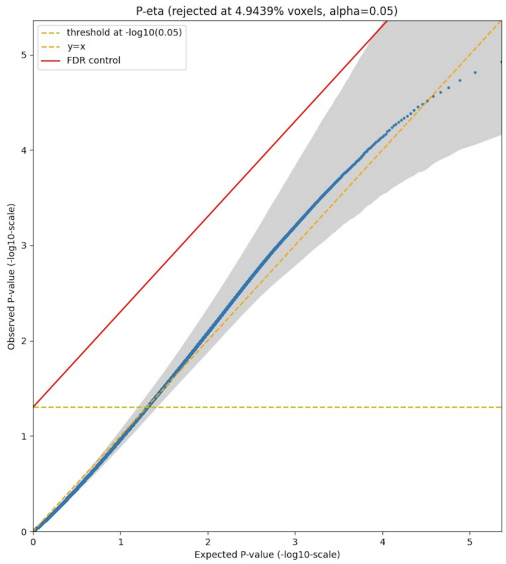
For datasets with an insufficient number of publications or foci, we recommend using the parametric bootstrap method instead, to avoid inverting the Fisher Information matrix (See 4.2.1.4 for details). Figure 4.4 presents the results of the same representative simulation setting as above, focusing on the more challenging scenario with a low underlying intensity function (10 foci sampled per publication on average). The PP-plot for both foci patterns (spatially homogeneous and with two bump signals) align closely with the $y = x$ line, with extreme p-values exhibiting only a slight conservative skew. These results indicate that the parametric bootstrap method is an effective



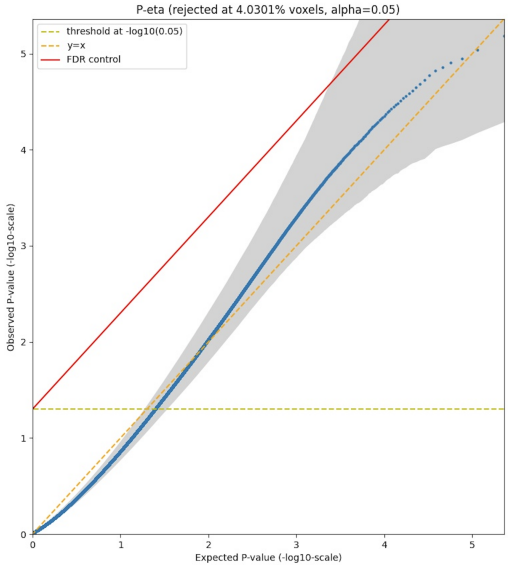
(a) Low underlying intensity function of spatial homogeneity



(b) Low underlying intensity function with two bump Gaussian signals



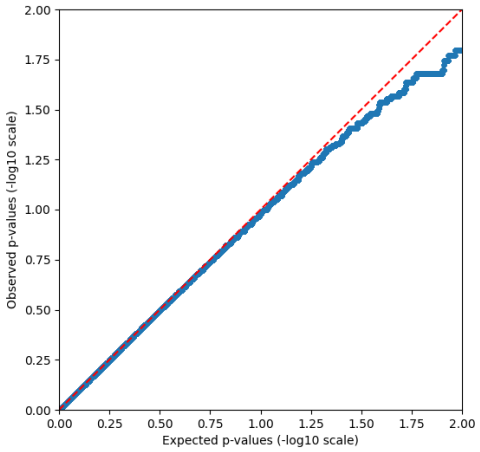
(c) High underlying intensity function of spatial homogeneity



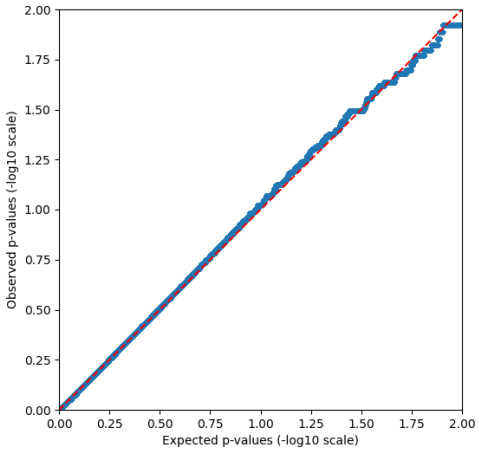
(d) High underlying intensity function with two bump Gaussian signals

Figure 4.3: PP-plots illustrating group comparison under two underlying intensity function settings: low (1,000 vs. 5,000 foci) or high (100,000 vs. 500,000 foci). Each setting includes both spatially homogeneous and two-bump signal configurations. The plots are based on P-values derived from statistical tests described in Equation 4.10.

alternative to parametric statistical tests for small datasets, although at the expense of increased computational time.



(a) Underlying intensity function of spatial homogeneity



(b) Underlying intensity function with two bump Gaussian signals

Figure 4.4: PP-plots illustrating group comparison under two different types of low underlying intensity functions (1,000 vs. 5,000 foci). These plots are generated from P-values obtained across 1,000 bootstrap samples

4.3.2 Real data results

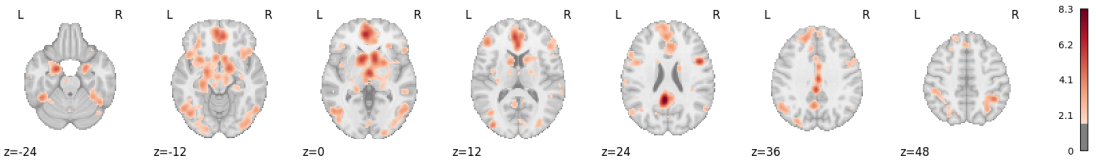
4.3.2.1 Model comparison

We evaluate the goodness of fit among two likelihood-based stochastic models (Poisson and NB model) by comparing their maximised log-likelihood values. Our analysis shows that CBMR using the NB model outperforms Poisson model on Cue Reactivity dataset, when comparing the maximised total log-likelihood values across multiple groups. This is not surprising, as the NB model accounts for the anticipated excess variance relative to the Poisson model at voxel level. Given the nested relationship between the Poisson and NB models (with the group-specific dispersion parameter $\alpha_g = 0$ for group g in the NB model), we also performed a Likelihood Ratio Test (LRT) to evaluate the trade-off between model sufficiency and complexity. The LRT results indicate that the null hypothesis – the simpler nested model (Poisson) is as good as the full model (NB) – is strongly rejected for the Cue Reactivity dataset,, with $p-$values less than 10^{-8}.

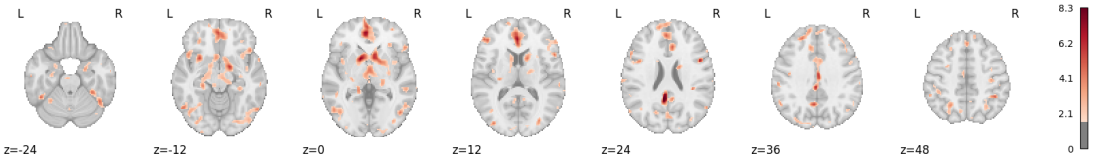
4.3.2.2 Analysis results

We have previously demonstrated the consistency of activation regions detected by ALE and the CBMR parametric inference method for single-group CBMR analysis [Yu et al., 2024]. In this section, we extend our investigation to datasets with multiple groups to assess whether this consistency persists, and to evaluate the similarity of activation regions identified by the CBMR inference using two methods for standard error estimation: Fisher information and a parametric bootstrap approach. Our analysis focuses on group-specific activation regions for each of the three groups and group-wise comparisons between any two groups within the Cue Reactivity dataset, which includes a total of $3,197$ foci [Hill-Bowen et al., 2021]. As the total count of foci exceeds the practical recommendation of at least 200 foci per group, both the parametric statistical test and the parametric bootstrap test are plausible for this dataset. Therefore, we conduct the subsequent analysis using both methods to confirm the consistency of the findings. For comparison, we present z-statistic values generated by the CBMR inference stage using ALE, the CBMR inference based on the parametric statistical tests described in Equation 4.10, and the parametric bootstrap method for all voxels significant at $\alpha = 0.05$ (uncorrected) in Figure 4.5-4.7 and Figure 4.8-B.2. To ensure comparable spatial resolution, we adopt the default FWHM determined by effect size in the Python package NiMARE [Salo et al., 2022].

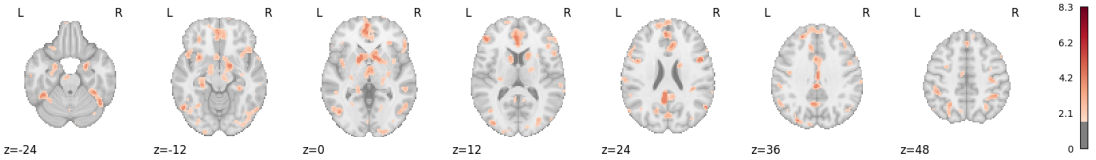
Figure 4.5 - Figure 4.7 demonstrate notable consistency in the detected activation regions (voxels with uncorrected significant $p-$values less than 0.05) across the three groups in the Cue Reactivity dataset. This consistency is particularly evident in the left cerebral cortex, frontal orbital cortex, insular cortex, and left and right accumbens. The observed activations in these regions during cue reactivity reflect the engagement of a complex neural network comprising multiple functional systems: reward processing and motivation, mediated by the nucleus accumbens and its dopaminergic projections; value-based decision making, supported by the orbitofrontal cortex, interoceptive awareness and conscious craving mediated by the insula; cognitive control, attention and emotional regulations, associated with various parts of prefrontal cortex; and learning and memory processes, involving cue-outcome associations encoded in the hippocampus and amygdala. However, slight differences in spatial specificity and smoothness are observed between the methods: ALE provides the smoothest activation regions and detects the largest extent of activation, likely because it is sensitive to the spatial convergence of reported coordinates across studies rather than voxelwise



(a) ALE activation map



(b) Z-statistics map generated by parametric statistical tests (see Equation 4.10)

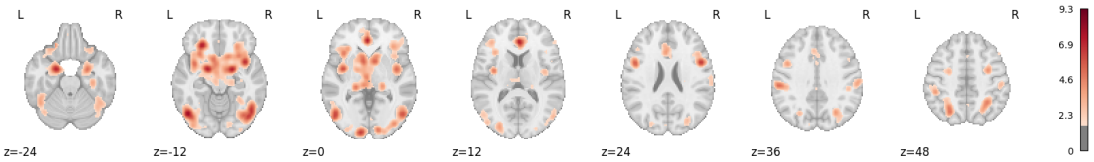


(c) Z-statistics map generated by parametric bootstrap method

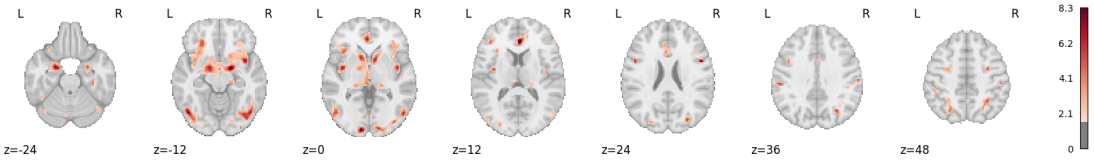
Figure 4.5: Comparison of activation regions for the **Drug** group using the ALE activation map, parametric statistical tests, and the parametric bootstrap method.

effect size magnitude or spatial specificity. In contrast, the parametric statistical tests and the parametric bootstrap method yield more stringent and localised activation regions. This is likely due to CBMR explicitly accounting for both within- and between-study variance, and often incorporate spatial heterogeneity and uncertainty more explicitly, leading to more conservative and spatially precise detection of significant voxels. Despite these differences, all methods exhibit high overall consistency, with only minimal differences in the location of detected activation regions. Notably, the Z-statistics in the maps generated by the parametric bootstrap method are often lower, due to the constraint imposed by the number of bootstrap samples (where the minimal achievable $p-$value is 0.001 with 10^3 bootstrap samples).

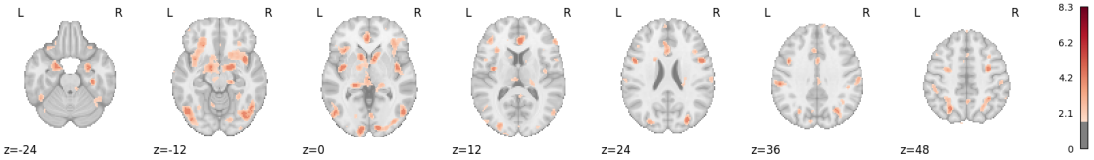
Figure 4.8 illustrates activation patterns observed in the group comparisons between the Drug and Natural groups within the Cue Reactivity dataset. Results from two additional group comparisons are presented in Figure B.1 and Figure B.2 in Appendix B.3). These figures highlight voxels with uncorrected significant $p-$values less than 0.05, demonstrating the reliability of CBMR inference when applied to real datasets, in comparison to kernel-based methods. Brain regions highlighted in red (indicating positive $z-$statistics values and corresponding to uncorrected significant $p-$values)



(a) ALE activation map



(b) Z-statistics map generated by parametric statistical tests (see Equation 4.10)

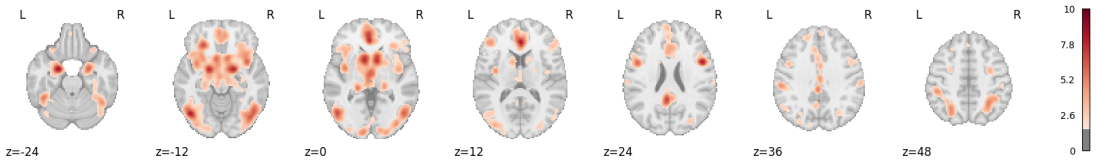


(c) Z-statistics map generated by parametric bootstrap method

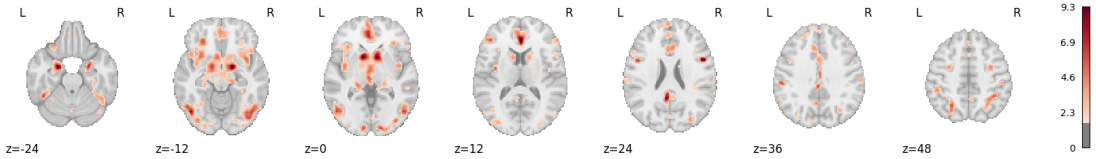
Figure 4.6: Comparison of activation regions for the **Natural** group using the ALE activation map, parametric statistical tests, and the parametric bootstrap method.

represent areas where one groups shows stronger activation than the other group. Conversely, regions highlighted in blue (indicating negative $z-$statistic values and corresponding to uncorrected significant $p-$values) denotes areas where the other group exhibits stronger activation. Figure 4.8 reveals strong consistency in findings between ALE subtraction analysis and CBMR inference, highlighting the stability and accuracy of CBMR even in the presence of group size imbalance.

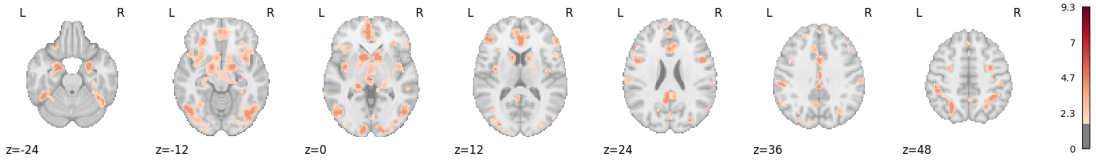
In the Cue Reactivity dataset, the (square root of) sample size and year of publication are considered as publication-level covariates to understand their global effects on group-wise activation intensity functions. Our CBMR analysis revealed that the activation intensity function increases globally by 8.1587% for each unit increase in the square root of sample size, and decreases globally by 0.5397% for each unit decrease in the year of publication. Additionally, we also performed hypothesis testing to evaluate whether these publication-level covariates have a significant effect (i.e., whether their regression coefficients are significantly different from zero). Under the null hypothesis that these publication-level covariates have no effect, we reject the null hypothesis for the square root of sample size at the 0.05 confidence level ($p = 1.1732 \times 10^{-9}$). However, we could not reject the null hypothesis for the year



(a) ALE activation map



(b) Z-statistics map generated by parametric statistical tests (see Equation 4.10)



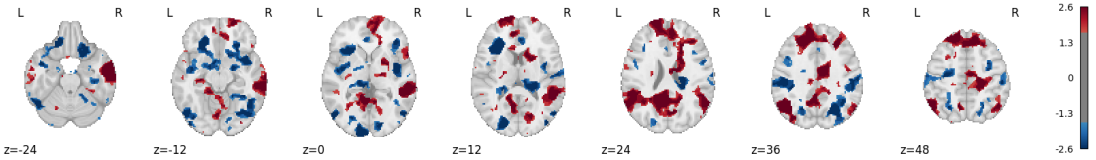
(c) Z-statistics map generated by parametric bootstrap method

Figure 4.7: Comparison of activation regions for the **Reward** group using the ALE activation map, parametric statistical tests, and the parametric bootstrap method.

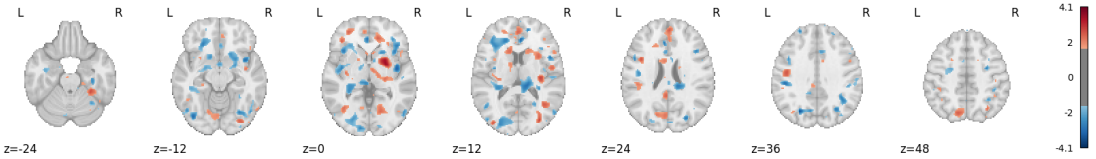
of publication ($p = 0.6681$). Additionally, leveraging the flexibility of the CBMR inference framework, we compared the effects of these two publication-level covariates. Under the null hypothesis that the effect of year of publication is stronger than that of the (square root) of sample size, we rejected the null hypothesis at 0.05 confidence level ($p = 5.1857 \times 10^{-5}$).

4.3.2.3 Computation time

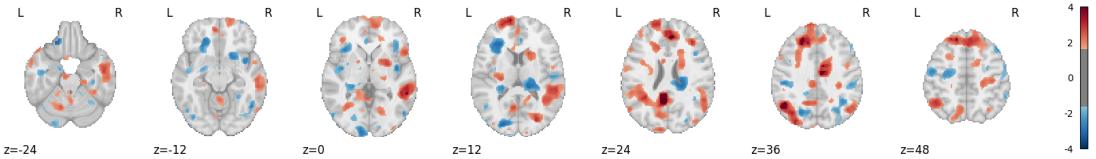
The computation time for CBMR multi-group analysis varies significantly between the inference stage using parametric statistical tests (as described in Equation 4.10) and parametric bootstrap method at voxel level. For large datasets with a sufficient number of publications and foci, where the numerical singularity of the Fisher Information matrix is not a concern, parametric statistical tests are more computationally efficient. These tests allow for flexible homogeneity or group comparison analyses with only a single meta-regression stage. On an NVIDIA GTX 1080 Graphics Card, this approach takes approximately 30 minutes to complete. However, for datasets with an insufficient number of publications or foci (fewer than 200 foci per group), the parametric bootstrap



(a) ALE subtraction analysis for group comparison



(b) Z-statistics map generated by parametric statistical tests (see Equation 4.10) for group comparison



(c) Z-statistics map generated by parametric bootstrap method for group comparison

Figure 4.8: Differences in activation regions between the **Drug** and **Natural** groups: results from ALE subtraction analysis, parametric statistical tests, and parametric the bootstrap method.

method becomes necessary during the inference stage. This method involves generating $1,000$ bootstrap samples, requiring repeated randomisation of foci locations or data re-generation and model refitting for each bootstrap samples, significantly increasing computational complexity. Despite implementing parallelisation to accelerate the process, running model refitting on 5 bootstrap samples in parallel on a single HPC cluster node with Intel(R) Xeon(R) Gold 6126 2.60Hz processors takes approximately 40 minutes. Achieving a comparable computational time to the parametric statistical test method would require around 200 HPC nodes. However, access to HPC resources with such computational capacity is often limited. Therefore, we recommend avoiding the parametric bootstrap method whenever possible and using parametric statistical tests for more computationally efficient analysis.

4.4 Discussion and conclusion

In this work, we have detailed and presented multi-group CBMR, a module implemented in the open source Python package NiMARE, designed for performing meta-regression and meta-inference on coordinate-based meta-analytic fMRI datasets. The meta-regression framework incorporates a spatial model based on spline parametrisation, where a roughness penalty is applied to regularise the smoothness of the spline basis functions. The meta-regression stage fits a generalised linear model with either Poisson or Negative Binomial (NB) distribution at the voxel level, and accommodates publication-level covariates such as sample size and year of publication. Our approach also provides two distinct inference frameworks based on the number of publications or foci in each group within the dataset: For datasets with a sufficient number of foci (above a threshold of 200 foci per group), we recommend a computationally efficient inference method based on parametric statistical tests at the voxel level. This approach is significantly more efficient than the previously proposed Bayesian spatial regression model, while retaining the flexibility and interpretability of hypothesis testing for either spatial homogeneity or group comparisons. For datasets with an insufficient number of publications or foci, we propose a parametric bootstrap method as alternative for more accurate inference. In this method, $p-$values are obtained by comparing observed values to the null bootstrap distribution. While inherently computationally intensive, as it requires repeated randomisation of foci locations or re-generation of data for thousands of bootstrap samples, this approach is necessary to address numerical issues related to inverting numerically singular Fisher Information matrices in small datasets. Through simulations on synthetic data under various experimental settings, we demonstrated that meta-inference outcomes based on parametric statistical tests are valid for datasets with sufficient a number of foci (high underlying intensity functions). However, for datasets with an insufficient number of foci (low underlying intensity functions), $p-$values tend to skew liberally. Despite these challenges, meta-inference outcomes based on parametric bootstrap method remain valid and accurate even under the most challenging simulation settings with insufficient foci. Using the Cue Reactivity dataset, we found that NB model is the preferred stochastic model, as indicated by model comparisons via Likelihood Ratio Test (LRT). The Poisson model, in contrast, cannot explain over-dispersion observed in foci counts. Meanwhile, we also compare the activation regions identified by both ALE and CBMR approaches, utilising both parametric statistical tests or parametric

bootstrap method. These comparisons validate the accuracy and robustness of CBMR inference framework, whether for spatial homogeneity or group comparisons.

There are a few limitations in our work. We employ the parametric bootstrap method for meta-inference on small datasets with insufficient foci count, which improves inference accuracy at the cost of increased computational time. In parametric bootstrap methods, the null distribution of a test statistic is generated by resampling data under the null hypothesis. While effective, this approach is computationally intensive and has limited precision, specifically, $1/N$ precision for N bootstrap samples, making it impossible to achieve finer precision for $p-$values. In future work, we might explore distributional approximations to address this limitation. Possible approaches include asymptotic approximations (e.g., normal or chi-squared distributions) to reduce the number of bootstrap samples required or data-driven methods (e.g., Gaussian or mixture distributions fitted to the observed data) to improve precision without additional computation, particularly for significant $p-$values below 0.05 [Bickel and Freedman, 1981, Hall, 2013]. However, these methods require careful considerations as they may introduce biases. $p-$values can be underestimated or overestimated if the true distribution of test statistics deviates from the assumed theoretical distribution. Moreover, these approaches rely on strong assumptions about the underlying distribution of the test statistic and might fall when applied to non-standard or complex models where the null distribution is unknown.

Another potential direction for future development is to ensure the CBMR framework more accessible to users without coding expertise or access to HPC clusters. Despite the significant effort invested in implementing CBMR as a module in the Python package NiMARE, its usage still requires basic coding knowledge, local environment setup, and familiarity with standard data preprocessing procedures. In contrast, platforms like Neurosynth Compose [Kent et al., 2024] allow users to perform neuroimaging meta-analyses entirely within a browser, avoiding the setup for local Python environment. Neurosynth Compose allows users to search and integrate data from thousands of neuroimaging publications in the Neurosynth dataset and perform fast computations in the cloud using automated analysis pipeline. As a free and open platform for neuroimaging meta-analyses, it eliminates technical barriers for broader accessibility. Our next step is to integrate the CBMR regression and inference pipeline into the Neurosynth Compose platform. This integration will make CBMR accessible directly through a browser interface. Additionally, we plan to explore strategies to accelerate

the parametric bootstrap method using cloud computing resources, ensuring efficient and scalable performance for computationally intensive tasks.

Furthermore, there is considerable potential for advancing the theoretical development of conducting meta-analyses using data from multiple sources. With the growing practice among researchers of sharing full statistical maps, it is becoming increasingly important to integrate additional information from both reported foci or full statistical maps (e.g., $p-$values or $t-$values), when available. Some researchers have proposed Markov melding as a fully Bayesian framework for joining probabilistic sub-models. In this method, evidence from different sources is specified in each sub-model, and sub-models are joined while preserving all information and uncertainty [Goudie et al., 2019]. This approach could enhance inferences derived from CBMR by integrating the magnitude of CBMR activation or even data from image-based meta-analytic results. Another promising avenue for future development involves using CBMR inference outcomes as weights to determine the contribution of voxel-wise statistics from individual publication to the synthesised meta-analytic results. A well-designed choice of voxel-wise weights could stabilise variance and control heterogeneity by ensuring that publications with greater variability contribute less to the overall meta-analysis. Since CBMR inference outcomes involve voxel-wise variation for each publication, they provide a data-driven approach for weighting. Future research will explore where these weights outperform existing methods based on inverse variance, sample size or effect size.

Chapter 5

Efficient Lesion Estimation Using a Spatial Poisson Process and a Scalable Approximate Model

In this chapter, we shift our focus from coordinate-based meta-analysis (CBMA) to the estimation of white matter hyperintensities (WMHs). To model their spatial associations, we propose a scalable voxel-wise generalise linear model (GLM) with a spatial component to capture spatial dependence across the brain. An efficient approximate factorisation method is introduced for scalable regression analysis on large-scale neuroimaging datasets. The model supports flexible statistical inference using Wald or Chi-square tests, allowing voxel-level testing of individual or combined risk factor effects. We also compared standard error estimates and find that the sandwich estimator provides better accuracy and robustness under model misspecification. The proposed approach is validated through simulation studies and applied to the UK Biobank dataset ($N = 13,677$).

5.1 Introduction

MRI is a non-invasive imaging technique to generate high-resolution, three-dimensional anatomical images of the brain. It's widely used for understanding brain structures and functions, as well as for the detection and diagnosis of various neurological diseases. White matter brain lesion refers to areas of damage or abnormalities in the brain's white matter, which typically appear as hyperintensity on T2-weighted and FLAIR MRI

scans [Wardlaw et al., 2013b]. The prevalence of white matter hyperintensities (WMH) is strongly associated with brain ageing and is more commonly observed in older adults, even in the absence of neurological disease. Ageing leads to progressive damage to small blood vessels in the brain, contributing to WHM formation. Additionally, other age-related risk factors (e.g., Hypertension, diabetes, high cholesterol) further accelerate these vascular changes, exacerbating white matter damage. Moreover, white matter integrity naturally declines with age, even in cognitively healthy individuals, leading to structural and functional alternations in brain connectivity [Griffanti et al., 2018]. In addition to ageing, non-ageing-related factors can also contribute to the occurrence of white matter brain lesions. For example, the total burden of WMH is greater in individuals with cerebrovascular risk (CVR) factors such as smoking, hypertension, hypercholesterolemia, diabetes, waist-to-hip ratio and the APOE-ϵ (apolypoprotein-E) status). Among these, hypertension typically emerges as the dominant risk factor, as it introduces arteriosclerosis, leading to thickening, stiffening and narrowing of small arteries. This process results in chronic hypoperfusion and ischaemic damage to white matter, which is highly vulnerable to ischaemia [Debette and Markus, 2010, Veldsman et al., 2020]. An increased WMH burden is strongly correlated with cognitive decline and higher risk of dementia: specifically, white matter is critical for efficient neural communication, and its deterioration can lead to slowed processing speed and executive dysfunction; Additionally, WMH is associated with memory deficits and an increased risk of dementia, as extensive white matter damage is associated with cognitive impairment and disease progression; Moreover, the presence of WMH often indicates underlying cerebrovascular disease [Cees De Groot et al., 2000, Prins et al., 2004].

Beyond the underlying causes and the resulting cognitive decline associated with WMH, their spatial location is another important clinical features. The distribution of WMH determines the neurological functions affected and provides valuable insights for diagnosis and prognosis. For example, the spatial distribution of white matter lesions is essential for clinicians to differentiate between various neurological disorders. Additionally, WMH exhibit significant variability in both spatial distribution and size. In older adults, lesions predominantly occur in the periventricular and subcortical white matter, contributing to age-related cognitive decline [De Leeuw et al., 2001]. While gender is not considered as a strong risk factor [Longstreth et al., 1996], CVR factors influence different vascular territories of the brain depending on the underlying pathology. CVR factors are also associated with fast progression of WMH, which

correlates with cognitive decline and executive dysfunction [Debette and Markus, 2010]. Lesion mapping is a powerful tool in clinical and cognitive neuroscience, as it enables the identification and localisation of brain lesions in relation to neurological and cognitive functions. Additionally, it accounts for spatially varying effects of risk factors, such as age, CVR factors and gender, providing insights into brain regions that are particularly vulnerable to damage from these factors. It is also crucial in monitoring disease progression and facilitating personalised treatment approaches by tailoring interventions based to lesion distribution and associated functional impairment [Moore et al., 2024].

5.1.1 Mass-Univariate Methods and Bayesian Methods

Researchers have developed a voxel-wise mass-univariate approach in which lesion probability is modelled as a function of clinical risk factors (e.g., age, CVR factors, gender and education etc.). In this method, logistic regression is fitted independently at each voxel to estimate voxel-wise lesion probability. However, this approach does not explicitly model the spatial dependence of neighbouring voxels, even though white matter brain lesions often exhibit a spatially clustered distribution rather than being randomly or uniformly distributed across the brain [Rostrup et al., 2012, Lampe et al., 2019]. Additionally, the potential limitation of small sample size and low incident response are not addressed, which can lead to convergence failure and unstable estimates. In some cases, complete separation may occur, giving rise to infinite coefficients. A possible solution is to use penalised logistic regression (e.g. LASSO, Ridge regression) to shrink extreme coefficient estimates and prevent overfitting. Alternatively, Firth's penalised likelihood estimation can be applied to correct bias in small-sample logistic regression and avoid complete separation [Firth, 1993, Kosmidis and Firth, 2021].

In contrast to mass-univariate GLMs at the voxel level, some Bayesian methods have addressed these limitations by incorporating spatially varying coefficients to model the spatial dependence between lesion location and subject-specific covariates. For example, Ge et al. [2014] proposed a Bayesian spatial generalised linear mixed model (BSGLMM) with a probit link function that accounts for the binary nature of the lesion data and includes spatially varying coefficients in Bayesian spatial model. In this model, these coefficients are treated as latent spatial processes, and jointly modelled using a multivariate pairwise difference prior model, a special instance of the multivariate

conditional autoregressive (MCAR) model. However, spatial smoothing priors on the parameters may induce bias by over-smoothing regression coefficients. Additionally, posterior approximation relies on sequential Markov chain Monte Carlo (MCMC) which is not scalable for large datasets, and requires parallel GPUs implementation [Kindalova et al., 2021]. To improve computational tractability, a Bayesian spatial generalised linear model with a structured spike-and-slab prior has been proposed to, leveraging a data augmentation approach [Menacher et al., 2024].

Previous neuroimaging research was limited by small and unrepresentative samples, due to the challenge of recruiting larger and more diverse participates. However, large-scale biomedical databases have addressed these limitations by providing researchers with easier access to extensive datasets. The UK Biobank [Sudlow et al., 2015] and the ABCD study [Casey et al., 2018] are examples of databases containing thousands of subjects with various MRI modalities. For instances, the UK Biobank collects and stores data from $500,000$ UK participants aged between $40-69$, including MRI brain scans. For brain, heart and full body MR imaging data within the UK Biobank, data is currently available for $40,000$ participates, with a target of $100,000$ [Sudlow et al., 2015]. Beyond neuroimaging data, the UK Biobank also provides genetics, health, activity monitor data and other clinical data, allowing researchers to explore risk factors for brain lesion mapping. As large datasets become more widely available, model scalability and computational efficiency are increasingly crucial for analysing these datasets effectively.

Motivated by the lack of spatial dependence modelling in mass-univariate voxel-wise GLMs and the high computational cost of Bayesian posterior inference, we propose a novel voxel-wise GLM with a spatial model to jointly capture spatial dependence across the brain, and an efficient approximate model with strong scalability. This approach provides both statistical interpretability and accuracy, while ensuring computational efficiency through model factorisation for dimension reduction.

5.1.2 Spatial model

The inclusion of spatial priors in Bayesian methods allows the identification of clustered brain lesion patterns while accounting for spatial dependencies among neighbouring locations. For example, in a Bayesian hierarchical spatial model [Xu et al., 2009], a spatial Poisson process prior is employed to model brain activation centres at

population level. Similarly, Cosman Jr et al. [2004] employ an Ising Markov random field (MRF) prior to detect neural activity in fMRI data. In BSGLMM [Ge et al., 2014], spatially varying coefficients are treated as latent spatial processes and jointly modelled with a multivariate pairwise difference prior to estimate brain lesion probability map. Likewise, in BLESS [Menacher et al., 2024], a continuous version of a spike-and-slab prior is applied to the spatially varying coefficients for lesion probability estimation. These studies collectively highlight the importance of spatial models in capturing spatial dependencies in neuroimaging data.

Here, we propose an explicit spatial model parametrised with a tensor product of cubic B-spline basis functions. This spatial model is chosen based on its previous success in modelling spatial coordinate-based neuroimaging data [Yu et al., 2024]. It generates smooth, stable and flexible spatial bases, with the level of spatial smoothness parametrised by the knots spacing: wider knots spacing generates fewer basis functions and greater smoothness, while closer knots spacing increases the number of basis functions, improving the model's ability to capture finer details. Using these pre-specified knots, cubic B-spline basis functions are constructed separately for the x, y and z directions. The coefficients of these basis functions are evaluated at each voxel location and used to construct spatial coefficient matrices along each axis. A three-dimensional coefficient matrix is then obtained by taking the tensor product of the three spatial coefficient matrices, after which weakly supported 3D B-spline bases near or outside the brain edges are removed.

Another important property of the three-dimensional spatial coefficient matrices is the partition of unity. Since each 1D B-spline basis function satisfies this property, their 3D tensor product naturally preserves it across dimensions. This ensures numerical stability by preventing the basis functions from growing arbitrarily large. We assert that the spatial parametrisation using spline bases efficiently captures spatial dependence. Additionally, alternative spatial models, such as Gaussian kernels and random Fourier features [Rahimi and Recht, 2007], would produce comparable results with minimal differences [Yu et al., 2024].

5.1.3 Parameter estimation

As a flexible generalisation of ordinary linear regression, the Generalised Linear Model (GLM) links the linear predictor to the response variable through a link function,

allowing the variance of each observations to be a function of its predicted value. Since the introduction of GLMs in 1972 [Nelder and Wedderburn, 1972], the iteratively re-weighted least square (IRLS) method remains popular for maximum likelihood estimation (MLE). Recognised for its reliability and computational efficiency, IRLS effectively addresses the complexity introduced by the non-linearity in GLMs. As a result, it has become the default methods for most Python and R packages for GLMs, including *MASS*, *mgcv*, *lme4* in R [Ripley et al., 2013, Wood and Wood, 2015, Bates et al., 2015], and *statsmodel* and *scikit-learn* in Python [Seabold and Perktold, 2010, Pedregosa et al., 2011]. These advancements, combined with improvements in computing, have enabled efficient parameter estimation for GLMs. The Fisher Scoring algorithm was introduced as a modification of Newton-Raphson optimisation, offering more stable and faster convergence, particularly on highly non-linear optimisation surfaces [McCullagh, 2019]. It also demonstrates better performance in small sample sizes and high-dimensional settings, as the expected Fisher information matrix facilitates smoother parameter updates and reduces numerical instability [Ripley, 2002]. Since the 1990s, penalisation has been integrated into GLM parameter estimation to improve stability and mitigate overfitting, particularly for high-dimensional data. Notable examples include the Lasso, Ridge and Elastic Net penalty terms [Tibshirani, 1996, Hoerl and Kennard, 1970, Zou and Hastie, 2005].

In brain lesion probability regression, parameter estimation is performed for a model that includes multiple risk factors as covariates (e.g., age, CVR factors and gender), each varying spatially. Polynomial terms (up to cubic) for each risk factor might also be considered in practice, leading to a substantial number of parameters. To optimise computational efficiency in large-scale, voxel-wise GLM analysis, it is crucial to vectorise computations across voxels while retaining shared model components. However, most existing GLM tools and software are not fully optimised for vectorised operations, especially when dealing with complex or large-scale data structures. For instance, while standard GLM implementation in R (*glm*) and Python (*statsmodel*) employ vectorisation to improve efficiency, they often rely on iterative methods that do not fully exploit vectorised parallel processing [Seabold and Perktold, 2010]. Moreover, voxel-wise neuroimaging analyses often require custom implementations to achieve optimal performance. As a result, the limited vectorisation in conventional GLM software can introduce significant computational bottlenecks, particularly in large-scale applications such as our model for brain lesion probability mapping.

Additionally, efforts should focus on reducing the dimensionality of variables to improve scalability, rather than using the product of the number of subjects and voxels as the dimension. With tens of thousands of subjects and hundreds of thousands of voxels per subject, this approach would likely prevent memory bottleneck and alleviates computational burden. In this work, we provide rigorous proofs on factorisation and simplification in the coefficient updating equations of voxel-wise GLMs for dimension reduction, along with Taylor expansion to approximate computational intractable terms.

As an efficient and fundamental approach for parameter estimation in GLMs, Maximum Likelihood Estimation (MLE) identifies the parameter values that maximise the probability of the observed data under the assumed model. It also ensures asymptotically normal parameter estimates, particularly for large samples, facilitating straightforward and accurate hypothesis testing and confidence interval estimation. An efficient numerical method for large-scale GLM estimation task is the Limited-memory Broyden-Fletcher-Goldfarb-Shanno (L-BFGS) algorithm, a quasi-Newton optimisation technique that approximates second-order information while maintaining computational efficiency [Liu and Nocedal, 1989]. L-BFGS offers several advantages over other parameter estimation algorithms: it efficiently manages memory by storing only a limited number of past gradients, achieves fast convergence by using an approximate inverse Hessian instead of computing it explicitly at each iteration, and demonstrates robust performance in high-dimensional settings such as neuroimaging analysis. Building on its effectiveness and efficiency in optimising voxel-wise GLMs for coordinate-based fMRI data [Yu et al., 2024], we will apply it in this work.

In practice, for high-dimensional or large-scale datasets such as the UK Biobank, computing and inverting the Hessian matrix or Fisher information matrix becomes computationally prohibitive, as its complexity scales poorly with the number of parameters. This limitation makes second-order optimisation methods, such as IRLS and Fisher scoring, impractical for large problems. As a result, gradient-based optimisation methods have been widely explored as alternative approaches. Unlike IRLS which explicitly solves for parameter updates, gradient-based methods iteratively adjust parameters in the direction of steepest descent. However, as first-order optimisation methods, gradient-based approaches often exhibit slow convergence and require careful parameter tuning, particularly when the Hessian or Fisher Information matrix is ill-conditioned, leading to numerical instability. To address this, preconditioners are often introduced to accelerate convergence. A preconditioner is a transformation

applied to the gradient that effectively rescales the optimisation problem. In this work, we propose an approximate, scalable gradient-based optimisation method and provide a practical recommendation for selecting an appropriate preconditioner to ensure numerical stability in large-scale GLM estimation.

5.1.4 Statistical Inference

In brain lesion probability map, the central objective is to estimate the voxel-wise probability distribution of lesion presence across the brain, conditioned on clinical risk factors (e.g., age, gender, hypertension). This estimation models how individual risk factor contributes to the spatial distribution and likelihood of brain lesions, contrasting to the lesion incidence rate at the baseline level. Specifically, the subject-specific brain lesion probability map is interpreted as a superposition of probability maps, each associated with an individual risk factor. By averaging these covariate-specific brain lesion probability maps across all subjects, we consistently perform significance-based GLM hypothesis tests to evaluate the effects of different risk factors at the voxel level, facilitate comparisons between risk factor effects, and enable groups comparisons.

When GLMs involve multiple hypotheses, such as comparing the effects of risk factors voxel-by-voxel, multiple comparison becomes a significant challenge. Uncorrected significance testing increases the risk of inflated Type I errors, particularly in neuroimaging data. For example, in localised tests across $228,484$ voxels (within an MNI152 2mm brain mask), even a 5% false positive rate could lead to a substantial number of false positives. Therefore, multiple testing corrections are applied in neuroimaging analyses by controlling either the family wise error rate (FWER), typically using the null maximum distribution [Westfall and Young, 1993], or the false false discovery rate (FDR) using the Benjamini-Hochberg (BH) procedure [Benjamini and Hochberg, 1995]. FWER correction provides stringent control against false positives, ensuring that results are highly reliable and reproducible. However, it can be overly conservative, reducing statistical power and making it difficult to detect true effects. In contrast, FDR correction balances Type I and Type II errors by allowing some false positives while maintaining statistical power. This approach enables the detection of more significant findings, while still controlling the overall rate of false discoveries among detected results, though it does not completely eliminate false positives.

When a GLM is misspecified and does not accurately capture the true underlying data generating process, standard maximum likelihood-based variance estimation may produce biased and inconsistent standard errors. In such cases, the sandwich estimator, also known as heteroscedasticity-consistent covariance matrix estimator (HCCME) or robust variance estimator, provides a more reliable alternative for inference [White, 1980]. This method does not rely on correct model specification and remains valid under heteroscedasticity or other forms of dependence in the residuals. By adjusting for unknown forms of variability and residual dependence, the sandwich estimator ensures asymptotically consistent standard error estimates, improving the robustness of hypothesis testing and confidence interval estimation [Freedman, 2006]. In applications such as spatial point process modelling of brain lesions, where data exhibits spatial dependence and potential deviations from assumed distributions (such as overdispersion in variance), the use of robust standard error estimation techniques like the sandwich estimator is critical for obtaining valid and reliable statistical inferences.

5.2 Methods

In this section, we provide an overview of the modelling pipeline for estimating brain lesion probability maps and analysing the effects of spatially varying covariates at the voxel level, along with a model factorisation to reduce dimensionality and create a scalable approximation of the original model. Each component is described in detail in Section 5.2.1 through Section 5.2.3.

5.2.1 Generic GLM structure

A Generalised linear model (GLM) with a Bernoulli distribution and a logit link function is commonly used for binary classification problems, such as predicting the probability of an event occurring. Inspired by this, brain lesion occurrence can be naturally modelled using a Bernoulli distribution, where each voxel for a given subject is represented as either 0 or 1, indicating the absence or presence of a lesion respectively. The logit function ensures that the predicted probability remains within the range $(0, 1)$. However, motivated by the previous success of the Poisson point process and the accuracy of Poisson approximation for low-rate Bernoulli data, we consider a Poisson model [Eisenberg et al., 1966]. We also provide proof in the Appendix C.1.1 of the

equivalence between these two distributions, when brain lesion counts are restricted to either 0 or 1 per voxel in each subject.

Assume there are N voxels for each of M subjects, and the brain lesion data at voxel j for subject i is represented by the voxelwise lesion count Y_{ij} (either 0 or 1). We define a subject-specific lesion count vector as $Y_i = [Y_{i1}, \cdots, Y_{iN}]^\top$. To model spatial variation, we construct a spatial design matrix $B(N \times P)$, parametrised by P cubic B-spline bases, and a covariates matrix $Z(M \times R)$, containing R spatially varying covariates for each of the M subjects. For estimating brain lesion probability map, the primary objective is the voxelwise brain lesion probability function for each subject i, which accounts for the effects of multiple spatially varying covariates. In this setting, the model for voxel j in subject i is expressed as

$$\log(\mu_{ij}) = \log[\mathbb{E}(Y_{ij})] = (Z_i^\top \otimes B_j^\top)\beta \tag{5.1}$$

where $\beta(PR \times 1)$ represents the regression coefficients of the spatially varying covariates, $Z_i^\top$ is the i^{th} row of covariates matrix Z, and $B_j^\top$ is the j^{th} row of spatial design matrix B. The estimated lesion probability is μ_{ij} for subject $i = 1, \cdots, M$ and voxels $j = 1, \cdots, N$, and the subject-specific lesion probability vector is defined as $\mu_i = [\mu_{i1}, \cdots, \mu_{iN}]^\top$. The GLM structure for all voxels in all M subjects is then

$$\log(\mu) = \log[\mathbb{E}(Y)] = X\beta = (Z \otimes B)\beta \tag{5.2}$$

where $\otimes$ represents Kronecker product and $X = Z \otimes B$ is the Kronecker product of Z and B, with size $(MN \times PR)$. The vectors $Y = [Y_1, Y_2, \cdots, Y_M]^\top$ and $\mu = [\mu_1, \mu_2, \cdots, \mu_M]^\top$ are of size $(M \times N)$, representing the voxelwise lesion count and the estimated voxelwise lesion probability for all of M subjects, respectively.

Here, the spatial design matrix B is formed as a tensor product of cubic B-spline coefficient matrices across the x, y, z axes: $B = B_x \otimes B_y \otimes B_Z$, where, for example $B_x(v_x \times n_x)$ represents the coefficients of v_x voxels evaluated at n_x B-spline bases. Columns corresponding to weakly supported bases are removed for efficiency and numerical stability. The covariate matrix $Z(M \times R)$ captures spatially varying covariates, which are standardised to have mean 0 and variance 1 before optimisation. In practice, an additional column of ones is included in both Z and B for capturing the overall mean.

Given the validity of the Poisson approximation as the stochastic component of this GLM, the log-likelihood function to be optimised is:

$$\begin{aligned} l(\theta) = l(\beta) &= \sum_{i=1}^{M}\sum_{j=1}^{N}[Y_{ij}\log(\mu_{ij}) - \mu_{ij} - \log(Y_{ij}!)] \\ &= Y^\top \log(\mu) - \mathbf{1}^\top \mu \end{aligned} \tag{5.3}$$

where vectorisation is applied, and the term $\log(Y_{ij}!)$ vanishes since Y_{ij} takes only the values 0 or 1.

The most computationally intensive stage of this modelling pipeline is the estimation of the unknown coefficients of spatially varying covariates, β, of size $PR \times 1$. A common approach for estimating these coefficients is Maximum Likelihood Estimation (MLE), based on the log-likelihood function across all subjects and voxels, as detailed in Equation 5.3. For optimisation, L-BFGS is an efficient algorithm particularly useful for large-scale and high dimensional settings [Liu and Nocedal, 1989]. L-BFGS approximates the inverse Hessian, providing faster convergence than gradient descent while requiring significantly less memory than the standard BFGS. As a result, we use L-BFGS in practice to optimise the full model and obtain the optimal estimates for the coefficients of the spatially varying covariates.

5.2.2 Scalable approximate model factorisation

Having derived the explicit log-likelihood function, we observe that it's a summation of the Poisson log-likelihood over all subjects and voxels, therefore, all variables involved in Equation 5.3 (i.e., Y and μ) have dimensions proportional to the product of the number of subjects and voxels. This can scale to multiple trillions, resulting in extremely high memory demands. Although L-BFGS is a memory efficient quasi-second-order optimisation method, it still struggles with such high-dimensional data. Therefore, to address implementation complexity and memory constraints, we introduce a model factorisation in this sections, ensuring that all involved parameters have dimensions no larger than either the number of subjects or the number of voxels, rather than their product.

Here, we propose an approximate scalable optimisation approach derived from the updating equation of IRLS, introducing computational modifications to improve efficiency. Building upon our previous success using GLMs with a Poisson model

and log link for estimating the intensity function in spatial point process data, where spatial and covariate effects are modelled separately for significant computational advantages [Yu et al., 2024], we adopt this established framework to obtain a robust initialisation for β in the current GLM,

$$\log \mathbb{E}(Y) = \eta = \eta^B + \eta^Z = B\xi + Z\gamma \tag{5.4}$$

where $\xi(P \times 1)$ represents regression coefficients corresponding to spatial effects, while $\gamma(R \times 1)$ represents regression coefficients for globally constant covariate effects. We interpret the spatial term $B\xi$ as a shared, log-transformed brain lesion intensity map η^B that reflects both spatial covariates effects and the underlying neurological structures common to brain lesion mechanism across all subjects. In contrast, the global effect term $Z\gamma$ captures subject-specific covariate effects, modulating the spatial log-transformed intensity map by a global scaling factor η^Z. Although this approach offers reduced flexibility in explicitly modelling the spatial variation of brain lesion intensity for each covariate, it substantially improves computational efficiency by separating the estimated voxelwise brain lesion intensity into spatial and global components: $\mu_{ij} = \mu_i^Z \mu_j^B = \exp(Z_i^\top \gamma) \cdot \exp(B_j^\top \xi)$.

For our GLM structure that accounts for spatially varying covariate effects, the updating equation for the Iteratively Re-weighted Least Square (IRLS) algorithm is given by,

$$\beta^{(k+1)} = \beta^{(k)} + (X^\top W^{(k)} X)^{-1} X^\top (Y - \mu^{(k)}) \tag{5.5}$$

Here, Y is the brain lesion data vector with dimensions $MN \times 1$, where each entry is binary (0 or 1), indicating the absence or presence of a brain lesion for each subject at each voxel location. $X = Z \otimes B$ (dimension $MN \times PR$) represents spatial parametrisation for each covariate. For iteration (k), the diagonal weight matrix $W^{(k)}$ is defined as $W^{(k)} = diag(\mu_{ij}^{(k)})$, where the diagonal elements $\mu_{ij}^{(k)}$ represent estimated mean values, indexed by $i = 1, \cdots, M$ and $j = 1, \cdots, N$. Additionally, the optimised spatial coefficient ξ obtained from the separable model is used as the initialisation for β.

Ideally, the Hessian of the log-likelihood $(X^\top W^{(k)} X)^{-1}$, should be computed using the updated brain lesion mean estimation μ in each iteration (k). However, this introduces a substantial computational burden since dimensions of both X and W involve MN. Moreover, the matrix inversion itself incurs an $O(n^3)$ computational complexity, making

it more computationally expensive. As a result, we propose a scalable approximate model factorisation that replaces variables with prohibitively high dimensions, MN, with variables of dimension at most either M or N in the IRLS updating equation 5.5. For the Hessian component $(X^\top W^{(k)} X)^{-1}$, rather than updating it at each iteration k, we fix it by setting its diagonal elements equal to the subject-specific estimated brain lesion intensity at the voxel level $\hat{\mu}_{ij} = \hat{\mu}_j^B \hat{\mu}_i^Z = \exp(B_j^\top \hat{\xi}) \exp(Z_i^\top \hat{\gamma})$, and keep these values unchanged throughout all iterations. Leveraging the separable structure of $\hat{\mu}$, we further simplify the weight matrix as a Kronecker product of two separable weight matrices corresponding to spatial effect and covariate effect respectively: $W = W_Z \otimes W_B$, where $W_Z = diag(\{\mu_i^Z\}_{i=1,\cdots,M})$ and $W_B = diag(\{\mu_j^B\}_{i=1,\cdots,N})$. Although using a fixed Hessian in the updating equation is less accurate compared to updating it in every iteration, this strategy substantially improves computational efficiency at the expense of precision. However, this algorithm will still converge to the optimal solution since the fixed Hessian is regarded as a preconditioner, transforming the IRLS updating equation 5.5 into a gradient descent algorithm. A preconditioner, in this context, is a transformation applied to the gradient that accelerates convergence and stabilise the optimisation process.

For the gradient and residue component $X^\top(Y - \mu^{(k)})$, we consider it essential to update this term at each iteration, which constitutes the primary computational challenge during optimisation. A conservative and reliable approach is to compute the exact update at each iteration while parallelising the computation by partitioning it into hundreds of independent blocks, each of which can be executed simultaneously. In practice, we achieve this using the open-source Python library Dask, which efficiently handles large-scale data by splitting tasks into smaller, manageable units executed concurrently, significantly improving computational efficiency and reducing memory usage [Rocklin, 2015]. A more aggressive but feasible alternative involves simplifying this component by employing a Taylor expansion of $\mu^{(k)}$ around its mean vector across subjects, represented as $\exp(\mathbf{1}_M \otimes \bar{\eta}^{(k)})$. Using properties of Kronecker products and the vectorisation-Kronecker product identity, we further simplify this expression to ensure that all involved parameters scale proportionally to either M or N, rather than their product MN. The simplified form is as follows,

$$X^\top(Y - \mu^{(k)}) = vec(Z^\top \tilde{Y} B) - (Z^\top \mathbf{1}_M) \otimes (B^\top \exp(\bar{\eta})^{(k)}) - [(Z^\top \tilde{Z}) \otimes (B^\top \tilde{B})]\beta^{(k)} \quad (5.6)$$

where $\tilde{Y}$ denotes the reshaped brain lesion data vector Y, transformed from dimension $(MN \times 1)$ to $(M \times N)$, and let $\tilde{\beta}$ represent the reshaped parameter $\beta^{(k)}$, converted from $(PR \times 1)$ to $(R \times P)$ accordingly. Define $\tilde{Z} = [I_M - \frac{1}{M}\mathbf{1}_M\mathbf{1}_M^\top]$ as the column-centred (demeaned) version of Z. Additionally, let $\tilde{B} = diag(\exp(\bar{\eta})^{(k)}B)$, which has dimension $N \times N$. For detailed proofs and further explanations, please refer to Appendix C.2.1.

A summary of this optimisation algorithm is provided in Algorithm 3. We assert that computing both Hessian and gradient terms block by block with parallelisation using the Python package Dask ensures an accurate update equation in each iteration, albeit at the cost of intensive computational complexity. This approach serves as our baseline method. Additionally, we investigate the use of a fixed Hessian as a preconditioner throughout all iterations and a Taylor expansion to approximate the gradient and residue term. Their accuracy is assessed by comparison with the baseline method and validation through corresponding PP-plots, with further details provided in Section 5.3.3.

Algorithm 3 Scalable approximate model factorisation with improved computational efficiency

Input: Brain lesion data vector Y, covariate matrix Z, spatial parametrisation matrix B, and parameters ξ and γ in the separable GLM,

Output: Estimated parameter vector β

while *Current* $l(\xi, \gamma)$ *and previous* $l_{prev}(\xi, \gamma)$ *differ by more than pre-defined tolerance (*$1e^{-10}$ *by default)* **do**

  Update the regression coefficients of the spatial effect (ξ) and the covariate regression coefficients (γ) using the L-BFGS algorithm in the separable GLM with a Poisson model and log link function: $\log(\mu) = \eta = B\xi + Z\gamma$.

Initialize result $\beta \leftarrow \hat{\xi}$ Save the estimated

while $\|\beta^{(k+1)} - \beta^{(k)}\| > tol$ *(*$1e^{-10}$ *by default)* ***AND*** *iteration_count* $<$ *max_iter (*300 *by default)* **do**

  if *preconditioner_mode = "exact"* **then**

    Compute $X^\top W^{(k)} X$ block-wise.

  else if *preconditioner_mode = "approximate"* **then**

    Use a fixed Hessian matrix $W^{(k)} = W_Z \otimes W_B = diag(\mu^Z) \otimes diag(\mu^B) = diag(\exp(Z\gamma)) \otimes diag(\exp(B\xi))$ so that $X^\top W^{(k)} X$ remains fixed throughout all iterations.

  if *gradient_mode = "exact"* **then**

    Compute $X^\top (Y - \mu^{(k)})$ block-wise.

  else if *gradient_mode = "approximate"* **then**

    Compute $X^\top (Y - \mu^{(k)})$ using a Taylor expansion around the point $\exp(\mathbf{1}_M \otimes \bar{\eta}^{(k)})$

    $$X^\top (Y - \mu^{(k)}) = vec(Z^\top \tilde{Y} B) - (Z^\top \mathbf{1}_M) \otimes (B^\top \exp(\bar{\eta})^{(k)}) - [(Z^\top \tilde{Z}) \otimes (B^\top \tilde{B})]\beta^{(k)}$$

  Update $\beta^{(k+1)}$ by computing

$$\beta^{(k+1)} = \beta^{(k)} + (X^\top W^{(k)} X)^{-1} X^\top (Y - \mu^{(k)})$$

return $\hat{\beta}$

5.2.3 Statistical inference

The final stage of this efficient lesion estimation pipeline involves conducting statistical inference to evaluate covariate effects at the voxel level. This includes testing their significance or evaluating whether their effects are identical through group comparisons between two groups. The analysis results are then presented as statistical maps in NIfTI format. For voxel-wise hypothesis testing in both cases, the mean estimated

lesion log-transformed contrast involving one or multiple covariates, $\bar{\eta}_C = (C \otimes B)\hat{\beta}$, or its exponential $\bar{\mu}_C$, is used to construct voxel-wise test statistics along with their standard errors. Here, C represents the contrast matrix that specifies the hypothesis tests, with further details and descriptions provided later in this section.

For testing whether the effect of a specific covariate is significant, or more generally, whether a flexible linear combination of multiple covariates is significant at the voxel level, we define a contrast matrix $C(S \times R)$ for R covariates and specify the voxel-wise null hypothesis $H_0 : (C \otimes B_j^\top)\hat{\beta} = \mathbf{0}_{S\times 1}$ for any voxel $j = 1, \cdots, N$. The test statistic at the voxel j is computed as:

$$[(C \otimes B_j^\top)\hat{\beta}]^\top [(C \otimes B_j^\top) Cov(\hat{\beta}) (C \otimes B_j^\top)^\top]^{-1} [(C \otimes B_j^\top)\hat{\beta}] \xrightarrow{D} \chi_S^2 \tag{5.7}$$

Here, the contrast matrix $C(S \times R)$ is normalised such that each row sums to 1, $B_j^\top(1 \times P)$ represents the j^{th} row of spatial parametrisation matrix, and $\hat{\beta}(PR \times 1)$ is the optimised regression coefficient, consequently, $Cov(\hat{\beta})$ has dimensions $PR \times PR$. For computational efficiency, we vectorise the test statistics across all voxels to accelerate the inference. The degrees of freedom for the statistical test are determined by the number of rows in the contrast matrix $C(S \times R)$. If there are multiple rows in C ($S > 1$), the null hypothesis H_0 holds only if the multiple hypothesis tests corresponding to each row of C are simultaneously satisfied. The corresponding p-values are then computed by approximating the test statistics using a Chi-square distribution.

In the scenarios where only a single hypothesis test is conducted ($S = 1$) at each voxel j, the statistical test simplifies to a Wald test. Consequently, the test statistics in Equation 5.7 can be written in the following simplified form,

$$\begin{aligned} W_j &= \frac{(C \otimes B_j^\top)\hat{\beta}}{(C \otimes B_j^\top) Cov(\hat{\beta}) (C \otimes B_j^\top)^\top} \\ \Rightarrow W &= \frac{(C \otimes B)\hat{\beta}}{(C \otimes B) Cov(\hat{\beta}) (C \otimes B)^\top} \end{aligned} \tag{5.8}$$

Here, W_j denotes the Wald test statistic at voxel j, and $W = [W_1, \cdots, W_j]^\top$ represents the vector of Wald test statistics across all voxels. The corresponding voxel-wise p-values, $P = [P_1, \cdots, P_N]^\top$ can be computed by assuming that the Wald test statistics W_j follow a standard normal distribution. Finally, p-value maps can be generated and

thresholded to control the false discovery rate (FDR) at 5% following the Benjamini-Hochberg (BH) procedure [Benjamini and Hochberg, 1995].

For testing whether the covariate effect estimations are independent of sample size, we extend the general form of voxel-wise test statistics from Equation 5.7 to group comparisons between groups with different number of subjects but identical underlying spatial covariate effects as a sanity check. For simplicity, we focus on comparing two groups, though this approach can be easily extended to comparisons involving more than two groups. The regression coefficients for group g and group g', denoted as $\hat{\beta}_g$ and $\hat{\beta}_{g'}$, are estimated independently using the maximum likelihood estimation (MLE) approach, as described in equation 5.3. Under the null hypothesis that $\bar{\eta}_g$ and $\bar{\eta}_{g'}$ are identical regardless of the group-specific sample sizes M_g and $M_{g'}$. By concatenating the group-specific regression coefficients as $\hat{\beta} = [\hat{\beta}_g, \hat{\beta}_{g'}]$, and defining a contrast matrix of dimension $1 \times 2R$ as $C = [1, \cdots, 1, -1, \cdots, -1]^\top$,the general form of voxel-wise test statistics from equation 5.7 becomes applicable in this scenario. Alternatively, since only two groups are involved, the Wald test statistics from equation 5.8 provides a more intuitive approach. It is important to note that $Cov(\hat{\beta})$ is a block diagonal matrix with $Cov(\hat{\beta}_g)$ and $Cov(\hat{\beta}_{g'})$ on the diagonal. The off-diagonal blocks are all zero because the group-specific regression coefficients are optimised independently, meaning there is no correlation between them.

The standard approach for estimating the covariance matrix of the regression coefficients $\hat{\beta}$ is to invert the Fisher Information matrix. Under general regularity conditions, the maximum likelihood estimator (MLE) $\hat{\beta}$ is asymptotically normally distributed around the true parameter value β, such that: $\hat{\beta} \sim N(\beta, [I(\beta)]^{-1})$. However, this standard maximum likelihood-based variance estimation lacks robustness and accuracy, particularly when the GLM is misspecified and does not accurately represent the underlying data-generating process. This issue arises in our application, where the standard maximum likelihood-based covariance estimation can be biased and might underestimate the variability of the regression coefficients. As a result, the sandwich estimator provides a more reliable and robust alternative, as it adjusts for unknown forms of variability and residual dependence. For our current GLM described in Equation 5.2, the sandwich estimator for the covariance matrix of regression coefficients is

given by

$$
\begin{aligned}
Cov(\hat{\beta}) &= (X^\top WX)^{-1}[\sum_{i=1}^{M} U_i U_i^\top](X^\top WX)^{-1} \\
&= (X^\top WX)^{-1}[\sum_{i=1}^{M} X_i^\top (Y_i - \mu_i)(Y_i - \mu_i)^\top X_i](X^\top WX)^{-1}
\end{aligned}
\tag{5.9}
$$

Here, $U_i = X_i^\top (Y_i - \mu_i)$, where $Y_i = [Y_{i1}, \cdots, Y_{iN}]$ and $\mu_i = [\mu_{i1}, \cdots, \mu_{iN}]$ represent the voxel-wise brain lesion count and estimated lesion intensity mean vectors, respectively. The meat term, $\sum_{i}^{M} U_i U_i^\top$, which estimates the variance of the estimating functions, accounts for the spatial correlation of voxels within each subject Y_i. In Section 5.3.2, we will present a detailed and thorough comparison of these two covariance estimation methods, evaluating their robustness and accuracy through PP-plots.

5.3 Simulation study

To quantitively evaluate and demonstrate the computational accuracy and efficiency of brain lesion probability estimation, as well as the validity of statistical inference, we conducted extensive simulations under various settings. In this section, we first describe the process of simulating lesion data where the ground truth is known, considering both homogeneous lesion intensity functions across space and those with bumps across 1D, 2D, and 3D spaces. We then perform a series of simulation studies to assess the performance of the full model, which is based on a spatial Poisson point process with a GLM structure, using small-scale datasets due to the computational constraints. The evaluation involves comparing parameter estimates, including their bias, variance and mean square error (MSE), against the known ground truth. In addition, we assess the validity of statistical inference for spatial covariate effects through group comparisons. Specifically, we compare two groups with identical underlying covariate effects but different sample sizes, using PP-plots to evaluate the accuracy and reliability of statistical inference.

Subsequently, we extend the model fitting to the scalable approximate model factorisation using the same small-scale simulated data. As outlined in Section 5.2.2, we proposed two approaches: utilising the Python library **Dask** for parallelisation or employing the **approximate** mode based on a separable GLM and a Taylor expansion for computing the preconditioner and gradient. We consider the method using Dask

for both preconditioner and gradient as the baseline, as it produces the exact IRLS updating equation at each iteration. To assess the overall model performance, we provide a detailed comparison against the baseline method, evaluating parameter estimates, computational time, and the validity of the corresponding statistical tests through PP-plots.

5.3.1 Data generation process

To evaluate the accuracy of lesion probability estimation and the robustness of statistical inference, we propose a data generation process in 1D, 2D and 3D spaces the is analogous to the spatial covariate effects of real lesions. We simulate lesion locations under two scenarios for background signals: either spatial homogeneity across the space or bump signals generated with a Gaussian probability intensity. In the latter case, a single bump is centred in the middle of the 1D, 2D or 3D setting. To ensure comparability, we scale the total lesion probability across spaces by a factor that results in an average lesion prevalence of 1%, aligning with real-world observations. To incorporate spatial covariate effects associated with subject-specific covariates, we introduce an additional lesion probability function on top of the background signals. This function is either spatially homogeneous or follows a bump signal patterns but differs from the background signal in terms of bandwidth or Gaussian variance. It is further modulated by one or more subject-specific covariate values, scaled by a pre-defined coefficient or coefficient vector. Since we standardise the covariate matrix to have a mean of 0 and a standard deviation of 1, some covariate values might be negative, leading to a negative signal being added to the background signal. For the settings where both the background signal and the covariate-associated signal contain bumps, we add a homogeneous background signal of 0.01 across the entire space. This prevents an excessive number of voxels from having numerically zero lesion probability, thereby significantly, thereby significantly improving numerical stability in statistical inference.

After determining the subject-specific spatial lesion probability functions, we generate the lesion occurrences at the voxel level for each subject. Each voxel is assigned a value of 1 (indicating presence) or 0 (indicating absence) based on a Bernoulli distribution, with the voxel-wise lesion probability as the mean parameter. To store the voxel-wise lesion counts for each subject, we construct a vector Y of dimension $MN \times 1$, where M represents the number of subjects and N represents the number of voxels. Given

that lesion prevalence in the simulated data is only 1%, majority elements of vector Y are zero, therefore, we use a sparse representation for Y to optimise storage efficiency in practice.

For the covariate Z, we use one-hot encoding for categorical risk factors, such as group names, to ensure numerical input and naturally categorise subjects into multiple groups. While for continuous covariates, we apply standardisation to achieve a mean of 0 and a standard deviation of 1, preventing large differences in magnitude and improving numerical stability. Additionally, to improve estimation accuracy, we include quadratic and cubic terms for the covariates of primary interest, increasing the flexibility of our GLM model structure. This adjustment is particularly important, as we have observed instances of overshoot and underestimation in practice, even in the simplest 1D homogeneous simulation experiments. For further details, please refer to the Appendix C.1.2

In each simulation setting, the spatial design matrix $X(N \times P)$ is constructed using cubic B-spline bases with knots spacing of either 10 or 5. To ensure numerical stability, B-spline with weak support, defined as having a total sum across all voxels below 0.1, are removed. An intercept column of ones is then added to X to account for the overall mean.

5.3.2 Regression and inference for probability estimation using the full model

Since the accuracy estimation including bias, standard deviation (Std) and mean square error (MSE), as well as inference validity evaluated by PP-plots, show similar patterns across 1D, 2D and 3D scenarios, we present only the results for the 3D setting. As the most challenging case, the 3D scenario rigorously evaluates the accuracy and robustness of our GLM framework. To reduce the computational burden of performing the full GLM analysis, including both regression and inference, we limit the 3D space to dimensions of $20 \times 20 \times 20$, totalling 8000 voxels. Table 5.1 presents the relative bias, standard deviation (Std) and mean squared error (MSE) of probability function estimation within a 3D space across $6,000$ subjects, divided into two groups ($2,000$ and $4,000$ subjects, respectively). Probability functions were modelled using a GLM framework that includes cubic polynomial covariate terms. The minimal differences

between the estimated probability and the actual underlying probability demonstrate that our GLM framework produces highly accurate probability estimations.

Table 5.1: Relative bias, standard deviation (Std) and mean squared error (MSE) for probability function estimation in 3D space with $8,000$ voxels across 6000 subjects. The background probability function is either spatially homogeneous or includes a Gaussian bump centred in the middle of the space. A covariate-associated probability function with a matching homogeneous or Gaussian-bump type (with the same centre but a different bandwidth or Gaussian variance) is added.

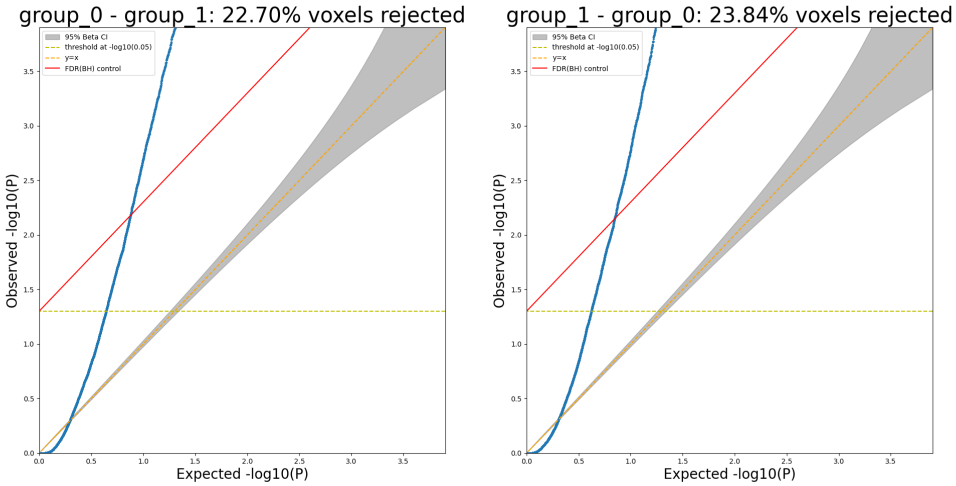
Underlying probability function	Rel.bias	Rel.std	Rel.MSE
Homogeneous	-8.9446×10^{-3}	3.7006×10^{-2}	1.0382×10^{-3}
Gaussian-bumped	-1.3359×10^{-2}	2.4864×10^{-2}	4.9351×10^{-4}

In order to validate the inference of our GLM framework, we conduct statistical tests between two groups with identical underlying probability function, as described by equation 5.7. The contrast matrix C, with dimensions $1 \times 2R$ is defined as $C = [1, \cdots, 1, -1, \cdots, -1]^\top$, and the regression coefficient vector is $\hat{\beta} = [\hat{\beta}_g, \hat{\beta}_{g'}]^\top$, constructed by concatenating group-specific regression coefficients. Since only a single hypothesis test is conducted at each voxel, we further simplify this hypothesis test to the Wald test, as shown in equation 5.8. We illustrate the comparison between the observed distribution of p-values and the expected uniform distribution under the null hypothesis using PP-plots. Derivations from the diagonal line $y = x$ indicate discrepancies, suggesting whether the observed data align with the assumed distribution. Specifically, we provides PP-plots for two scenarios in figure 5.1 and 5.2: spatially homogeneous signals and Gaussian-bump signals with a centrally located bump. We estimate the standard errors of regression coefficients using either the Fisher information or the sandwich estimator. The p-values are computed based on inference on the log-transformed, group-specific Poisson mean estimate vector η_g, which represents the log-linear response in GLM, without applying further approximations (such as delta methods) to obtain the Poisson mean estimate vector μ. The grey shaded region indicates the 95% confidence interval estimated using the beta distribution. The yellow dashed line denotes the 0.05 significance level, with voxels below this line indicating significant p-values and rejection of the null hypothesis. The red line represents the false discovery rate (FDR) control threshold obtained using the Benjamini-Hochberg (BH) procedure. From figures 5.1 and 5.2, we observe that the PP-plots exhibit a significant liberal skew for both the spatially homogeneous signals and the centrally Gaussian-bump signals when the standard errors of the regression

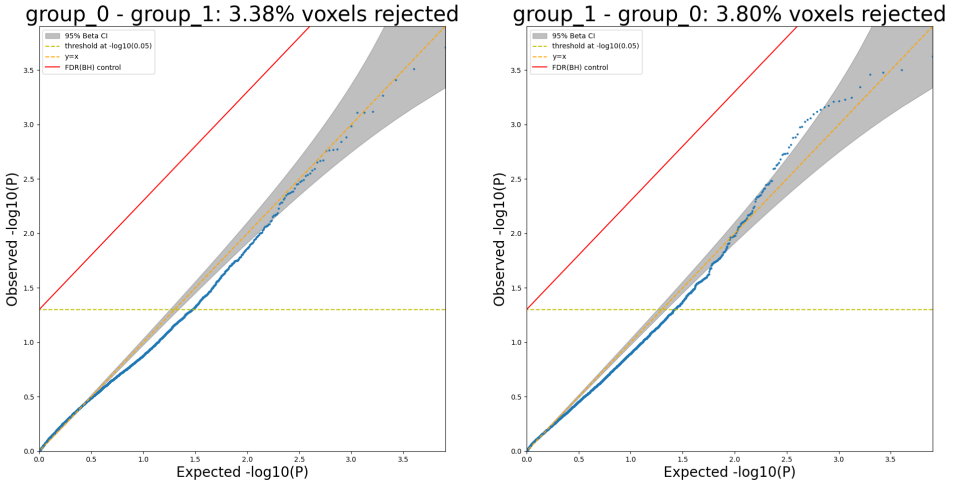
coefficients $\hat{\beta}$ are estimated using the inverse of Fisher information. In contrast, the PP-plots based on standard errors estimated via the sandwich estimator align closely with the diagonal $y = x$ line in both scenarios. This demonstrates that the sandwich estimator provides a more robust and accurate estimation, particularly when the GLM is misspecified, as it accounts for spatial dependence between neighbouring voxels. Additionally, despite the unequal sample sizes between the two groups, the PP-plots for both group_0 minus group_1 and group_1 minus group_0 comparisons exhibit similar behaviour. This further supports the robustness of our GLM inference framework, even in the presence of substantial group imbalance.

5.3.3 Regression and inference for probability estimation using the scalable approximate model factorisation

Given the extremely high dimensionality of variables in regression for real-world brain lesion applications, which scales with the number of subjects and voxels, we propose a scalable approximate model factorisation to improve computational efficiency. Specifically, in the update equation 5.5 for regression coefficients, we approximate the gradient and residue component by applying a Taylor expansion of $\mu^{(k)}$ around its mean at each iteration k. For the preconditioner component, we estimate it using the optimised probability estimation $\hat{\mu}$ obtained from a separable model fit and keep it fixed throughout all iterations. As a results, it is essential to assess the accuracy and validity of these approximations by comparing their regression outcomes with those obtained by the exact regression using the full update equation. Table 5.2 compares the accuracy of the scalable approximate model factorisation with that of baseline method based on the exact regression, evaluated using relative bias, standard deviation (Std), and mean square error (MSE). Minimal differences are observed when combining the exact gradient with the approximate preconditioner, while significant discrepancies are found when both the gradient and the preconditioner are approximated. This suggests that the Taylor approximation around the mean of $\mu^{(k)}$ might lack precision, particularly for subjects whose probability estimates deviate significantly from the mean, as the covariate-dependent probability functions can vary considerably. This issue requires further investigation, for example, by exploring more precise and efficient alternatives for gradient approximation. In this work, however, we adopt the exact gradient in combination with the approximate preconditioner to ensure accurate regression outcomes.

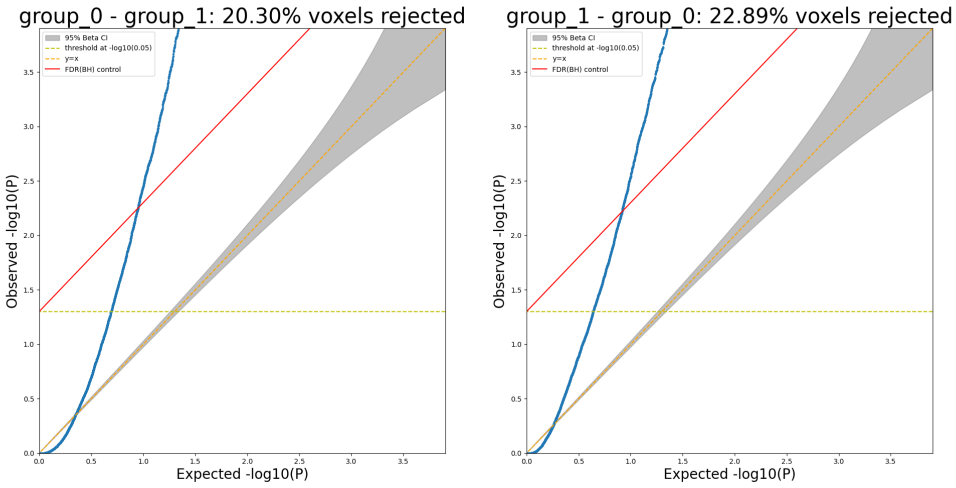


(a) PP-plot with standard error estimated using Fisher Information

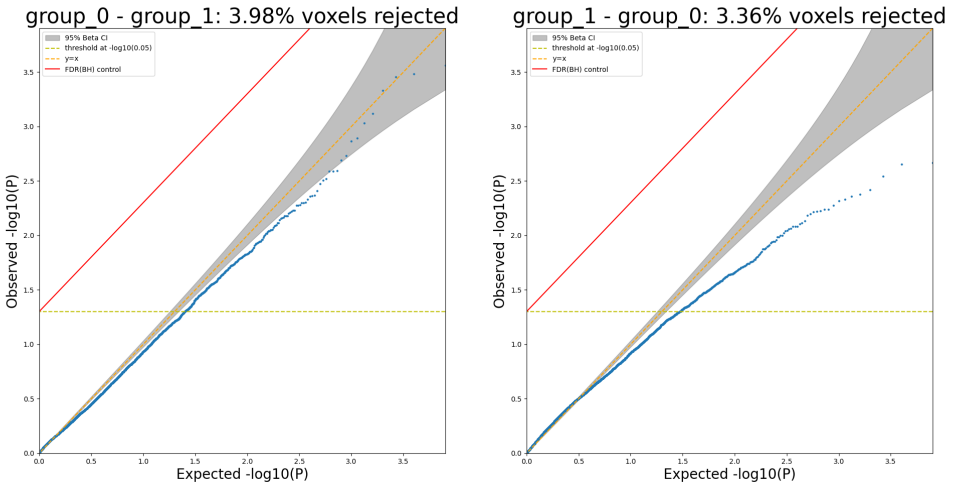


(b) PP-plot with standard error estimated using the sandwich estimator

Figure 5.1: PP-plots (uncorrected p-values on a $-log_{10}$ scale) for spatially homogeneous signals in the hypothesis testing of group comparisons between two groups with different sample sizes (group_0 with 2000 subjects and group_1 with 4000 subjects), using standard errors estimated with Fisher Information or the sandwich estimator.



(a) PP-plot with standard error estimated using Fisher Information



(b) PP-plot with standard error estimated using the sandwich estimator

Figure 5.2: PP-plots (uncorrected p-values on a $-log_{10}$ scale) for Gaussian-bumped signals located at the centre of the 3D space, illustrating the hypothesis testing for group comparisons between two groups with different sample sizes (group_0 with 2000 subjects and group_1 with 4000 subjects), using standard errors estimated with Fisher Information or the sandwich estimator.

Table 5.2: Relative bias, standard deviation (Std), mean squared error (MSE) and computation time for probability function estimation in a 3D space with $8,000$ voxels across 6000 subjects. Results are shown for two methods: (1) a scalable approximation to the preconditioner; (2) a scalable approximation to the preconditioner combined with a Taylor approximation to the gradient. Both methods are compared against a baseline using the exact update equation. The underlying probability function is either spatially homogeneous or includes a Gaussian bump centred in the middle of the space.

Prob	Gradient	Preconditioner	Rel.bias	Rel.std	Rel.mse	computation time
Homo	exact	exact	-	-	-	-
Homo	exact	approx	$1.1277e^{-2}$	$1.8255e^{-2}$	$1.1346e^{-2}$	1109.3968s
Homo	approx	approx	4.3270	1.0445	19.2479	803.7243s
Bump	exact	exact	-	-	-	-
Bump	exact	approx	$1.8537e^{-2}$	$2.1667e^{-2}$	$4.7039e^{-2}$	1132.4808s
Bump	approx	approx	7.6704	8.9551	116.6792	819.6692s

We further evaluate the validity of the scalable approximate model factorisation, using the exact gradient combined with the approximate preconditioner, through PP-plots presented in Appendix C.2.2. Figure C.4 and C.5 exhibits similar patterns to the previous PP-plots, regardless of whether the underlying probability is spatially homogeneous or includes a Gaussian bump at the centre of the space. Specifically, the PP-plots based on standard errors estimated using the sandwich estimator align closely with the diagonal $y = x$ line, supporting the validity of the inference. In contrast, PP-plots based on standard errors estimated via the Fisher Information exhibit a noticeable liberal skew, indicating that this approach may lead to substantially biased standard error estimates.

5.4 UK Biobank Application

5.4.1 Dataset description and pre-processing steps

The UK Biobank is a large-scale biomedical database that has been recruiting predominantly healthy individuals since 2006. Medical imaging has significant potential for early disease prediction but is often hindered by the difficulty and expense of acquiring datasets before symptom onset. This has motivated UK Biobank to initiate an imaging enhancement study in 2014, aiming to acquire high-quality, consistently

obtained imaging data while continuously tracking health outcomes. The imaging extension aims to scan $100,000$ participants, with imaging modalities including MRI of brain, heart and body, low-dose X-ray bone and joint scans, and carotid artery ultrasounds. As of early 2020, imaging data for approximately $45,000$ participants had been collected. This extensive imaging dataset is designed to facilitate early disease detection and enhance our understanding of various health conditions. By integrating imaging data with comprehensive health records, UK Biobank provides a valuable resource for researchers worldwide [Miller et al., 2016]. Our goal is to understand how various risk factors influence the incidence of brain lesions at specific voxel location, their potential clinical significance, and the associated cognitive and neurological consequences. To demonstrate the scalability and accuracy of our efficient approximate model, we apply it to a subset of the UK Biobank data [Miller et al., 2016]. The dataset consists of $N = 13,677$ subjects, each with a brain lesion mask segmented using the Brain Intensity Abnormality Classification Algorithm (BIANCA) algorithm [Griffanti et al., 2016] and subject-specific risk factors, including age, gender, head size and cardiovascular risk (CVR) factors. The CVR score, which ranges from 0 to 8, is derived from summing six categorical variables as detailed in [Veldsman et al., 2020]. These variables represent hypertension, hypercholesterolemia, smoking, diabetes, waist-to-hip ratio and the APOE-ϵ (apolypoprotein-E) status. Specially, smoking and APOE-ϵ are ordinal (0, 1 or 2), while the other four factors are binary (1 indicating presence). While this summed score offers a simple and interpretable representation of the overall CVR burden, it assumes additivity and treats ordinal risk factors as if they were on an interval scale.

To improve computational efficiency, each subject's brain lesion mask, originally derived from 1mm T2 FLAIR images, was registered to the MNI space at a 2mm resolution. This resulted in each lesion mask being represented in a 3D space with dimensions of $91 \times 109 \times 91$ voxels, totalling of $228,483$ within-brain voxels. To focus our analysis on relevant regions, we refined the brain lesion masks through the following steps:

- **Exclusion of low-incidence voxels**: Voxels with lesion occurrences below an empirical threshold of 10^{-3} across all subjects were excluded to focus on areas with significant lesion presence.
- **Morphological operations**: We applied morphological operations to smooth the binary lesion mask. Specifically, we retained only the largest connected

clusters, removing small and isolated clusters below a size threshold of 20 voxels.

- **Cerebrospinal Fluid (CSF) exclusion**: Voxels where the CSF probability exceeded 50% were removed from the brain lesion mask to avoid including non-brain tissues.

These steps resulted in a refined brain mask containing a total of $14,807$ voxels. The central object of interest is to model the influence of age and CVR factors on the probability of lesion incidence, while controlling for potential confounding variables that might also affect lesion probability, including gender and the head size scaling factor. In our dataset, 52.90% subjects are female ($N = 7,235$) and 47.10% are male ($N = 6,442$), with a mean age of 62.92 years ($\pm$7.39 years).

Before model estimation, we performed pre-processing steps on the risk factors. To analyse the spatially varying, subject-specific risk factors, we constructed a covariate matrix $Z(M \times R)$, where each column corresponds to a specific risk factor. Each risk factor was standardised to have a mean of 0 and a standard deviation of 1 to ensure numerical comparability and improve numerical stability. This standardisation prevents differences in measurement scales from disproportionately affecting the results. For the spatial matrix $B(N \times P)$, parametrised by spline, we selected a spline spacing of either 10mm ($P = 889$) or 20mm ($P = 297$) along each spatial dimension. We also removed weakly supported B-spline bases at the edges of the lesion mask, specifically, those for which the sum of all 3D B-spline basis coefficients is below a threshold of 0.1. This thresholding leverages the partition of unity property of spline parametrisation. In practice, we included an intercept column of ones in both the covariate matrix Z and the spatial matrix B to capture the overall mean during regression.

To avoid numerical issues such as overflow during regression, we initialised the spatial parameter vector $\beta(PR \times 1)$ to zero, except for the element corresponding to the intercept column of ones in $Z \otimes B$, which was set to the log-transformed voxelwise mean incidence of lesion count. Additionally, the matrices B and Z were rescaled to further reduce the risk of overflow during optimisation.

5.4.2 Estimation accuracy and computational efficiency

Although we have demonstrated the importance of incorporating quadratic and cubic terms to enhance model flexibility in Appendix C.1.2, Figure 5.3 reveals a

predominantly linear relationship between age and the mean empirical lesion probability across the 100 voxels with the highest lesion incidence rates, and the difference between the linear fit and the cubic polynomial fit is minimal. Thus, including quadratic and cubic age terms in regression analyses of risk factors with potential spatially varying effects might be unnecessary for the UK Biobank dataset. This observation is further supported by the square-root transformed fitted lesion probabilities shown in Figure 5.4 for the linear fit and Figure C.3 in Appendix C.1.3 for the cubic polynomial fit. Minimal differences are observed between the highlighted regions produced by regressions models including only a linear age terms compared to those incorporating quadratic and cubic age terms. Additionally, the consistency between the empirical lesion probabilities derived from $13,677$ UK Biobank subjects and the fitted lesion probabilities produced by our proposed scalable approximate model is demonstrated in Figure 5.4, supporting the validity and accuracy of our approach. We also observe that a spline spacing of 20mm results in overly smoothed estimates, which fail to capture finer details in lesion probability patterns. In contrast, a finer spline spacing of 10mm more effectively represents localised variations. While 10mm might not be optimal, we recognise the trade-off between spatial precision and computation efficiency. Even slightly smaller spacing would significantly increase the number of spline bases, leading to a substantial increase in computational demands. Ideally, a stepwise evaluation is necessary to determine the optimal spline spacing that balances precision and efficiency, this represents a future research direction. For the present work, however, we adopt a 10mm spline spacing and include only the linear term of the age risk factor in all subsequent regression and inference analyses.

Table 5.3: Comparison of computational time for regression using the proposed scalable approximate model on 13,677 UK Biobank subjects, with either 20mm or 10mm spline spacing, and using either a linear or cubic polynomial term for the age covariate.

B-spline spacing	Polynomial order for age covariate	Computational time
20mm	Linear	39 mins 2 seconds
20mm	Cubic	47 mins 34 seconds
10mm	Linear	67 mins 17 seconds
10mm	Cubic	77 mins 52 seconds

According to the computational time comparison for regression using our proposed scalable approximate model on $13,677$ UK Biobank subjects (Table 5.3), we find that employing an update equation with an approximate preconditioner and exact gradient at each iteration demonstrates strong scalability with respect to both the

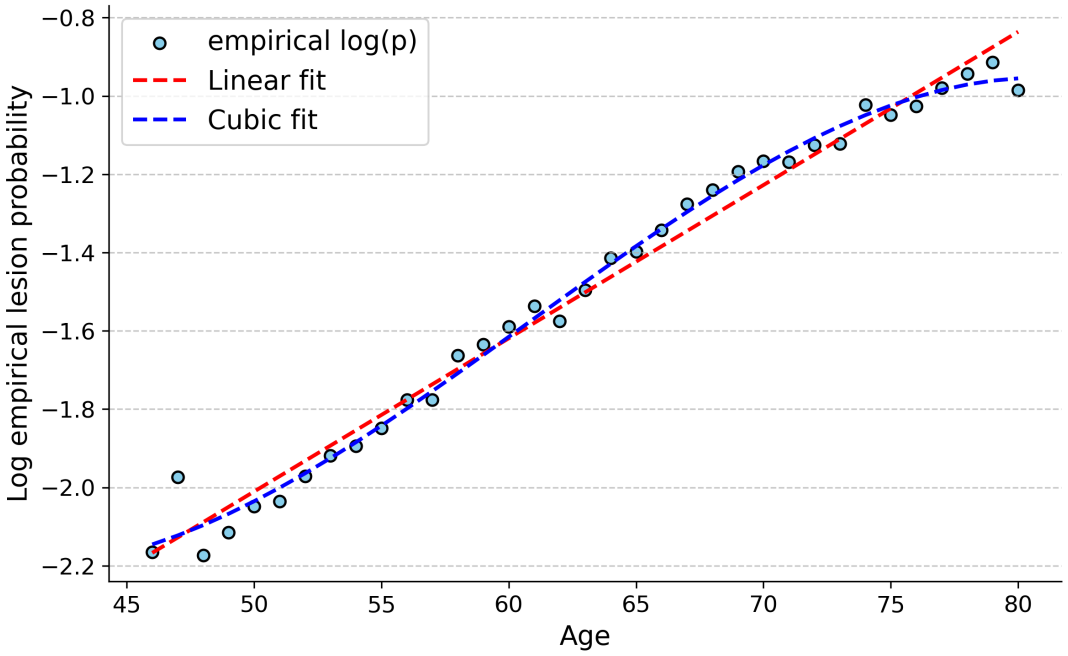
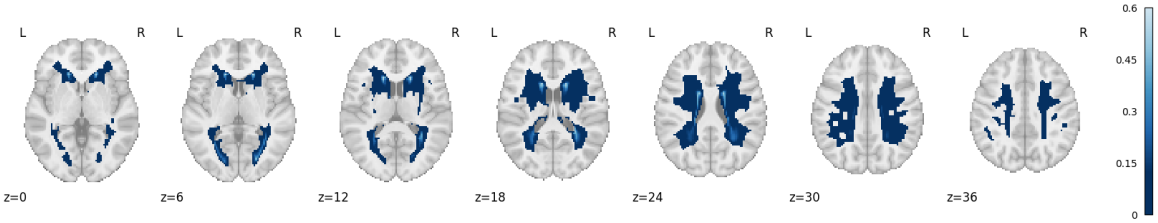
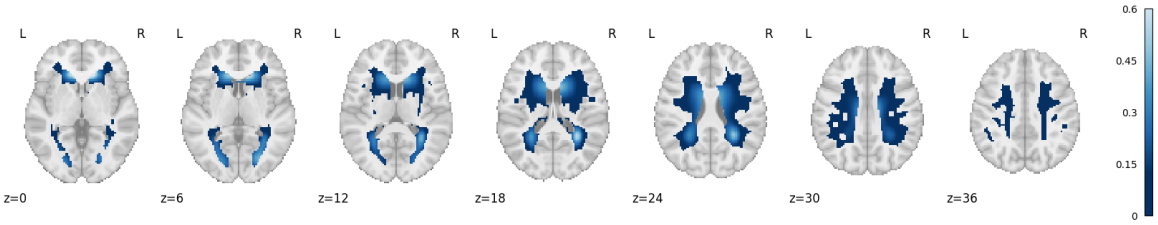


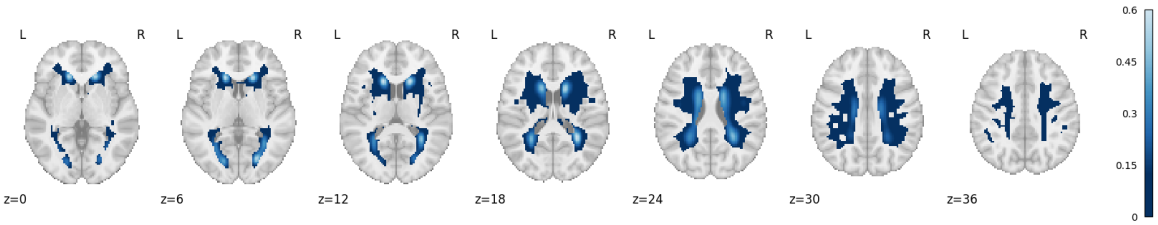
Figure 5.3: Scatter plot showing log-transformed empirical lesion probabilities at the 100 voxels with the highest empirical lesion probabilities among 13,677 UK Biobank subjects, plotted against subjects' ages ($46 - 80$ years). The dashed red line represents the fitted linear relationship between age and log-transformed empirical lesion probability across these voxels, while the dashed blue line represents the fitted cubic polynomial relationship across ages.



(a) Empirical lesion probability $\sqrt{p}$



(b) Fitted lesion probability $\sqrt{\hat{p}}$: linear age covariate, 20mm spline spacing



(c) Fitted lesion probability $\sqrt{\hat{p}}$: linear age covariate, 10mm spline spacing

Figure 5.4: Comparison between square-root transformed empirical lesion probability ($\sqrt{p}$) and model-fitted lesion probabilities ($\sqrt{\hat{p}}$) across 13,677 UK Biobank participants. Lesion probabilities are fitted using spline spacing of either 20 or 10mm, incorporating a linear age covariate. Other covariates, including sex, head size and CVR factors, are modelled linearly in all analyses.

number of subjects and the dimensionality of the brain lesion data. While the computational time does not differ significantly between models using a linear versus a cubic term for the age covariate, a substantial difference is observed between B-spline spacings of 20mm and 10mm, which correspond to 297 and 889 spatial spline bases, respectively. Nevertheless, the proposed approximate model remains significantly more computationally efficient than the full model, making it practical for large-scale data applications while still delivering accurate regression outcomes.

Table 5.4: Estimated lesion probability, relative risk (RR), and risk difference (RD) for two risk factors (age and CVR factor) at peak coordinates within 6 clusters, obtained by thresholding the estimated probability map at 0.1. RR and RD were calculated based on comparison between mean age 62.72 ± 5 years, and CVR factor 1.65 ± 0.5. PV WM represents periventricular white matter.

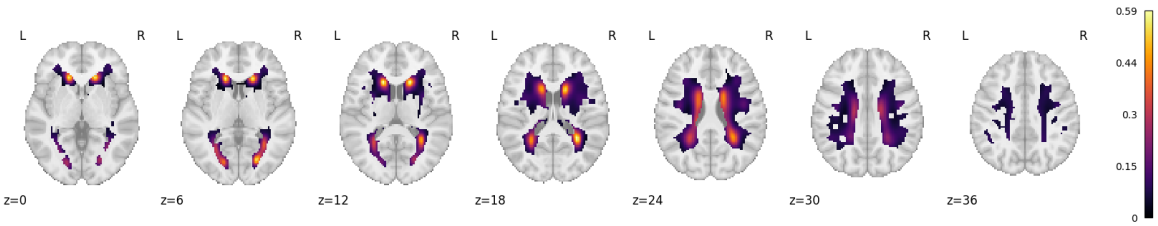
Location name	MNI152 coordinate	$\hat{p}$	$\mathrm{RR}_{\mathrm{age}}$	$\mathrm{RD}_{\mathrm{age}}$	$\mathrm{RR}_{\mathrm{CVR}}$	$\mathrm{RD}_{\mathrm{CVR}}$
L anterior PV WM	$(-16, 22, 12)$	0.3190	1.5949	0.1898	1.0761	0.0243
R anterior PV WM	$(18, 24, 10)$	0.3072	1.6158	0.1892	1.0735	0.0226
R posterior PV WM	$(30, -48, 18)$	0.2987	1.8602	0.2570	1.0100	0.0297
R posterior PV WM	$(22, -76, 6)$	0.2882	1.0864	0.0249	1.0072	0.0021
L posterior PV WM	$(-28, -50, 18)$	0.1481	2.4588	0.2161	1.1496	0.0222
L posterior PV WM	$(-20, -78, 4)$	0.1755	1.1873	0.0329	1.0148	0.0026

Another appealing feature of the regression framework based on the proposed approximate model is its interpretability. In particular, a relative risk (RR) map visualises lesion probability relative to a baseline setting or reference level, where all risk factors are fixed at their population mean. From this baseline, the RR map then provides a spatial representation for understanding how each individual risk factor independently modifies lesion incidence probability across different brain regions, while all other factors remain at their mean values. The relative risk (RR) at a given voxel j is typically defined as,

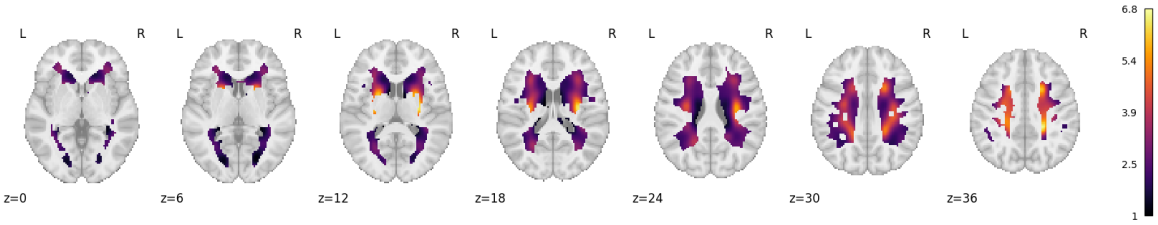
$$\mathrm{RR}_j = \frac{\text{Fitted lesion probability at voxel } j \text{ (with covariate)}}{\text{Baseline lesion probability at voxel } j \text{ (reference)}}$$

In contrast, the absolute risk difference (RD) map visualises the absolute difference in lesion probability relative to the baseline setting or reference level, where all risk factors are fixed at their population mean. The absolute risk difference (RD) at a given voxel j is typically defined as

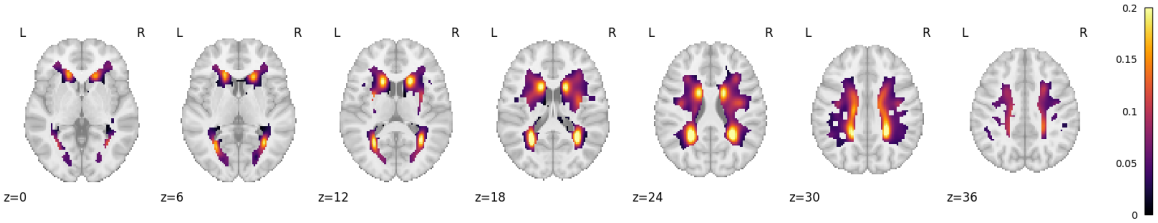
$$\begin{aligned}\mathrm{RD}_j = &\text{ Fitted lesion probability at voxel } j \text{ (with covariate)} \\ &- \text{Baseline lesion probability at voxel } j \text{ (reference)}\end{aligned}$$



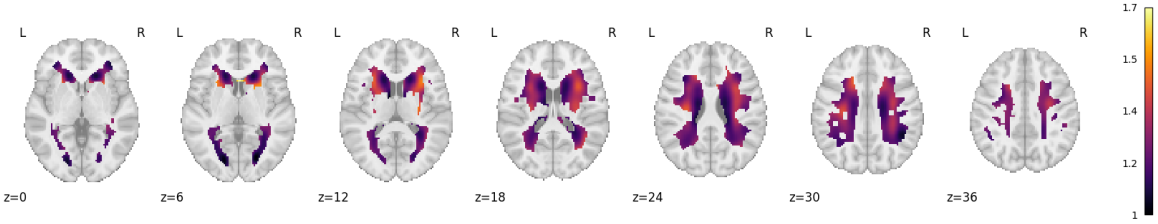
(a) Square-root transformed fitted lesion probability ($\sqrt{\bar{p}}$) at the reference level, evaluated at the mean age of 62.92 years and mean CVR factor of 1.65



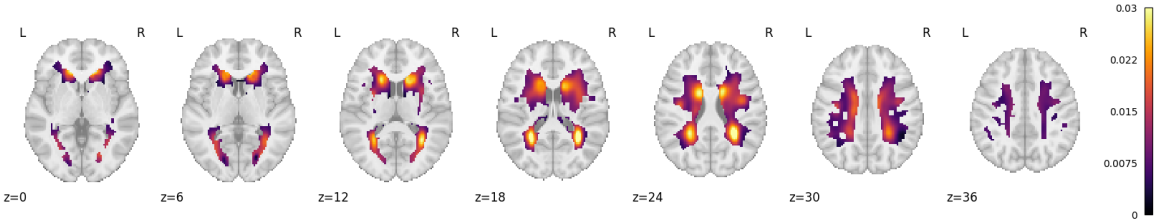
(b) Relative risk map $\exp(10 \cdot \tilde{\beta})$ for the age risk factor, representing the effect of a 10-year increase in age



(c) Absolute risk difference map ($\sqrt{\hat{p}} - \sqrt{\bar{p}}$) for the age risk factor, representing the change in risk between ages 67.72 years (mean age+5 years) and 57.72 years (mean age−5 years).



(d) Relative risk map $\exp(\tilde{\beta})$ for the CVR risk factor, representing the effect of a one-unit increase in CVR



(e) Absolute risk difference map ($\sqrt{\hat{p}} - \sqrt{\bar{p}}$) for the CVR risk factor, representing the change in risk between a CVR factor of 2.15 (mean CVR+0.5) and 1.15 (mean CVR−0.5)

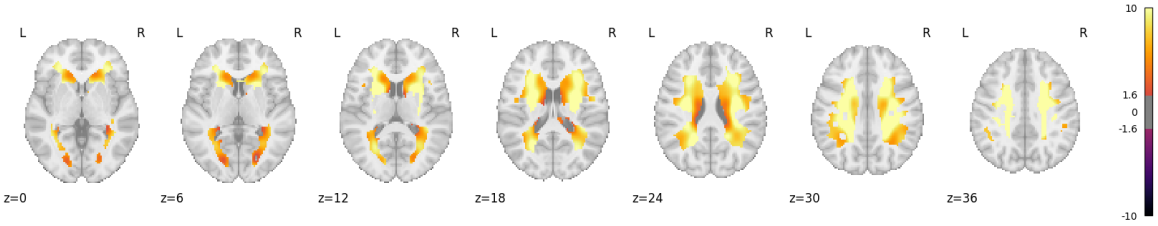
Figure 5.5: Relative risks map for two risk factors, age and CVR, estimated using the regression with the proposed scalable approximate model on 13,677 UK Biobank subjects.

Figure 5.5 presents the relative risk (RR) maps and absolute risk difference (RD) maps for two specific risk factors: age and CVR factor. In these analyses, all other risk factors are fixed at their population mean. The figure also includes the square-root transformed fitted lesion probability at the reference level. Table 5.4 summarises the estimated lesion probability, RR and RD for age and CVR factor, at the peak coordinates within 6 clusters, obtained by thresholding the estimated probability map at 0.1. RR and RD for each risk factor were calculated by comparing ages of 67.72 years (mean age +5 years) versus 57.72 years (mean age −5 years), and CVR factor of 2.15 (mean CVR +0.5 unit) compared to 1.15 (mean CVR −0.5 unit). A 10-year increase in age around the mean is associated with a substantially stronger risk difference compared to a one-unit increase in the CVR factor around its mean score. In both cases, all other risk factors remain consistent at their population mean. Here, a one-unit increase in the CVR factor represent either developing a new cardiovascular disease (e.g., hypertension, hypercholesterolemia or diabetes) or one-unit increase in the level of smoking or APOE-ϵ status. Specifically, the average relative risk increases by approximately 184.61% per 10-year age increase, whereas a one-unit increase in the CVR factor results in an average increase of 22.83%. Despite the difference in magnitude, both risk factors exhibit consistent patterns of lesion increase around the lateral ventricles. Additionally, we observe an unusually strong association between a 10−year increase in age and relative risk in the white matter at $z = 36$, as highlighted in Figure 5.5b. This may be attributed to the low lesion incidence in this regions at the reference level, as illustrated in Figure 5.5a, rather than indicating an exceptionally high sensitivity to age-rated changes.

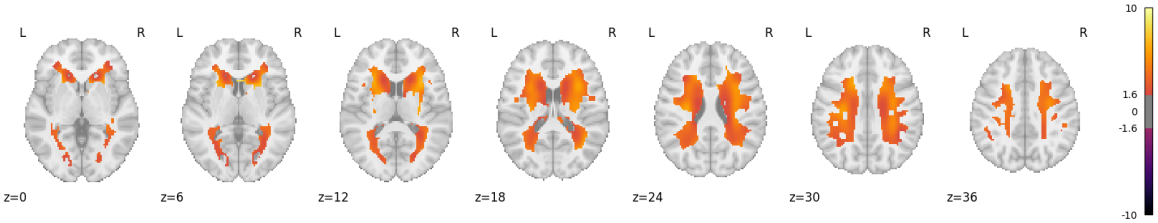
5.4.3 Statistical inference

To investigate the spatially varying effects of risk factors (such as age and the CVR factor) on lesion incidence probability, we perform voxel-wise statistical inference using scalable approximate model factorisation. As demonstrated in our simulation studies (Appendix C.2.2), standard error estimates obtained by inverting the Fisher information matrix tend to be underestimated, particularly under model misspecification. In contrast, the sandwich estimator provides more robust and accurate standard error estimates. Therefore, we adopt the sandwich estimator for voxel-level inference in the following analysis.

As demonstrated in the covariance estimation of β using the sandwich estimator in Equation 5.9, the bread term $(X^\top WX)^{-1}$ corresponds to the inverse of the observed Hessian of the log-likelihood. In practice, the diagonal matrix $W = diag(\hat{\mu}_{ij})$, of size $MN \times MN$, for $i = 1, \cdots, M$ and $j = 1, \cdots, N$, is often numerically unstable due to many elements $\hat{\mu}_{ij}$ being close to zero. To address this instability, we employ singular value decomposition (SVD) to decompose the symmetric matrix $X^\top WX$ as $U\Sigma V^\top$, instead of directly computing its inverse or using eigenvalue decomposition. We then apply a threshold (default: 10^{-8}) to the singular values, replacing any value below the threshold with the threshold itself. The inverse is subsequently computed as $U'\Sigma^{-1}U'^\top$, where $U' = \frac{1}{2}(U + V)$, to improve the numerical stability and precision of the SVD-based inversion.



(a) Z-score map showing the effect of the age risk factor



(b) Z-score map showing the effect of the CVR risk factor

Figure 5.6: Significance maps (uncorrected Z-scores) for two risk factors, age and CVR, fitted via regression using the proposed scalable approximate model on 13,677 UK Biobank subjects. The significance maps assess voxel-wise significance of each risk factor's effect at a 5% significance level.

Figure 5.6 and Figure C.6 in Appendix C.3.1 presents the uncorrected and FDR-corrected Z-scores (at the 5% level) for the risk factors (age and CVR factor) with spatially varying effects at the voxel level. We observed a strong positive association between brain lesion incidence and the age risk factor across the entire white matter brain lesion mask. However, the p-values were less significant in regions around the caudate nucleus and the cerebral cortex. As these structures are primarily composed of gray matter, the reduced statistical significance is therefore neuroanatomically consistent with the absence of WMHs in these regions. These apparently "significant"

gray-matter voxels are likely attributable to methodological artefacts rather than true pathology. Principal sources of such false positives include:

- Segmentation error of BIANCA: Due to weak intensity contrasts between deep gray matter and adjacent periventricular white matter on T2 FLAIR image, BIANCA may apply overly conservative criteria, inadvertently affecting the accuracy of WMH segmentation.
- Resolution/partial volume effects: Lesion masks segmented at 1mm resolution in T2 FLAIR space are subsequently resampled to 2mm MNI template for analysis. This down-sampling and interpolation process can lead to the partial volume effect, causing lesions to appear to cross tissue boundaries, resulting in gray-matter voxels inheriting partial WMH labels.
- Non-linear registration inaccuracies: When brains with significant atrophy or pronounced ventricular enlargement are warped to a standardised template, structures like the caudate nucleus and cortical mantle may become misaligned. This can lead to white-matter signals being displaced into adjacent gray-matter regions.
- Spatial smoothing of statistical maps: The smoothing process applied to statistical images can artificially extend lesion signals across tissue boundaries, resulting in spurious significance within adjacent gray-matter regions.

A similar patterns was observed in the Z-score maps for the CVR factor, although the overall association was weaker compared to age. Furthermore, the p-values for the effect of CVR factor were not significant at the 5% level in some regions around the caudate nucleus, likely due to the same methodological artifacts described above.

5.5 Discussion

In this work, we proposed a novel voxel-wise generalised linear model (GLM) with an integrated spatial model to jointly capture spatial dependence across the brain. Model flexibility is determined by spline spacing and the polynomial orders of the risk factors, which together control the smoothness of the spatial effect. An efficient approximate model factorisation ensures strong scalability, making the approach

practical for large-scale datasets while significantly reducing computational time. We validated the accuracy of this approximation by conducting a systematic comparison with the results obtained from the full model. We also derived relative risk maps to characterise the effect of each risk factor at the voxel level. In addition, we developed a robust and accurate inference method that provides standard error estimates using the sandwich estimator. This allows for voxel-wise hypothesis testing to assess the statistical significance of risk factor effects and can be extended to more flexible tests, such as group comparisons or comparisons of different risk factor effects at the voxel level. This approach is significantly more efficient than the previously proposed Bayesian spatial regression model. For comparison, using a dataset of $13,677$ UK Biobank subjects, the Bayesian spatial spike-and-slab regression model (BLESS, [Menacher et al., 2024]) required approximately varying between 13 hours 57 minutes to 21 hours 42 minutes for sample sizes ranging from 500 to $5,000$, although the computational hardware was not specified in the paper. In contrast, our proposed efficient approximate model factorisation, with 10mm B-spline spacing and only linear term of age risk factor, involved a total computation time of approximately 4671.55 seconds (less than 1 hour 18 minutes) on an Intel Xeon Gold 6340R CPU. Moreover, as our method is grounded in a generalised linear model, it offers strong interpretability and is more accessible for users without access to advanced computation resources such as GPUs, especially when compared to inference methods relying on spatial posterior functions.

There are a few limitations in our work. One of the main drawbacks is that the model does not guarantee that the predicted lesion probability remain strictly within the valid range of $[0, 1]$. We employed a generalised linear model (GLM) with a log link function and a Poisson distribution to approximate the low-rate Bernoulli distribution at the voxel level. The estimated Poisson mean λ_{ij} for subject i at voxel j is then used to compute the lesion probability, which is truncated at 1, thereby disregarding the possibility of predicted values exceeding this upper bound. As a result, our approximation effectively behaves as a truncated Poisson model on the interval $[0, 1]$. Although this happens very rarely in the context of low-rate brain lesion probabilities, this discrepancy might still lead to model misspecification and inference errors, particularly in regions with relatively high lesion incidence. Another limitation is the numerical instability in estimating the covariance matrix $X^\top WX$ of the regression coefficient β, whether computed via the Fisher information or the sandwich estimator. This issue is particularly pronounced in brain lesion settings,

where the estimated lesion probabilities are close to zero for most within-brain voxels. Currently, we address this by using truncated SVD, retaining only the dominant components with singular values above a pre-specified threshold (10^{-8} by default). In future work, we plan to investigate practical guideline for selecting this threshold, as well as explore alternative approaches, such as incorporating ridge regularisation.

Another potential direction for future development is to explore further acceleration of the efficient approximate model factorisation. Currently, the preconditioner is fixed based on the mean intensity estimation from a separable model fit and remains unchanged throughout all iterations. Meanwhile, the exact gradient, as the most computationally intensive component, must be computed at each step. Using an approximate gradient based on a Taylor expansion around the mean intensity either leads to divergence or convergence to values that deviate significantly from the regression estimates obtained with the full model. We believe this issue is driven by subjects whose estimated lesion probabilities deviate substantially from the mean, likely due to spatially varying covariate effects. Future research will investigate alternative gradient approximation methods that outperform the Taylor expansion in accuracy while remaining computationally efficient to be evaluated at each iteration.

Chapter 6

Conclusions and Future Direction

6.1 Conclusions

Throughout this thesis, we explored statistical modelling and inference frameworks that combine efficiency with strong scalability for neuroimaging applications.

In Chapter 1, we discussed the importance of accurately modelling spatial dependence across brain regions, as well as the trade-off between model complexity and computational feasibility. To contextualise our work, in Chapter 2, we provided a brief overview of neuroimaging fundamentals, statistical modelling in neuroimaging, inference methods specific to neuroimaging, meta-analysis approaches and white matter hyperintensities.

In Chapter 3, we introduced a meta-regression framework that incorporates a spatial model for CBMA data. The main innovation of this work is the use of spline parametrisation to model the smooth spatial distribution of activation foci, combined with a GLM that includes different variants of voxelwise or publication-wise statistical distribution as its stochastic component. We demonstrated that this approach offers a computationally efficient alternative to existing Bayesian spatial regression models, while delivering comparable results. Moreover, this work highlights the presence of over-dispersion in CBMA data, highlighting the necessity of considering the negative binomial (NB) model within the meta-regression framework to adequately account for this characteristic.

To extend this framework to multi-group settings and enable group comparisons,

we introduced multi-group coordinate-based meta-regression (CBMR) in Chapter 4, implemented as a module within the open-source Python package NiMARE. By leveraging a roughness penalty to regularise the smoothness of the spline basis functions in the spatial model, and applying a parametric bootstrap method for datasets with an insufficient number of publications or foci, we ensure that inference results remain valid and reliable. Our theoretical analysis supports the use of model factorisation to simplify the likelihood function during the iterative optimisation of maximum likelihood estimates (MLEs). Practical experiments further validated the accuracy and robustness of our approach for CBMA group comparisons. This work represents a significant step toward building an efficient, scalable, and flexible meta-regression and inference framework for multi-group CBMA data.

Similar to CBMA, lesion mapping presents common methodological challenges, such as accurately modelling spatial dependence across the brain while ensuring computational efficiency and scalability. In Chapter 5, we demonstrated the broader applicability of the spatial modelling framework by applying similar statistical principles to voxel-wise lesion probability mapping using structural MRI data. An efficient approximate model factorisation ensures strong scalability, making the approach practical for large-scale datasets and significantly reducing computational time. Additionally, we developed a robust and accurate inference method that provides standard error estimates via the sandwich estimator. This work bridges the gap between computationally intensive Bayesian spatial regression models and other existing methods that lack scalability for large-scale datasets.

6.2 Future Directions

Having highlighted our research contributions, we now delve into potential future directions and open questions that merit further investigation.

While significant advancements have been made in CBMA methodologies [Eickhoff et al., 2016, Müller et al., 2018], several important avenues remain unexplored. One promising direction is the integration of richer metadata, such as task paradigms, demographic variables (e.g., age, sex), clinical characteristics and imaging acquisition parameters [Maumet et al., 2016]. The inclusion of such auxiliary information would facilitate more nuanced and context-sensitive meta-analyses, allowing researchers to disentangle task-specific, population-specific or disease-specific activation patterns and

to systematically investigate potential sources of heterogeneity across publications. For example, leveraging structured metadata could facilitate stratified analyses or meta-regressions that explore how cognitive tasks interact with demographic or clinical factors, potentially revealing different activation profiles that remain obscured within traditional CBMA framework, which often neglect these contextual variables [Fox et al., 2005].

Moreover, the adoption of hierarchical or multi-level modelling frameworks represents a critical step forward in addressing the complex variability inherent in neuroimaging publications [Wager et al., 2007]. By explicitly modelling variance components at multiple levels, such as within-publication variability, between-publication heterogeneity and higher-level group effects, these approaches can improve statistical power, reduce bias and produce more generalisable inferences. Hierarchical models also provide a principled mechanism for incorporating publication-level covariates, accounting for clustered data structures, and manage unbalanced datasets with varying sample sizes or numbers of foci. Future developments in this area could involve Bayesian hierarchical spatial models [Montagna et al., 2018] or mixed-effects GLMs tailored for CBMA, which would better capture spatial dependencies and cross-publication variability in large-scale neuroimaging meta-analyses. Furthermore, advances in variational inference and Markov Chain Monte Carlo (MCMC) techniques could facilitate scalable estimation procedures, making these complex models computationally feasible for large-scale neuroimaging meta-analyses.

Additionally, the integration of CBMA with IBMA approaches offers a promising avenue to enhance statistical power and spatial precision by combining peak coordinate data with voxel-wise statistical information [Salimi-Khorshidi et al., 2009, Nichols et al., 2017]. To further advance such integrative frameworks, some researchers have proposed Markov melding as a fully Bayesian framework for joining probabilistic sub-models. This method allows evidence from different sources to be specified within each sub-model, which are then coherently joined while preserving both information and associated uncertainty [Goudie et al., 2019]. Such hybrid strategies, leveraging both data types and advanced probabilistic modelling, can effectively overcome the limitations inherent in relying solely on sparse coordinate data, thereby enabling more comprehensive, flexible and precise meta-analytic inferences.

In lesion mapping, a key avenue for development is the integration of multimodal neuroimaging data into existing analytical frameworks. Traditional approaches have

predominantly relied on structural MRI to identify lesion locations and assess their associations with behavioural or clinical outcomes [Bates et al., 2003, Rorden et al., 2009]. However, this emphasis on structural damage alone, but overlooks the broader impact of lesions on brain connectivity and function, particularly through mechanisms such as diaschisis and network-level disruptions [Price, 2001, Carrera and Tononi, 2014]. Incorporating additional modalities, such as diffusion-weighted imaging (DWI) to assess white matter integrity [Thiebaut de Schotten et al., 2015], functional MRI (fMRI) to capture alterations in functional connectivity [Reber et al., 2021], and perfusion imaging to evaluate cerebral blood flow dynamics [Hillis, 2016], can provide a more comprehensive understanding of lesion-induced disturbances. This multimodal integration allow researchers to move beyond lesion location alone, providing deeper insights into how structural disconnections and functional reorganisation contribute to clinical deficits and influence recovery trajectories [Foulon et al., 2018]

Another critical direction involves the incorporation of spatial priors and anatomical constraints into lesion mapping analyses. Leveraging prior knowledge regarding vascular territories [Pexman et al., 2001, Tatu et al., 2012], functional parcellations [Glasser et al., 2016], and anatomical connectivity [Thiebaut de Schotten et al., 2014, Foulon et al., 2018] can enhance both statistical power and interpretability. Spatial Bayesian models, including those based on Gaussian Markov random fields [Bates et al., 2003, Chen et al., 2008] or spatial Poisson processes [Kang et al., 2011], represent promising frameworks for embedding such biologically informed constraints.

Additionally, a growing area of interest is the formalisation of causal inference within lesion mapping [Yokoyama et al., 2014, Sperber, 2020]. Whereas traditional analyses predominantly identify correlational associations between lesion locations and behavioural deficits [Rorden and Karnath, 2004], these associations may be confounded by vascular architecture or lesion volume [Sperber and Karnath, 2017]. Moving towards causal interpretations necessitates the adoption of robust frameworks such as counterfactual modelling [Pearl, 2009, Hernán and Robins, 2020], instrumental variable techniques [Angrist et al., 1996, Baiocchi et al., 2014], and Bayesian causal networks [Ramsey et al., 2010, Mumford and Ramsey, 2014]. These methodologies provide a more principled foundation for inferring causality, thereby improving both the translational relevance and clinical utility of lesion mapping findings [Kraemer et al., 1997].

In summary, the future of spatial statistical frameworks for CBMA and lesion mapping

applications will be driven by methodological advancements, including meta-data and multi-model integration, advanced modelling techniques, hierarchical or multi-level framework and causal inference approaches. Effectively addressing these challenges will be crucial for advancing our understanding of brain-behaviour relationships and enhancing the clinical utility of these analytical methods.

Appendix A

Appendix for Neuroimaging Meta Regression for Coordinate Based Meta Analysis Data with a Spatial Model

A.1 Detailed derivation of stochastic models

A.1.1 Poisson model

We assert that the sum of two independent Poisson random variables is also Poisson. Let $X \sim \text{Poi}(\lambda_1)$ and $Y \sim \text{Poi}(\lambda_2)$ be two independent random variables, and $Z = X + Y$, then,

$$\begin{aligned} P(Z = n) = P(X + Y = N) &= \sum_{k=-\infty}^{\infty} P(X = k)P(Y = n - K) \\ &= \sum_{k=0}^{n} P(X = k)P(Y = n - k) \\ &= \sum_{k=0}^{n} e^{-\lambda_1} \frac{\lambda_1^k}{k!} e^{-\lambda_2} \frac{\lambda_2^{n-k}}{(n-k)!} \\ &= e^{-(\lambda_1+\lambda_2)} \sum_{k=0}^{n} \frac{\lambda_1^k \lambda_2^{n-k}}{k!(n-k)!} \\ &= \frac{e^{-(\lambda_1+\lambda_2)}}{n!} \sum_{k=0}^{n} \frac{n!}{k!(n-k)!} \lambda_1^k \lambda_2^{n-k} \\ &= \frac{e^{-(\lambda_1+\lambda_2)}}{n!} (\lambda_1 + \lambda_2)^n \end{aligned} \tag{A.1}$$

Therefore, $Z = X + Y \sim \text{Poi}(\lambda_1 + \lambda_2)$ is also a Poisson variable. The conclusion can be extended further: the sum of multiple Poisson random variables $(\text{Poi}(\lambda_i), i = 1, \cdots, n)$ also follows a Poisson distribution, with the parameter $\lambda = \sum_{i=1}^{n} \lambda_i$.

Hence, under the assumption of independence of counts across publications, we believe that the likelihood function is exactly the same if we model the voxelwise total foci count over publications (with length-N) instead of voxelwise foci count for each publication (with expanded length-(NM)). This reformulation can simplify the computation of the log-likelihood function and reduce the dimensionality of statistics (never larger than M or N in dimension).

A.1.2 Negative Binomial (Poisson-Gamma) Model

In this section, we describe the formulation of the NB distribution in more detail. Based on the assumption of NB (Poisson-Gamma) model, there is a single parameter α, which indicates variance in excess of the Poisson model. For voxel j in publication i, the voxelwise mean of intensity λ_{ij} follows a Gamma distribution with mean $\mathbb{E}(\lambda_{ij}) = \mu_{ij}$ and variance $\text{Var}(\lambda_{ij}) = \alpha\mu_{ij}^2$

$$\lambda_{ij} \sim \text{Gamma}(\alpha^{-1}, \frac{\alpha^{-1}}{\mu_{ij}}) \Rightarrow \mathbb{E}(\lambda_{ij}) = \mu_{ij}, \text{Var}(\lambda_{ij}) = \alpha\mu_{ij}^2$$

And $Y_{ij}|\lambda_{ij}$ follows a Poisson distribution with conditional mean $\mathbb{E}(Y_{ij}|\lambda_{ij}) = \lambda_{ij}$

$$Y_{ij}|\lambda_{ij} = \text{Poisson}(\lambda_{ij}) \Rightarrow P(Y_{ij}|\lambda_{ij} = k) = \frac{\lambda_{ij}^k e^{-\lambda_{ij}}}{k!}$$

which gives rise to marginal probability of Y_{ij}

$$\begin{aligned}
P(Y_{ij} = y_{ij}) &= \int_{\lambda_{ij}} P(Y_{ij}|\lambda_{ij})P(\lambda_{ij})d\lambda_{ij} = \int_{\lambda_{ij}=0}^{\infty} \frac{\lambda_{ij}^{y_{ij}} e^{-\lambda_{ij}}}{y_{ij}!} \frac{(\frac{1}{\alpha\mu_{ij}})^{\frac{1}{\alpha}}}{\Gamma(\frac{1}{\alpha})} \lambda_{ij}^{\frac{1}{\alpha}-1} e^{-\frac{\lambda_{ij}}{\alpha\mu_{ij}}} d\lambda_{ij} \\
&= \frac{\frac{1}{\alpha\mu_{ij}}^{\frac{1}{\alpha}}}{y_{ij}!\Gamma(\frac{1}{\alpha})} \int_{\lambda_{ij}=0}^{\infty} \lambda_{ij}^{y_{ij}} e^{-\lambda_{ij}} \lambda_{ij}^{\frac{1}{\alpha}-1} e^{-\frac{\lambda_{ij}}{\alpha\mu_{ij}}} d\lambda_{ij} = \frac{\frac{1}{\alpha\mu_{ij}}^{\frac{1}{\alpha}}}{y_{ij}!\Gamma(\frac{1}{\alpha})} \frac{\Gamma(y_{ij} + \frac{1}{\alpha})}{(\frac{1}{\alpha\mu_{ij}} + 1)^{y_{ij}+\frac{1}{\alpha}}} \\
&= \frac{\Gamma(y_{ij} + \frac{1}{\alpha})}{\Gamma(y_{ij} + 1)\Gamma(\frac{1}{\alpha})} \left(\frac{\frac{1}{\alpha\mu_{ij}}}{\frac{1}{\alpha\mu_{ij}} + 1}\right)^{\frac{1}{\alpha}} \left(\frac{1}{\frac{1}{\alpha\mu_{ij}} + 1}\right)^{y_{ij}} \\
&= \frac{\Gamma(y_{ij} + \alpha^{-1})}{\Gamma(y_{ij} + 1)\Gamma(\alpha^{-1})} \left(\frac{1}{1 + \alpha\mu_{ij}}\right)^{\alpha^{-1}} \left(\frac{\alpha\mu_{ij}}{1 + \alpha\mu_{ij}}\right)^{y_{ij}}
\end{aligned}$$

which satisfies the mathematical form of probability density function of the NB model, $Y_{ij} \sim NB(\alpha^{-1}, \frac{\mu_{ij}}{\alpha^{-1}+\mu_{ij}})$, with mean $\mathbb{E}[Y_{ij}] = \mu_{ij}$ and variance $\mathbb{V}(Y_{ij}) = \mu_{ij} + \alpha\mu_{ij}^2$.

A.1.3 Moment Matching Approach

For the purpose of approximating the sum of multiple independent NB random variables, we approximate a sum of NB variates with a NB distribution by moment matching (mean and variance). Suppose the voxelwise count in each individual publication is $Y_{ij} \sim \text{NB}(\alpha^{-1}, \frac{\mu_{ij}}{\mu_{ij}+\alpha^{-1}})$, and α is a global dispersion parameter. Using the independence of publications at voxel j,

$$\begin{cases} \mathbb{E}(Y_{\bullet,j}) = \sum_{i=1}^{M} \mathbb{E}(Y_{ij}) = \sum_{i=1}^{M} \mu_{ij} \\ \mathbb{V}(\mathbb{E}(Y_{\bullet,j})) = \sum_{i=1}^{M} \text{Var}(Y_{ij}) = \sum_{i=1}^{M} \mu_{ij} + \sum_{i=1}^{M} \alpha\mu_{ij}^2 \end{cases}$$

To ensure that the proposed NB distribution ($Y_{\bullet,j} \sim NB(r', p')$) matches the mixture of NB distributions, with regard to both mean and variance, we need

$$\begin{cases} \mathbb{E}(Y_{\bullet,j}) = \sum_{i=1}^{M} \mu_{ij} \\ \text{Var}(Y_{\bullet,j}) = \sum_{i=1}^{M} \mu_{ij} + \sum_{i=1}^{M} \alpha\mu_{ij}^2 \end{cases} \Rightarrow \begin{cases} p' = \frac{\sum_{i=1}^{M} \mu_{ij}^2}{\alpha^{-1} \sum_{i=1}^{M} \mu_{ij} + \sum_{i=1}^{M} \mu_{ij}^2} \\ r' = \frac{(\sum_{i=1}^{M} \mu_{ij})^2}{\alpha \sum_{i=1}^{M} \mu_{ij}^2} \end{cases}$$

Therefore, the approximated NB distribution of the sum of foci count at voxel j is, $Y_{\bullet,j} \sim \text{NB}\left(\frac{(\sum_{i=1}^{M} \mu_{ij})^2}{\alpha \sum_{i=1}^{M} \mu_{ij}^2}, \frac{\sum_{i=1}^{M} \mu_{ij}^2}{v \sum_{i=1}^{M} \mu_{ij} + \sum_{i=1}^{M} \mu_{ij}^2}\right)$, with excess variance in the NB approximation α',

$$\frac{1}{\alpha'} = \frac{(\sum_{i=1}^{M} \mu_{ij})^2}{\alpha \sum_{i=1}^{M} \mu_{ij}^2} \Rightarrow \alpha' = \frac{\sum_{i=1}^{M} \mu_{ij}^2}{(\sum_{i=1}^{M} \mu_{ij})^2} \alpha$$

.

A.1.4 Evaluating the effectiveness of moment matching approach

To evaluate the effectiveness of the moment matching approach and to substantiate its application to CBMR with the NB model, we now incorporate simulation experiments to provide empirical evidence supporting this application.

We conducted a univariate simulation with $N = 10000$ publications, where the true rate is homogeneous but the dispersion parameter is shared. Specifically, each publication i has mean $\mu_i = \mu_0 \cdot \frac{10i}{N}$, $\mu_0 = 10^{-3}$, and variance $\mu_i + \alpha\mu_i^2$, $\alpha = 0.5$, generating Y_i, $i = 1, \cdots, N$,

$$Y_i \sim NB(\mu_i, \alpha), \quad \mu_i = \mu_0 \cdot \frac{10i}{N} \tag{A.2}$$

Our approximate approach for modelling this data is to assume

$$\sum_i Y_i \sim NB(\mu, \alpha') \tag{A.3}$$

where $\mu = \sum_i \mu_i = \mu_0 \cdot \frac{10(N+1)}{2}$, and α' is given by $(\sum_i \mu_i^2)/(\sum_i \mu_i)^2\alpha$. Like in our full CBMR model, here we use Maximum Likelihood (ML) to estimate μ and α, and we construct large-sample standard errors for μ using Fisher's information. Using 1000 Monte Carlo (MC) realisations, we compute the MC standard deviation of $\hat{mu}$ and compare it to the large sample standard errors (inverse Fisher's information). Our analysis revealed that, with 1000 Monte Carlo realisations, the MC standard deviation of $\hat{\mu}$ is 7.1000. In comparison, the standard error, estimated using the Fisher information from the log-likelihood function, is 7.0780. The relative bias is 0.3098%, which strongly supports the accuracy of the standard error estimates for $\hat{\mu}$ in the moment matching approach.

The figure A.1 presented below further justifies the accuracy of the moment matching approach. As demonstrated, the estimate of the mean sum across all publications ($\mu = \sum_i \mu_i$) can be effectively parameterised using μ_0 alone. It is evident that the log-likelihood functions of the moment matching approach and the exact log-likelihood have consistent shapes, although they are on different value scales. The Maximum likelihood estimates (MLEs) for both methods are remarkably close to their true value of 10^{-3}. This indicates that the MLE for the mean sum of the moment matched Negative Binomial (NB) distribution is precise.

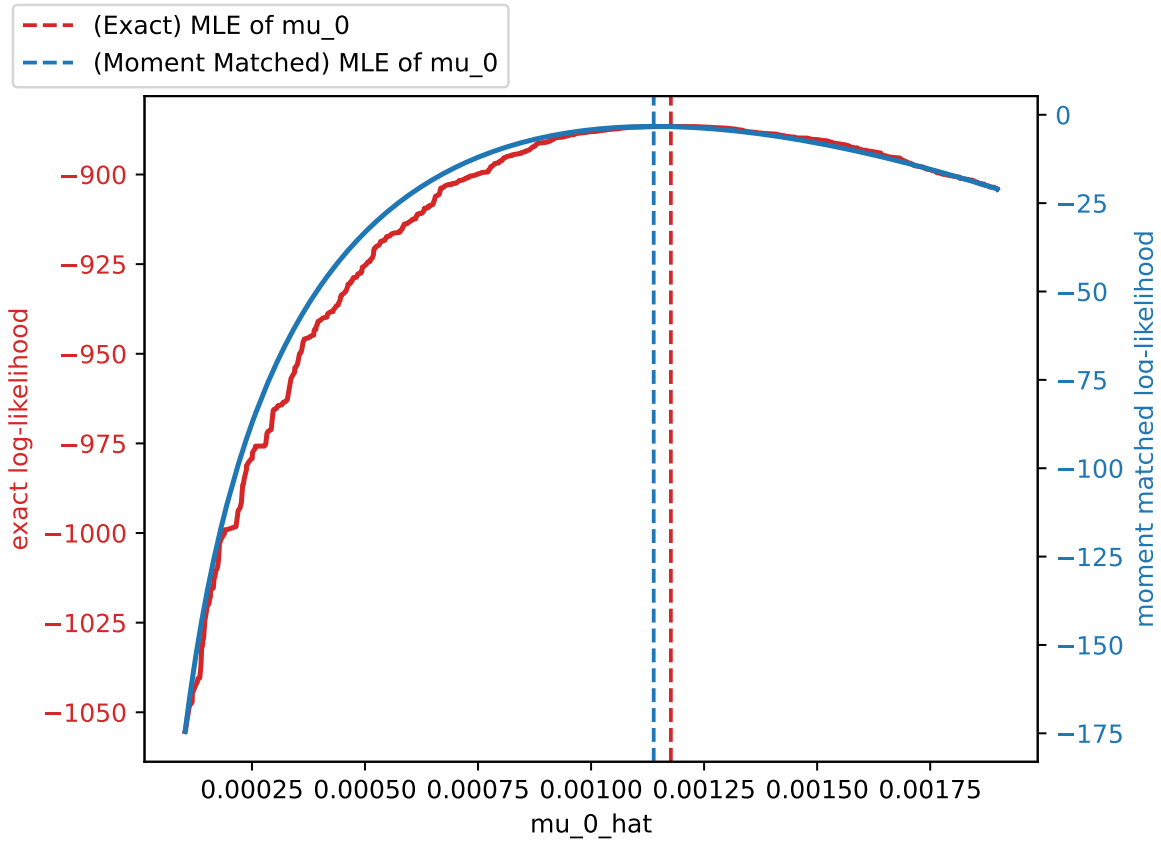


Figure A.1: Exact log-likelihood functions of NB distribution and its moment matching approximation

A.1.5 Two-stage hierarchy Poisson-Gamma model

In this section, we propose a two-stage hierarchy Poisson-Gamma model, which regards the random (Gamma) effect as a latent characteristic of each publication, instead of independent voxelwise effects. The name "two-stage hierarchy Poisson-Gamma model" comes from the modelling procedure: consider the clustered count data Y_{ij}, $i = 1, \cdots, M$ (number of publications), $j = 1, \cdots, N$ (number of voxels). Draw λ_i from a Gamma distribution with mean 1 and variance α.

$$\lambda_i \sim \text{Gamma}(\alpha^{-1}, \alpha^{-1}) \Rightarrow \mathbb{E}(\lambda_i) = 1, \mathbb{V}(\lambda_i) = \alpha,$$

$$f(\lambda_i) = \frac{\frac{1}{\alpha}^{\alpha^{-1}}}{\Gamma(\alpha^{-1})} \lambda_i^{\alpha^{-1}-1} e^{-\alpha^{-1}\lambda_i}$$

For each publication i, for each voxel j, $Y_{ij}|\lambda_i$ are drawn from a Poisson distribution with mean $\lambda_i \mu_{ij}$, where μ_{ij} is the spatial mean parameterized by some β (B-spline

basis coefficients).

$$Y_{ij}|\lambda_i \sim \text{Poisson}(\lambda_i\mu_{ij}) \Rightarrow P(Y_{ij}|\lambda_i = k) = \frac{(\lambda_i\mu_{ij})^k e^{-\lambda_i\mu_{ij}}}{k!}$$

Therefore, the marginal probability of the foci count Y_{ij} is,

$$\begin{aligned}
P(Y_{ij} = y_{ij}) &= \int_{\lambda_i} P(Y_{ij}|\lambda_i)P(\lambda_i)d\lambda_i = \int_{\lambda_i=0}^{\infty} \frac{(\mu_{ij}\lambda_i)^{y_{ij}} e^{-\mu_{ij}\lambda_i}}{y_{ij}!} \frac{\frac{1}{\alpha}^{\alpha^{-1}}}{\Gamma(\alpha^{-1})} \lambda_i^{\alpha^{-1}-1} e^{-\alpha^{-1}\lambda_i} d\lambda_i \\
&= \frac{\mu_{ij}^{y_{ij}} \frac{1}{\alpha}^{\alpha^{-1}}}{\Gamma(\alpha^{-1})y_{ij}!} \int_{\lambda_i=0}^{\infty} \lambda_i^{y_{ij}+\alpha^{-1}-1} e^{-\lambda_i(\mu_{ij}+\alpha^{-1})} d\lambda_i = \frac{\mu_{ij}^{y_{ij}} \frac{1}{\alpha}^{\alpha^{-1}}}{\Gamma(\alpha^{-1})y_{ij}!} \frac{\Gamma(y_{ij}+\alpha^{-1})}{(\mu_{ij}+\alpha^{-1})^{y_{ij}+\alpha^{-1}}} \\
&= \frac{\Gamma(y_{ij}+\alpha^{-1})}{\Gamma(y_{ij}+1)\Gamma(\alpha^{-1})} \left(\frac{\mu_{ij}}{\mu_{ij}+\alpha^{-1}}\right)^{y_{ij}} \left(\frac{\alpha^{-1}}{\mu_{ij}+\alpha^{-1}}\right)^{\alpha^{-1}}
\end{aligned}$$

Therefore, the marginal distribution of foci count follows a NB distribution, $Y_{ij} \sim \text{NB}(\alpha^{-1}, \frac{\mu_{ij}}{\mu_{ij}+\alpha^{-1}})$, with mean $\mathbb{E}(Y_{ij}) = \mu_{ij}$ and variance $\text{Var}(Y_{ij}) = \mu_{ij} + \alpha\mu_{ij}^2$.

A.1.6 Covariance structure in Clustered NB model

The two-stage hierarchical clustered NB model also introduces a covariance structure between foci within a publication. Specifically, the covariance of the number of count in voxel j and voxel j' (Y_{ij} and $Y_{ij'}$) in publication i, modelled by the clustered NB model is,

$$\begin{aligned}
\mathbb{E}[Y_{ij}Y_{ij'}] &= \mathbb{E}_{\lambda_i}\left[\mathbb{E}[Y_{ij}Y_{ij'}|\lambda_i]\right] = \mathbb{E}_{\lambda_i}\left[\mathbb{E}[Y_{ij}|\lambda_i]\mathbb{E}[Y_{ij'}|\lambda_i]\right] = \mathbb{E}_{\lambda_i}\left[(\lambda_i\mu_{ij})(\lambda_i\mu_{ij'})\right] \\
&= \mu_{ij}\mu_{ij'} \int_{\lambda_i} \lambda_i^2 f(\lambda_i)d\lambda_i = \mu_{ij}\mu_{ij'} \int_{\lambda_i} \frac{\frac{1}{\alpha}^{\alpha^{-1}}}{\Gamma(\alpha^{-1})} \lambda_i^{\alpha^{-1}+1} e^{-\alpha^{-1}\lambda_i} d\lambda_i \\
&= \mu_{ij}\mu_{ij'} \frac{\frac{1}{\alpha}^{\alpha^{-1}}}{\Gamma(\alpha^{-1})} \alpha^{\alpha^{-1}+1}\alpha \int_{\lambda_i} (\frac{1}{\alpha}\lambda_i)^{(\alpha^{-1}+1)} e^{\alpha^{-1}\lambda_i} d(\alpha^{-1}\lambda_i) \\
&= \mu_{ij}\mu_{ij'}\alpha^2 \frac{1}{\Gamma(\alpha^{-1})}\Gamma(\alpha^{-1}+2) = \mu_{ij}\mu_{ij'}\alpha^2\alpha^{-1}(\alpha^{-1}+1) \\
&= (1+\alpha)\mu_{ij}\mu_{ij'}
\end{aligned}$$

For different voxels j and j' within a same publication i,

$$\text{Cov}(Y_{ij}Y_{ij'}) = \mathbb{E}[Y_{ij}Y_{ij'}] - \mathbb{E}[Y_{ij}]\mathbb{E}[Y_{ij'}] = (1+\alpha)\mu_{ij}\mu_{ij'} - \mu_{ij}\mu_{ij'} = \alpha\mu_{ij}\mu_{ij'}$$

While for different publications i and i',

$$\text{Cov}(Y_{ij}Y_{i'j'}) = 0$$

Now, we will look at the total log-likelihood function of the clustered NB model, using dependence between publications. Let $Y_{i,\cdot} = \sum_{j=1}^{N} Y_{ij}$ be the sum of foci within publication i regardless of location. The joint probability for the number of counts Y_{ij} $(\forall j = 1, \cdots, N)$ from the i^{th} publication is,

$$f(Y_{i1}, Y_{i2}, \cdots, Y_{iN}) = \int_{\lambda_i} f(Y_{i1}, Y_{i2}, \cdots, Y_{iN}|\lambda_i) f(\lambda_i) d\lambda_i = \int_{\lambda_i} \prod_{j=1}^{N} f(Y_{ij}|\lambda_i) f(\lambda_i) d\lambda_i$$

$$= \int_{\lambda_i} \prod_{j=1}^{N} \frac{(\mu_{ij}\lambda_i)^{Y_{ij}} e^{-\mu_{ij}\lambda_i}}{Y_{ij}!} \frac{\frac{1}{\alpha}^{\alpha^{-1}}}{\Gamma(\alpha^{-1})} \lambda_i^{\alpha^{-1}-1} e^{-\alpha^{-1}\lambda_i} d\lambda_i$$

$$= \frac{\frac{1}{\alpha}^{\alpha^{-1}} \prod_{j=1}^{N} \mu_{ij}^{Y_{ij}}}{\Gamma(\alpha^{-1}) \prod_{j=1}^{N} Y_{ij}!} \int_{\lambda_i} \exp\{-\lambda_i(\alpha^{-1} + \sum_{j=1}^{N} \mu_{ij})\} \lambda_i^{\sum_{j=1}^{N} Y_{ij} + \alpha^{-1} - 1} d\lambda_i$$

$$= \frac{\frac{1}{\alpha}^{\alpha^{-1}} \prod_{j=1}^{N} \mu_{ij}^{Y_{ij}}}{\Gamma(\alpha^{-1}) \prod_{j=1}^{N} Y_{ij}!} \int_{\lambda_i} \exp\{-\lambda_i(\alpha^{-1} + \mu_{i,\cdot})\} \lambda_i^{Y_{i,\cdot} + \alpha^{-1} - 1} d\lambda_i$$

$$= \frac{\frac{1}{\alpha}^{\alpha^{-1}} \prod_{j=1}^{N} \mu_{ij}^{Y_{ij}}}{\Gamma(\alpha^{-1}) \prod_{j=1}^{N} Y_{ij}!} \frac{1}{(\alpha^{-1} + \mu_{i,\cdot})^{Y_{i,\cdot} + \alpha^{-1} - 1}} \frac{1}{\alpha^{-1} + \mu_{i,\cdot}} \cdot$$

$$\int_{\lambda_i} \exp\{-\lambda_i(\alpha^{-1} + \mu_{i,\cdot})\} [\lambda_i(\alpha^{-1} + \mu_{i,\cdot})]^{Y_{i,\cdot} + \alpha^{-1} - 1} d[\lambda_i(\alpha^{-1} + \mu_{i,\cdot})]$$

$$= \frac{\prod_{j=1}^{N} \mu_{ij}^{Y_{ij}}}{\Gamma(\alpha^{-1}) \prod_{j=1}^{N} Y_{ij}!} \frac{\frac{1}{\alpha}^{\alpha^{-1}}}{(\alpha^{-1} + \mu_{i,\cdot})^{Y_{i,\cdot} + \alpha^{-1}}} \Gamma(Y_{i,\cdot} + \alpha^{-1})$$

$$= \frac{\Gamma(Y_{i,\cdot} + \alpha^{-1}) \frac{1}{\alpha}^{\alpha^{-1}}}{\Gamma(\alpha^{-1}) \prod_{j=1}^{N} Y_{ij}!} (\alpha^{-1} + \mu_{i,\cdot})^{-(Y_{i,\cdot} + \alpha^{-1})} \exp[\sum_{j=1}^{N} Y_{ij} \log(\mu_{ij})]$$

It gives rise to the log-likelihood function for publication i,

$$\log f(Y_{i1}, \cdots, Y_{iN}) = \alpha^{-1} \log(\alpha^{-1}) + \log \Gamma(Y_{i,\cdot} + \alpha^{-1}) - \log \Gamma(\alpha^{-1}) - \sum_{j=1}^{N} \log Y_{ij}!$$

$$- (Y_{i,\cdot} + \alpha^{-1}) \log(\alpha^{-1} + \mu_{i,\cdot}) + \sum_{j=1}^{N} Y_{ij} \log(\mu_{ij})$$

Therefore, using the independence between publication i and i' ($i \neq i'$), the log-likelihood of $Y_{ij}(\forall i = 1, \cdots, M, j = 1, \cdots, N)$ across all publications and voxels is,

$$\begin{aligned} l(\beta, \alpha) &= \sum_{i=1}^{M} \log[f(Y_{i1}, Y_{i2}, \cdots, Y_{iN})] \\ &= M\alpha^{-1}\log(\alpha^{-1}) - M\log\Gamma(\alpha^{-1}) + \sum_{i=1}^{M}\log\Gamma(Y_{i,\bullet} + \alpha^{-1}) - \sum_{i=1}^{M}\sum_{j=1}^{N}\log Y_{ij}! \\ &- \sum_{i=1}^{M}(Y_{i,\bullet} + \alpha^{-1})\log(\alpha^{-1} + \mu_{i,\bullet}) + \sum_{i=1}^{M}\sum_{j=1}^{N} Y_{ij}\log(\mu_{ij}) \end{aligned}$$

Using the assumption that $Y_{ij} = 0$ or 1, so that $\log(Y_{ij}!) = 0$ and $\mu_{ij} = \mu_j^X \mu_i^Z$,

$$\begin{aligned} l(\beta, \alpha) &= M\alpha^{-1}\log(\alpha^{-1}) - M\log\Gamma(\alpha^{-1}) + \sum_{i=1}^{M}\log\Gamma(Y_{i,\bullet} + \alpha^{-1}) - \sum_{i=1}^{M}(Y_{i,\bullet} + \alpha^{-1})\log(\alpha^{-1} + \mu_{i,\bullet}) + \\ &= M\alpha^{-1}\log(\alpha^{-1}) - M\log\Gamma(\alpha^{-1}) + \sum_{i=1}^{M}\log\Gamma(Y_{i,\bullet} + \alpha^{-1}) - \sum_{i=1}^{M}(Y_{i,\bullet} + \alpha^{-1})\log(\alpha^{-1} + \mu_{i,\bullet}) \\ &+ \left(\sum_{j=1}^{N}\sum_{i=1}^{M} Y_{ij}\right)\left(\sum_{k=1}^{P} X_{jk}\beta_k\right) \\ &= M\alpha^{-1}\log(\alpha^{-1}) - M\log\Gamma(\alpha^{-1}) + \sum_{i=1}^{M}\log\Gamma(Y_{i,\bullet} + \alpha^{-1}) - \sum_{i=1}^{M}(Y_{i,\bullet} + \alpha^{-1})\log(\alpha^{-1} + \mu_{i,\bullet}) \\ &+ \left[\sum_{j=1}^{N} Y_{\bullet,j}\sum_{k=1}^{P} X_{jk}\beta_k\right] \\ &= M\alpha^{-1}\log(\alpha^{-1}) - M\log\Gamma(\alpha^{-1}) + \sum_{i=1}^{M}\log\Gamma(Y_{i,\bullet} + \alpha^{-1}) - \sum_{i=1}^{M}(Y_{i,\bullet} + \alpha^{-1})\log(\alpha^{-1} + \mu_{i,\bullet}) + \end{aligned}$$

A.2 Deterministic model

A.2.1 Model factorisation: Poisson model

We consider model factorisation to replace the full $(MN) - vector$ of foci counts with sufficient statistics (the dimension of which is not larger than M or N). Following the

total log-likelihood function in Equation (3.5),

$$
\begin{aligned}
l &= \sum_{i=1}^{M}\sum_{j=1}^{N}[Y_{ij}\log(\mu_{ij}) - \mu_{ij} - \log(y_{ij}!)] = \sum_{i=1}^{M}\sum_{j=1}^{N} Y_{ij}\log\mu_{ij} - \sum_{i=1}^{M}\sum_{j=1}^{N}\mu_{ij} - 0 \\
&= \left(\sum_{i=1}^{M}\sum_{j=1}^{N} Y_{ij}\right)\left(\sum_{k=1}^{P} X_{jk}\beta_k + \sum_{s=1}^{R} Z_{is}\gamma_s\right) - \sum_{i=1}^{M}\sum_{j=1}^{N}\mu_j^X\mu_i^Z \\
&= \left[\sum_{j=1}^{N} Y_{\cdot,j}\sum_{k=1}^{P} X_{jk}\beta_{gk}\right] + \sum_{i=1}^{M} Y_{i,\cdot}\sum_{s=1}^{R} Z_{is}\gamma_s - \left[\sum_{j=1}^{N}\mu_j^X\right]\left[\sum_{i=1}^{M}\mu_i^Z\right] \\
&= \left[\sum_{j=1}^{N} Y_{\cdot,j}\log\mu_j^X\right] + \sum_{i=1}^{M} Y_{i,\cdot}\log\mu_i^Z - \left[\mathbf{1}^\top\mu^X\right]\left[\mathbf{1}^\top\mu^Z\right] \\
&= Y_{\cdot,\cdot}^\top\log(\mu^X) + Y_{\cdot,\cdot}^\top\log(\mu^Z) - \left[\mathbf{1}^\top\mu^X\right]\left[\mathbf{1}^\top\mu^Z\right]
\end{aligned}
\tag{A.4}
$$

A.2.2 Model factorisation: NB model

Following the log-likelihood function in Equation (3.7),

$$
\begin{aligned}
l(\beta,\alpha) = \sum_{i=1}^{M}\sum_{j=1}^{N} &\big[\log\Gamma(Y_{ij}+\alpha^{-1}) - \log\Gamma(Y_{ij}+1) - \log\Gamma(\alpha^{-1}) \\
&+ Y_{ij}\log(\alpha\mu_{ij}) - (Y_{ij}+\alpha^{-1})\log(1+\alpha\mu_{ij}) \\
&= \sum_{i=1}^{M}\sum_{j=1}^{N}\left[\{\sum_{k=0}^{Y_{ij}-1}\log(k+\alpha^{-1})\} - \log\Gamma(Y_{ij}+1) + Y_{ij}\log(\alpha\mu_{ij}) - (Y_{ij}+\alpha^{-1})\log(1+\alpha\mu_{ij})\right] \\
&= \sum_{i=1}^{M}\sum_{j=1}^{N}\left[\{\sum_{k=0}^{Y_{ij}-1}\log(k+\alpha^{-1})\} - \log\Gamma(Y_{ij}+1) + Y_{ij}\log(\alpha) + Y_{ij}\log(\mu_{ij}) - (Y_{ij}+\alpha^{-1})\log(\right. \\
&= \left(\sum_{i=1}^{M}\sum_{j=1}^{N} Y_{ij}\log(\alpha^{-1}) - \sum_{i=1}^{M}\sum_{j=1}^{N}\log(1)\right) + \left(\sum_{i=1}^{M}\sum_{j=1}^{N} Y_{ij}\right)\log(\alpha) \\
&+ \sum_{i=1}^{M}\sum_{j=1}^{N} Y_{ij}\left(\sum_{k=1}^{P} X_{jk}\beta_k + \sum_{s=1}^{R} Z_{is}\gamma_s\right) - \sum_{i=1}^{M}\sum_{j=1}^{N}(Y_{ij}+\alpha^{-1})\log(1+\alpha\mu_{ij})
\end{aligned}
\tag{A.5}
$$

Here, the last term $\sum_{j=1}^{N}(Y_{ij}+\alpha^{-1})\log(1+\alpha\mu_{ij})$, is impractical to simplify, therefore, we consider a moment matching method similar to that in Appendix A.1.3, $Y_{\cdot,j} \sim$

$\mathrm{NB}(r'_j, p'_j)$ where

$$r'_j = \alpha^{-1}\frac{(\sum_{i=1}^{M}\mu_{ij})^2}{\sum_{i=1}^{M}\mu_{ij}^2} = \alpha^{-1}\frac{(\mu_j^X\sum_{i=1}^{M}\mu_i^Z)^2}{\sum_{i=1}^{M}(\mu_j^X\mu_i^Z)^2} = \alpha^{-1}\frac{(\mu_j^X)^2(\sum_{i=1}^{M}\mu_i^Z)^2}{\sum_{i=1}^{M}(\mu_j^X\mu_i^Z)^2}$$

$$p'_j = \frac{\sum_{i=1}^{M}\mu_{ij}^2}{\alpha^{-1}\sum_{i=1}^{M}\mu_{ij} + \sum_{i=1}^{M}\mu_{ij}^2} = \frac{\sum_{i=1}^{M}(\mu_j^X\mu_i^Z)^2}{\alpha^{-1}\sum_{i=1}^{M}(\mu_j^X\mu_i^Z) + \sum_{i=1}^{M}(\mu_j^X\mu_i^Z)^2} = \frac{\sum_{i=1}^{M}(\mu_j^X\mu_i^Z)^2}{\alpha^{-1}\mu_j^X\sum_{i=1}^{M}\mu_i^Z + \sum_{i=1}^{M}(\mu_j^X\mu_i^Z)^2} \tag{A.6}$$

And the parameter α' of excess variance in the NB approximation is

$$\frac{1}{\alpha'} = \frac{1}{\alpha}\frac{(\sum_{i=1}^{M}\mu_{ij})^2}{\sum_{i=1}^{M}\mu_{ij}^2} \Rightarrow \alpha' = \frac{\sum_{i=1}^{M}\mu_{ij}^2}{(\sum_{i=1}^{M}\mu_{ij})^2}\alpha = \frac{\sum_{i=1}^{M}(\mu_j^X\mu_i^Z)^2}{(\mu_j^X)^2(\sum_{i=1}^{M}\mu_i^Z)^2}\alpha \tag{A.7}$$

A.2.3 Model factorisation: clustered NB model

Following the total log-likelihood function in Appendix A.1.6, we incorporate the effect of publication-level covariates into the Clustered NB model,

$$
\begin{aligned}
l(\beta,\alpha) &= Mv\log(v) - M\log\Gamma(v) + \sum_{i=1}^{M}\log\Gamma(Y_{i,\cdot}+v) - \sum_{i=1}^{M}(Y_{i,\cdot}+v)\log(v+\mu_{i,\cdot}) + \sum_{i=1}^{M}\sum_{j=1}^{N}Y_{ij}\log(\\
&= Mv\log(v) - M\log\Gamma(v) + \sum_{g=1}^{B}\sum_{i\in I_g}\log\Gamma(Y_{i,\cdot}+v) - \sum_{g=1}^{B}\sum_{i\in I_g}(Y_{i,\cdot}+v)\log(v+\mu_{i,\cdot}) \\
&+ \sum_{g=1}^{B}\left(\sum_{j=1}^{N}\sum_{i\in I_g}Y_{ij}\right)\left(\sum_{k=1}^{P}X_{jk}\beta_{g(i)k} + \sum_{s=1}^{R}Z_{is}\gamma_s\right) \\
&= Mv\log(v) - M\log\Gamma(v) + \sum_{g=1}^{B}\sum_{i\in I_g}\log\Gamma(Y_{i,\cdot}+v) - \sum_{g=1}^{B}\sum_{i\in I_g}(Y_{i,\cdot}+v)\log(v+\mu_{i,\cdot}) \\
&+ \sum_{g=1}^{B}\left[\sum_{j=1}^{N}Y_{gj}\sum_{k=1}^{P}X_{jk}\beta_{g(i)k}\right] + \sum_{i=1}^{M}Y_{i,\cdot}\sum_{s=1}^{R}Z_{is}\gamma_s \\
&= Mv\log(v) - M\log\Gamma(v) + \sum_{g=1}^{B}\sum_{i\in I_g}\log\Gamma(Y_{i,\cdot}+v) - \sum_{g=1}^{B}\sum_{i\in I_g}(Y_{i,\cdot}+v)\log(v+\mu_{i,\cdot}) \\
&+ \sum_{g=1}^{B}Y_g^{\top}\log(\mu_g^X) + Y_{\cdot}^{\top}\log(\mu^Z)
\end{aligned}
\tag{A.8}
$$

A.2.4 Using IRLS to Optimize the Quasi-Poisson Model

Previously, we optimised the regression coefficients for likelihood-based models (e.g., Poisson, NB and clustered NB models) using Fisher scoring or L-BFGS algorithm. However, for Quasi-likelihood models (e.g., Quasi-Poisson model), where exact likelihood functions are computationally infeasible, we use the Iteratively Reweighted Least Squares (IRLS) method to iteratively determine the optimal regression coefficients.

An ordinary one-parameter exponential family of density functions can be written as,

$$
f_\mu(y) = e^{\eta y - \psi(\mu)}\cdot[dG(y)] \tag{A.9}
$$

Here μ is the expectation parameter, $\mu = \int_{-\infty}^{\infty} yf(y)dG(y)$; y is the natural statistic; η is the natural or canonical parameter, a monotone function of μ; $\psi(\mu)$ is a normalizing

function, chosen to make the density integrate to 1. $G(y)$ is the *carrier measure* for the exponential family so that $Pr_\mu\{A\} = \int_A f_\mu(y)dG(y)$ for measurable sets A.

The Poisson model belongs to the exponential family, as its probability density function can be written as,

$$\begin{aligned} f_\mu(y) &= \frac{\mu^y}{y!}e^{-\mu} \\ &= \exp[y\log\mu - \mu - \log(y!)] \end{aligned} \tag{A.10}$$

where $\eta = \log(\mu)$ and $\psi(\mu) = \mu + \log(y!)$.

The double exponential families include an extra parameter θ to allow for over-dispersion, so that $Var(y) = \frac{\mathbb{E}(y)}{\theta}$. The probability distribution function can be written as,

$$\tilde{f}_{\mu,\theta}(y) = c(\mu.\theta)\theta^{\frac{1}{2}}\{f_\mu(y)\}^\theta\{f_y(y)\}^{1-\theta}[dG(y)] \tag{A.11}$$

The constant $c(\mu,\theta)$ is defined to make $\int_{-\infty}^{\infty}\tilde{f}_{\mu,\theta}(y)dG(y) = 1$. Therefore, the probability of Quasi-Poisson with unknown parameters μ and θ is written as,

$$\begin{aligned} \tilde{f}_{\mu,\theta}(y) &= c(\mu,\theta)\theta^{\frac{1}{2}}\left[\exp(y\log\mu - \mu - \log(y!))\right]^\theta\left[\exp(y\log y - y - \log(y!))\right]^{1-\theta} \\ &= c(\mu,\theta)\theta^{\frac{1}{2}}\exp\{\theta[y\log\mu - \mu - \log(y!)] + (1-\theta)[y\log y - y - \log(y!)]\} \end{aligned} \tag{A.12}$$

As a discrete distribution, we can choose a maximum count data n, compute the probability $\tilde{f}_{\mu,\theta}(y)$ of possible count data $y = 1, 2, \cdots, n$ and scale up the sum to 1.

The updating equation for the k^{th} iteration of IRLS is given by,

$$\hat{\beta}^{[j+1]} = (X^T W^{[j]} X)^{-1} X^T W^{[j]} \xi^{[j]} \tag{A.13}$$

where $\xi^{[j]} = \eta^{[j]} + (W^{[j]})^{-1}(y - \mu^{[j]})$ and link function $\eta^{[k]} = g(\mu^{[k]}) = X\beta^{[k]}$. $W^{[j]}$ is a diagonal matrix with elements,

$$\frac{\left[\frac{\partial g^{-1}(\eta_1^{[j]})}{\partial \eta_1^{[j]}}\right]^2}{v(\mu_1^{[j]})}, \cdots, \frac{\left[\frac{\partial g^{-1}(\eta_n^{[j]})}{\partial \eta_n^{[j]}}\right]^2}{v(\mu_n^{[j]})} \tag{A.14}$$

For Quasi-Poisson model,

$$W = diag(\frac{\mu_1^2}{\theta\mu_1}, \cdots, \frac{\mu_n^2}{\theta\mu_n}) = diag(\frac{\mu_1}{\theta}, \cdots, \frac{\mu_n}{\theta}) \tag{A.15}$$

and Equation A.13 can be written as,

$$\hat{\beta}^{[j+1]} = \beta^{[j]} + (X^\top W^{[j]} X)^{-1} X^\top (y - \mu^{[j]}) \tag{A.16}$$

A.2.5 Using the Delta Method to Estimate the Standard Errors of η^X and μ^X

In the test of homogeneity to identify activation regions, the standard error for β (regression coefficients) can be asymptotically estimated from the inverse of the observed Fisher Information matrix. Additionally, the standard error of the linear response η_{ij}^X ($\eta^X = X\beta$), can be estimated using the delta method.

By definition, the optimal regression coefficients $\hat{\beta}$ converges in probability to its true value β, and a central limit theorem can be applied to obtain asymptotic normality,

$$\sqrt{n}(\beta - \hat{\beta}) \xrightarrow{D} N(0, \Sigma) \tag{A.17}$$

where n is the number of observations and Σ is a (symmetric positive semi-definite) covariance matrix.

$$\begin{aligned} Var(\hat{\eta}^X) &= Var(X\hat{\beta}) \\ &= XCov(\hat{\beta})X^\top = X\Sigma X^\top \end{aligned} \tag{A.18}$$

Since keeping only the first two terms of the Taylor series, and using vector notation for the gradient, we can estimate μ^X as

$$\begin{aligned} \hat{\mu}^X &= \exp(\hat{\eta}^X) = \exp(\eta^X) + \nabla \exp(\eta^X)(\beta - \hat{\beta}) \\ &= \exp(\eta^X) + \operatorname{diag}(\exp(\eta^X))(\eta^X - \hat{\eta}^X) \\ \operatorname{Var}(\mu^X) &= \operatorname{Var}\left(\exp(\eta^X) + \operatorname{diag}(\exp(\eta^X))(\eta^X - \hat{\eta}^X)\right) \\ &= \operatorname{Var}(\exp(\eta^X)) + \operatorname{Var}(\operatorname{diag}(\exp(\eta^X)) \cdot (\eta^X - \hat{\eta}^X)) \\ &= \operatorname{Var}(\exp(\eta^X)) + \operatorname{Var}(\operatorname{diag}(\exp(\eta^X)) \cdot \eta^X) - \operatorname{Var}(\operatorname{diag}(\exp(\eta^X)) \cdot \hat{\eta}^X) \\ &= \operatorname{Var}(\operatorname{diag}(\exp(\eta^X)) \cdot \hat{\eta}^X) \\ &= \operatorname{diag}(\exp(\eta^X)) \operatorname{Var}(\hat{\eta}^X) \operatorname{diag}(\exp(\eta^X)) \\ &= \operatorname{diag}(\exp(\eta^X)) X\Sigma X^\top \operatorname{diag}(\exp(\eta^X)) \end{aligned} \tag{A.19}$$

The delta method therefore implies that

$$\sqrt{n}(\mu^X - \hat{\mu}^X) \xrightarrow{D} N\left[0, \operatorname{diag}(\exp(\eta^X)) X\Sigma X^\top \operatorname{diag}(\exp(\eta^X))\right]$$

A.3 Simulation studies to validate the spatial design matrix in CBMR

A.3.1 Cubic B-spline basis and Gaussian kernel basis functions

To rigorously evaluate the robustness of CBMR with a spatial B-spline basis matrix, we have conducted simulation experiments to demonstrate the effectiveness of the CBMR approach. These experiments are conducted in 2D settings, with either homogeneous intensity over the space or two bump signals located at the top-left and bottom-right corners of the image (these bump signals are constructed using Gaussian distributions, on the basis of background noise). The spatial design matrix is generated either with a cubic B-spline basis or a Gaussian kernel basis.

The specific setups are as follows:

- 2D simulation with homogeneous intensity: the intensity values are 0.01, 0.1 and 1.
- 2D simulation with bump signals: generated with two Gaussian distributions at the top-left and bottom-right corners of the image, combined with background noise. The expected numbers of foci per experiment are 2, 20, 200, respectively.

We then evaluate the mean, bias and mean square error (MSE) of the intensity estimates, denoted as $\hat{\mu}$ (averaged across space), as well as the relative difference in maximised log-likelihood (ML) between spatial design matrix generated with a B-spline basis or a Gaussian kernel, in each simulation scenario (see Tables A.1 and A.2 for details).

Therefore, we believe that we've substantiated the effectiveness of spatial design matrices created using either a cubic B-spline basis or a Gaussian kernel in the estimation of spatial intensity within the CBMR framework, and the difference between these two approaches is found to be minimal.

Intensity Intensity	Spatial matrix	Bias($\hat{\mu}$)	Std($\hat{\mu}$)	MSE($\hat{\mu}$)	Rel. diff in ML
0.01 0.01	B-spline basis Gaussian kernel	6.6516×10^{-6} 5.5857×10^{-6}	2.1370×10^{-2} 2.2222×10^{-2}	4.5666×10^{-6} 4.9388×10^{-6}	0.0402%
0.1 0.1	B-spline basis Gaussian kernel	-5.2330×10^{-6} -5.9101×10^{-6}	6.7805×10^{-3} 7.0259×10^{-3}	4.5978×10^{-5} 4.9366×10^{-5}	−0.0007%
1.0 1.0	B-spline basis Gaussian kernel	7.2501×10^{-6} 8.2326×10^{-6}	2.136×10^{-2} 2.2054×10^{-2}	4.5628×10^{-4} 4.8639×10^{-4}	−0.0002%

Table A.1: Bias, Std, MSE and relative difference in maximised log-likelihood (ML) for homogeneous spatial intensity

Expected n_foci	Spatial matrix	Bias($\hat{\mu}$)	Std($\hat{\mu}$)	MSE($\hat{\mu}$)	Rel.diff in ML
2 2	B-spline basis Gaussian kernel	$-6.1176e^{-6}$ $-5.6949e^{-6}$	$1.0586e^{-4}$ $9.1947e^{-5}$	$1.1244e^{-8}$ $8.4867e^{-9}$	1.0890%
20 20	B-spline basis Gaussian kernel	$-6.2624e^{-7}$ $-3.5496e^{-8}$	$2.4268e^{-4}$ $2.5395e^{-4}$	5.88964^{-8} $6.4492e^{-8}$	0.08738%
200 200	B-spline basis Gaussian kernel	$-1.29598e^{-5}$ $-1.1154e^{-5}$	$5.9853e^{-4}$ $7.8658e^{-4}$	$3.5840e^{-7}$ $6.1883e^{-7}$	1.4398%

Table A.2: Bias, Std, MSE and relative bias of difference in maximised log-likelihood values for inhomogeneous spatial intensity with two bump signals

A.3.2 Sensitivity analysis on knots locations, numbers and degree of B-spline basis

In all experiments described in Section 4.3, we have consistently used a cubic B-spline basis with a knot spacing of 10 voxels, equivalent to 20 mm. (Henceforth, we will refer knot spacing only using voxel units.) In this section we demonstrate that this choice is supported by evidence from experiments, not just prior knowledge or pragmatic considerations. We will explore how variations in the number of knots, their locations, and the degree of the B-spline basis affect the downstream meta-regression approach. This analysis will allow us to provide practical recommendations for parameter selection in future applications of the CBMR approach.

We perform sensitivity analysis using both simulated settings and real datasets. The specific setups are as follows:

- Simulation: We create an underlying intensity function, which is the sum of the following two components:
  - An intensity function that is the sum of two scaled two Gaussian probability density functions with centres at $(25, 25, 25)$ and $(65, 65, 65)$ in voxel space, with covariance matrix $5 \cdot I_3$ (i.e. standard deviation $\sqrt{5}$, FWHM ≈ 5.3). We used a scale factor of 5, which produced an average of 8.44 foci locations per publication within the brain mask.
  - a spatial homogeneous background intensity of 10^{-6} across the entire brain image.

  Using this intensity function we simulated data at each voxel according to a Negative Binomial distribution, with inflation factor $\alpha = 0.5$.

  This procedure is repeated 100 times to compare the relative mean, standard deviation (SD) and root mean squared error (RMSE) of the estimated intensity for the following experiment settings. Additionally, we evaluate the relative difference in maximised log-likelihood values.

  We consider the following variations:

  - Baseline experiment: Spatial design matrix uses a cubic B-spline basis with knots spacing of 10 voxels (as in body of the paper);

B-spline basis	1/2 int. shift	Rel.bias($\hat{\mu}$)	Rel.std($\hat{\mu}$)	RMSE($\hat{\mu}$)	Rel. diff in ML
Cubic	**No**	4.3941%	13.8096%	$5.6967e^{-6}$	−2.6157%
Quadratic	No	4.1127%	12.2380%	$5.3711e^{-6}$	−2.5617%
Cubic	Yes	4.7631%	12.7943%	$5.9858e^{-6}$	−2.4767%
Quadratic	Yes	4.4988%	14.2543%	$5.3105e^{-6}$	−2.5986%

Table A.3: Relative Mean, relative SD, RMSE and relative bias of difference in maximised log-likelihood values for the simulated dataset, with spline spacing of 5 voxels. The baseline experiment results are highlighted in bold.

  - Quadratic vs Cubic B-spline basis: Design matrix uses a quadratic B-spline bases, while keeping the knot configurations from the baseline experiment.
  - Half interval shift: To understand the impact of knot locations, we shift all knot locations by half an interval to the right, keeping the other knot and spline configurations unchanged from the baseline experiment.
  - Knot spacing: We vary knot spacing (from 4 voxels to 40 voxels) and analyse a data sufficiency index (described below) to provide practical recommendations for the future application of the CBMR approach.

- Real datasets: We run CBMR with a Negative Binomial (NB) model on 20 cognitive datasets (details provided in Table 3.1) across various knot spacings (from 4 voxels to 40 voxels). We use the CBMR-estimated intensity with a knot spacing of 10 voxels as the baseline experiment, and compare the relative bias, standard deviation, difference in maximised log-likelihood values and RMSE of estimated intensity functions generated by other knot spacings.

For simulated data, since the actual underlying intensity function is known, we compare the CBMR-estimated intensities across various experiment settings with the actual underlying intensity function. The results are summarised in Table A.3. Our findings indicate that both cubic and quadratic B-spline bases, regardless of whether the knot locations are original or shifted to the right by half an interval, provide comparable levels of relative bias (ranging between 4.1127% and 4.7631%), relative standard deviation (ranging between 12.2380% and 14.2543%) and similar level of RMSE (ranging between 5.3105×10^{-6} and 5.9858×10^{-6}). (We examined the spatial variation in bias, and found it generally occurred in background area where a simulated dataset had no foci and our method estimated a negligible intensity, below the true

B-spline basis	knot spacing	Rel.bias($\hat{\mu}$)	Rel.SD($\hat{\mu}$)	RMSE($\hat{\mu}$)	Rel.diff in ML
Cubic	4	4.3364%	6.8069%	$5.9917e^{-6}$	−3.0709%
Cubic	**5**	4.3941%	13.8096%	$5.6967e^{-6}$	−2.6157%
Cubic	7.5	3.2286%	10.2068%	$3.9609e^{-6}$	−0.9849%
Cubic	10	1.2322%	10.7362%	$3.4287e^{-6}$	−0.3988%
Cubic	15	−0.6630%	4.7229%	$3.7970e^{-6}$	0.1300%
Cubic	20	−4.7553%	5.6380%	$5.1880e^{-6}$	0.5380%
Cubic	30	−10.0698%	7.0024%	$9.1770e^{-6}$	2.4265%
Cubic	40	−13.5894%	4.4585%	$1.0088e^{-5}$	2.77314%

Table A.4: Relative mean, SD, RMSE and relative bias of difference in maximised log-likelihood values of CBMR results across various knot spacings of cubic B-spline bases in the simulated dataset. The baseline experiment results are highlighted in bold.

10^{-6} background intensity.) We also compared the maximised log-likelihood values to those evaluated with true μ values. This comparison also shows small variations. These findings indicate that the knot locations and the degree of the B-spline basis are not significant factor influencing the estimated intensity function of the CBMR model.

Additionally, to provide practical guidelines for parameter selection in future applications of the CBMR approach, we introduce a data sufficiency index. This index will help to identify the minimum foci contributions per basis element required to ensure the effective functioning of the CBMR method. Since our basis is a partition of unity, when we project the foci onto the basis functions and sum over voxels, the total foci count is $\sum_{jk} X_{jk} Y_{\bullet,j}$, recalling that $Y_{\bullet,j}$ is sum count over publications at voxel j and X is the $(N \times P)$ is the spatial design matrix. Thus we can consider the total contribution to each basis element in this simplified setting by

$$\left[\sum_j X_{j1} Y_{\bullet,j}, \sum_j X_{j2} Y_{\bullet,j}, \ldots, \sum_j X_{jP} Y_{\bullet,j}\right],$$

and summarise this P-vector by the maximum total foci contribution to any one basis element

$$\max_{k=1,\ldots,P} \sum_j X_{jk} Y_{\bullet,j}.$$

B-spline basis	knot spacing	1/2 interval shift	Max total foci contribution
Cubic	4.0	No	13.7408
Cubic	5.0	**No**	18.7957
Cubic	7.5	No	45.3228
Cubic	10.0	No	73.8215
Cubic	15.0	No	182.0279
Cubic	20.0	No	286.8437
Cubic	30.0	No	375.7031
Cubic	40.0	No	362.0541

Table A.5: Maximum of total foci contribution per basis element for the Cue Reactivity dataset, with the baseline experiment in bold.

Using the maximum total foci contribution per basis element as the data sufficiency index, we calculated and compared this index across various knot spacings for one specific dataset, the Cue Reactivity, in Table A.5. This analysis reveals that the maximum total foci contribution per basis element increases almost monotonically with wider spline spacings of the B-spline basis, reaching maximum levels at a spacing of 30 voxels, beyond which no significant increases were observed. Additionally, we compared the relative bias, standard deviation (SD), difference in maximised log-likelihood values and RMSE in Table A.5. It indicates that all cubic B-spline basis functions with knot spacings between 4 and 20 voxels are capable of producing very accurate estimations of the intensity function. Among these, a spline spacing of 15 voxels is the best option, giving rise to the lowest relative bias, standard deviation, difference in maximised likelihood value and RMSE. Previously, our practical experience with the CBMR model indicated that a spline spacing of 10 voxels in datasets with low foci counts can lead to singularity issues in Fisher information and result in inaccurate standard errors. As a result, here we explore how to determine the optimal spline spacing for datasets of different sizes to avoid numerical singularity issues as guided by our data sufficiency index.

In table A.8, we used the intensity estimation from a spline spacing of 10 voxels as the reference (since the true underlying intensity function is unknown) and compared the relative bias, standard deviation, and RMSE of the CBMR with the NB model across various spline spacings between 4 and 40. We observed that the dissimilarity (in terms of relative bias and RMSE) increases as the spline spacing diverges from the reference with spline spacing of 10 voxels, while the relative standard deviation decreases as the spline spacing widens. This indicates that spline spacing significantly

influences intensity estimation in real datasets, with larger spline spacings giving rise to less variation in intensity estimation across different voxel locations. Additionally, Figure A.2 and Figure A.3, present the results from 100 experiments conducted the CBMR with the NB model for each of the 20 cognitive datasets (details provided in Table 3.1) where these datasets were categorised into four groups according to their total foci counts. We calculated the rate of convergence failures over 100 realisations and compared these rates across various spline spacings and computed data sufficiency indiex. (Each optimisation starts with random sampling of β's; here we only use one initialisation for each realisation, though in practice we re-initialise on convergence failure for real data analyses).

Our analysis revealed that different groups of real datasets require different spline spacings (or data sufficiency indices), as outlined below:

- For datasets with fewer than 500 foci counts, we recommend a spline spacing of 30 voxels or a data sufficiency index greater than 20, as larger spline spacings are associated with a higher data sufficiency index which helps to avoid non-convergence.
- For datasets with foci counts ranging from 500 and 1500, a spline spacing of 20 voxels or a data sufficiency index greater than 20 is recommended. At this spacing, the rates of failure to converge have been controlled at a very low level, and no further reductions in failure rates were observed with larger spline spacings.
- For datasets containing between 1500 and 4000 foci, we recommend a spline spacing of 15 voxels or a data sufficiecy index more than 65. At this level, the rate of failures decreased to zero, indicating successful convergence in all 100 experiments.
- For datasets with foci counts more than 4000, a spline spacing of 10 voxels or a data sufficiency index more than 65 is recommended. For these large datasets, there is sufficient foci contribution per basis element even with larger spline spacing, and a spline spacing of 10 voxels helps avoid numerical singularity in Fisher information.

We have also compared the maximised log-likelihood for each of the 20 cognitive datasets (details provided in Table 3.1) across 100 runs of the CBMR with the NB

Dataset		knot spacing							
		4	5	7.5	10	15	20	30	40
1	Bias	85.87%	65.5%	38.48%	0.00%	31.83%	44.53%	52.55%	55.04%
	SD	176.33%	137.61%	104.13%	87.33%	75.78%	63.78%	61.67%	54.83%
	MSE	$7.0767e^{-5}$	$5.4990e^{-5}$	$4.0051e^{-5}$	$3.1499e^{-5}$	$2.9655e^{-5}$	$2.8073e^{-5}$	$2.9245e^{-5}$	$2.8045e$
2	Bias	191.6%	188.56%	140.02%	0.00%	126.30%	143.2%	153.68%	159.95%
	SD	3868.61%	3681.34%	1894.89%	728.63%	224.51%	171.22%	122.16%	124.82%
	MSE	$1.1887e^{-3}$	$1.1313e^{-3}$	$5.8314e^{-4}$	$2.2362e^{-4}$	$7.9070e^{-5}$	$6.8522e^{-5}$	$6.0274e^{-5}$	$6.2290e$
3	Bias	191.42%	160.58%	74.21%	0.00%	55.01%	68.62%	75.34%	82.44%
	SD	1582.41%	920.24%	211.92%	130.73%	96.93%	78.91%	69.44%	75.32%
	MSE	$5.2983e^{-4}$	$3.1051e^{-4}$	$7.4638e^{-5}$	$4.3454e^{-5}$	$3.7049e^{-5}$	$3.4764e^{-5}$	$3.4059e^{-5}$	$3.7122e$
4	Bias	166.95%	130.33%	63.06%	0.00%	43.94%	50.54%	57.60%	60.04%
	SD	896.18%	461.70%	166.37%	116.22%	87.21%	78.23%	70.80%	67.83%
	MSE	$1.7039e{-}3$	$8.9665e^{-4}$	$3.3253e^{-4}$	$2.1724e^{-4}$	$1.8251e^{-4}$	$1.7404e^{-4}$	$1.7054e^{-4}$	$1.6925e$
5	Bias	102.27%	77.41%	41.89%	0.00%	36.68%	48.8%	57.64%	61.57%
	SD	264.64%	190.12%	136.72%	115.62%	98.71%	86.79%	80.33%	75.94%
	MSE	$1.4387e^{-4}$	$1.0409e^{-4}$	$7.2508e^{-5}$	$5.8621e^{-5}$	$5.3398e^{-5}$	$5.0497e^{-5}$	$5.0148e^{-5}$	$4.9595e$
6	Bias	98.49%	71.38%	39.93%	0.00%	33.78%	44.92%	51.31%	54.61%
	SD	228.12%	161.59%	118.08%	99.14%	83.03%	71.05%	67.46%	59.95%
	MSE	$1.1348e^{-4}$	$8.0678e^{-5}$	$5.6927e^{-5}$	$4.5275e^{-5}$	$4.0936e^{-5}$	$3.8389e^{-5}$	$3.8707e^{-5}$	$3.7036e$
7	Bias	154.98%	116.96%	60.53%	0.00%	43.95%	58.60%	72.66%	76.76%
	SD	730.71%	381.54%	190.41%	137.53%	115.32%	94.82%	81.46%	71.76%
	MSE	$2.7209e^{-4}$	$1.4536e^{-4}$	$7.2779e^{-5}$	$5.0098e^{-5}$	$4.4953e^{-5}$	$4.0603e^{-5}$	$3.9760e^{-5}$	$3.8275e$
8	Bias	80.82%	61.42%	38.33%	0.00%	32.18%	41.22%	50.46%	53.56%
	SD	196.23%	163.07%	135.44%	109.80%	96.08%	88.78%	81.64%	79.00%
	MSE	$7.3694e^{-5}$	$6.0508e^{-5}$	$4.8879e^{-5}$	$3.8129e^{-5}$	$3.5185e^{-5}$	$3.3990e^{-5}$	$3.3329e6{-}5$	$3.3143e$
9	Bias	197.88%	185.54%	93.98%	0.00%	64.04%	75.64%	83.72%	85.85%
	SD	2189.71%	1484.66%	290.07%	143.11%	91.46%	71.98%	63.13%	53.846%
	MSE	$9.9500e^{-4}$	$6.7712e^{-4}$	$1.3799e^{-4}$	$6.4767e^{-5}$	$5.0529e^{-5}$	$4.7252e^{-5}$	$4.7453e^{-5}$	$4.5860e$
10	Bias	175.07%	129.23%	61.81%	0.00%	47.85%	57.55%	65.39%	64.84%
	SD	924.02%	413.72%	149.45%	103.79%	74.75%	63.66%	60.01%	50.78%
	MSE	$3.8582e^{-4}$	$1.7781e^{-4}$	$6.6347e^{-5}$	$4.2580e^{-5}$	$3.6412e^{-5}$	$3.5208e^{-5}$	$3.6410e^{-5}$	$3.3786e$
11	Bias	100.42%	74.67%	41.29%	0.00%	36.24%	51.44%	67.23%	74.56%
	SD	302.45%	221.58%	167.48%	145.78%	130.8%	111.18%	95.27%	78.36%
	MSE	$1.5210e^{-4}$	$1.1160e^{-4}$	$8.2328e^{-5}$	$6.9579e^{-5}$	$6.4781e^{-5}$	$5.8466e^{-5}$	$5.5654e^{-5}$	$5.1627e$
12	Bias	59.51%	46.86%	31.61%	0.00%	25.32%	34.54%	43.39%	46.43%
	SD	133.43%	118.22%	102.98%	86.22%	78.66%	71.56%	66.92%	62.73%
	MSE	$8.1195e^{-5}$	$7.0674e^{-5}$	$5.9869e^{-5}$	$4.7917e^{-5}$	$4.5925e^{-5}$	$4.4162e6{-}5$	$4.4325e^{-5}$	$4.3373e$
13	Bias	198.52%	192.38%	121.30%	0.00%	80.00%	91.32%	98.83%	99.61%
	SD	2693.68%	2340.74%	585.23%	216.71%	116.21%	89.57%	78.86%	71.72%
	MSE	$4.5113e^{-4}$	$3.9228e^{-4}$	$9.9824e^{-5}$	$3.6196e^{-5}$	$2.3564e^{-5}$	$2.1364e^{-5}$	$2.1117e^{-5}$	$2.0501e$
14	Bias	147.03%	147.06%	141.87%	0.00%	177.53%	185.28%	189.74%	192.12%
	SD	5499.95%	5500.63%	5164.26%	2567.68%	382.94%	201.42%	129.23%	148.33%
	MSE	$1.4306e^{-3}$	$1.4307e^{-3}$	$1.3433e^{-3}$	$6.6762e^{-4}$	$1.0975e^{-4}$	$7.1158e^{-5}$	$5.9691e^{-5}$	$6.3110e$
15	Bias	61.14%	48.01%	30.33%	0.00%	25.52%	36.86%	45.54%	50.12%
	SD	135.25%	120.18%	105.26%	95.58%	87.24%	78.39%	73.93%	66.47%
	MSE	$7.7930e^{-5}$	$6.7951e^{-5}$	$5.7513e^{-5}$	$5.0183e^{-5}$	$4.7728e^{-5}$	$4.5481e^{-5}$	$4.5590e^{-5}$	$4.3711e$
16	Bias	106.78%	79.12%	43.33%	0.00%	35.58%	50.5%	62.55%	65.18%
	SD	292.45%	200.44%	143.39%	117.93%	101.4%	86.01%	78.87%	68.37%
	MSE	$1.0351e^{-4}$	$7.1642e^{-5}$	$4.9800e^{-5}$	$3.9206e^{-5}$	$3.5726e^{-5}$	$3.3160e^{-5}$	$3.3467e^{-5}$	$3.1405e$

17	Bias	198.0%	193.39%	115.88%	0.00%	74.11%	85.73%	95.32%	98.57%
	SD	2549.71%	2239.09%	471.09%	172.02%	96.78%	76.27%	63.3%	58.98%
	MSE	$5.1769e^{-4}$	$4.5495e^{-4}$	$9.8206e^{-5}$	$3.4822e^{-5}$	$2.4675e^{-5}$	$2.3228e^{-5}$	$2.3163e^{-5}$	$2.3253e^{-5}$
18	Bias	104.83%	77.16%	43.33%	0.00%	38.09%	59.29%	72.81%	78.34%
	SD	340.22%	234.88%	187.97%	163.09%	145.22%	119.39%	106.13%	93.61%
	MSE	$1.6885e^{-4}$	$1.1726e^{-4}$	$9.1489e^{-5}$	$7.7354e^{-5}$	$7.1205e^{-5}$	$6.3224e^{-5}$	$6.1041e^{-5}$	$5.7894e^{-5}$
19	Bias	183.14%	169.13%	108.53%	0.00%	91.59%	108.5%	126.61%	130.17%
	SD	3166.52%	2614.67%	1225.12%	610.3%	356.0%	283.14%	231.17%	207.65%
	MSE	$1.0340e^{-3}$	$8.5412e^{-4}$	$4.0093e^{-4}$	$1.9895e^{-4}$	$1.1983e^{-4}$	$9.8843e^{-5}$	$8.5919e^{-5}$	$7.9892e^{-5}$
20	Bias	163.83%	138.05%	73.84%	0.00%	62.3%	87.26%	96.93%	109.48
	SD	1318.59%	935.59%	345.83%	253.51%	196.2%	152.56%	138.69%	110.45%
	MSE	$1.2851e^{-3}$	$9.1470e^{-4}$	$3.4203e^{-4}$	$2.4520e^{-4}$	$1.9910e^{-4}$	$1.6999e^{-4}$	$1.6366e^{-4}$	$1.5041e^{-4}$

Table A.8: Relative bias, standard deviation (SD) and relative MSE for each of the 20 cognitive datasets (details provided in Table 3.1), with a spline spacing of 10 voxels as reference (as the underlying intensity function is unknown).

model, using different spacings of cubic B-spline bases and a reference of spline spacing of 10 voxels. As depicted in Figure A.4, maximised log-likelihood values decrease with larger B-spline basis knot spacings. These findings supports the conclusion that B-spline bases with smaller knot spacings are able to capture finer details and can produce more accurate estimations of intensity function. Although higher maximised log-likelihood values are associated with smaller spline spacing (less than the reference of 10 voxels), they mostly occur in datasets with relatively small foci counts. This could be linked to numerical singularity in estimating Fisher Information matrix and reduced accuracy in standard error estimates. Additionally, for any specific dataset, the relatively high maximised log-likelihood values associated with smaller spline knot spacings might due to overfitting.

A.4 Statistical inference and generalised linear hypothesis testing

A.4.1 Contrasts on regression coefficient of publication-level covariates

To investigate the effects of publication-level covariates (e.g., sample size, year of publication) on activation intensity estimation, we conduct generalised linear hypoth-

esis testing on the regression coefficients γ. For every publication-level covariate $\gamma_r, \forall r = 1, \cdots, s$,

- $H_0 : C_\gamma \gamma = C_\gamma[\gamma_1, \gamma_2, \cdots, \gamma_s]^T = \mathbf{0}^T_{m\times 1}$ where C_γ is the contrast matrix of size $m \times s (m \leq s)$
- $H_1 : C_\gamma \gamma = C_\gamma[\gamma_1, \gamma_2, \cdots, \gamma_s]^T \neq \mathbf{0}^T_{m\times 1}$

The covariance of the regression coefficient γ, $\text{Cov}_\gamma = \text{Cov}([\gamma_1, \gamma_2, \cdots, \gamma_s]^T)$ is approximated from the inverse of the Fisher Information matrix. According to the asymptotic normality of the maximum likelihood estimator,

$$\begin{aligned} \hat{\gamma} - \gamma &\xrightarrow{D} N\left(\mathbf{0}^T_{s\times 1}, \text{Cov}_\gamma\right) \\ C_\gamma(\hat{\gamma} - \gamma) &\xrightarrow{D} N\left(\mathbf{0}^T_{s\times 1}, C_\gamma \text{Cov}_\gamma C^T_\gamma\right) \\ C_\gamma \hat{\gamma} &\xrightarrow{D} N\left(C_\gamma \gamma, C_\gamma \text{Cov}_\gamma C^T_\gamma\right) \end{aligned}$$

Since a quadratic form of normal distribution has a Chi-square distribution,

$$(C_\gamma \hat{\gamma})^T (C_\gamma \text{Cov}_\gamma C^T_\gamma)^{-1} (C_\gamma \hat{\gamma}) \xrightarrow{D} \chi^2_m$$

for example, the contrast matrix $C_\gamma = [1, 0]$ or $[0, 1]$ is for testing if the regression coefficient of the 1^{st} or 2^{nd} publication-level covariate is zero.

A.4.2 PP-plots of spatial homogeneity tests for each 20 meta-analytic datasets

Previously, we displayed only the PP-plots of spatial homogeneity tests for four representative datasets in Section 4.3.1. Here, we will include all PP-plots for 20 meta-analytic datasets in Figure A.5.

A.4.3 Likelihood-based comparison between Poisson, NB and clustered NB model

To demonstrate the likelihood-based comparison between the Poisson, NB and clustered NB model, we plot the maximised log-likelihood and AIC for each of the 20 meta-analytic datasets in Figure A.6 and Figure A.7. We also conduct a Likelihood ratio

test (LRT) to evaluate the trade-off between model sufficiency and complexity. Here, we only list the p-values of the LRT between the Poisson and clustered NB model in Table A.9, as $p < 10^{-8}$ for the LRT between the Poisson and NB model for each of the 20 meta-analytic datasets.

Table A.9: p-values of Likelihood Ratio test between Poisson and clustered NB model

Dataset	p-value	Dataset	p-value	Dataset	p-value
Social Processing	$p < 10^{-8}$	PTSD	$p < 10^{-8}$	Substance Use	$p < 10^{-8}$
Dementia	$p < 10^{-8}$	Cue Reactivity	$p < 10^{-8}$	Emotion Regulation	$p < 10^{-8}$
Decision Making	$p < 10^{-8}$	Reward	$p < 10^{-8}$	Sleep Deprivation	$p < 10^{-8}$
Naturalistic	$p < 10^{-8}$	Problem Solving	$p < 10^{-8}$	Emotion	$p < 10^{-8}$
Cannabis Use	1	Nicotine Use	$p < 10^{-8}$	Frontal Pole CBP	$p < 10^{-8}$
Face Perception	$p < 10^{-8}$	Nicotine Administration	0.99	Executive Function	$p < 10^{-8}$
Finger Tapping	0.99	n-Back	$p < 10^{-8}$		

A.4.4 Effect of publication-level covariates

Here, we investigate the effect of publication-wise (square root) sample size and year of publication (after centring and standardisation) on each of the 20 meta-analytic datasets. Under the null hypothesis that regression coefficient of each covariate is not distinguishable from 0 ($\gamma_i = 0$ for $i = 1, 2$), we conduct hypothesis testing and summarise the $Z-$score and $p-$value in Table A.10.

A.5 Comparison of ALE and CBMR activation maps

In addition to the comparison shown in Figure 3.5 in Section 3.4.3, which represents uncorrected activation maps generated using CBMR and ALE for the Cue Reactivity dataset, Figure A.8 to Figure fig. A.16 also provide comparisons of uncorrected activation maps produced by CBMR and ALE for other coordinate-based meta-analytic datasets with varying numbers of reported foci.

A.6 Comparison with Bayesian LGCP regression

To validate the accuracy of intensity estimation and the detected activation regions generated by our CBMR approach, we reached out to the authors of the Bayesian

Table A.10: Hypothesis testing on the effect of two publication-level covariates on 20 meta-analytic datasets

	(Square root) sample size		Year of publication	
Dataset	$Z-$score	p-value	$Z-$score	p-value
Social Processing	10.9053	$p < 10^{-8}$	0.4164	0.6771
PTSD	2.8789	0.004	0.6029	0.5466
Substance Use	4.3887	1.1404×10^{-5}	6.8398	$p < 10^{-8}$
Dementia	20.7177	$p < 10^{-8}$	-1.3985	0.1620
Cue Reactivity	6.1454	$p < 10^{-8}$	-0.6880	0.4915
Emotion Regulation	6.8934	$p < 10^{-8}$	-3.9588	7.5329×10^{-5}
Decision Making	4.1104	3.9499×10^{-5}	0.1060	0.9156
Reward	-0.1228	0.9022	-	-
Sleep Deprivation	12.8765	$p < 10^{-8}$	0.4201	0.6744
Naturalistic	1.7038	0.0884	0.5395	0.5896
Problem Solving	4.3079	1.6485×10^{-5}	2.2789	0.0227
Cannabis Use	3.5915	3.2878×10^{-4}	2.2117	0.0270
Nicotine Use	5.0024	5.6631×10^{-7}	3.1836	0.0015
Frontal Pole CBP	5.5190	3.4101×10^{-8}	7.4040	$p < 10^{-8}$
Face Perception	3.4090	6.5212×10^{-4}	5.1018	3.3651×10^{-7}
Nicotine Administration	1.4594	0.1445	-1.0516	0.2930
Executive Function	1.6989	0.0932	0.5047	0.6138
Finger Tapping	-	-	0.1764	0.8600
n-Back	1.4616	0.1439	0.1239	0.9014

log-Gaussian Cox Process regression (LGCP; as detailed in [Samartsidis et al., 2019]). LGCP is a fully Bayesian random-effect meta-regression model capable of estimating activation intensity through a simulation-based approximation of the posterior using Markov Chain Monte Carlo (MCMC) methods, and it also accounts for publication-wise heterogeneity, similar to our CBMR approach. After obtaining the source code for their method, we applied it to the Cue Reactivity dataset for comparative analysis.

In the absence of p-value maps in the Bayesian LGCP approach, we chose to compare the estimated intensity maps generated by both the CBMR and LGCP methods, as presented in Figure A.17a and figure A.17b below (both are thresholded at 2×10^{-5}). Our analysis indicates significant consistency in the activation regions identified by these two approaches, particularly in the left cerebral cortex, frontal orbital cortex, insular cortex, and left and right accumbens. While we noticed that the activation regions identified in the CBMR intensity map appeared more isolated. However, these regions appeared more cohesive in the p-value maps when we controlled the false discovery rate (FDR) using the BH method (see Figure 3.6).

Although the LGCP model is a robust Bayesian meta-regression model that includes random-effect terms to address publication-wise heterogeneity, it's mathematically

complex, and its MCMC algorithm requires approximately 30 hours of computational time on an NVIDIA Tesla K20c GPU card, in contrast to approximately 537.52 seconds (approx 9 minutes) required for the CBMR with the NB model (tested on an Intel Xeon Gold 6340R CPU) for the Cue Reactivity dataset. Therefore, we believe our CBMR approach offers a computationally efficient alternative to the LGCP model, while still achieving comparable accuracy.

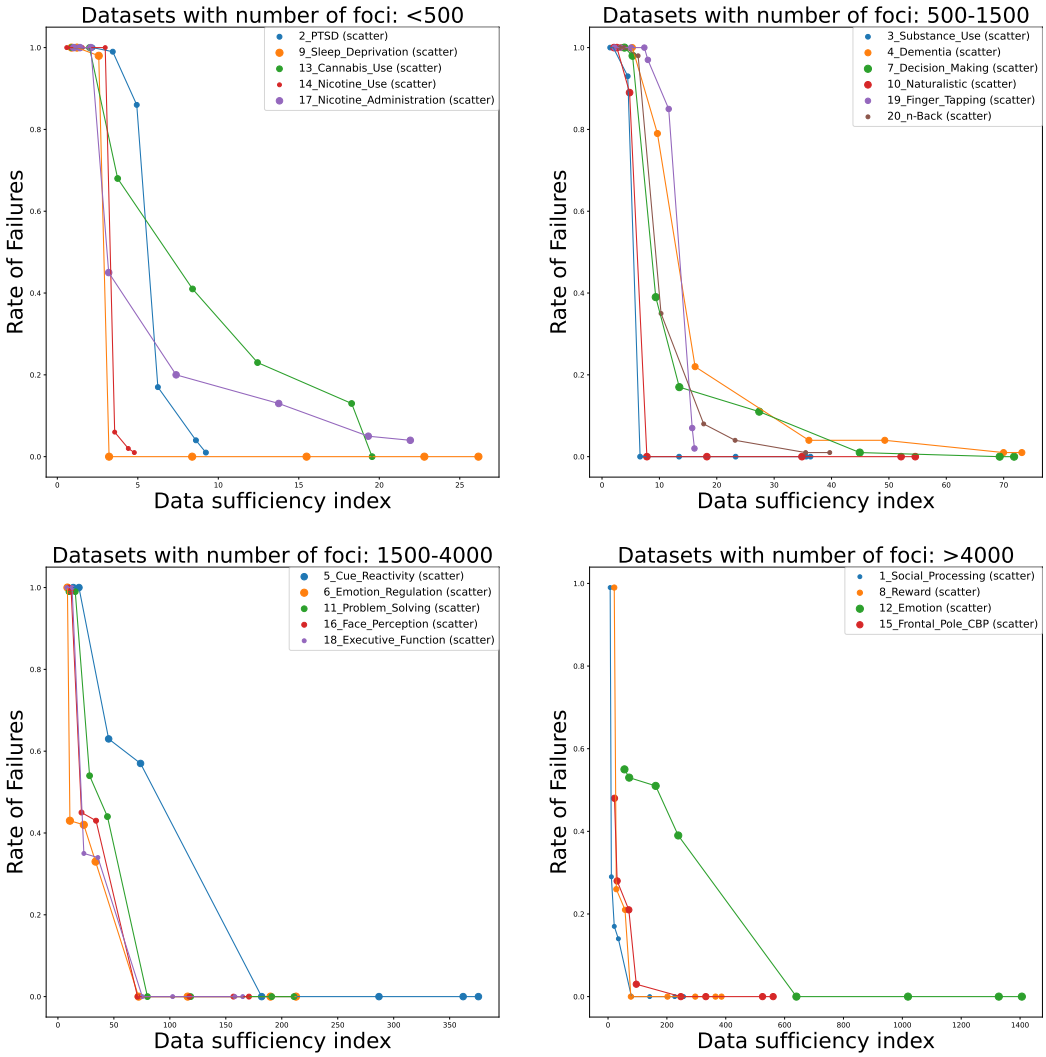


Figure A.2: Rate of failure to converge in 100 experiments for each of the 20 datasets (details provided in Table 3.1) across various data sufficiency index, categorised into 4 groups according to their total foci counts.

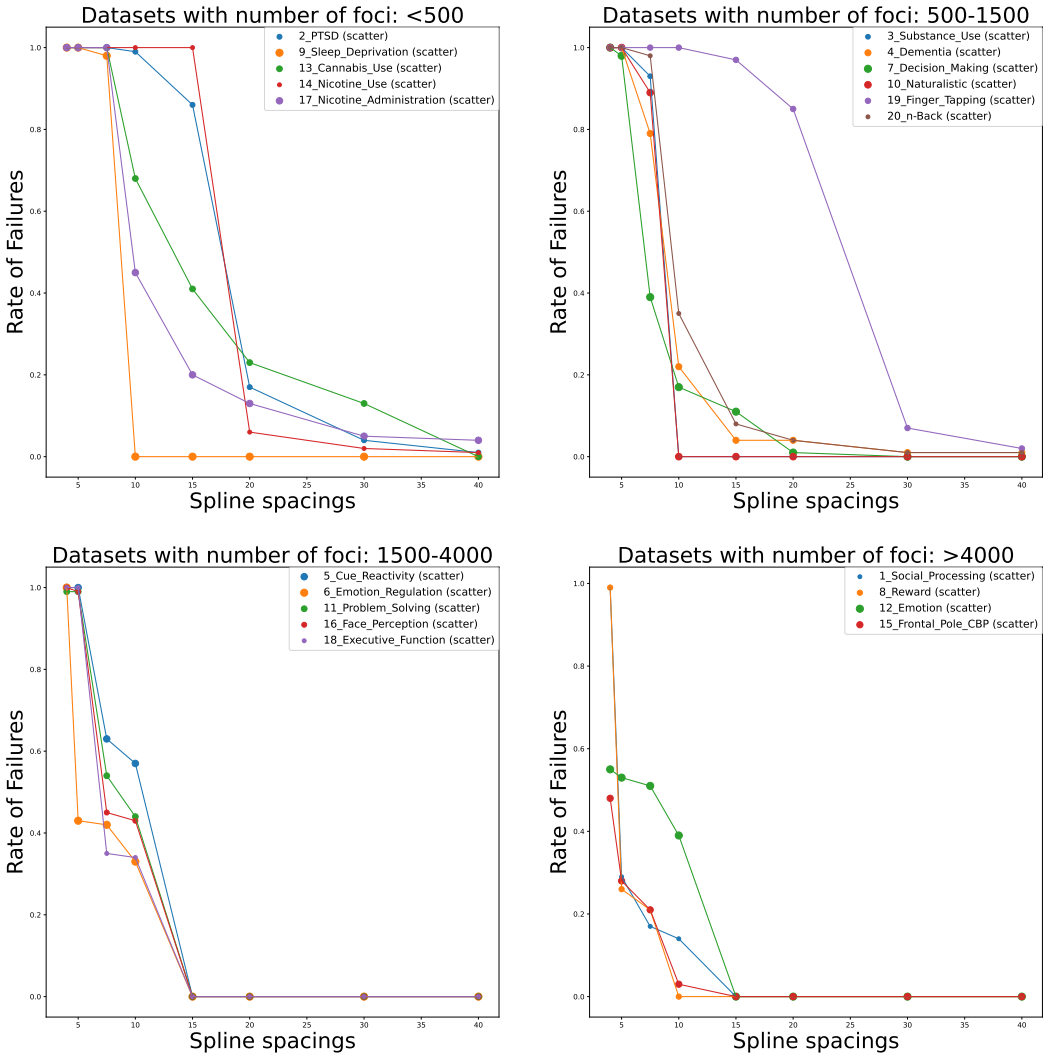


Figure A.3: Rate of failure to converge in 100 experiments for each of the 20 datasets (details provided in Table 3.1) across various spline spacings, categorised into 4 groups according to their total foci counts.

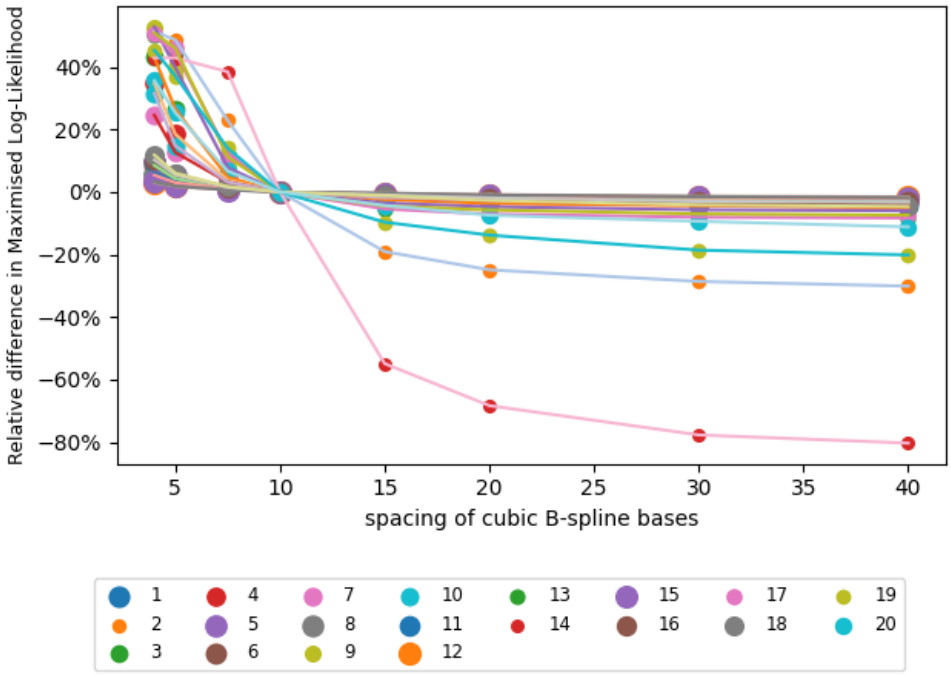


Figure A.4: Relative difference in maximised log-likelihood for each of the 20 cognitive datasets (details provided in Table 3.1), with a spline spacing of 10 voxels as the reference. The size of scatters reflect the total foci counts for each dataset.

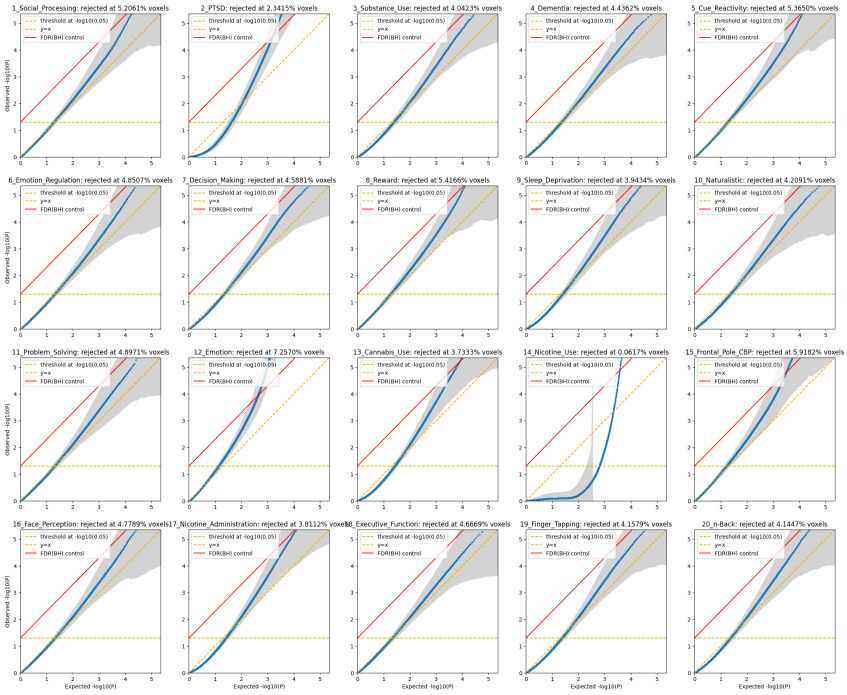


Figure A.5: P-P plot of p-value (under $-\log_{10}$ scale) with all of 20 meta-analytic datasets, estimated by CBMR with NB model without publication-level covariates, sampled with model-based approach.

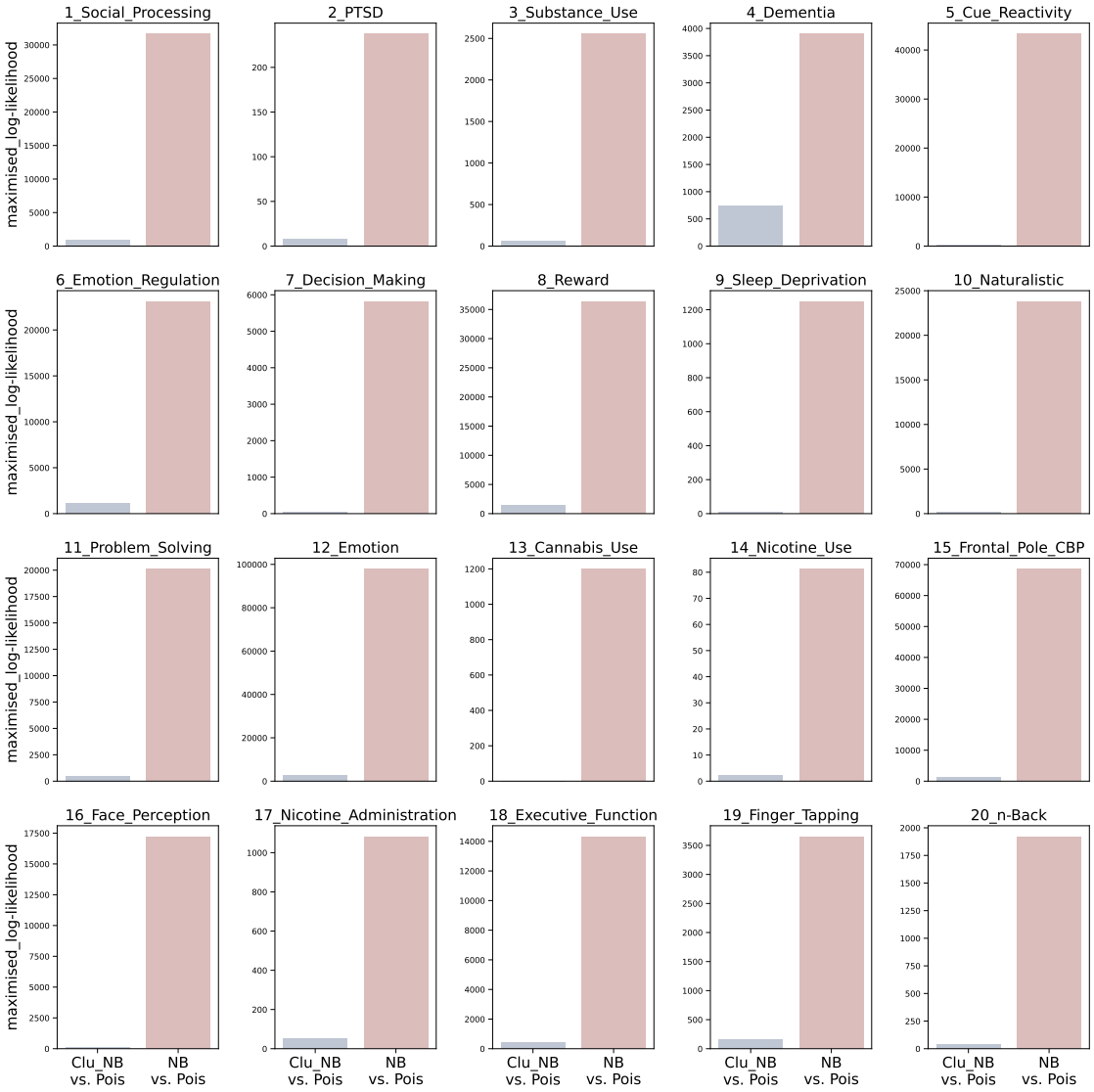


Figure A.6: Likelihood-based comparison of CBMR with Poisson, NB and clustered NB models (difference in maximised log-likelihood values, with Poisson model as the reference). We found that the maximised log-likelihood value of NB model is always the highest, while for some datasets, the difference of maximised log-likelihood values between clustered NB and Poisson model is negligible, therefore, the existence of excess variance has been justified among CBMA data.

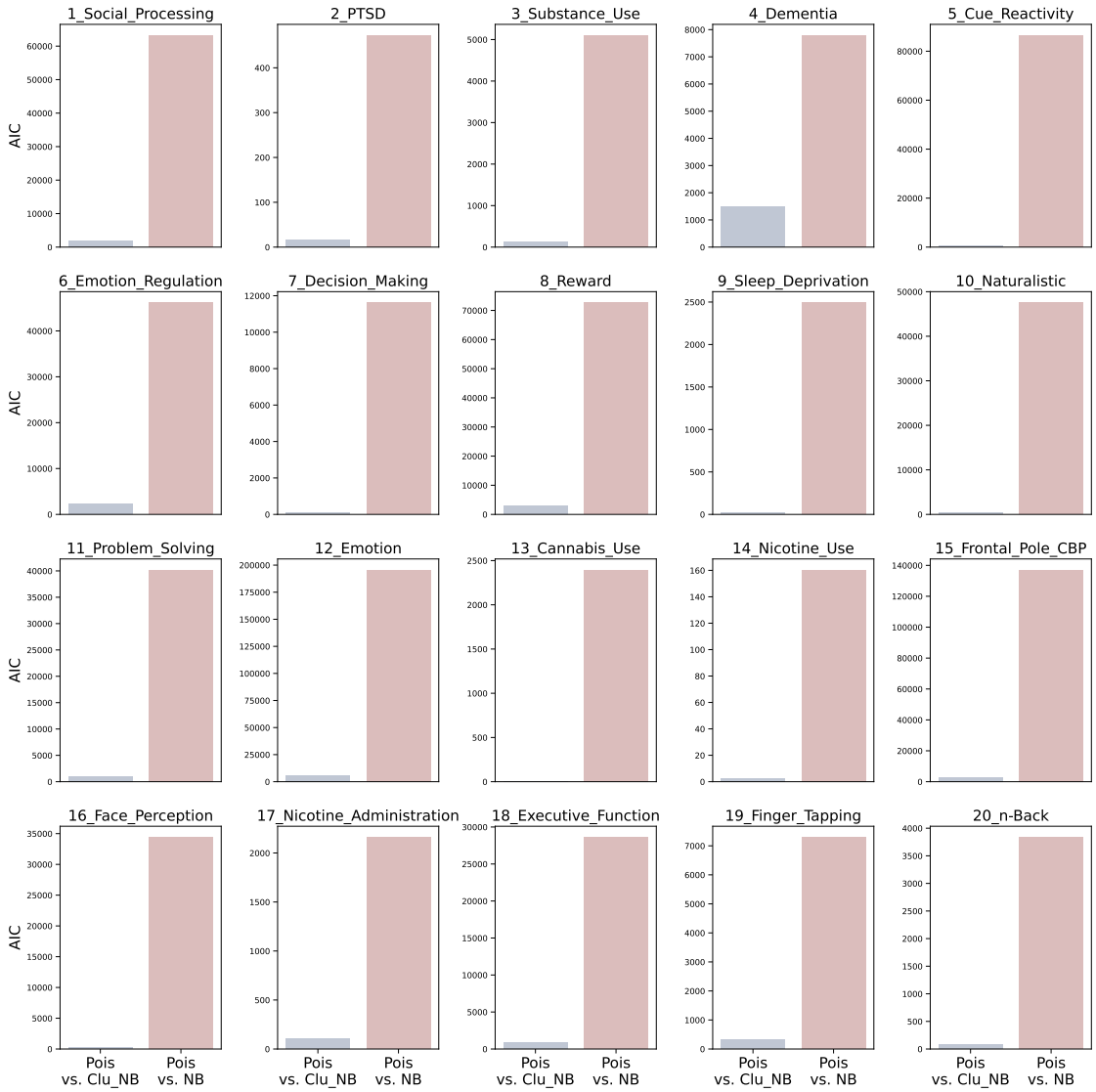
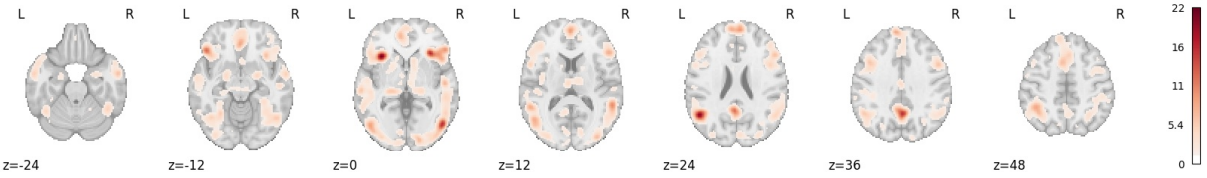
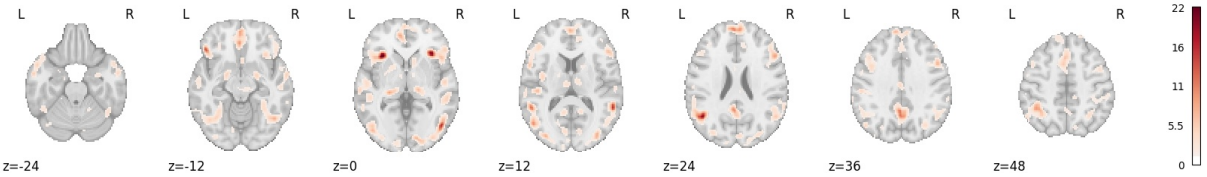


Figure A.7: Likelihood-based comparison of CBMR with Poisson, NB and clustered NB models (difference in AIC, with Poisson model as the reference). We found that the AIC of NB model is always the smallest, while for some datasets, the difference of AIC between clustered NB and Poisson model is negligible, therefore, the least information loss is observed in NB model.

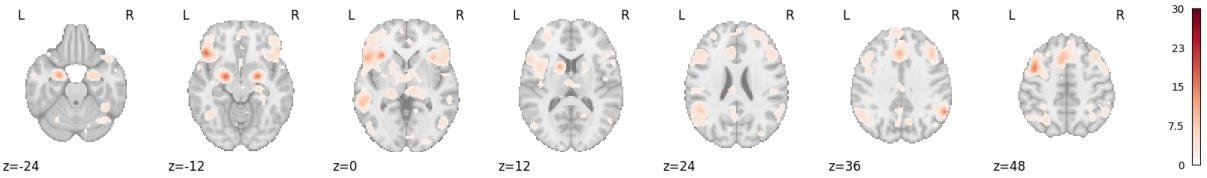


(a) $-log_{10}(p)$ map generated by ALE

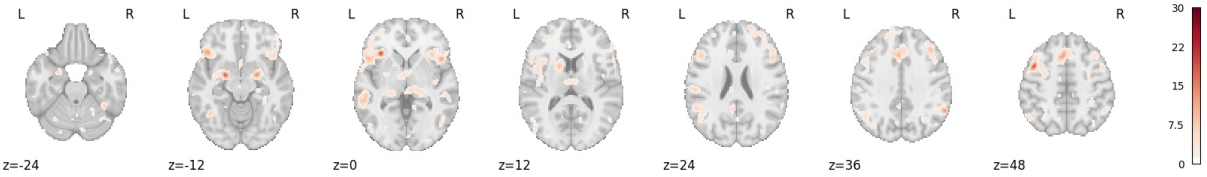


(b) $-log_{10}(p)$ map generated by CBMR (with NB model)

Figure A.8: Activation maps (for significant uncorrected p-values, $p \leq 5\%$, displayed as $-\log_{10} p$) generated by ALE (with FWHM=14) and CBMR (with NB model) for the Social Processing dataset (599 experiments, 4,934 foci). Axial slices are presented at $z = -24, -12, 0, 12, 24, 36, 48$.

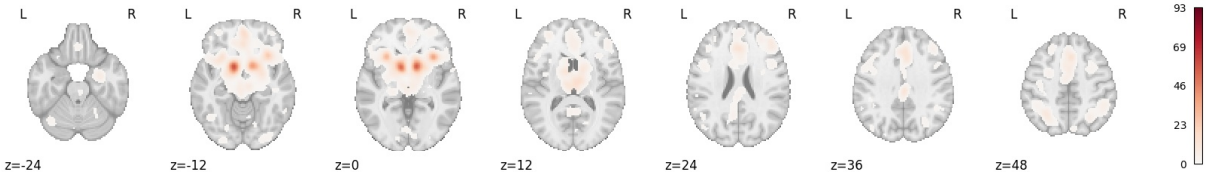


(a) $-log_{10}(p)$ map generated by ALE

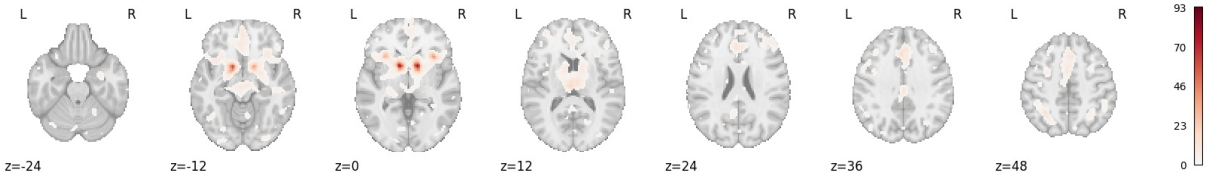


(b) $-log_{10}(p)$ map generated by CBMR (with NB model)

Figure A.9: Activation maps (for significant uncorrected p-values, $p \leq 5\%$, displayed as $-\log_{10} p$) generated by ALE (with FWHM=14) and CBMR (with NB model) for the Emotion Regulation dataset (338 experiments, 3,543 foci). Axial slices are presented at $z = -24, -12, 0, 12, 24, 36, 48$.

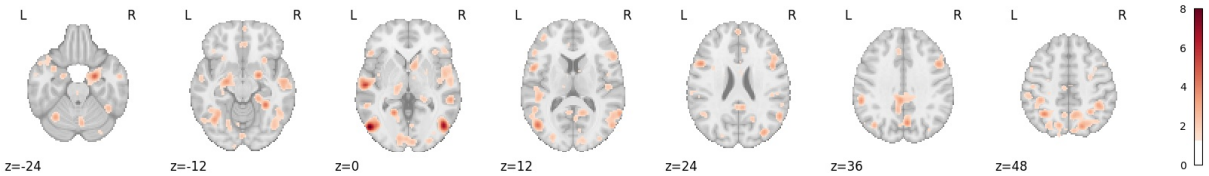


(a) $-log_{10}(p)$ map generated by ALE

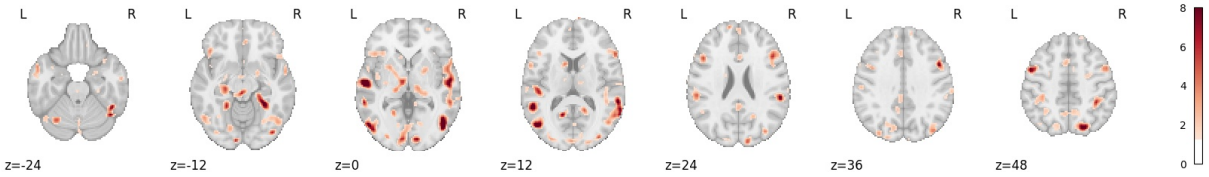


(b) $-log_{10}(p)$ map generated by CBMR (with NB model)

Figure A.10: Activation maps (for significant uncorrected p-values, $p \leq 5\%$, displayed as $-\log_{10} p$) generated by ALE (with FWHM=14) and CBMR (with NB model) for the Reward dataset (850 experiments, 6,791 foci). Axial slices are presented at $z = -24, -12, 0, 12, 24, 36, 48$.

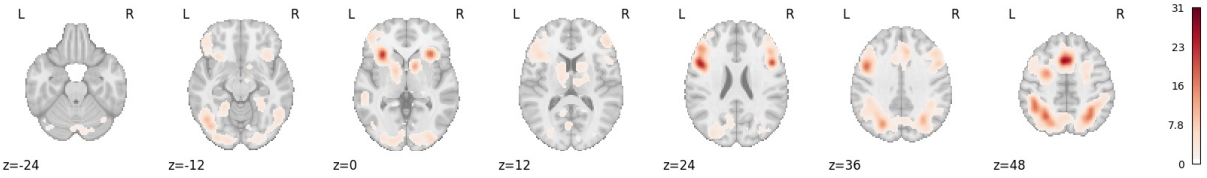


(a) $-log_{10}(p)$ map generated by ALE

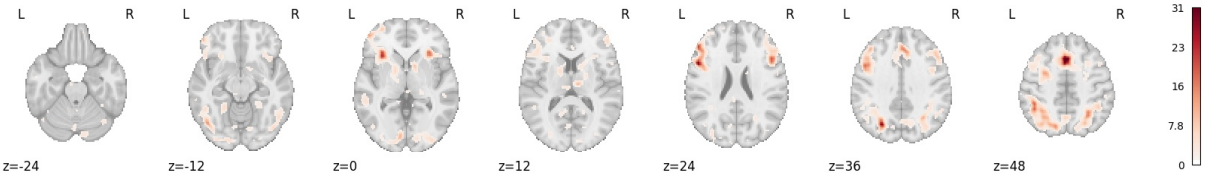


(b) $-log_{10}(p)$ map generated by CBMR (with NB model)

Figure A.11: Activation maps (for significant uncorrected p-values, $p \leq 5\%$, displayed as $-\log_{10} p$) generated by ALE (with FWHM=14) and CBMR (with NB model) for the Naturalistic dataset (122 experiments, 1,220 foci). Axial slices are presented at $z = -24, -12, 0, 12, 24, 36, 48$.

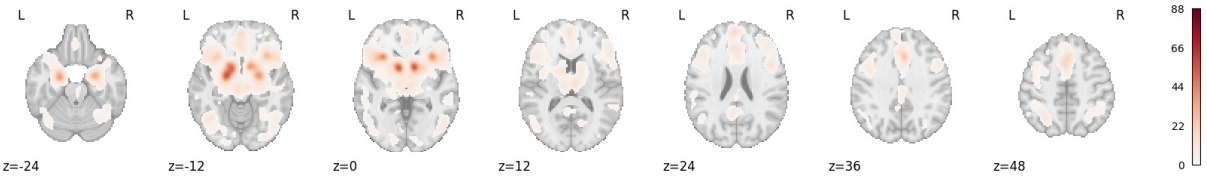


(a) $-log_{10}(p)$ map generated by ALE

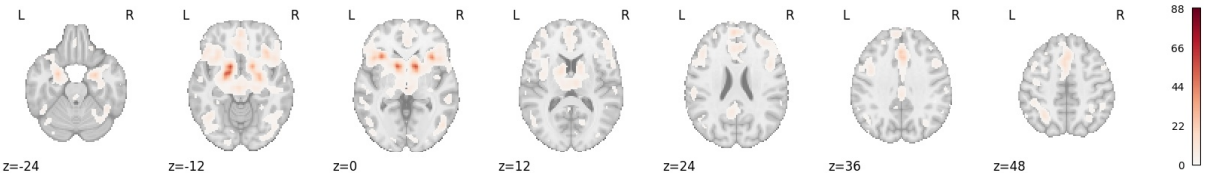


(b) $-log_{10}(p)$ map generated by CBMR (with NB model)

Figure A.12: Activation maps (for significant uncorrected p-values, $p \leq 5\%$, displayed as $-\log_{10} p$) generated by ALE (with FWHM=14) and CBMR (with NB model) for the Problem Solving dataset (282 experiments, 3,043 foci). Axial slices are presented at $z = -24, -12, 0, 12, 24, 36, 48$.

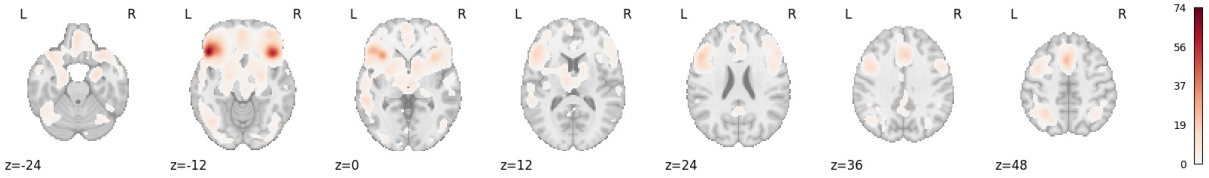


(a) $-log_{10}(p)$ map generated by ALE

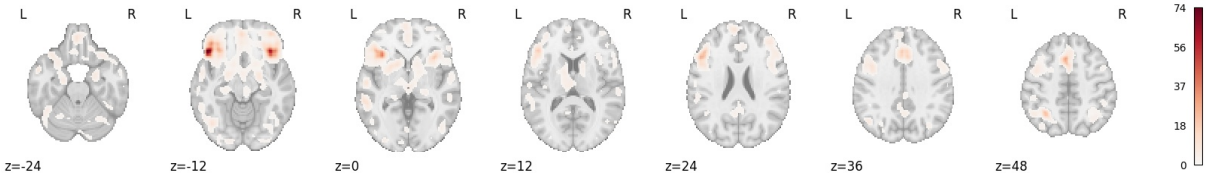


(b) $-log_{10}(p)$ map generated by CBMR (with NB model)

Figure A.13: Activation maps (for significant uncorrected p-values, $p \leq 5\%$, displayed as $-\log_{10} p$) generated by ALE (with FWHM=14) and CBMR (with NB model) for the Emotion dataset (1,738 experiments, 22,038 foci). Axial slices are presented at $z = -24, -12, 0, 12, 24, 36, 48$.

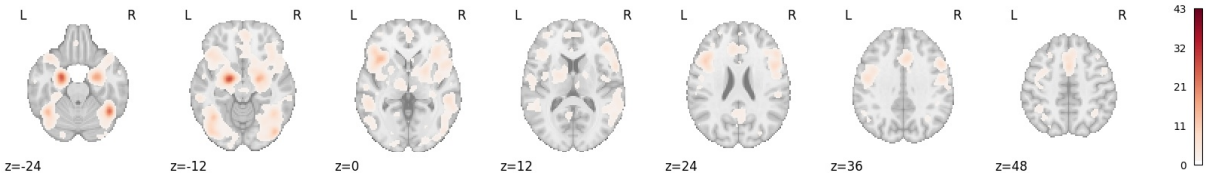


(a) $-log_{10}(p)$ map generated by ALE

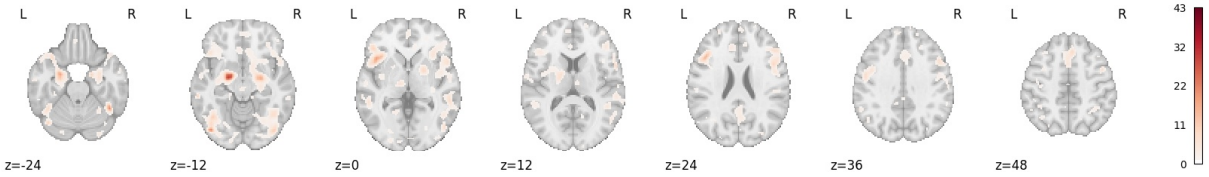


(b) $-log_{10}(p)$ map generated by CBMR (with NB model)

Figure A.14: Activation maps (for significant uncorrected p-values, $p \leq 5\%$, displayed as $-\log_{10} p$) generated by ALE (with FWHM=14) and CBMR (with NB model) for the Frontal Pole CBP dataset (795 experiments, 9,525 foci). Axial slices are presented at $z = -24, -12, 0, 12, 24, 36, 48$.

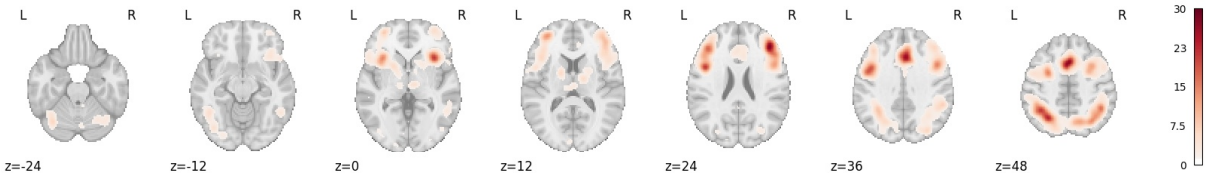


(a) $-log_{10}(p)$ map generated by ALE

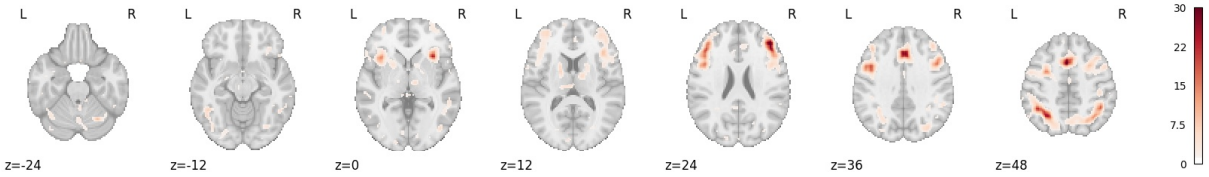


(b) $-log_{10}(p)$ map generated by CBMR (with NB model)

Figure A.15: Activation maps (for significant uncorrected p-values, $p \leq 5\%$, displayed as $-\log_{10} p$) generated by ALE (with FWHM=14) and CBMR (with NB model) for the Face Perception dataset (385 experiments, 2,920 foci). Axial slices are presented at $z = -24, -12, 0, 12, 24, 36, 48$.

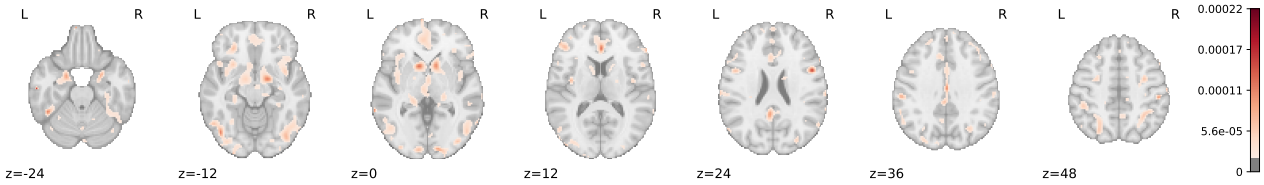


(a) $-log_{10}(p)$ map generated by ALE

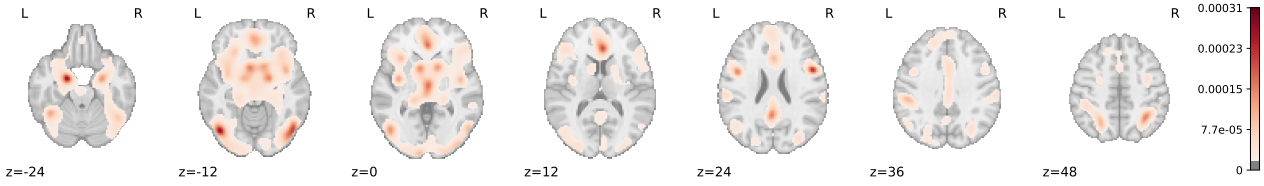


(b) $-log_{10}(p)$ map generated by CBMR (with NB model)

Figure A.16: Activation maps (for significant uncorrected p-values, $p \leq 5\%$, displayed as $-\log_{10} p$) generated by ALE (with FWHM=14) and CBMR (with NB model) for the Executive Function dataset (243 experiments, 2,629 foci). Axial slices are presented at $z = -24, -12, 0, 12, 24, 36, 48$.



(a) Estimated Intensity for the Cue Reactivity Dataset Using CBMR with a NB Model



(b) Estimated Intensity for the Cue Reactivity Dataset Using Bayesian LGCP approach

Figure A.17: Estimated intensity maps for the Cue Reactivity Dataset with CBMR (with NB model) and LGCP approaches

Appendix B

Appendix for CBMR: Meta Regression and Inference for Coordinate Based Meta Analysis Data Across Multiple Groups

B.1 Roughness penalty of spline bases

In the CBMR framework, we used spline parametrisation to construct the spatial model. However, in practice, we found that the curvature of the spline basis function often becomes excessively large, adversely impacting numerical stability and subsequent stages of CBMR inference. To address this issue, we introduced a penalty term for the B-spline basis to regularise the roughness of the spline basis functions. Following the definition of B-spline, we have

$$f_x(x) = \sum_{i=1}^{I} \alpha_i a_i(x),\ f_y(y) = \sum_{j=1}^{J} \beta_j b_j(y),\ f_z(z) = \sum_{k=1}^{K} \delta_k c_k(z) \tag{B.1}$$

as the B-spline curves on x, y, z direction, where $\alpha_i, \beta_j, \delta_k$ are parameters and a_i, b_j, c_k are known B-spline basis functions. As we construct the coefficient matrix C of $3-$dimensional B-spline bases by taking tensor product of coefficient matrices along x, y, z direction ($C = C_x \otimes C_y \otimes C_z$), the $3-$dimensional B-spline bases $d(x, y, z)$ is

given by the product of marginal B-spline basis in x, y, z direction,

$$\begin{aligned} f_{xyz}(x,y,z) &= \sum_{s=1}^{S} \gamma_s d_s(x,y,z) \\ &= \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{i=k}^{K} \gamma_{ijk} a_i(x) b_j(y) c_k(z) \end{aligned} \tag{B.2}$$

Suppose that each marginal B-spline basis (f_x, f_y, f_z) has an associated function (J_x, J_y, J_z) that measures wiggliness, an example of a penalty function is the cubic spline penalty,

$$\begin{aligned} J_x &= \lambda_x \int_x |f_x''(x)|^2 dx = \lambda_x \int_x [\sum_{i=1}^{I} \alpha_i a_i''(x)]^2 dx \\ &= \lambda_x \int_x \alpha^\top a''(x) a''(x)^\top \alpha dx = \lambda_x \cdot \alpha^\top (\int_x a''(x) a''(x)^\top dx) \alpha \end{aligned} \tag{B.3}$$

where $(I \times 1)-$vector of weights is $\alpha = [\alpha_1, \cdots, \alpha_I]^\top$ and $(I \times 1)-$vector of cubic B-spline bases is $a(x) = [a_1(x), \cdots, a_I(x)]^\top$, therefore, $a''(x) = [a_1''(x), \cdots, a_I''(x)]^\top$ represents the vector of second derivatives of the basis functions, and each element is first order since $a_i(x)$ is a cubic B-spline basis. For the penalty term, let $d_{ijk}(x,y,z)$ denotes $3-$dimension B-spline bases after taking tensor product of B-spline bases on each of x, y, z dimension, similarly,

$$\begin{aligned} J_{xyz}(f_{xyz}) &= \lambda_{xyz} \int_{x,y,z} |f_{xyz}''(x,y,z)|^2 dxdydz = \int_{x,y,z} [\sum_{ijk} \gamma_{ijk} d_{ijk}''(x,y,z)]^2 dxdydz \\ &= \lambda_{xyz} \int_{x,y,z} \gamma^\top d_{ijk}''(x,y,z) d_{ijk}''(x,y,z)^\top \gamma dxdydz \\ &= \lambda_{xyz} \gamma^\top (\int_{x,y,z} d_{ijk}''(x,y,z) d_{ijk}''(x,y,z)^\top dxdydz) \gamma \\ &= \lambda_{x,y,z} \gamma^\top S_{xyz} \gamma \end{aligned} \tag{B.4}$$

Here, γ is a $IJK \times 1$ vector, and $d''(x,y,z) = [d_{111}(x,y,z), d_{112}(x,y,z), \cdots, d_{IJK}(x,y,z)]^\top$ also has dimension $IJK \times 1$, if there are I, J, K B-spline bases on x, y, z dimension respectively. $S_{xyz} = \int_{x,y,z} d''(x,y,z) d''(x,y,z)^\top dxdydz$ is an $IJK \times IJK$ matrix, and its ijk^{th} element,

$$\int_{x,y,z} d''_{ijk}(x,y,z)d''_{i'j'k'}(x,y,z)dxdydz = \int_{x,y,z} a''_i(x)b''_j(y)c''_k(z)a''_{i'}(x)b''_{j'}(y)c''_{k'}(z)dxdydz$$
$$= (\int_x a''_i(x)a''_{i'}(x)dx)(\int_y b''_j(y)b''_{j'}(y)dy)(\int_z c''_k(z)c''_{k'}(z)dz) \tag{B.5}$$

If we define $I \times I$ matrix $A'' = \int_x a''(x)a''(x)^\top dx$ with ii'^{th} element equal to $\int_x a''_i(x)a''_{i'}(x)dx$. Similarly, we define $B'' = \int_y b''(y)b''(y)^\top dy$ and $C'' = \int_z c''(z)c''(z)^\top dz$, therefore, Equation B.5 becomes,

$$\int_{x,y,z} d''_{ijk}(x,y,z)d''_{i',j',k'}(x,y,z)dxdydz = A''_{ii'}B''_{jj'}C''_{kk'} \tag{B.6}$$

B.2 Log-likelihood function in CBMR regression

B.2.1 Model factorisation: Poisson model

In meta-regression, we employ model factorisation to replace the full $(M_g N)-$vector of foci counts for each group g with sufficient statistics, reducing the dimensionality to no more than M_g or N). Leveraging the model structure and desirable property of the Poisson process – where the sum of multiple Poisson random variables is also Poisson – we simplify the total log-likelihood function for a dataset containing G groups as follows:

$$
\begin{aligned}
l(\theta) = l(\beta_1, \cdots, \beta_G, \gamma) &= \sum_{g=1}^{G} l(\beta_g, \gamma) \\
&= \sum_{g=1}^{G} \sum_{i=1}^{M_g} \sum_{j=1}^{N} [Y_{ij} \log(\mu_{ij}) - \mu_{ij} - \log(Y_{ij}!)] \\
&= \sum_{g=1}^{G} [\sum_{i=1}^{M_g} \sum_{j=1}^{N} Y_{ij} \log(\mu_{ij}) - \sum_{i=1}^{M_g} \sum_{j=1}^{N} \mu_{ij} - 0] \\
&= \sum_{g=1}^{G} [(\sum_{i=1}^{M_g} \sum_{j=1}^{N} Y_{ij})[\log(\mu_{gj}^X) + \log(\mu_i^Z)] - \sum_{i=1}^{M_g} \sum_{j=1}^{N} \mu_{gj}^X \mu_i^Z] \\
&= \sum_{g=1}^{G} [\sum_{j=1}^{N} Y_{gj} \log(\mu_{gj}^X) + \sum_{i=1}^{M_g} Y_t \log(\mu_i^Z) - [\sum_{j=1}^{N} \mu_{gj}^X] \cdot [\sum_{i=1}^{M_g} \mu_i^Z]] \\
&= \sum_{g=1}^{G} \sum_{j=1}^{N} Y_{gj} \log(\mu_{gj}^X) + \sum_{i=1}^{M} Y_t \log(\mu_i^Z) - \sum_{g=1}^{G} [1^\top \mu_g^X][1^\top \mu_g^Z] \\
&= \sum_{g=1}^{G} Y_g^\top \log(\mu_g^X) + Y_t^\top \log(\mu^Z) - \sum_{g=1}^{G} [1^\top \mu_g^X][1^\top \mu_g^Z]
\end{aligned}
\tag{B.7}
$$

B.2.2 Model factorisation: Negative Binomial (NB) model

Similarly, we apply model factorisation with a Negative Binomial model to ensure that the dimensionality of each statistic are does not exceed N or M_g for group g. However, since the Negative Binomial is not technically a spatial point process, the sum of multiple NB variables does not follow a Negative Binomial distribution. Consequently, we employ a moment matching approach – matching the first two moments (mean and variance) – to approximate the distribution of the sum of multiple independent NB variables with another Negative Binomial distribution.

$$
\begin{aligned}
l(\theta) &= l(\beta_1, \cdots, \beta_G, \alpha'_1, \cdots, \alpha'_G, \gamma) = \sum_{g=1}^{G} l(\beta_g, \alpha'_g, \gamma) \\
&= \sum_{g=1}^{G}\sum_{i=1}^{M}\sum_{j=1}^{N} \left[\log\Gamma(Y_{ij}+\alpha_g^{-1}) - \log\Gamma(Y_{ij}+1) - \log\Gamma(\alpha_g^{-1}) + Y_{ij}\log(\alpha_g\mu_{ij}) - (Y_{ij}+\alpha_g^{-1})\log(\right. \\
&= \sum_{g=1}^{G}\sum_{i=1}^{M}\sum_{j=1}^{N} \left[\{\sum_{k=0}^{Y_{ij}-1}\log(k+\alpha_g^{-1})\} - \log\Gamma(Y_{ij}+1) + Y_{ij}\log(\alpha_g) + Y_{ij}\log(\mu_{ij}) - (Y_{ij}+\alpha_g^{-1})\,\mathrm{lo}\right. \\
&= \sum_{g=1}^{G}[\left(\sum_{i=1}^{M}\sum_{j=1}^{N} Y_{ij}\log(\alpha^{-1}) - \sum_{i=1}^{M}\sum_{j=1}^{N}\log(1)\right) + \left(\sum_{i=1}^{M}\sum_{j=1}^{N} Y_{ij}\right)\log(\alpha) \\
&+ \sum_{i=1}^{M}\sum_{j=1}^{N} Y_{ij}\left(\sum_{k=1}^{P} X_{jk}\beta_k + \sum_{s=1}^{R} Z_{is}\gamma_s\right) - \sum_{i=1}^{M}\sum_{j=1}^{N}(Y_{ij}+\alpha^{-1})\log(1+\alpha\mu_{ij})]
\end{aligned}
\tag{B.8}
$$

According to moment matching approach, for voxel j, we assume $Y_{gj} = \sum_{i=1}^{M_g} Y_{ij} \sim NB(r'_{gj}, p'_{gj})$, where

$$
\begin{aligned}
r'_{gj} &= \alpha_g^{-1}\frac{(\sum_{i=1}^{M_g}\mu_{ij})^2}{\sum_{i=1}^{M_g}\mu_{ij}^2} = \frac{\mu_{gj}^2}{\alpha_g\sum_{i=1}^{M_g}\mu_{ij}^2} \\
p'_{gj} &= \frac{\sum_{i=1}^{M_g}\mu_{ij}^2}{\alpha_g^{-1}\sum_{i=1}^{M_g}\mu_{ij} + \sum_{i=1}^{M_g}\mu_{ij}^2} = \frac{\sum_{i=1}^{M_g}\mu_{ij}^2}{\alpha_g^{-1}\mu_{gj} + \sum_{i=1}^{M_g}\mu_{ij}^2}
\end{aligned}
\tag{B.9}
$$

with corresponding excess variance for each group g,

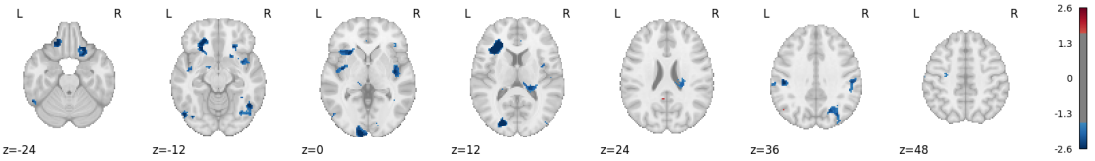
$$
\alpha'_g = \alpha_g\frac{\sum_{i=1}^{M_g}\mu_{ij}^2}{(\sum_{i=1}^{M_g}\mu_{ij})^2} = \alpha_g\frac{\sum_{i=1}^{M_g}\mu_{ij}^2}{\mu_{gj}^2} \tag{B.10}
$$

Therefore, the simplified NB log-likelihood function is given by,

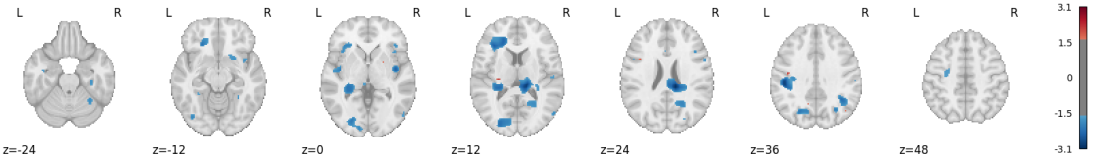
$$
\begin{aligned}
l(\theta) &= l(\beta_1, \cdots, \beta_G, \alpha'_1, \cdots, \alpha'_G, \gamma) = \sum_{g=1}^{G} l(\beta_g, \alpha'_g, \gamma) \\
&= \sum_{g=1}^{G}\sum_{j=1}^{N}\left[\log\Gamma(Y_{gj}+r'_{gj}) - \log\Gamma(Y_{gj}+1) - \log\Gamma(r'_{gj}) + r'_{gj}\log(1-p'_{gj}) + Y_{gj}\log(p'_{gj})\right]
\end{aligned}
\tag{B.11}
$$

B.3 Group-Wise Comparison of Activation Regions

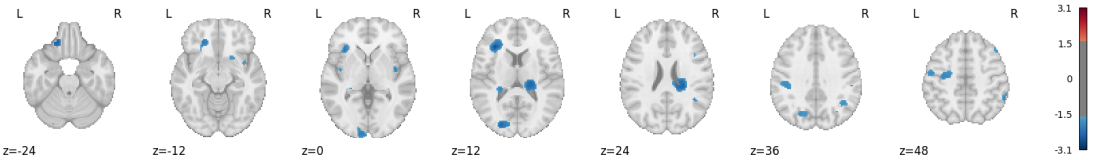
Below, we present group comparisons for two additional pairs within the Cue Reactivity dataset: Drug versus Reward, and Natural versus Reward.



(a) ALE subtraction analysis for comparison of activation regions between Drug-Neutral and Reward-Neutral groups

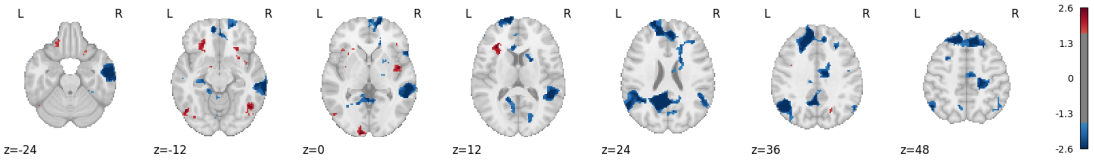


(b) Z-statistics map generated by parametric statistical tests (see Equation 4.10) for group comparison

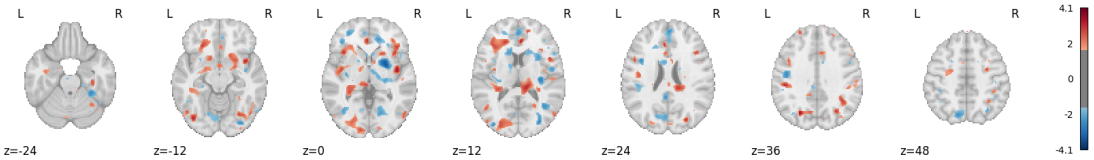


(c) Z-statistics map generated by parametric bootstrap method for group comparison

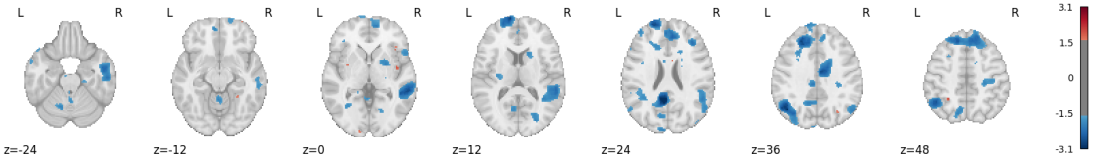
Figure B.1: Differences in activation regions between the **Drug** and **Reward** groups: results from ALE subtraction analysis, parametric statistical tests, and the parametric bootstrap method.



(a) ALE subtraction analysis for group comparison



(b) Z-statistics map generated by parametric statistical tests (see Equation 4.10) for group comparison



(c) Z-statistics map generated by parametric bootstrap method for group comparison

Figure B.2: Differences in activation regions between the **Natural** and **Reward** groups: results from ALE subtraction analysis, parametric statistical tests, and the parametric bootstrap method.

Appendix C

Appendix for Efficient Lesion Estimation Using a Spatial Poisson Process and a Scalable Approximate Model

C.1 Generic GLM structure

C.1.1 Poisson approximation for low-rate Bernoulli distributions

We assert that Poisson approximation is equivalent to the low-rate Bernoulli model, when the brain lesion counts are restricted to either 0 or 1 per voxel in each subject. Let Y_{ij} represent the brain lesion count at voxel j for subject i, where $Y_{ij} \in \{0, 1\}$. Let p_{ij} represent the probability of brain lesion at voxel j for subject i, let λ_{ij} represent the Poisson mean parameter at the same voxel, where $p_{ij}, \lambda_{ij} \in (0, 1)$, $\eta_{ij} = \log(\mu_{ij})$ is the log-transformed probability. The odds ratio (i.e., the ratio of brain lesion presence to absence) of the GLM with Bernoulli distribution and logit link function is given by

$$\begin{aligned} p_{ij} &= \frac{1}{1 + \exp(-\eta_{ij})} \\ \text{ratio}_{\text{Bernoulli}} &= \frac{P(Y_{ij} = 1)}{P(Y_{ij} = 0)} = \frac{p_{ij}}{1 - p_{ij}} = \frac{1/(1 + \exp(-\eta_{ij}))}{\exp(-\eta_{ij})/(1 + \exp(-\eta_{ij}))} = \exp(\eta_{ij}) \end{aligned} \tag{C.1}$$

Now, if we restrict the Poisson distribution to only $Y = 0$ and $Y = 1$ events in the GLM with log link, we have to normalize these probabilities so they sum to 1,

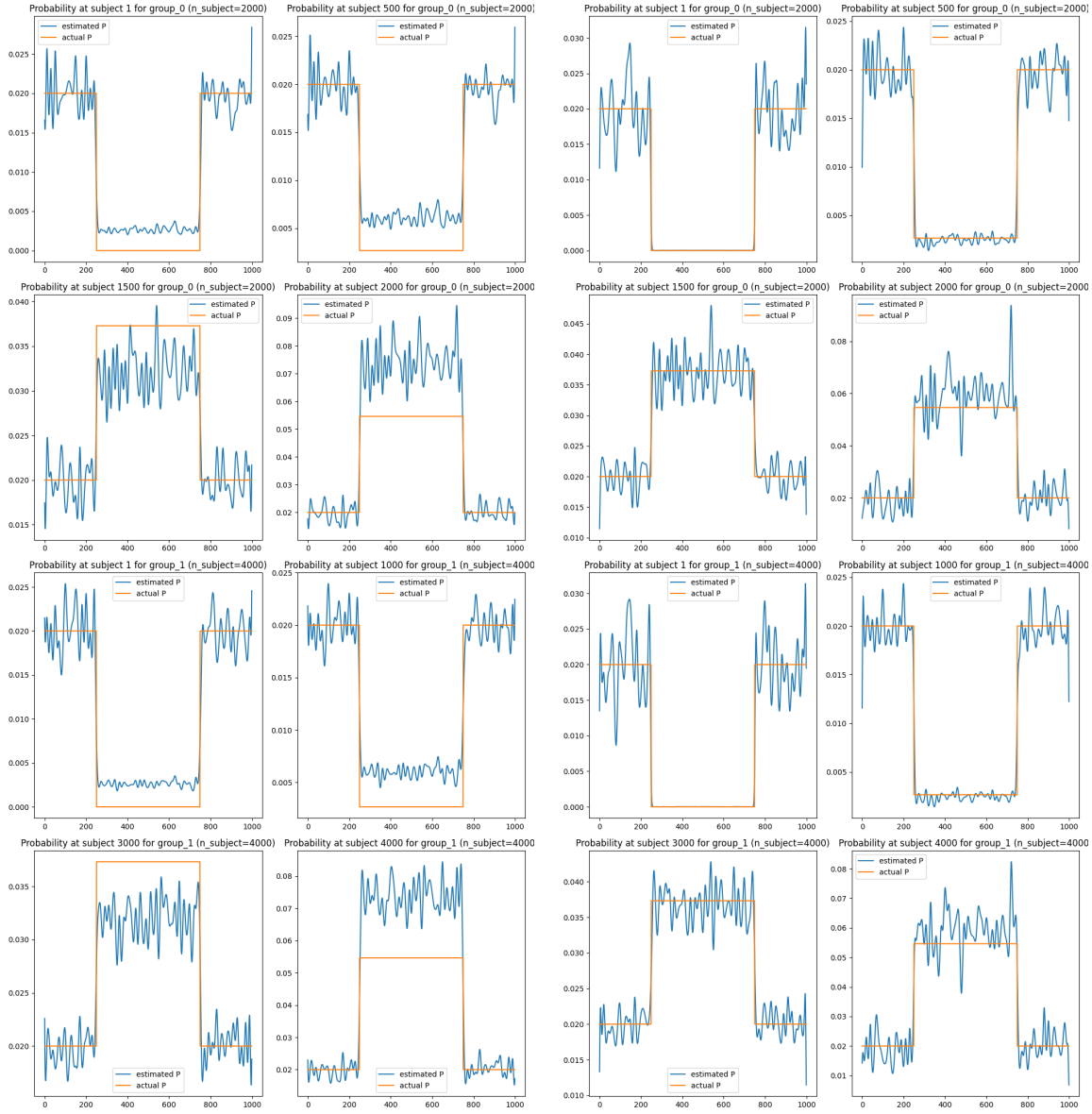
$$
\begin{aligned}
\lambda_{ij} &= \exp(\eta_{ij}) \\
P(Y_{ij}=0) &= \frac{P(Y_{ij}=0)}{P(Y_{ij}=0)+P(Y_{ij}=1)} = \frac{e^{-\lambda_{ij}}}{e^{-\lambda_{ij}}+\lambda_{ij}e^{-\lambda_{ij}}} \\
P(Y_{ij}=1) &= \frac{P(Y_{ij}=1)}{P(Y_{ij}=0)+P(Y_{ij}=1)} = \frac{\lambda_{ij}e^{-\lambda_{ij}}}{e^{-\lambda_{ij}}+\lambda_{ij}e^{-\lambda_{ij}}} \\
\Rightarrow \text{ratio}_{\text{Poisson}} &= \frac{P(Y_{ij}=1)}{P(Y_{ij}=0)} = \lambda_{ij} = \exp(\eta_{ij})
\end{aligned}
\tag{C.2}
$$

Therefore, we have justified that a GLM with log link and Poisson distribution restricted to $\{0,1\}$ has the same probability ratio as a GLM with logit link and Bernoulli distribution with parameter p_{ij} at voxel j for subject i, making the Poisson approximation equivalent to Bernoulli distribution.

C.1.2 Incorporation of quadratic and cubic terms in the covariate matrix

The most basic covariate matrix Z (of size $M \times R$) contains R columns, each corresponding to a covariate in the generic GLM framework defined by: $\log(\mu) = \log[\mathbb{E}(Y)] = X\beta = (Z \otimes B)\beta$. However, in practice, we observed that using a standard Z covariate matrix consisting only linear covariate terms may not provide sufficient flexibility to accurately capture the intensity estimation μ, primarily because $\mu = \exp[(Z \otimes B)\beta]$ implied an exponential relationship. Consequently, to improve model flexibility, we extended the covariate matrix Z by including quadratic and cubic polynomial terms of the covariates. We illustrate the comparison between lesion probability estimations derived from models with only linear terms versus those including up to cubic terms, in both the homogeneous settings (Figure C.1) and the Gaussian-bumped setting (Figure C.2) below, using four subject-specific probabilities selected based on quantiles.

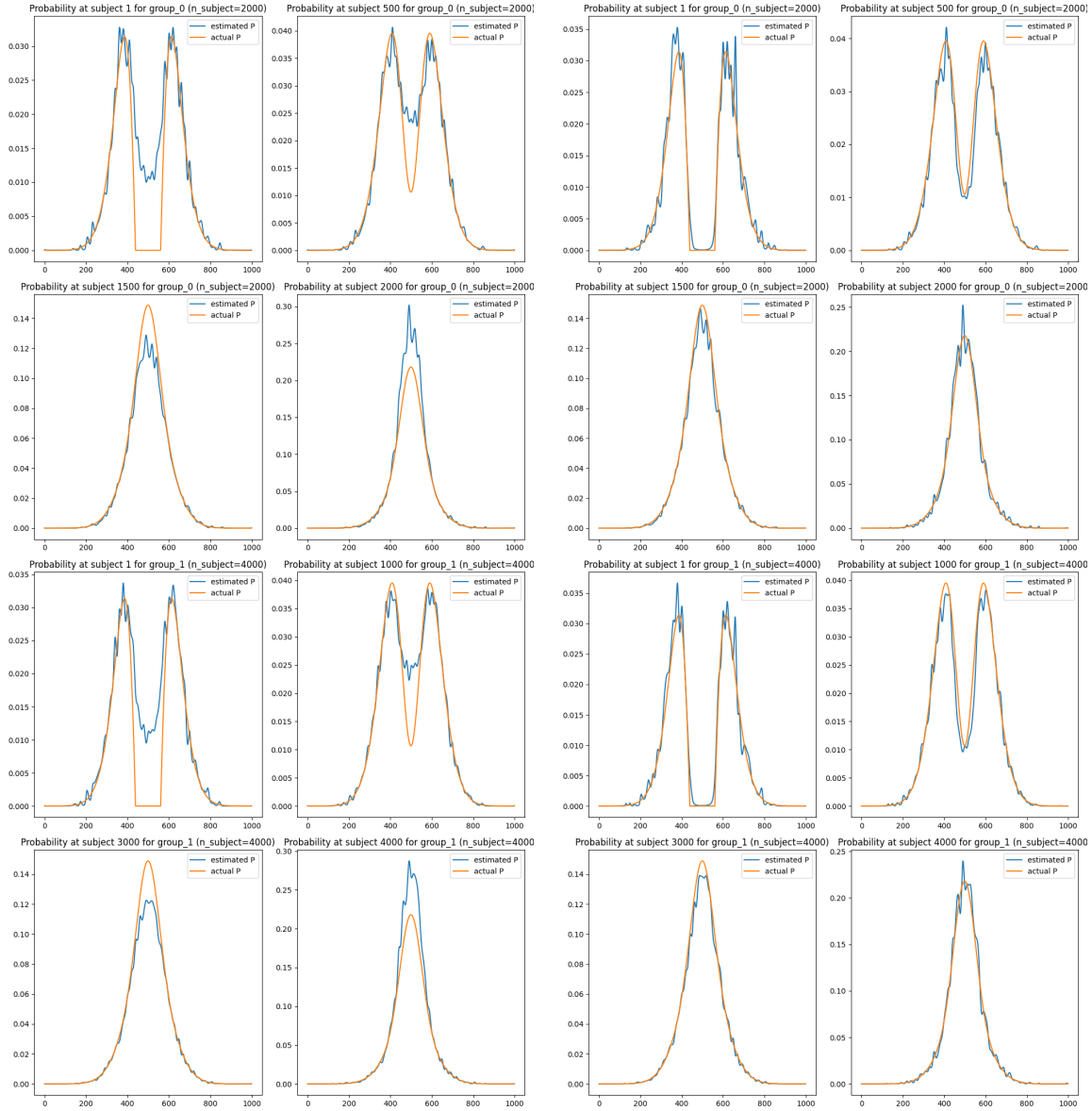
We observed overshoots and under-estimation when using a GLM with only linear term of covariates. In contrast, the GLM including up to cubic polynomial terms provides probability estimates that closely match the actual underlying probability function, thus justifying the inclusion of cubic polynomial terms in the GLM framework.



(a) GLM including only linear terms of covariates.

(b) GLM including cubic polynomial terms covariates.

Figure C.1: Estimation of brain lesion probabilities in a 1D setting with homogeneous background signals and homogeneous covariate-associated intensities for two groups containing 2000 and 4000 subjects, respectively. The Generalized Linear Model (GLM) was fitted using either linear (Figure C.1a) or cubic polynomial (Figure C.1b) terms of the covariates. The comparison between actual (orange line) and estimated (blue line) lesion probabilities is presented.



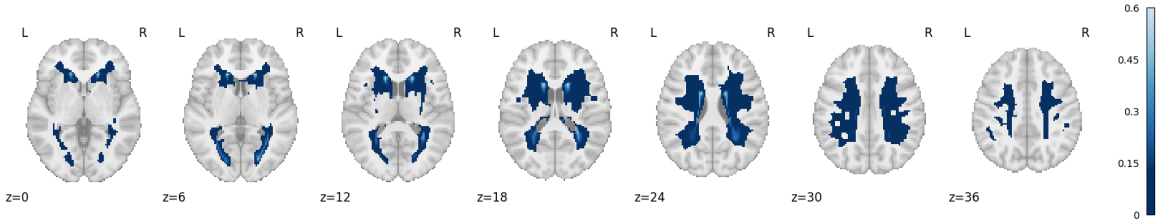
(a) GLM including only linear terms of covariates.

(b) GLM including cubic polynomial terms covariates.

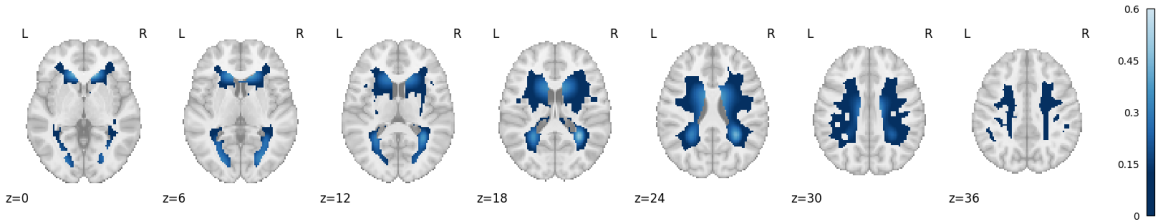
Figure C.2: Estimation of brain lesion probabilities in a 1D setting with Gaussian-bumped background signals and Gaussian-bumped covariate-associated intensities for two groups containing 2000 and 4000 subjects, respectively. The Generalized Linear Model (GLM) was fitted using either linear (Figure C.2a) or cubic polynomial (Figure C.2b) terms of the covariates. The comparison between actual (orange line) and estimated (blue line) lesion probabilities is presented.

C.1.3 Modelling age effects with linear or cubic terms in UK Biobank data

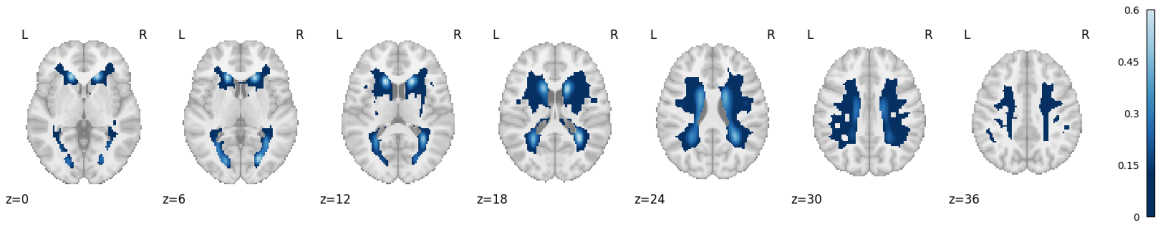
We have demonstrated that the difference between the linear fit and the cubic polynomial fit is minimal when modelling the log-transformed empirical lesion probabilities at the 100 voxels with the highest incidence among 13,677 UK Biobank subjects, with respect to age ($46 - 80$ years), as shown in Figure 5.3. Nonetheless, we present the fitted lesion probabilities using a cubic polynomial of the age covariate below in Figure C.3, for comparison with the linear fit shown in Figure 5.4



(a) Empirical lesion probability $\sqrt{p}$



(b) Estimated lesion probability $\sqrt{\hat{p}}$: cubic age covariate, 20mm spline spacing



(c) Estimated lesion probability $\sqrt{\hat{p}}$: cubic age covariate, 10mm spline spacing

Figure C.3: Comparison between square-root transformed empirical lesion probability ($\sqrt{p}$) and model-fitted lesion probabilities ($\sqrt{\hat{p}}$) across 13,677 UK Biobank participants. Lesion probabilities are fitted using spline spacing of either 20 or 10mm, incorporating a cubic polynomial for the age covariate. Other covariates, including sex, head size and CVR factors, are modelled linearly in all analyses.

C.2 Model factorisation

C.2.1 Scalable approximate multivariate modelling of binary image data

Suppose there are M subjects indexed by $i = 1, \cdots, M$, N voxels indexed by $j = 1, \cdots, N$, P basis elements indexed by $p = 1, \cdots, P$ and R regression coefficient indexed by $r = 1, \cdots, R$. Let $B(N \times P)$ denote the spatial matrix of cubic B-spline bases, $Z(M \times R)$ the covariate matrix, and $Y(MN \times 1)$ the vector of data for all subjects and voxels. The log-linear regression is,

$$\log \mathbb{E}(Y) = \eta = X\beta \tag{C.3}$$

where $X = Z \otimes B$ is a matrix with dimension $(MN \times PR)$, the regression coefficient β has dimension $PR \times 1$. We propose a scalable approximate model based on the following steps:

- By leveraging the computational efficiency of the separable model (Yu et al. [2024]) described in Section 5.2.2, we obtain a separable fit for brain lesion intensity at the voxel level, expressed as $\hat{\mu}_{ij} = \hat{\mu}_i^Z \hat{\mu}_j^B$,
- Given the extremely high dimensionality of the weight matrix $W^{(k)} = diag(\mu_{ij}^{(k)})$, where $i = 1, \cdots, M, j = 1, \cdots, N$, as well as the substantial computational burden of inverting $(X^\top W^{(k)} X)$ in each iteration k of IRLS, we fix the weights for all future iteration as $W = (\hat{\mu}_{ij})$. For iteration k, we compute

$$\begin{aligned} \mu^{(k)} &= \exp\left[(Z \otimes B)\beta^{(k)}\right] \\ \beta^{(k+1)} &= \beta^{(k)} + (X^\top W X)^{-1} X^\top (Y - \mu^{(k)}) \\ W &= W_Z \otimes W_B \end{aligned} \tag{C.4}$$

  where $W_Z = diag(\{\mu_i^Z\}_{i=1,\cdots,M})$, $W_B = diag(\{\mu_j^B\}_{j=1,\cdots,N})$. Hereafter, we refer $(X^\top W X)^{-1}$ as pre-conditioner and $X^\top (Y - \mu^{(k)})$ as gradient.

Using properties of Kronecker product, we simplify the computation of $(X^\top W X)^{-1}$

(dimension: $PR \times PR$) as follows:

$$\begin{aligned} W^{1/2}X &= (W_Z^{1/2} \otimes W_B^{1/2})(Z \otimes B) = (W_Z^{1/2} Z) \otimes (W_B^{1/2} B) \\ (X^\top W X)^{-1} &= [(W^{1/2}X)^\top W^{1/2} X]^{-1} = \{[(W_Z^{1/2} Z)^\top \otimes (W_B^{1/2} B)^\top][(W_Z^{1/2} Z) \otimes (W_B^{1/2} B)]\}^{-1} \\ &= \{[Z^\top W_Z Z] \otimes [B^\top W_B B]\}^{-1} \\ &= [Z^\top W_Z Z]^{-1} \otimes [B^\top W_B B]^{-1} \end{aligned} \tag{C.5}$$

where $[Z^\top W_Z Z]^{-1}$ has dimension $R \times R$, $[B^\top W_B B]^{-1}$ has dimension $P \times P$. Since we have computed $W = diag(\hat{\mu}_{ij})$, the preconditioner $(X^\top W X)^{-1}$ only needs to be computed once using Equation C.5 and remains fixed throughout all iterations.

Afterwards, we compute the mean of log linear response of GLM: $\eta = X\beta = (Z \otimes B)\beta$ and the mean of intensity estimation from the GLM for iteration (k), using the following approach:

$$\begin{aligned} \bar{\eta}^{(k)} &= \frac{1}{M}(\mathbf{1}_M^\top \otimes I_N)(Z \otimes B)\beta^{(k)} \\ &= \frac{1}{M}(\mathbf{1}_M^\top Z) \otimes B\beta^{(k)} \\ \bar{\mu}^{(k)} &= \frac{1}{M}(\mathbf{1}_M^\top \otimes I_N) \exp[(Z \otimes B)\beta^{(k)}] \end{aligned} \tag{C.6}$$

where $(\mathbf{1}_M^\top Z) \otimes B$ has dimension $N \times PR$, $\bar{\eta}^{(k)}$ and $\bar{\mu}^{(k)}$ has dimension $N \times 1$.

As for the gradient and residue term $X^\top(Y - \mu^{(k)})$ in IRLS, updating it at each iteration is unavoidable. Therefore, we employ a Taylor expansion of $\mu^{(k)}$ around the point $\exp(\mathbf{1}_M \otimes \bar{\eta}^{(k)})$ as follows

$$\begin{aligned} \exp[(Z \otimes B)\beta^{(k)}] &\approx \exp(\mathbf{1}_M \otimes \bar{\eta}^{(k)}) + [(Z \otimes B\beta^{(k)} - \mathbf{1}_M \otimes \bar{\eta}) \odot \exp(\mathbf{1}_M \otimes \bar{\eta}^{(k)})] \\ &= \mathbf{1}_M \otimes \exp(\bar{\eta}^{(k)}) + [Z \otimes B\beta^{(k)} - \frac{1}{M}\mathbf{1}_M \otimes (\mathbf{1}_M^\top Z) \otimes B\beta^{(k)}] \odot [\mathbf{1}_M \otimes \exp(\bar{\eta}^{(k)})] \\ &= \mathbf{1}_M \otimes \exp(\bar{\eta}^{(k)}) + [(Z - \frac{1}{M}\mathbf{1}_M \otimes (\mathbf{1}_M^\top Z)) \otimes B\beta^{(k)}] \odot [\mathbf{1}_M \otimes \exp(\bar{\eta}^{(k)})] \\ &= \mathbf{1}_M \otimes \exp(\bar{\eta}^{(k)}) + [(I_M - \frac{1}{M}\mathbf{1}_M\mathbf{1}_M^\top)Z \otimes B\beta^{(k)}] \odot [\mathbf{1}_M \otimes \exp(\bar{\eta}^{(k)})] \end{aligned} \tag{C.7}$$

At iteration k, the gradient term $X^\top(Y - \mu^{(k)})$ can be further reduced dimension as

follows,

$$\begin{aligned} X^\top(Y-\mu^{(k)}) &= X^\top Y - X^\top \mu^{(k)} \\ &= (Z^\top \otimes B^\top)Y - (Z^\top \otimes B^\top)\exp[(Z\otimes B)\beta^{(k)}] \\ &= vec(Z^\top \tilde{Y} B) - (Z^\top \otimes B^\top)\{\mathbf{1}_M \otimes \exp(\bar{\eta}^{(k)}) + [(I_M - \frac{1}{M}\mathbf{1}_M\mathbf{1}_M^\top)Z \otimes B\beta^{(k)}] \\ &\odot [\mathbf{1}_M \otimes \exp(\bar{\eta})^{(k)}]\} \ (*) \end{aligned} \tag{C.8}$$

- Trick 1: Let $\tilde{Y}(M\times N)$ represent the reshaped data vector $Y(MN\times 1)$, and use $vec(Z^\top\tilde{Y}B) = (Z^\top\otimes B^\top)Y$ to avoid Kronecker product computation,
- Trick 2: Define $\tilde{Z} = [I_M - \frac{1}{M}\mathbf{1}_M\mathbf{1}_M^\top]Z$ as the column-centred (demeaned) version of Z.

$$\begin{aligned} (*) &= vec(Z^\top\tilde{Y}B) - (Z^\top\mathbf{1}_M)\otimes(B^\top\exp(\bar{\eta})^{(k)}) - (Z^\top\otimes B^\top)[(\tilde{Z}\otimes B)\beta^{(k)}]\odot[\mathbf{1}_M\otimes\exp(\bar{\eta})^{(k)}] \\ &= vec(Z^\top\tilde{Y}B) - (Z^\top\mathbf{1}_M)\otimes(B^\top\exp(\bar{\eta})^{(k)}) - (Z^\top\otimes B^\top)[vec(\tilde{Z}\tilde{\beta}^{(k)}B^\top)\odot(\mathbf{1}_M\otimes\exp(\bar{\eta})^{(k)})] \\ &= vec(Z^\top\tilde{Y}B) - (Z^\top\mathbf{1}_M)\otimes(B^\top\exp(\bar{\eta})^{(k)}) - (Z^\top\otimes B^\top)vec[\tilde{Z}\tilde{\beta}^{(k)}B^\top diag(\exp(\bar{\eta})^{(k)})] \end{aligned} \tag{C.9}$$

- Trick 3: Let $\tilde{\beta}^{(k)}(R\times P)$ represent the reshaped parameter $\beta^{(k)}(PR\times 1)$, then $(\tilde{Z}\otimes B)\beta^{(k)} = vec(\tilde{Z}\tilde{\beta}^{(k)}B^\top)$,
- Trick 4: $vec[\tilde{Z}\tilde{\beta}^{(k)}B^\top diag(\exp(\bar{\eta}))] = vec(\tilde{Z}\tilde{\beta}^{(k)}B^\top)\odot(\mathbf{1_M}\otimes\exp(\bar{\eta}))$ because row-wise multiplication with a vector is equivalent to matrix multiplication of diagonal matrix (with the vector on the diagonal),
- Trick 5: Let $\tilde{B} = diag(\exp(\bar{\eta})^{(k)}B)$ has dimension $N\times N$,

$$\begin{aligned} (*) &= vec(Z^\top\tilde{Y}B) - (Z^\top\mathbf{1}_M)\otimes(B^\top\exp(\bar{\eta})^{(k)}) - (Z^\top\otimes B^\top)vec[\tilde{Z}\tilde{\beta}^{(k)}\tilde{B}^\top] \\ &= vec(Z^\top\tilde{Y}B) - (Z^\top\mathbf{1}_M)\otimes(B^\top\exp(\bar{\eta})^{(k)}) - (Z^\top\otimes B^\top)[\tilde{Z}\otimes\tilde{B}\beta^{(k)}] \\ &= vec(Z^\top\tilde{Y}B) - (Z^\top\mathbf{1}_M)\otimes(B^\top\exp(\bar{\eta})^{(k)}) - [(Z^\top\tilde{Z})\otimes(B^\top\tilde{B})]\beta^{(k)} \end{aligned} \tag{C.10}$$

- Trick 6: $(\tilde{Z}\otimes\tilde{B})\beta^{(k)} = vec[\tilde{Z}\tilde{\beta}^{(k)}\tilde{B}^\top]$, which converts back to Kronecker product for future simplification.

Here, $Z^\top \tilde{Z}$ has dimension $R \times R$, $B^\top \tilde{B}$ has dimension $P \times P$. The vectors $vec(Z^\top \tilde{Y} B)$, $(Z^\top \mathbf{1}_M) \otimes (B^\top \exp(\bar{\eta})^{(k)})$ and $[(Z^\top \tilde{Z}) \otimes (B^\top \tilde{B})]\beta^{(k)}$ all have dimensions $PR \times 1$. Thus, the updating equation $\beta^{(k+1)} = \beta^{(k)} + (X^\top W X)^{-1} X^\top (Y - \mu^{(k)})$ has been simplified to only involve terms without the prohibitively large dimension MN. Additionally, the dimensions P and R, corresponding to the number of spatial bases and covariates respectively, are typically moderate and manageable in size. Thus, this approximate model offers strong scalability, remaining computationally efficient even with an increase number of subjects.

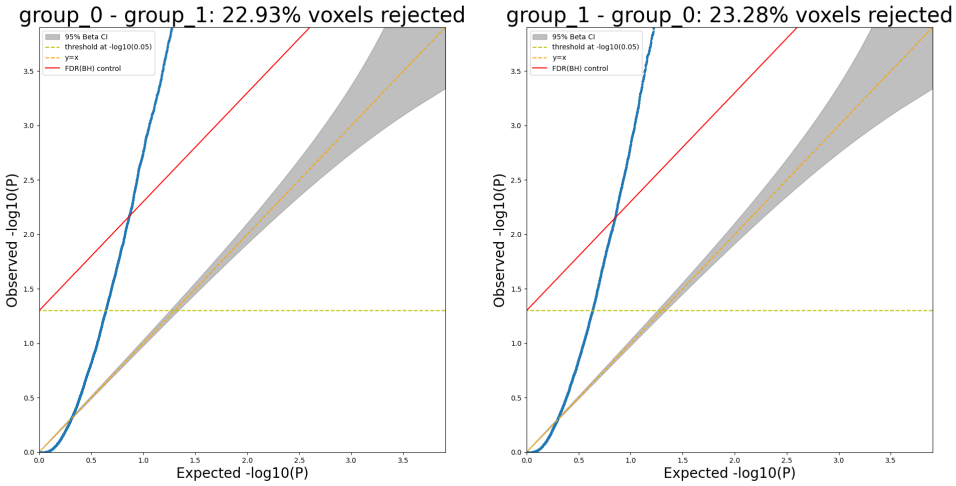
C.2.2 Assessing inference validity in the scalable approximate model factorisation

We further evaluate the validity of the scalable approximate model factorisation using PP-plots, where the exact gradient is combined with the approximate preconditioner at each iteration. Derivations from the diagonal line $y = x$ indicate discrepancies, while alignment with the diagonal suggests that the inference outcomes are valid.

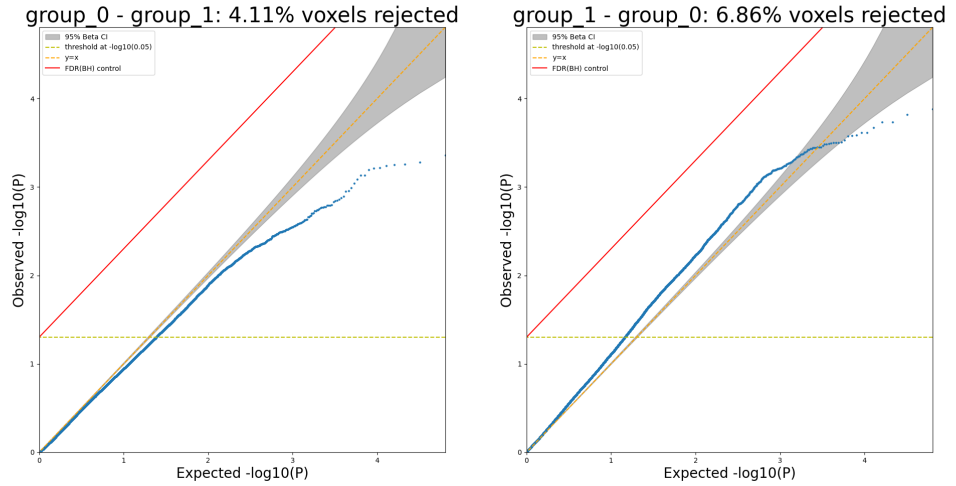
C.3 UK Biobank Application

C.3.1 Statistics maps for inference using UK Biobank data

Figure C.6 presents the corrected significance maps, showing FDR-corrected Z-scores at the 5% level

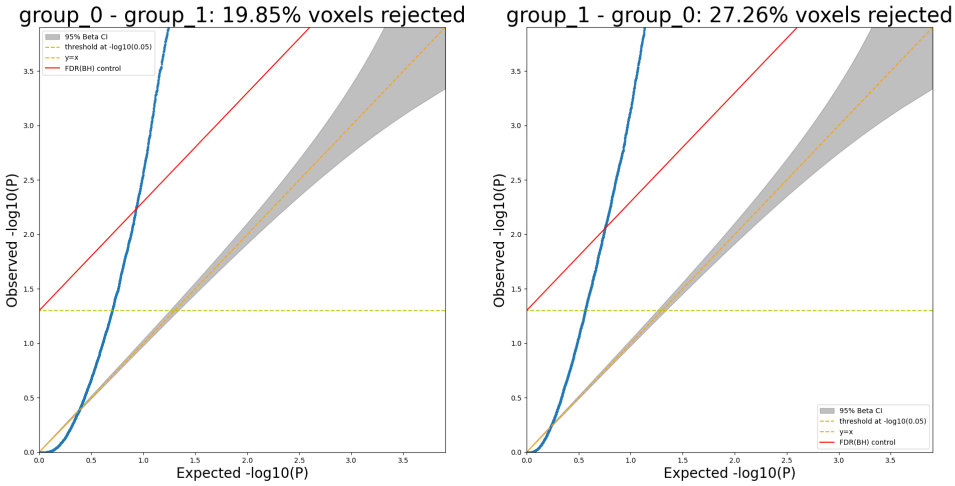


(a) PP-plot with standard error estimated using Fisher Information

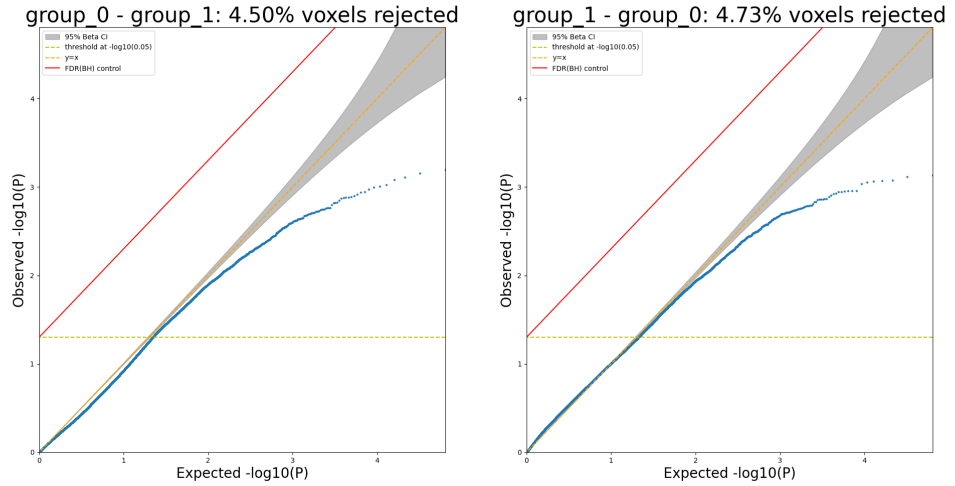


(b) PP-plot with standard error estimated using the sandwich estimator

Figure C.4: PP-plots (uncorrected p-values on a $-log_{10}$ scale) for spatially homogeneous signals, generated using the scalable approximate model with exact gradient and approximate preconditioner, are shown for hypothesis testing of group comparisons between two groups with different sample sizes (group_0 with 2000 subjects and group_1 with 4000 subjects). Standard errors are estimated using either Fisher Information or the sandwich estimator.

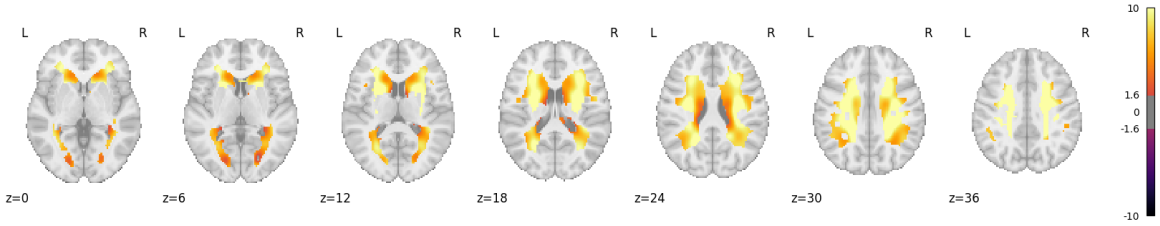


(a) PP-plot with standard error estimated using Fisher Information

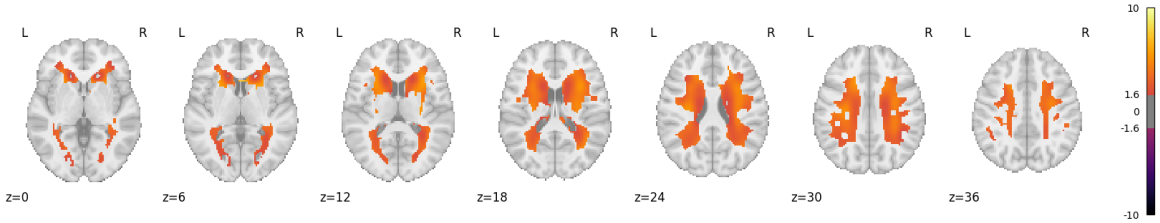


(b) PP-plot with standard error estimated using the sandwich estimator

Figure C.5: PP-plots (uncorrected p-values on a $-log_{10}$ scale) for Gaussian-bumped signals located at the centre of the 3D space, using the scalable approximate model with exact gradient and approximate preconditioner. These plots illustrate hypothesis testing for group comparisons between two groups with different sample sizes (group_0 with 2000 subjects and group_1 with 4000 subjects), using standard errors estimated with Fisher Information or the sandwich estimator.



(a) Z-score map showing the effect of the age risk factor



(b) Z-score map showing the effect of the CVR risk factor

Figure C.6: Significance maps (FDR-corrected Z-scores at the 5% level) for two risk factors, age and CVR, fitted via regression using the proposed scalable approximate model on 13,677 UK Biobank subjects. The significance maps assess voxel-wise significance of each risk factor's effect at a 5% significance level.

Bibliography

Hirotogu Akaike. Information theory and an extension of the maximum likelihood principle. In *Selected papers of hirotugu akaike*, pages 199–213. Springer, 1998.

Hirotugu Akaike. A new look at the statistical model identification. *IEEE transactions on automatic control*, 19(6):716–723, 1974.

James H Albert and Siddhartha Chib. Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association*, 88(422):669–679, 1993.

J.D. Angrist, G.W. Imbens, and D.B. Rubin. Identification of causal effects using instrumental variables. *Journal of the American Statistical Association*, 91(434): 444–455, 1996.

John Ashburner and K Friston. Multimodal image coregistration and partitioning—a unified framework. *Neuroimage*, 6(3):209–217, 1997.

John Ashburner and Karl J Friston. Nonlinear spatial normalization using basis functions. *Human brain mapping*, 7(4):254–266, 1999.

Sylvain Baillet. Magnetoencephalography for brain electrophysiology and imaging. *Nature neuroscience*, 20(3):327–339, 2017.

M. Baiocchi, J. Cheng, and D.S. Small. Instrumental variable methods for causal inference. *Statistics in Medicine*, 33(13):2297–2340, 2014.

Ole Barndorff-Nielsen and GF Yeo. Negative binomial processes. *Journal of Applied Probability*, 6(3):633–647, 1969.

Peter J Basser, James Mattiello, and Denis LeBihan. Mr diffusion tensor spectroscopy and imaging. *Biophysical journal*, 66(1):259–267, 1994.

Mark E Bastin, Jonathan D Clayden, Alison Pattie, Iona F Gerrish, Joanna M Wardlaw, and Ian J Deary. Diffusion tensor and magnetization transfer mri measurements of periventricular white matter hyperintensities in old age. *Neurobiology of aging*, 30 (1):125–136, 2009.

Dougla Bates. Fitting linear mixed-effects models using lme4. *arXiv preprint arXiv:1406.5823*, 2014.

Douglas Bates, Martin Maechler, Ben Bolker, Steven Walker, Rune Haubo Bojesen Christensen, Henrik Singmann, Bin Dai, Gabor Grothendieck, Peter Green, and Maintainer Ben Bolker. Package 'lme4'. *convergence*, 12(1):2, 2015.

Elizabeth Bates, Stephen M Wilson, Ayse Pinar Saygin, Frederic Dick, Martin I Sereno, Robert T Knight, and Nina F Dronkers. Voxel-based lesion–symptom mapping. *Nature neuroscience*, 6(5):448–450, 2003.

Betsy Jane Becker. Using results from replicated studies to estimate linear models. *Journal of Educational Statistics*, 17(4):341–362, 1992.

Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, 57(1):289–300, 1995.

Peter J Bickel and David A Freedman. Some asymptotic theory for the bootstrap. *The annals of statistics*, 9(6):1196–1217, 1981.

Alberto Bizzi, Valeria Blasi, Andrea Falini, Paolo Ferroli, Marcello Cadioli, Ugo Danesi, Domenico Aquino, Carlo Marras, Dario Caldiroli, and Giovanni Broggi. Presurgical functional mr imaging of language and motor functions: validation with intraoperative electrocortical mapping. *Radiology*, 248(2):579–589, 2008.

Michael Borenstein, Larry V Hedges, Julian PT Higgins, and Hannah R Rothstein. *Introduction to meta-analysis*. John wiley & sons, 2021.

F Dubois Bowman. Spatiotemporal models for region of interest analyses of functional neuroimaging data. *Journal of the American Statistical Association*, 102(478): 442–453, 2007.

Alexander Bowring, Camille Maumet, and Thomas E Nichols. Exploring the impact of analysis software on task fmri results. *Human brain mapping*, 40(11):3362–3384, 2019.

Mark A Brown and Richard C Semelka. *MRI: basic principles and applications*. John Wiley & Sons, 2011.

Katherine S Button, John PA Ioannidis, Claire Mokrysz, Brian A Nosek, Jonathan Flint, Emma SJ Robinson, and Marcus R Munafò. Power failure: why small sample size undermines the reliability of neuroscience. *Nature reviews neuroscience*, 14(5): 365–376, 2013.

Richard B Buxton. Dynamic models of bold contrast. *Neuroimage*, 62(2):953–961, 2012.

Clare Bycroft, Colin Freeman, Desislava Petkova, Gavin Band, Lloyd T Elliott, Kevin Sharp, Allan Motyer, Damjan Vukcevic, Olivier Delaneau, Jared O'Connell, et al. The uk biobank resource with deep phenotyping and genomic data. *Nature*, 562 (7726):203–209, 2018.

Richard H Byrd, Peihuang Lu, Jorge Nocedal, and Ciyou Zhu. A limited memory algorithm for bound constrained optimization. *SIAM Journal on scientific computing*, 16(5):1190–1208, 1995.

Roberto Cabeza and Lars Nyberg. Imaging cognition ii: An empirical review of 275 pet and fmri studies. *Journal of cognitive neuroscience*, 12(1):1–47, 2000.

Maria Eugenia Caligiuri, Paolo Perrotta, Antonio Augimeri, Federico Rocca, Aldo Quattrone, and Andrea Cherubini. Automatic detection of white matter hyperintensities in healthy aging and pathology using magnetic resonance imaging: a review. *Neuroinformatics*, 13:261–276, 2015.

Adrian Colin Cameron and Pravin K Trivedi. *Regression analysis of count data*. Number 53. Cambridge university press, 2013.

Bob Carpenter, Andrew Gelman, Matthew D Hoffman, et al. Stan: A probabilistic programming language. *Journal of Statistical Software*, 76(1), 2017.

Emmanuel Carrera and Giulio Tononi. Diaschisis: past, present, future. *Brain*, 137 (9):2408–2422, 2014.

George Casella and Roger L. Berger. *Statistical Inference*. Duxbury Press, 2nd edition, 2002. ISBN 9780534243128.

Betty Jo Casey, Tariq Cannonier, May I Conley, Alexandra O Cohen, Deanna M Barch, Mary M Heitzeg, Mary E Soules, Theresa Teslovich, Danielle V Dellarco, Hugh Garavan, et al. The adolescent brain cognitive development (abcd) study: imaging acquisition across 21 sites. *Developmental cognitive neuroscience*, 32:43–54, 2018.

Jan Cees De Groot, Frank-Erik De Leeuw, Matthijs Oudkerk, Jan Van Gijn, Albert Hofman, Jellemer Jolles, and Monique MB Breteler. Cerebral white matter lesions and cognitive function: the rotterdam scan study. *Annals of Neurology: Official Journal of the American Neurological Association and the Child Neurology Society*, 47(2):145–151, 2000.

Rong Chen, Argye E Hillis, Mikolaj Pawlak, and Edward H Herskovits. Voxelwise bayesian lesion-deficit analysis. *Neuroimage*, 40(4):1633–1642, 2008.

D Louis Collins, Peter Neelin, Terrence M Peters, and Alan C Evans. Automatic 3d intersubject registration of mr volumetric data in standardized talairach space. *Journal of computer assisted tomography*, 18(2):192–205, 1994.

Eric R Cosman Jr, John W Fisher III, and William M Wells III. Exact map activity detection in fmri using a glm with an ising spatial prior. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 703–710. Springer, 2004.

Robert W Cox. Afni: software for analysis and visualization of functional magnetic resonance neuroimages. *Computers and Biomedical research*, 29(3):162–173, 1996.

A. Datta, J. Zhang, and T. D. Johnson. Spatial bayesian variable selection and group lasso for lesion-symptom mapping. *NeuroImage*, 206:116317, 2020.

Anthony Christopher Davison and David Victor Hinkley. *Bootstrap methods and their application*. Number 1. Cambridge university press, 1997.

FE De Leeuw, Jan Cees de Groot, E Achten, Matthijs Oudkerk, LMP Ramos, R Heijboer, A Hofman, J Jolles, J Van Gijn, and MMB Breteler. Prevalence of cerebral white matter lesions in elderly people: a population based magnetic resonance imaging study. the rotterdam scan study. *Journal of Neurology, Neurosurgery & Psychiatry*, 70(1):9–14, 2001.

FE de Leeuw, JC de Groot, E Achten, M Oudkerk, LM Ramos, R Heijboer, A Hofman, J Jolles, J van Gijn, and MMB Breteler. Prevalence of cerebral white matter lesions in elderly people: a population based magnetic resonance imaging study. the rotterdam scan study. *Journal of Neurology, Neurosurgery Psychiatry*, 70(1):9–14, 2001.

Stefan Debener, Markus Ullsperger, Markus Siegel, and Andreas K Engel. Single-trial eeg–fmri reveals the dynamics of cognitive function. *Trends in cognitive sciences*, 10(12):558–563, 2006.

Stéphanie Debette and HS20660506 Markus. The clinical importance of white matter hyperintensities on brain magnetic resonance imaging: systematic review and meta-analysis. *Bmj*, 341, 2010.

Charles DeCarli, Evan Fletcher, Vincent Ramey, Danielle Harvey, and William J Jagust. Anatomical mapping of white matter hyperintensities (wmh) exploring the relationships between periventricular wmh, deep wmh, and total wmh burden. *Stroke*, 36(1):50–55, 2005a.

Charles DeCarli, Joseph Massaro, Danielle Harvey, John Hald, Mats Tullberg, Rhoda Au, Alexa Beiser, Ralph D'Agostino, and Philip A Wolf. Measures of brain morphology and infarction in the framingham heart study: establishing what is normal. *Neurobiology of aging*, 26(4):491–510, 2005b.

Rebecca DerSimonian and Nan Laird. Meta-analysis in clinical trials. *Controlled clinical trials*, 7(3):177–188, 1986.

Ilan Dinstein, David J Heeger, Lauren Lorenzi, Nancy J Minshew, Rafael Malach, and Marlene Behrmann. Unreliable evoked responses in autism. *Neuron*, 75(6):981–991, 2012.

Annette J Dobson and Adrian G Barnett. *An introduction to generalized linear models*. Chapman and Hall/CRC, 2018.

Jérôme Dockès, Russell A Poldrack, Romain Primet, Hande Gözükan, Tal Yarkoni, Fabian Suchanek, Bertrand Thirion, and Gaël Varoquaux. Neuroquery, comprehensive meta-analysis of human brain mapping. *elife*, 9:e53385, 2020.

Andrew T Drysdale, Logan Grosenick, Jonathan Downar, Katharine Dunlop, Farrokh Mansouri, Yue Meng, Robert N Fetcho, Benjamin Zebley, Desmond J Oathes,

Amit Etkin, et al. Resting-state connectivity biomarkers define neurophysiological subtypes of depression. *Nature medicine*, 23(1):28–38, 2017.

Bradley Efron. Bootstrap methods: another look at the jackknife. In *Breakthroughs in statistics: Methodology and distribution*, pages 569–593. Springer, 1992.

Bradley Efron. *Large-Scale Inference: Empirical Bayes Methods for Estimation, Testing, and Prediction*. Cambridge University Press, 2010. ISBN 9780521192491. doi: 10.1017/CBO9780511761362.

Bradley Efron and Robert J Tibshirani. *An introduction to the bootstrap*. Chapman and Hall/CRC, 1994.

Simon B Eickhoff, Angela R Laird, Christian Grefkes, Ling E Wang, Karl Zilles, and Peter T Fox. Coordinate-based activation likelihood estimation meta-analysis of neuroimaging data: A random-effects approach based on empirical estimates of spatial uncertainty. *Human brain mapping*, 30(9):2907–2926, 2009.

Simon B Eickhoff, Danilo Bzdok, Angela R Laird, Florian Kurth, and Peter T Fox. Activation likelihood estimation meta-analysis revisited. *Neuroimage*, 59(3):2349–2361, 2012.

Simon B Eickhoff, Thomas E Nichols, Angela R Laird, Felix Hoffstaedter, Katrin Amunts, Peter T Fox, Danilo Bzdok, and Claudia R Eickhoff. Behavior, sensitivity, and power of activation likelihood estimation characterized by massive empirical simulation. *Neuroimage*, 137:70–85, 2016.

Herbert B Eisenberg, Randolph RM Geoghagen, and John E Walsh. A general use of the poisson approximation for binomial events, with application to bacterial endocarditis data. *Biometrics*, pages 74–82, 1966.

Walid I Essayed, Fan Zhang, Prashin Unadkat, G Rees Cosgrove, Alexandra J Golby, and Lauren J O'Donnell. White matter tractography for neurosurgical planning: A topography-based review of the current state of the art. *NeuroImage: Clinical*, 15: 659–672, 2017.

Oscar Esteban, Christopher J Markiewicz, Ross W Blair, Craig A Moodie, A Ilkay Isik, Asier Erramuzpe, James D Kent, Mathias Goncalves, Elizabeth DuPre, Madeleine Snyder, et al. fmriprep: a robust preprocessing pipeline for functional mri. *Nature methods*, 16(1):111–116, 2019.

Franz Fazekas, John B Chawluk, Abass Alavi, Howard I Hurtig, and Robert A Zimmerman. Mr signal abnormalities at 1.5 t in alzheimer's dementia and normal aging. *American journal of roentgenology*, 149(2):351–356, 1987.

David A Feinberg and Essa Yacoub. The rapid development of high speed, resolution and precision in fmri. *Neuroimage*, 62(2):720–725, 2012.

Massimo Filippi, Maria A Rocca, Olga Ciccarelli, Nicola De Stefano, Nikos Evangelou, Ludwig Kappos, Alex Rovira, Jaume Sastre-Garriga, Mar Tintorè, Jette L Frederiksen, et al. Mri criteria for the diagnosis of multiple sclerosis: Magnims consensus guidelines. *The Lancet Neurology*, 15(3):292–303, 2016.

Stanley Finger. *Origins of neuroscience: a history of explorations into brain function*. Oxford University Press, 2001.

David Firth. Bias reduction of maximum likelihood estimates. *Biometrika*, 80(1): 27–38, 1993.

Bruce Fischl. Freesurfer. *Neuroimage*, 62(2):774–781, 2012.

Ronald Aylmer Fisher. Statistical methods for research workers. In *Breakthroughs in statistics: Methodology and distribution*, pages 66–70. Springer, 1970.

Chris Foulon, Leonardo Cerliani, Serge Kinkingnéhun, Richard Levy, Charlotte Rosso, Marika Urbanski, Emmanuelle Volle, and Michel Thiebaut de Schotten. Advanced lesion symptom mapping analyses and implementation as bcbtoolkit. *Gigascience*, 7(3):giy004, 2018.

John Fox. *Applied regression analysis and generalized linear models*. Sage publications, 2015.

Peter T Fox, Angela R Laird, Sarabeth P Fox, P Mickle Fox, Angela M Uecker, Michelle Crank, Sandra F Koenig, and Jack L Lancaster. Brainmap taxonomy of experimental design: description and evaluation. *Human brain mapping*, 25(1): 185–198, 2005.

David A Freedman. On the so-called "huber sandwich estimator" and "robust standard errors". *The American Statistician*, 60(4):299–302, 2006.

Jerome Friedman, Trevor Hastie, and Rob Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of statistical software*, 33 (1):1, 2010.

Karl J. Friston. *Classical and Bayesian Inference in Neuroimaging: Applications*. Statistical Parametric Mapping: The Analysis of Functional Brain Images. Academic Press, 2002.

Karl J Friston, Andrew P Holmes, Keith J Worsley, J-P Poline, Chris D Frith, and Richard SJ Frackowiak. Statistical parametric maps in functional imaging: a general linear approach. *Human brain mapping*, 2(4):189–210, 1994a.

Karl J Friston, Steven Williams, Robert Howard, Richard SJ Frackowiak, and Robert Turner. Movement-related effects in fmri time-series. *Magnetic resonance in medicine*, 35(3):346–355, 1996.

Karl J Friston, EORNA Zarahn, O Josephs, Richard NA Henson, and Anders M Dale. Stochastic designs in event-related fmri. *Neuroimage*, 10(5):607–619, 1999.

K.J. Friston, A.P. Holmes, K.J. Worsley, J.B. Poline, C.D. Frith, and R.S.J. Frackowiak. Assessing the significance of focal activations using their spatial extent. *Human Brain Mapping*, 1(3):210–220, 1994b. doi: 10.1002/hbm.460010306.

Chris D. Frith and Uta Frith. Social cognition in humans. *Current Biology*, 17(16): R724–R732, 2007. doi: 10.1016/j.cub.2007.05.068.

Anna Fry, Thomas J Littlejohns, Cathie Sudlow, Nicola Doherty, Ligia Adamska, Tim Sprosen, Rory Collins, and Naomi E Allen. Comparison of sociodemographic and health-related characteristics of uk biobank participants with those of the general population. *American journal of epidemiology*, 186(9):1026–1034, 2017.

Anne Gallagher, Christine Bulteau, David Cohen, and Jacques L Michaud. *Neurocognitive Development: Normative Development*. Elsevier, 2019.

Tian Ge, Nicole Müller-Lenke, Kerstin Bendfeldt, Thomas E Nichols, and Timothy D Johnson. Analysis of multiple sclerosis lesions via spatially varying coefficients. *The annals of applied statistics*, 8(2):1095, 2014.

Christopher R Genovese, Nicole A Lazar, and Thomas Nichols. Thresholding of statistical maps in functional neuroimaging using the false discovery rate. *Neuroimage*, 15(4):870–878, 2002.

Pedro Geoffroy and Govinda Weerakkody. A poisson-gamma model for two-stage cluster sampling data. *Journal of Statistical Computation and Simulation*, 68(2): 161–172, 2001.

Walter R Gilks, Sylvia Richardson, and David Spiegelhalter. *Markov chain Monte Carlo in practice*. CRC press, 1995.

Matthew F Glasser, Timothy S Coalson, Emma C Robinson, Carl D Hacker, John Harwell, Essa Yacoub, Kamil Ugurbil, Jesper Andersson, Christian F Beckmann, Mark Jenkinson, et al. A multi-modal parcellation of human cerebral cortex. *Nature*, 536(7615):171–178, 2016.

Gary H Glover. Deconvolution of impulse response in event-related bold fmri1. *Neuroimage*, 9(4):416–429, 1999.

Krzysztof J Gorgolewski, Gael Varoquaux, Gabriel Rivera, Yannick Schwarz, Satrajit S Ghosh, Camille Maumet, Vanessa V Sochat, Thomas E Nichols, Russell A Poldrack, Jean-Baptiste Poline, et al. Neurovault. org: a web-based repository for collecting and sharing unthresholded statistical maps of the human brain. *Frontiers in neuroinformatics*, 9:8, 2015.

Robert JB Goudie, Anne M Presanis, David Lunn, Daniela De Angelis, and Lorenz Wernisch. Joining and splitting models with markov melding. *Bayesian analysis*, 14 (1):81, 2019.

Alida A Gouw, Wiesje Maria van der Flier, Elisabeth CW van Straaten, Leonardo Pantoni, Antonio J Bastos-Leite, Domenico Inzitari, Timo Erkinjuntti, Lars-Olof Wahlund, C Ryberg, Reinhold Schmidt, et al. Reliability and sensitivity of visual scales versus volumetry for evaluating white matter hyperintensity progression. *Cerebrovascular diseases*, 25(3):247–253, 2008.

Alida A Gouw, Alexandra Seewann, Wiesje M Van Der Flier, Frederik Barkhof, Annemieke M Rozemuller, Philip Scheltens, and Jeroen JG Geurts. Heterogeneity of small vessel disease: a systematic review of mri and histopathology correlations. *Journal of Neurology, Neurosurgery & Psychiatry*, 82(2):126–135, 2011.

Ludovica Griffanti, Giovanna Zamboni, Aamira Khan, Linxin Li, Guendalina Bonifacio, Vaanathi Sundaresan, Ursula G Schulz, Wilhelm Kuker, Marco Battaglini, Peter M Rothwell, et al. Bianca (brain intensity abnormality classification algorithm): A new tool for automated segmentation of white matter hyperintensities. *Neuroimage*, 141:191–205, 2016.

Ludovica Griffanti, Mark Jenkinson, Sana Suri, Enikő Zsoldos, Abda Mahmood, Nicola Filippini, Claire E Sexton, Anya Topiwala, Charlotte Allan, Mika Kivimäki, et al. Classification and characterization of periventricular and deep white matter hyperintensities on mri: a study in older adults. *Neuroimage*, 170:174–181, 2018.

Faith M Gunning-Dixon and Naftali Raz. The cognitive correlates of white matter abnormalities in normal aging: a quantitative review. *Neuropsychology*, 14(2):224, 2000.

Peter Hall. *The bootstrap and Edgeworth expansion*. Springer Science & Business Media, 2013.

Larry V Hedges. Distribution theory for glass's estimator of effect size and related estimators. *journal of Educational Statistics*, 6(2):107–128, 1981.

RNA Henson, C Buechel, O Josephs, and KJ Friston. The slice-timing problem in event-related fmri. *NeuroImage*, 9:125, 1999.

M.A. Hernán and J.M. Robins. *Causal Inference: What If*. Chapman Hall/CRC, 2020.

Lauren D Hill-Bowen, Michael C Riedel, Ranjita Poudel, Taylor Salo, Jessica S Flannery, Julia A Camilleri, Simon B Eickhoff, Angela R Laird, and Matthew T Sutherland. The cue-reactivity paradigm: An ensemble of networks driving attention and cognition when viewing drug and natural reward-related stimuli. *Neuroscience & Biobehavioral Reviews*, 130:201–213, 2021.

Argye E. Hillis. Incorporating perfusion imaging into stroke management. *Stroke*, 47 (4):e77–e80, 2016.

Arthur E Hoerl and Robert W Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, 1970.

Joseph B Hopfinger, Christian Büchel, Andrew P Holmes, and Karl J Friston. A study of analysis parameters that influence the sensitivity of event-related fmri analyses. *Neuroimage*, 11(4):326–333, 2000.

Oliver D Howes and Shitij Kapur. The dopamine hypothesis of schizophrenia: version iii—the final common pathway. *Schizophrenia bulletin*, 35(3):549–562, 2009.

Peter J. Huber. The behavior of maximum likelihood estimates under nonstandard conditions. *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, 1:221–233, 1967.

John E Hunter and Frank L Schmidt. *Methods of meta-analysis: Correcting error and bias in research findings*. Sage, 2004.

José Eugenio Iglesias, Cheng-Yi Liu, Paul M Thompson, and Zhuowen Tu. Robust brain extraction across datasets and comparison with publicly available methods. *IEEE transactions on medical imaging*, 30(9):1617–1634, 2011.

John PA Ioannidis. Why most published research findings are false. *PLoS medicine*, 2 (8):e124, 2005.

Clifford R Jack, David S Knopman, William J Jagust, Leslie M Shaw, Paul S Aisen, Michael W Weiner, Ronald C Petersen, and John Q Trojanowski. Hypothetical model of dynamic biomarkers of the alzheimer's pathological cascade. *The Lancet Neurology*, 9(1):119–128, 2010.

Mark Jenkinson and Michael Chappell. *Introduction to neuroimaging analysis*. Oxford University Press, 2018.

Robert I Jennrich and PF Sampson. Newton-raphson and related algorithms for maximum likelihood variance component estimation. *Technometrics*, 18(1):11–17, 1976.

Peter Jezzard and Robert S Balaban. Correction for geometric distortion in echo planar images from b0 field variations. *Magnetic resonance in medicine*, 34(1):65–73, 1995.

Oliver Josephs, Robert Turner, and Karl Friston. Event-related f mri. *Human brain mapping*, 5(4):243–248, 1997.

Eric R Kandel. The molecular biology of memory storage: a dialogue between genes and synapses. *Science*, 294(5544):1030–1038, 2001.

Jian Kang, Timothy D Johnson, Thomas E Nichols, and Tor D Wager. Meta analysis of functional neuroimaging data via bayesian spatial point processes. *Journal of the American Statistical Association*, 106(493):124–134, 2011.

Jian Kang, Thomas E Nichols, Tor D Wager, and Timothy D Johnson. A bayesian hierarchical spatial point process model for multi-type neuroimaging meta-analysis. *The annals of applied statistics*, 8(3):1800, 2014.

James Kent, Nicolas Lee, Julio Peraza, Taylor Salo, Katherine Bottenhorn, Jerome Dockès, Ross Blair, Kendra Oudyk, Yifan Yu, Thomas Nichols, et al. 141. neurosynth compose: A free an open platform for precise large-scale neuroimaging meta-analysis. *Biological Psychiatry*, 95(10):S156–S157, 2024.

H. Kim, S. R. Das, W. Zhang, and B. B. Avants. Bayesian spatial modeling of lesion-symptom mapping data. *NeuroImage*, 233:117919, 2021.

Petya Kindalova, Ioannis Kosmidis, and Thomas E Nichols. Voxel-wise and spatial modelling of binary lesion masks: Comparison of methods with a realistic simulation framework. *NeuroImage*, 236:118090, 2021.

Jens Kleesiek, Gregor Urban, Arnaud Hubert, Dominik Schwarz, Klaus H Maier-Hein, Martin Bendszus, and Andreas Biller. Deep mri brain extraction: A 3d convolutional neural network for skull stripping. *NeuroImage*, 129:460–469, 2016.

Ioannis Kosmidis and David Firth. Jeffreys-prior penalty, finiteness and shrinkage in binomial-response generalized linear models. *Biometrika*, 108(1):71–82, 2021.

H.C. Kraemer, G.T. Wilson, C.G. Fairburn, and W.S. Agras. Mediators and moderators of treatment effects in randomized clinical trials. *Archives of General Psychiatry*, 54(10):877–883, 1997.

Sofia Ira Ktena, Sarah Parisot, Enzo Ferrante, Martin Rajchl, Matthew Lee, Ben Glocker, and Daniel Rueckert. Metric learning with spectral graph convolutions on brain connectivity networks. *NeuroImage*, 169:431–442, 2018.

Hugo J Kuijf, J Matthijs Biesbroek, Jeroen De Bresser, Rutger Heinen, Simon Andermatt, Mariana Bento, Matt Berseth, Mikhail Belyaev, M Jorge Cardoso, Adria Casamitjana, et al. Standardized assessment of automatic segmentation of white matter hyperintensities and results of the wmh segmentation challenge. *IEEE transactions on medical imaging*, 38(11):2556–2568, 2019.

Angela R Laird, Jack J Lancaster, and Peter T Fox. Brainmap. *Neuroinformatics*, 3 (1):65–77, 2005.

Leonie Lampe, Rui Zhang, Frauke Beyer, Sebastian Huhn, Shahrzad Kharabian Masouleh, Sven Preusser, Pierre-Louis Bazin, Matthias L Schroeter, Arno Villringer, and A Veronica Witte. Visceral obesity relates to deep white matter hyperintensities via inflammation. *Annals of neurology*, 85(2):194–203, 2019.

Jerald F Lawless. Negative binomial and mixed poisson regression. *The Canadian Journal of Statistics/La Revue Canadienne de Statistique*, pages 209–225, 1987.

Nicole Lazar. *The statistical analysis of functional MRI data*. Springer Science & Business Media, 2008.

Kung-Yee Liang and Scott L. Zeger. Longitudinal data analysis using generalized linear models. *Biometrika*, 73(1):13–22, 1986.

Geert Litjens, Thijs Kooi, Babak Ehteshami Bejnordi, Arnaud Arindra Adiyoso Setio, Francesco Ciompi, Mohsen Ghafoorian, Jeroen Awm Van Der Laak, Bram Van Ginneken, and Clara I Sánchez. A survey on deep learning in medical image analysis. *Medical image analysis*, 42:60–88, 2017.

Chenghao Liu, Zhizheng Zhuo, Liying Qu, Ying Jin, Tiantian Hua, Jun Xu, Guirong Tan, Yuna Li, Yunyun Duan, Tingting Wang, et al. Deepwmh: A deep learning tool for accurate white matter hyperintensity segmentation without requiring manual annotations for training. *Science Bulletin*, 69(7):872–875, 2024.

Dong C Liu and Jorge Nocedal. On the limited memory bfgs method for large scale optimization. *Mathematical programming*, 45(1):503–528, 1989.

Rosario Lobo, Katherine Bottenhorn, Afra Toma, Megan Hare, Donisha Smith, Alexandra Moor, Isis Cowan, Javier Valdes, Jessica Bartley, Taylor Salo, Emily Boeving, Brianna Pankey, Michael Riedel, Matthew Sutherland, Erica Musser, and Angela Laird. Neural systems underlying rdoc social constructs: An activation likelihood estimation meta-analysis. 2021.

Nikos K Logothetis. What we can do and what we cannot do with fmri. *Nature*, 453 (7197):869–878, 2008.

Nicholas T Longford. A fast scoring algorithm for maximum likelihood estimation in unbalanced mixed models with nested random effects. *Biometrika*, 74(4):817–827, 1987.

WT Longstreth, TA Manolio, A Arnold, GL Burke, N Bryan, C Jungreis, PL Enright, D O'Leary, and LP Fried. Clinical correlates of white matter findings on cranial magnetic resonance imaging of 3301 elderly people: the cardiovascular health study. *Stroke*, 27(8):1274–1282, 1996.

Steven J Luck. *An introduction to the event-related potential technique*. MIT press, 2014.

Alexander Selvikvåg Lundervold and Arvid Lundervold. An overview of deep learning in medical imaging focusing on mri. *Zeitschrift fuer medizinische Physik*, 29(2): 102–127, 2019.

David J Madden, Julia Spaniol, Matthew C Costello, Barbara Bucur, Leonard E White, Roberto Cabeza, Simon W Davis, Nancy A Dennis, James M Provenzale, and Scott A Huettel. Cerebral white matter integrity mediates adult age differences in cognitive performance. *Journal of cognitive neuroscience*, 21(2):289–302, 2008.

Pauline Maillard, Evan Fletcher, Samuel N Lockhart, Alexandra E Roach, Bruce Reed, Dan Mungas, Charles DeCarli, and Owen T Carmichael. White matter hyperintensities and their penumbra lie along a continuum of injury in the aging brain. *Stroke*, 45(6):1721–1726, 2014.

Scott Marek, Brenden Tervo-Clemmens, Finnegan J Calabro, David F Montez, Benjamin P Kay, Alexander S Hatoum, Meghan Rose Donohue, William Foran, Ryland L Miller, Timothy J Hendrickson, et al. Reproducible brain-wide association studies require thousands of individuals. *Nature*, 603(7902):654–660, 2022.

Camille Maumet, Tibor Auer, Alexander Bowring, Gang Chen, Samir Das, Guillaume Flandin, Satrajit Ghosh, Tristan Glatard, Krzysztof J Gorgolewski, Karl G Helmer, et al. Sharing brain mapping statistical results with the neuroimaging data model. *Scientific data*, 3(1):1–15, 2016.

Peter McCullagh. *Generalized linear models*. Routledge, 2019.

Geoffrey J McLachlan and Thriyambakam Krishnan. *The EM algorithm and extensions*. John Wiley & Sons, 2008.

Donald W McRobbie, Elizabeth A Moore, Martin J Graves, and Martin R Prince. *MRI from Picture to Proton*. Cambridge university press, 2017.

Anna Menacher, Thomas E Nichols, Chris Holmes, and Habib Ganjgahi. Bayesian lesion estimation with a structured spike-and-slab prior. *Journal of the American Statistical Association*, 119(545):66–80, 2024.

Michal Mikl, Radek Mareček, Petr Hluštík, Martina Pavlicová, Aleš Drastich, Pavel Chlebus, Milan Brázdil, and Petr Krupa. Effects of spatial smoothing on fmri group inferences. *Magnetic resonance imaging*, 26(4):490–503, 2008.

Karla L Miller, Fidel Alfaro-Almagro, Neal K Bangerter, David L Thomas, Essa Yacoub, Junqian Xu, Andreas J Bartsch, Saad Jbabdi, Stamatios N Sotiropoulos, Jesper LR Andersson, et al. Multimodal population brain imaging in the uk biobank prospective epidemiological study. *Nature neuroscience*, 19(11):1523–1536, 2016.

Silvia Montagna, Tor Wager, Lisa Feldman Barrett, Timothy D Johnson, and Thomas E Nichols. Spatial bayesian latent factor regression modeling of coordinate-based meta-analysis data. *Biometrics*, 74(1):342–353, 2018.

Douglas C Montgomery, Elizabeth A Peck, and G Geoffrey Vining. *Introduction to linear regression analysis*. John Wiley & Sons, 2021.

Margaret J Moore, Nele Demeyere, Chris Rorden, and Jason B Mattingley. Lesion mapping in neuropsychological research: A practical and conceptual guide. *Cortex*, 170:38–52, 2024.

Veronika I Müller, Edna C Cieslik, Angela R Laird, Peter T Fox, Joaquim Radua, David Mataix-Cols, Christopher R Tench, Tal Yarkoni, Thomas E Nichols, Peter E Turkeltaub, et al. Ten simple rules for neuroimaging meta-analysis. *Neuroscience & Biobehavioral Reviews*, 84:151–161, 2018.

J.A. Mumford and J.D. Ramsey. Bayesian networks for fmri: A primer. *NeuroImage*, 86:573–582, 2014.

Erwin Neher and Bert Sakmann. Single-channel currents recorded from membrane of denervated frog muscle fibres. *Nature*, 260(5554):799–802, 1976.

John Ashworth Nelder and Robert WM Wedderburn. Generalized linear models. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 135(3): 370–384, 1972.

Andrew Y Ng. Feature selection, l1 vs. l2 regularization, and rotational invariance. In *Proceedings of the twenty-first international conference on Machine learning*, page 78. ACM, 2004.

Thomas E. Nichols and Satoru Hayasaka. Controlling the familywise error rate in functional neuroimaging: a comparative review. *Statistical Methods in Medical Research*, 12(5):419–446, 2003. doi: 10.1191/0962280203sm341ra.

Thomas E Nichols and Andrew P Holmes. Nonparametric permutation tests for functional neuroimaging: a primer with examples. *Human brain mapping*, 15(1): 1–25, 2002.

Thomas E Nichols, Samir Das, Simon B Eickhoff, Alan C Evans, Tristan Glatard, Michael Hanke, Nikolaus Kriegeskorte, Michael P Milham, Russell A Poldrack, Jean-Baptiste Poline, et al. Best practices in data analysis and sharing in neuroimaging using mri. *Nature neuroscience*, 20(3):299–303, 2017.

Seiji Ogawa, Tso-Ming Lee, Alan R Kay, and David W Tank. Brain magnetic resonance imaging with contrast dependent on blood oxygenation. *proceedings of the National Academy of Sciences*, 87(24):9868–9872, 1990.

MRCP O'Sullivan, Derek K Jones, PE Summers, RG Morris, SCR Williams, and HS Markus. Evidence for cortical "disconnection" as a mechanism of age-related cognitive decline. *Neurology*, 57(4):632–638, 2001.

Leonardo Pantoni. Cerebral small vessel disease: from pathogenesis and clinical characteristics to therapeutic challenges. *The Lancet Neurology*, 9(7):689–701, 2010.

Yudi Pawitan. *In All Likelihood: Statistical Modelling and Inference Using Likelihood*. Oxford University Press, 2001. ISBN 9780198507659.

J. Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, 2nd edition, 2009.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830, 2011.

William D Penny, Nelson J Trujillo-Barreto, and Karl J Friston. Bayesian fmri time series analysis with spatial priors. *NeuroImage*, 24(2):350–362, 2005.

Steven E Petersen and Michael I Posner. The attention system of the human brain: 20 years after. *Annual review of neuroscience*, 35(1):73–89, 2012.

Steven E Petersen and Olaf Sporns. Brain networks and cognitive architectures. *Neuron*, 88(1):207–219, 2015.

Jeffrey R Petrella, Lubdha M Shah, Katy M Harris, Allen H Friedman, Timothy M George, John H Sampson, Joseph S Pekala, and James T Voyvodic. Preoperative functional mr imaging localization of language and motor areas: effect on therapeutic decision making in patients with potentially resectable brain tumors. *Radiology*, 240 (3):793–802, 2006.

JH Warwick Pexman, Philip A Barber, Michael D Hill, Robert J Sevick, Andrew M Demchuk, Mark E Hudon, William Y Hu, and Alastair M Buchan. Use of the alberta stroke program early ct score (aspects) for assessing ct scans in patients with acute stroke. *American Journal of Neuroradiology*, 22(8):1534–1542, 2001.

Russell A Poldrack, Jeanette A Mumford, and Thomas E Nichols. *Handbook of functional MRI data analysis*. Cambridge University Press, 2011.

Russell A Poldrack, Deanna M Barch, Jason P Mitchell, Tor D Wager, Anthony D Wagner, Joseph T Devlin, Chad Cumba, Oluwasanmi Koyejo, and Michael P Milham. Toward open sharing of task-based fmri data: the openfmri project. *Frontiers in neuroinformatics*, 7:12, 2013.

Jonathan D Power, Alexander L Cohen, Steven M Nelson, Gagan S Wig, Kelly Anne Barnes, Jessica A Church, Alecia C Vogel, Timothy O Laumann, Fran M Miezin, Bradley L Schlaggar, et al. Functional network organization of the human brain. *Neuron*, 72(4):665–678, 2011.

Jonathan D Power, Kelly A Barnes, Abraham Z Snyder, Bradley L Schlaggar, and Steven E Petersen. Spurious but systematic correlations in functional connectivity mri networks arise from subject motion. *Neuroimage*, 59(3):2142–2154, 2012.

Cathy J Price. Functional imaging studies of aphasia: a review of the literature. *Neurocase*, 7(6):419–428, 2001.

Niels D Prins, Ewoud J van Dijk, Tom den Heijer, Sarah E Vermeer, Peter J Koudstaal, Matthijs Oudkerk, Albert Hofman, and Monique MB Breteler. Cerebral white matter lesions and the risk of dementia. *Archives of neurology*, 61(10):1531–1534, 2004.

J Radua, D Mataix-Cols, Mary L Phillips, W El-Hage, DM Kronhaus, N Cardoner, and S Surguladze. A new meta-analytic method for neuroimaging studies that combines reported peak coordinates and statistical parametric maps. *European psychiatry*, 27(8):605–611, 2012.

Ali Rahimi and Benjamin Recht. Random features for large-scale kernel machines. *Advances in neural information processing systems*, 20, 2007.

Marcus E Raichle. Behind the scenes of functional brain imaging: a historical and physiological perspective. *Proceedings of the National Academy of Sciences*, 95(3): 765–772, 1998.

Marcus E Raichle. A paradigm shift in functional brain imaging. *Journal of Neuroscience*, 29(41):12729–12734, 2009.

J.D. Ramsey, S.J. Hanson, C. Hanson, and et al. Six problems for causal inference from fmri. *NeuroImage*, 49(2):1545–1558, 2010.

Calyampudi Radhakrishna Rao, Calyampudi Radhakrishna Rao, Mathematischer Statistiker, Calyampudi Radhakrishna Rao, and Calyampudi Radhakrishna Rao. *Linear statistical inference and its applications*, volume 2. Wiley New York, 1973.

Joshua Reber, Scott Gaudino, Chia-Lin Wang, and Bin He. Functional connectivity reorganization after stroke predicts motor recovery. *Neurology*, 97(8):e757–e771, 2021.

Martin Reuter, M Dylan Tisdall, Abid Qureshi, Randy L Buckner, André JW van der Kouwe, and Bruce Fischl. Head motion during mri acquisition reduces gray matter volume and thickness estimates. *Neuroimage*, 107:107–115, 2015.

Brian Ripley, Bill Venables, Douglas M Bates, Kurt Hornik, Albrecht Gebhardt, David Firth, and Maintainer Brian Ripley. Package 'mass'. *Cran r*, 538:113–120, 2013.

Brian D Ripley. *Modern applied statistics with S*. springer, 2002.

Matthew Rocklin. Dask: Parallel computation with blocked algorithms and task scheduling. In Kathryn Huff and James Bergstra, editors, *Proceedings of the 14th Python in Science Conference*, pages 130 – 136, 2015.

C. Rorden and H.O. Karnath. Using human brain lesions to infer function: a relic from a past era in the fmri age? *Nature Reviews Neuroscience*, 5:813–819, 2004.

Chris Rorden, Julius Fridriksson, and Hans-Otto Karnath. An evaluation of traditional and novel tools for lesion behavior mapping. *Neuroimage*, 44(4):1355–1362, 2009.

Egill Rostrup, Alida A Gouw, Hugo Vrenken, Elisabeth CW van Straaten, Stefan Ropele, Leonardo Pantoni, Domenico Inzitari, Frederik Barkhof, Gunhild Waldemar, LADIS Study Group, et al. The spatial distribution of age-related white matter changes as a function of vascular risk factors—results from the ladis study. *Neuroimage*, 60(3):1597–1607, 2012.

Håvard Rue, Sara Martino, and Nicolas Chopin. Approximate bayesian inference for latent gaussian models by using integrated nested laplace approximations. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 71(2):319–392, 2009.

Gholamreza Salimi-Khorshidi, Stephen M Smith, John R Keltner, Tor D Wager, and Thomas E Nichols. Meta-analysis of neuroimaging data: a comparison of image-based and coordinate-based pooling of studies. *Neuroimage*, 45(3):810–823, 2009.

Taylor Salo, Tal Yarkoni, Thomas E Nichols, Jean-Baptiste Poline, Murat Bilgel, Katherine L Bottenhorn, Dorota Jarecka, James D Kent, Adam Kimbler, Dylan M Nielson, et al. Nimare: neuroimaging meta-analysis research environment. *NeuroLibre*, 1(1):7, 2022.

Pantelis Samartsidis, Silvia Montagna, Thomas E Nichols, and Timothy D Johnson. The coordinate-based meta-analysis of neuroimaging data. *Statistical science: a review journal of the Institute of Mathematical Statistics*, 32(4):580, 2017.

Pantelis Samartsidis, Claudia R Eickhoff, Simon B Eickhoff, Tor D Wager, Lisa Feldman Barrett, Shir Atzil, Timothy D Johnson, and Thomas E Nichols. Bayesian log-gaussian cox process regression: with applications to meta-analysis of neuroimaging working memory studies. *Journal of the Royal Statistical Society. Series C, Applied statistics*, 68(1):217, 2019.

Pantelis Samartsidis, Silvia Montagna, Angela R. Laird, Peter T. Fox, Timothy D. Johnson, and Thomas E. Nichols. Estimating the prevalence of missing experiments in a neuroimaging meta-analysis. *Research Synthesis Methods*, 11(6):866–883, nov 2020a. ISSN 1759-2879. doi: 10.1002/jrsm.1448. URL https://onlinelibrary.wiley.com/doi/10.1002/jrsm.1448.

Pantelis Samartsidis, Silvia Montagna, Angela R Laird, Peter T Fox, Timothy D Johnson, and Thomas E Nichols. Estimating the prevalence of missing experiments in a neuroimaging meta-analysis. *Research synthesis methods*, 11(6):866–883, 2020b.

Theodore D Satterthwaite, Daniel H Wolf, James Loughead, Kosha Ruparel, Mark A Elliott, Hakon Hakonarson, Ruben C Gur, and Raquel E Gur. Impact of in-scanner head motion on multiple measures of functional connectivity: relevance for studies of neurodevelopment in youth. *Neuroimage*, 60(1):623–632, 2012.

Paul Schmidt, Christian Gaser, Milan Arsic, Dorothea Buck, Annette Förschler, Achim Berthele, Muna Hoshi, Rüdiger Ilg, Volker J Schmid, Claus Zimmer, et al. An automated tool for detection of flair-hyperintense white-matter lesions in multiple sclerosis. *Neuroimage*, 59(4):3774–3783, 2012.

R Schmidt, F Fazekas, P Kapeller, H Schmidt, and H-P Hartung. Mri white matter hyperintensities: three-year follow-up of the austrian stroke prevention study. *Neurology*, 53(1):132–132, 1999.

Gideon Schwarz. Estimating the dimension of a model. *The annals of statistics*, pages 461–464, 1978.

Skipper Seabold and Josef Perktold. Statsmodels: econometric and statistical modeling with python. *SciPy*, 7(1), 2010.

George AF Seber and Alan J Lee. *Linear regression analysis*. John Wiley & Sons, 2003.

Terrence J Sejnowski, Patricia S Churchland, and J Anthony Movshon. Putting big data to good use in neuroscience. *Nature neuroscience*, 17(11):1440–1441, 2014.

David F Shanno. Conditioning of quasi-newton methods for function minimization. *Mathematics of computation*, 24(111):647–656, 1970.

Jun Shao and Dongsheng Tu. *The jackknife and bootstrap*. Springer Science & Business Media, 2012.

Gordon M Shepherd. *Foundations of the neuron doctrine*. Oxford University Press, 2015.

Ronald Sladky, Karl J Friston, Jasmin Tröstl, Ross Cunnington, Ewald Moser, and Christian Windischberger. Slice-timing effects and their correction in functional mri. *Neuroimage*, 58(2):588–594, 2011.

John G Sled, Alex P Zijdenbos, and Alan C Evans. A nonparametric method for automatic correction of intensity nonuniformity in mri data. *IEEE transactions on medical imaging*, 17(1):87–97, 1998.

Stephen M Smith. Fast robust automated brain extraction. *Human brain mapping*, 17 (3):143–155, 2002.

Stephen M. Smith and Thomas E. Nichols. Threshold-free cluster enhancement: Addressing problems of smoothing, threshold dependence and localisation in cluster inference. *NeuroImage*, 44(1):83–98, 2009.

Stephen M Smith, Peter T Fox, Karla L Miller, David C Glahn, P Mickle Fox, Clare E Mackay, Nicola Filippini, Kate E Watkins, Roberto Toro, Angela R Laird, et al. Correspondence of the brain's functional architecture during activation and rest. *Proceedings of the national academy of sciences*, 106(31):13040–13045, 2009.

Stephen M Smith, Karla L Miller, Gholamreza Salimi-Khorshidi, Matthew Webster, Christian F Beckmann, Thomas E Nichols, Joseph D Ramsey, and Mark W Woolrich. Network modelling methods for fmri. *Neuroimage*, 54(2):875–891, 2011.

Stephen M Smith, Diego Vidaurre, Christian F Beckmann, Matthew F Glasser, Mark Jenkinson, Karla L Miller, Thomas E Nichols, Emma C Robinson, Gholamreza Salimi-Khorshidi, Mark W Woolrich, et al. Functional connectomics from resting-state fmri. *Trends in cognitive sciences*, 17(12):666–682, 2013.

C. Sperber and H.O. Karnath. Impact of correction factors in human brain lesion-behavior inference. *Human Brain Mapping*, 38(3):1693–1701, 2017.

Christoph Sperber. Rethinking causality and data complexity in brain lesion-behaviour inference and its implications for lesion-behaviour modelling. *Cortex*, 126:49–62, 2020.

Samuel A. Stouffer, Edward A. Suchman, Leland C. DeVinney, Shirley A. Star, and Robin M. Williams Jr. *The American Soldier: Adjustment During Army Life*, volume 1. Princeton University Press, 1949.

Cathie Sudlow, John Gallacher, Naomi Allen, Valerie Beral, Paul Burton, John Danesh, Paul Downey, Paul Elliott, Jane Green, Martin Landray, et al. 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):e1001779, 2015.

Laurent Tatu, Thierry Moulin, Fabrice Vuillier, and Julien Bogousslavsky. Arterial territories of the human brain. *Stroke syndromes*, pages 329–343, 2012.

Michel Thiebaut de Schotten, Francesco Tomaiuolo, Marilena Aiello, Sheila Merola, Massimo Silvetti, Francesca Lecce, Paolo Bartolomeo, and Fabrizio Doricchi. Damage to white matter pathways in subacute and chronic spatial neglect: a group study and 2 single-case studies with complete virtual "in vivo" tractography dissection. *Cerebral cortex*, 24(3):691–706, 2014.

Michel Thiebaut de Schotten, Charlotte Foulon, and Parashkev Nachev. From phineas gage and monsieur leborgne to h.m.: Revisiting disconnection syndromes. *Cortex*, 66:6–17, 2015.

Simon G Thompson and Julian PT Higgins. How should meta-regression analyses be undertaken and interpreted? *Statistics in medicine*, 21(11):1559–1573, 2002.

Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 58(1):267–288, 1996.

Peter E Turkeltaub, Guinevere F Eden, Karen M Jones, and Thomas A Zeffiro. Meta-analysis of the functional neuroanatomy of single-word reading: method and validation. *Neuroimage*, 16(3):765–780, 2002.

Nicholas J Tustison, Brian B Avants, Philip A Cook, Yuanjie Zheng, Alexander Egan, Paul A Yushkevich, and James C Gee. N4itk: improved n3 bias correction. *IEEE transactions on medical imaging*, 29(6):1310–1320, 2010.

Kamil Uğurbil. Magnetic resonance imaging at ultrahigh fields. *IEEE transactions on biomedical engineering*, 61(5):1364–1379, 2014.

Martijn P van den Heuvel and Olaf Sporns. A cross-disorder connectome landscape of brain dysconnectivity. *Nature reviews neuroscience*, 20(7):435–446, 2019.

David C Van Essen, Stephen M Smith, Deanna M Barch, Timothy EJ Behrens, Essa Yacoub, Kamil Ugurbil, Wu-Minn HCP Consortium, et al. The wu-minn human connectome project: an overview. *Neuroimage*, 80:62–79, 2013.

Frank Van Overwalle. Social cognition and the brain: a meta-analysis. *Human Brain Mapping*, 30(3):829–858, 2009. doi: 10.1002/hbm.20547.

Michele Veldsman, Petya Kindalova, Masud Husain, Ioannis Kosmidis, and Thomas E Nichols. Spatial distribution and cognitive impact of cerebrovascular risk-related white matter hyperintensities. *NeuroImage: Clinical*, 28:102405, 2020.

Jay M Ver Hoef and Peter L Boveng. Quasi-poisson vs. negative binomial regression: how should we model overdispersed count data? *Ecology*, 88(11):2766–2772, 2007.

Sandra Vieira, Walter HL Pinaya, and Andrea Mechelli. Using deep learning to investigate the neuroimaging correlates of psychiatric and neurological disorders: Methods and applications. *Neuroscience & Biobehavioral Reviews*, 74:58–75, 2017.

Tor D Wager, Martin Lindquist, and Lauren Kaplan. Meta-analysis of functional neuroimaging data: current and future directions. *Social cognitive and affective neuroscience*, 2(2):150–158, 2007.

Joanna M Wardlaw, Colin Smith, and Martin Dichgans. Mechanisms of sporadic cerebral small vessel disease: insights from neuroimaging. *The Lancet Neurology*, 12 (5):483–497, 2013a.

Joanna M Wardlaw, Eric E Smith, Geert J Biessels, Charlotte Cordonnier, Franz Fazekas, Richard Frayne, Richard I Lindley, John T O'Brien, Frederik Barkhof, Oscar R Benavente, et al. Neuroimaging standards for research into small vessel disease and its contribution to ageing and neurodegeneration. *The Lancet Neurology*, 12(8):822–838, 2013b.

Van J Wedeen, Ruopeng P Wang, Jeremy D Schmahmann, Thomas Benner, Wen-Yih Isaac Tseng, Guangping Dai, Deepak N Pandya, Patric Hagmann, Helen D'Arceuil, and Alex J de Crespigny. Diffusion spectrum magnetic resonance imaging (dsi) tractography of crossing fibers. *Neuroimage*, 41(4):1267–1277, 2008.

Junhao Wen, Elina Thibeau-Sutre, Mauricio Diaz-Melo, Jorge Samper-González, Alexandre Routier, Simona Bottani, Didier Dormont, Stanley Durrleman, Ninon Burgos, Olivier Colliot, et al. Convolutional neural networks for classification of alzheimer's disease: Overview and reproducible evaluation. *Medical image analysis*, 63:101694, 2020.

Peter H Westfall and S Stanley Young. *Resampling-based multiple testing: Examples and methods for p-value adjustment*, volume 279. John Wiley & Sons, 1993.

Halbert White. A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica*, 48(4):817–838, 1980.

Halbert White. Maximum likelihood estimation of misspecified models. *Econometrica*, 50(1):1–25, 1982.

Samuel S Wilks. The large-sample distribution of the likelihood ratio for testing composite hypotheses. *The annals of mathematical statistics*, 9(1):60–62, 1938.

Anderson M Winkler, Gerard R Ridgway, Matthew A Webster, Stephen M Smith, and Thomas E Nichols. Permutation inference for the general linear model. *Neuroimage*, 92:381–397, 2014.

Choong-Wan Woo, Liane Schmidt, Anjali Krishnan, Marieke Jepma, Mathieu Roy, Martin A Lindquist, Lauren Y Atlas, and Tor D Wager. Quantifying cerebral contributions to pain beyond nociception. *Nature communications*, 8(1):14211, 2017.

Simon Wood and Maintainer Simon Wood. Package 'mgcv'. *R package version*, 1(29): 729, 2015.

M. W. Woolrich, S. Jbabdi, B. Patenaude, et al. Bayesian analysis of neuroimaging data in fsl. *NeuroImage*, 45:S173–S186, 2009.

Keith J Worsley, Sean Marrett, Peter Neelin, Alain C Vandal, Karl J Friston, and Alan C Evans. A unified statistical approach for determining significant signals in images of cerebral activation. *Human brain mapping*, 4(1):58–73, 1996.

Keith J Worsley, Chien Heng Liao, John Aston, V Petre, GH Duncan, F Morales, and Alan C Evans. A general statistical analysis for fmri data. *Neuroimage*, 15(1):1–15, 2002.

Lei Xu, Timothy D Johnson, Thomas E Nichols, and Derek E Nee. Modeling inter-subject variability in fmri activation location: a bayesian hierarchical spatial model. *Biometrics*, 65(4):1041–1051, 2009.

Tal Yarkoni, Russell A Poldrack, Thomas E Nichols, David C Van Essen, and Tor D Wager. Large-scale automated synthesis of human functional neuroimaging data. *Nature methods*, 8(8):665–670, 2011.

Y. Yokoyama, Y.H. Mah, M. Husain, and et al. Lesion analysis for causal inference in neuropsychology. *Neuron*, 84:605–618, 2014.

Yifan Yu, Rosario Pintos Lobo, Michael Cody Riedel, Katherine Bottenhorn, Angela R Laird, and Thomas E Nichols. Neuroimaging meta regression for coordinate based meta analysis data with a spatial model. *Biostatistics*, page kxae024, 2024.

Yu Ryan Yue, Martin A Lindquist, and Ji Meng Loh. Meta-analysis of functional neuroimaging data using bayesian nonparametric binary regression. *The Annals of Applied Statistics*, pages 697–718, 2012.

Xiang Zhou, Siyu Lee, and Liang Sun. Robust inference with misspecified models: Theory and practice. *Annual Review of Statistics and Its Application*, 7:341–368, 2020.

Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 67(2): 301–320, 2005.