On the Reproducibility and Interpretability of Group-Level Task-fMRI Results

by

Alexander Bowring

St Catherine's College

Submitted to the University of Oxford

for the degree of

Doctor of Philosophy

Nuffield Department of Population Health

October 2019

Abstract

On the Reproducibility and Interpretability of Group-Level Task-fMRI Results

Alexander Bowring
St Catherine's College, University of Oxford
Submitted for the degree of Doctor of Philosophy
Michaelmas Term, 2019

In this thesis, we aim to address two topical issues at the forefront of task-based functional magnetic resonance imaging (fMRI). The first of these is a growing apprehension within the field about the reproducibility of findings that make up the neuroimaging literature. To confront this, we assess how the choice of software package for analyzing fMRI data can impact the final group-level results of a neuroimaging study. We reanalyze data from three published task-fMRI studies within the three most widely-used neuroimaging software packages – AFNI, FSL, and SPM – and then apply a range of comparison methods to gauge the scale of variability across the results. While qualitatively we find similarities, our quantitative assessment methods discover considerable differences between the final statistical images obtained with each package. Ultimately, we conclude that exceedingly weak effects may not generalize across fMRI analysis software.

In the second part of this work we shift our attention to the analytical methods applied for fMRI inference. Here, we seek to overcome limitations with the traditional statistical approach, where for sufficiently large data sizes current methods determine universal activation across the brain, rendering the results as uninterpretable. We extend on a method proposed by *Sommerfeld, Sain, and Schwartzman (2018)* *(SSS)* to develop spatial Confidence Sets (CSs) on clusters found in thresholded raw blood-oxygen-level-dependent (BOLD) effect size maps. The CSs give statements on the locations where raw effect sizes exceed, and fall short of, a purposeful *non-zero* threshold. We propose several theoretical and practical implementation advancements to the original method formulated in *SSS*, delivering a procedure with superior performance in sample sizes as low as $N = 60$. We validate the method with 3D Monte Carlo simulations that resemble fMRI data. We then compute CSs for the Human Connectome Project (HCP) working memory task contrast images, illustrating the brain regions that show a reliable %BOLD for a given %BOLD threshold.

In the final part of this thesis, we develop the CSs to operate on standardized Cohen's d effect size images. We derive the statistical properties of the Cohen's d estimator to motivate three algorithms for computing Cohen's d CSs, including a novel method based on normalizing the distribution of Cohen's d. With intensive 3D Monte Carlo simulations, we find that two of these methods can be effectively applied to fMRI data. We compute Cohen's d CSs on the HCP data, and by comparing the CSs with results obtained from a standard testing procedure, exemplify the improved localization of effects that can be gained by using the Confidence Sets.

Acknowledgments

Declarations

I, Alexander Bowring, hereby declare that 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 in these, or any other universities. This thesis is the result of my own work and includes nothing which is the outcome of work done in collaboration, except where specifically indicated in the text.

- The work presented in Chapter 3 has been published in the *Human Brain Mapping* journal, *Bowring, Maumet, and Nichols (2019a)*. This work was presented at the *Organization for Human Brain Mapping* (OHBM) Annual Meetings in 2017 and 2018. At the OHBM 2018 Annual Meeting, this work was the recipient of an oral presentation and a Merit Abstract Award.
- The work presented in Chapter 4 has been published in the *NeuroImage* journal *Bowring, Telschow, Schwartzman, and Nichols (2019b)*. This work was presented at the OHBM Annual Meeting in 2017, where it was the recipient of an oral presentation.
- The work presented in Chapter 5 will shortly be submitted for publication.

Alexander Bowring
September 2019

Contents

CHAPTER 1

Introduction

Since its inception at the end of the twentieth century, functional magnetic resonance imaging (fMRI) has experienced a meteoric rise to become the primary tool for human brain mapping. While many forms of the technique exist, introduction of the particular method based on the blood-oxygen-level-dependent (BOLD) effect has ultimately been the catalyst in elevating fMRI to such stature within the neuroimaging community. Taking advantage of the magnetic properties of oxygen-rich red blood cells, BOLD fMRI measures changes in blood oxygenization alongside cerebral blood flow and volume as a proxy to identify brain areas where elevated neuronal activity has occurred in response to a stimulus. While the relationship between the BOLD effect and neuronal activity is complex and remains controversial, it is the unique attributes of BOLD fMRI – in particular, its capacity for non-invasive recording of signals across the entire brain at a high spatial resolution – that set the technique apart from other scanning methods.

However, BOLD fMRI is also a noisy process. The magnetic resonance (MR) signals researchers set out to measure during a scanning session are corrupted by artefacts from both the imaging hardware and the physiology of the participant. Examples of scanner noise include inhomogeneities of the magnetic field that can cause spatial distortion or blurring in the MR image and scanner drift characterized by tem-

poral degradation of the MR signal. Physiological noise induced by subject motion, respiration, and heartbeat exacerbate the problem.

Because of the low signal-to-noise ratio, researchers must apply a series of statistical techniques to find meaning in the data. This usually entails carrying out a number of preprocessing, modelling and analysis steps that together constitute the fMRI processing pipeline. The fundamental objectives of preprocessing are to standardize brain locations across participants, to apply methods ensuring that the data conform to statistical assumptions required for analysis, and to reduce the influence of the noise artefacts present in the data. This is achieved by conducting a number of steps, including slice-timing correction, motion correction, normalization, registration of the functional data to an anatomical template, and spatial smoothing.

For task-based fMRI, a mass-univariate approach is utilized to model the data. During the scanning session, functional data are acquired in the form of voxels – cubic intensity units that partition the brain comparable to the way in which pixels partition a computer screen. Each voxel's time-series is considered independently within the general linear model framework as a combination of signal components. To evaluate the effect of an experimental task condition relative to a baseline condition, hypothesis testing is performed at each voxel to compute a statistical parametric map of statistic values. Here, the behaviour of the signal under the null-hypothesis of no activation is estimated using either a parametric approach, appealing to the body of mathematics known as Random Field Theory, or a nonparametric approach, where permutation methods are applied to estimate the null-distribution directly from the data. Finally, the statistical parametric map is thresholded to localize brain function.

While we have provided a brief overview of the fMRI analysis pipeline, it is notable that there is not a general consensus as to how each particular analysis step should be carried out. Consequently, researchers have the freedom to make many choices during an analysis, such as how much smoothing is applied to the data, or how the haemodynamic response of blood flow to active neuronal tissues is mod-

elled. However, this 'methodological plurality' comes with a drawback. While conceptually similar, two different analysis pipelines applied on the same dataset may not produce the same scientific results; choice of analysis pipeline, operating system, and even software version can influence the final research outcomes of a study. Because of this, the high analytic flexibility associated with fMRI has been pinpointed as a key factor that can lead to distorted and irreproducible results (Hong et al., 2019; Ioannidis, 2005; Wager et al., 2009).

The degree to which varying methodological decisions can lead to discrepancies in observed results has been investigated extensively. Choices for each individual procedure in the analysis pipeline (for example, head-motion regression (Lund et al., 2005), temporal filtering (Skudlarski et al., 1999), and autocorrelation correction (Woolrich et al., 2001)) alongside the order in which these procedures are conducted (Carp, 2013) can all deeply influence the final determined areas of brain activation. In perhaps the most comprehensive of such studies (Carp, 2012a), a single publicly available fMRI dataset was analyzed using over 6,000 different pipelines, generating 34,560 unique thresholded activation images. These results displayed a substantial degree of flexibility in both the sizes and locations of significant activation. In combination, these examples of research shape a sombre picture for the possibility of study reproducibility.

Alongside issues concerned with the flexibility of the analysis workflow, the statistical procedures carried out for fMRI inference have also come under intense scrutiny in recent times. Because statistical tests are conducted at each brain voxel independently, the p-value used to threshold the statistical parametric map is corrected to account for the large number of simultaneous comparisons being carried out and limit the expected number of voxels falsely declared as significant. This is almost always done using a false discovery rate correction procedure (Benjamini and Hochberg, 1995) or a Bonferonni correction to limit the familywise error rate of making at least one significant finding.

The importance of such statistical correction methods were made prominent within the neuroimaging community using a humorous example, where one author identified significant activation in the brain of a dead salmon after applying inference with an uncorrected p-value (Bennett et al., 2009). However, in recent times they have been a source of major controversy. In 2016, Eklund, Nichols, and Knutsson (2016) discovered that many fMRI software packages were incorrectly carrying out the multiple-correction procedures for clusterwise inference, inflating the false-positive rate to beyond 40%. The implications of this study brought into question the validity of thousands of published fMRI results, leading to updates in the main fMRI analysis packages (Cox et al., 2017a) and a re-evaluation of how clusterwise inference should be applied (Flandin and Friston, 2019; Mueller et al., 2017; Cox et al., 2017b).

Nevertheless, a number of conceptual limitations remain with the fMRI statistical approach to inference. Specifically, there is a considerable amount of information that is *not* captured when applying inference using cluster-size. In this setting, a cluster-level p-value only conveys information about a cluster's spatial extent under the null-hypothesis. Since no information is provided regarding the statistical significance of each voxel comprising a significant cluster, the most we can say is that significant activation has occurred *somewhere* inside the cluster (Woo et al., 2014). An implication of this is that when a large, sprawling cluster covers many anatomical regions, the precise spatial specificity of the activation is in fact poor. A related problem of cluster inference is that no information is provided about the spatial variation of significant clusters. For example, if an fMRI study was to be repeated many times with new sets of subjects there would be variation in the size and shape of clusters found, yet the current statistical results have no way to characterize this variability.

A more pressing issue, perhaps, stems from an age-old paradox caused by the 'fallacy of the null-hypothesis' (Rozeboom, 1960). The paradox is that while statistical models conventionally assume mean-zero noise, in reality all sources of noise will never cancel, and therefore improvements in experimental design will eventu-

ally lead to statistically significant results. Thus, the null-hypothesis will, eventually, *always* be rejected (Meehl, 1967). The recent availability of ambitious, large-sample studies (e.g Human Connectome Project (HCP), N = 1,200; UK Biobank, N = 30,000 and counting) have exemplified this problem. Analysis of high-quality fMRI data acquired under optimal noise conditions has been shown to display almost universal activation across the entire brain after hypothesis testing, even with stringent correction (Gonzalez-Castillo et al., 2012). For these reasons, there is an increased urgency for methods that can provide meaningful inference to interpret all significant effects.

In this work, we make contributions in two topical areas currently challenging the field of task-based fMRI: Firstly, the need for further transparency as to the degree in which the body of work comprising the fMRI literature is reproducible. Secondly, the need for further methods to improve current inference practices carried out within the field. To end this section, we summarize our main contributions before providing an outline of the organization of this thesis:

1. While we have discussed a number of studies exploring how analysis decisions can influence the results of an fMRI investigation, for all of these studies the fundamental choice of software package through which the analysis was conducted remained constant. This is despite a vast array of packages that are used throughout the neuroimaging literature, the most popular of which are AFNI, FSL and SPM. In Chapter 3 we comprehensively assess how the choice of software package can impact the final results of a neuroimaging analysis. We reanalyze three published task-fMRI studies within AFNI, FSL and SPM, and quantify several aspects of variability between the three package's group-level statistical maps. Our findings suggest that exceedingly weak effects may not generalize across software. We are unaware of any comparable exercise in the literature.

2. In carrying out the software comparison exercise, we implement a range of quantitative methods for the novel application of comparing fMRI statistical maps. These include Dice statistics, for assessing differences in the *locations* of activation determined in each software's thresholded statistical results, Bland-Altman plots, for assessing differences between the *magnitude* of statistic values in each software's unthresholded results, Euler Characteristics, for assessing differences in the topological properties of each software's activation profile, and Neurosynth analyses, for assessing differences in the anatomical regions associated to each software's activation pattern. We believe these methods are generalizable and hope they may benefit any further comparison of neuroimaging results.

3. In Chapter 4, we develop an inference method originally proposed for application on geospatial data in Sommerfeld, Sain, and Schwartzman (2018) (*SSS*) to create spatial Confidence Sets (CSs) on clusters found in fMRI percentage BOLD effect size maps. While currently used hypothesis testing methods indicate where the null, i.e. an effect size of zero, can be rejected, this form of inference allows for statements about anatomical regions where effect sizes have exceeded, and fallen short of, a meaningful *non-zero* threshold, such as areas where a BOLD change of 2.0% has occurred.

4. We make a number of theoretical and implementation advancements to the *SSS* method for computing CSs that improve the method's finite-sample performance in the context of neuroimaging. In particular, we propose a combination of the Wild t-Bootstrap method and the use of Rademacher variables for multiplication of the bootstrapped residuals, which we find substantially improves performance of the method in moderate sample sizes. We also develop a linear interpolation method for computing the boundary over which the bootstrap is applied, and a novel approach for assessing the empirical coverage of the CSs

that reduces upward bias in how the simulation results are measured.

5. In Chapter 5, we make further theoretical developments to the CSs for application on standardized, Cohen's d effect size images. By deriving the statistical properties of the Cohen's d estimator, we motivate three separate algorithms to obtain Cohen's d CSs. One of these methods is based on a novel procedure to normalize the distribution of the sample Cohen's d. By testing the three algorithms using intensive 3D Monte Carlo simulations, we conclude that two of the methods may perform particularly well on task-fMRI Cohen's d effect size maps.

The remainder of this thesis is organized into five chapters.

Chapter 2 is dedicated to presenting the context of this work and providing background on the current methodological procedures carried out for analysis of task-fMRI data, with a particular emphasis on the statistical inference methods relevant to this thesis.

In **Chapter 3**, we assess the analytic variability of group-level task-fMRI results under the choice of software package through which the analysis is conducted. We reanalyze three published task-fMRI studies whose data has been made publicly available, attempting to replicate the original analysis procedures within each software package. We then make a number of qualitative and quantitative comparisons to assess the similarity of our results.

In **Chapter 4**, we develop the inference method originally proposed in *SSS* to create spatial Confidence Sets on clusters found in fMRI percentage BOLD effect size maps. We summarize the theory in *SSS* before detailing our proposed modifications. We then carry out intensive Monte Carlo simulations to investigate the performance of the CSs on synthetic 3D signals representative of clusters found in fMRI effect size maps. Finally, we illustrate the method by computing CSs on 80 subject's percentage BOLD data from the Human Connectome Project working memory task.

In **Chapter 5**, we make further theoretical developments to adapt the CSs for application to Cohen's d effect size images. We derive the finite and asymptotic properties of the Cohen's d estimator before formalizing three separate algorithms to compute Cohen's d CSs. We assess the performance of the three methods using Monte Carlo simulations similarly to the previous chapter. Finally, we provide a demonstration of the three methods on Cohen's d effect size maps from the Human Connectome Project dataset, comparing the Cohen's d CSs to results obtained using a traditional statistical inference procedure.

In **Chapter 6** we conclude this thesis, summarizing our contributions and providing discussion of possibilities for future work.

CHAPTER 2

Background

In this chapter, we provide the context that forms the basis of our research. We begin by presenting a broad overview of the study of brain function, before narrowing down to the specific field of task-based functional magnetic resonance imaging (t-fMRI) that will be the main focus of study in this thesis. Here, we describe each of the preprocessing and modelling components of a typical t-fMRI analysis pipeline. Finally, we give an in-depth discussion of the state-of-the-art procedures used for subject- and group-level t-fMRI inference that are of particular relevance to the remaining chapters of this work.

2.1 The Study of Brain Function

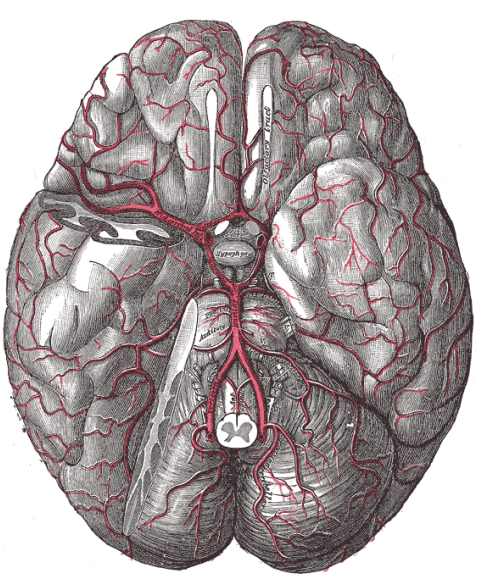


Figure 2.1: An illustration showing the arteries at the base of the human brain. Reprinted from Anatomy (1918).

The human brain, the central organ of the human nervous system, has been described as one of the most complex structures in the known universe. Made up of approximately 86 billion neurons (Azevedo et al., 2009), where neuronal interaction occurs continuously via trillions of synaptic networks to form intricate and dynamic neural networks, the myriad of processes taking place inside the brain at any given time make the study of brain function an intimidating challenge. Nevertheless, our understanding of the brain has come along way from our ancient Egyptian ancestors, who believed that the heart was the source of human intelligence, and for whom the practice of drilling a hole into the skull was regarded as a solution to cure a headache (Adelman and Others, 1987; Mohamed, 2014).

Remarkably, much of this progress has come in the last century alone. A number of key developments within this time-frame include: confirmation of the neuron doctrine, the concept that the nervous system is a collection of discrete individual cells, postulated by Santiago Ramon y Cajal at the end of the 19th century

and demonstrated in the 1950s thanks to the development of electron microscopy (López-Muñoz et al., 2006); the first evidence of neuroplasticity, the ability for the brain's structure to change during an individual's lifetime (Diamond et al., 1964; Bennett et al., 1964); and the emergence of neuroimaging techniques such as electroencephalography (EEG), positron emission tomography (PET), and magnetic resonance imaging (MRI). The toils of this scientific endeavour are now translating into concrete advancements influencing a wide variety of aspects concerned with population health. Neuroscience research is beginning to find applications in the clinical setting to advance our understanding of neurodevelopmental and neurodegenerative disorders and generate novel therapies to treat and prevent such diseases. Brain imaging has been used to localize the source of neurological impairment for diseases such as epilepsy (Stacey and Litt, 2008), and neuroengineering techniques based on our capability to stimulate neural circuits are implemented to treat Parkinson's disease (Kalia et al., 2013) and dystonia (Fox and Alterman, 2015). Structural- and functional-MRI are being explored to determine biomarkers for diagnosis of Alzheimer's disease *prior* to symptom onset (Sperling et al., 2014; McEvoy et al., 2009), alongside providing information about the role of different brain regions in human behaviour that can contribute to an improved prognosis and patient response to therapy (Matthews et al., 2006).

Modern neuroscience can be dissected into many major branches, each subfield taking a specific slant to studying the nervous system. It is therefore perhaps unsurprising that in isolation, the phrase 'the study of brain function' is rather vague. Brain function can manifest itself in ways that can be observed using a variety of different measurements, whether that be with a molecular, chemical, structural, or functional approach (Hargreaves and Klimas, 2012). Different modalities of MRI are employed to evaluate specific properties of the brain that characterize whichever approach is taken. For instance, looking at brain function from an anatomical perspective, voxel-based morphometry (VBM) could be used to measure differences in

local concentrations of brain tissue in order to assess changes in grey matter volume (Mechelli et al., 2005). Additionally, one could apply diffusion tensor imaging (DTI) to map white matter tractography in the brain (Alexander et al., 2007; Soares et al., 2013). From a functional outlook, resting state fMRI (rs-fMRI) determines that spatially remote brain areas are functionally connected when each region's BOLD response is temporally correlated in the absence of an explicit task (Lee et al., 2013). On the other hand, task-based fMRI (t-fMRI) measures spatio-temporal changes in the BOLD signal between task-stimulated and control states to find brain regions that are activated in the presence of a stimulus (Glover, 2011).

Each imaging method and modality does not live inside a vacuum, and recent work within the field has provided further insight of the interdependence between different approaches to examining brain function. One example of this is in the study of resting state networks, which explores how distinct sets of brain regions can reveal temporally correlated activation patterns when the brain is at rest. While resting state networks have been most widely investigated using rs-fMRI techniques (e.g. Smith et al., 2009; Lee et al., 2012; Moussa et al., 2012), more recently the same correlation patterns have been independently detected using EEG and magnetoencephalography (MEG) (Brookes et al., 2011; Fomina et al., 2015). This work not only demonstrates how utilization of numerous tools can further our understanding of resting state mechanisms, but also suggests a direct relationship between the electro-physiological signals recorded with MEG and the BOLD fluctuations associated with fMRI. Similarly, other recent efforts have shown that the functional response to a cognitive task measured with t-fMRI may be able to be predicted by connectivity features from the same individual's brain at rest (Parker Jones et al., 2017; Tavor et al., 2016). This research signals towards an innate functional signature that defines our behaviour, while also providing potential clinical solutions to obtain t-fMRI data from patients who are unable to perform the specific task of interest.

In the context of this thesis, we will study brain function from a functional

perspective, primarily focussed on task-based fMRI.

2.2 Blood-oxygen-level-dependent (BOLD) functional Magnetic Resonance Imaging (fMRI)

Whereas structural MRI is concerned with the anatomy of the brain, functional MRI (fMRI) measures dynamic changes in blood flow in order to ultimately make inference on neuronal activation. This is possible due to the intrinsic relationship between local neuronal activity and subsequent changes in cerebral blood flow (CBF), a biological phenomenon known as neurovascular coupling. An increased supply of oxygen is carried by haemoglobin in red blood cells to provide energy to active neurons, and it is the magnetic properties of the haemoglobin that MRI takes advantage of. Specifically, as deoxygenated haemoglobin is more magnetic than oxygenated haemoglobin, MRI uses haemoglobin as an endogenous contrast agent from which to source the signal. Neurovascular coupling induces inhomogeneities in the local magnetic field due to a decreased concentration of deoxygenated haemoglobin that lead to a detectable change in the MR signal.

The complete chain of events linking neuronal activity to a change in MRI signal is referred to as the blood-oxygen-level-dependent (BOLD) effect, and this type of imaging is known as BOLD fMRI. Proof of concept of the BOLD effect was first provided in Ogawa, Lee, Kay, and Tank (1990), and the first use of BOLD fMRI for human brain mapping was carried out in 1992 (Bandettini et al., 1992; Kwong et al., 1992; Ogawa et al., 1992), leading to a large uptake of the method that has continued to this day. There are a number of alternative approaches for brain imaging, such as functional Arterial Spin Labelling (fASL), which uses magnetically labelled arterial blood water to quantify changes in CSF. While fASL can offer some advantages over fMRI, and changes in CSF measured with this technique are more closely tied to neuronal activation than the BOLD signal, fASL suffers from a much lower signal-to-noise ratio

that has consequently made fMRI the preferred imaging modality of choice.

2.2.1 Physiology of the BOLD Response

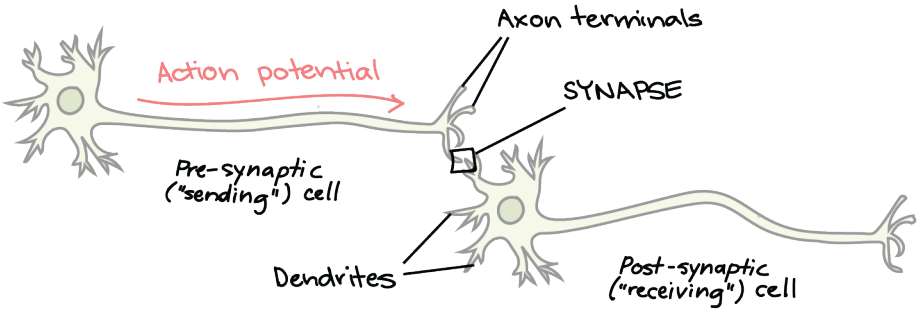


Figure 2.2: A schematic of the interaction between neurons. Image reused from Khan Academy[1] (CC BY-NC-SA 3.0 US).

Neuronal interaction transpires via a system of electrical and chemical activity. To send out information, an individual neuron – the pre-synaptic cell – emits an electrical signal known as an action potential, for the purpose of stimulating another target neuron – the post-synaptic cell. The action potential travels along the axon of the sending cell, and is transmitted to the receiving cell at the synapse. Information is delivered from the output branches (or, axon terminals) of the sending cell, across the synapse, to the input branches (or, dendrites) of the receiving cell, involving the release of chemical neurotransmitters alongside a number of other cellular processes. This may stimulate or inhibit the firing of action potentials at the target cell to communicate with other neurons, eventually leading to a configuration of neurons collectively processing and responding to information.

The electrical and chemical processes involved in neuronal activation require energy, which drives the neurovascular coupling. Blood vessels that flow into the capillaries pervading the neuronal tissue dilate and the rate of CBF increases to regulate a greater supply of oxygen and nutrients to localized regions of active neurons. Overall, the increases in CBF and cerebral blood volume (CBV) are many orders of

[1] All Khan Academy content is available for free at www.khanacademy.org.

magnitude greater than the increases in oxygen extraction ($CMRO_2$) caused by the neuronal activation. Thus, there is an overall net increase of oxygenated haemoglobin, and an increase in the BOLD signal. The expected BOLD response generated from a brief stimulus is characterized quantitatively by the Haemodynamic Response Function (HRF), encompassing the individual changes in CBF, CBV and $CMRO_2$ induced by neuronal stimulation.

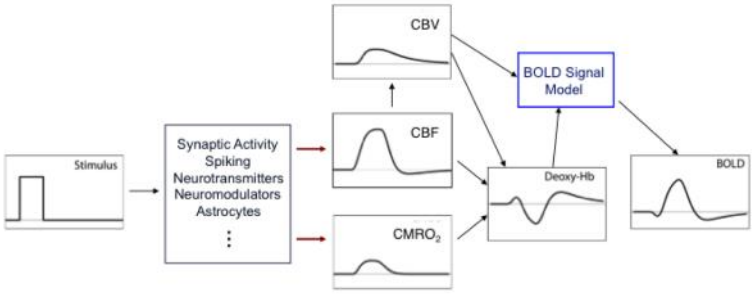


Figure 2.3: The current BOLD signal model. In the presence of a stimulus, changes in biological parameters such as CBF, CBV and $CMRO_2$ influence the final observed BOLD response. Reprinted from Buxton (2012), with permission from Elsevier.

2.3 Task-based functional Magnetic Resonance Imaging (t-fMRI)

The ultimate goal of a task-based functional magnetic resonance imaging (t-fMRI) experiment is to understand the brain regions that are responsive to a particular task or stimulus the researcher has chosen to investigate. Explicitly, the researcher seeks to detect brain areas whose BOLD time series data is correlated to the task the participant is instructed to perform in the scanner. Researchers can choose from a wide range of possible tasks to explore how the brain processes in a variety of circumstances. For example, a cognitive task may be chosen to gain insight into how the brain processes decision-making or recognition, while a physiological task may be used to see how the brain reacts to a stimulus intended to cause pain or arousal, or how the brain functions when participants are told to hold their breath. In general, the experimenter is only limited in choice of task by the constraints that the task must

be able to be conducted within the scanner, and that the task should not involve any sort of head movement which could corrupt the signal.

The MR signal measured in the scanner is noisy, and the haemodynamic response induced by a stimulus only causes fractional changes in the BOLD response, typically of around one percent. Therefore, in order to increase the signal-to-noise ratio (SNR) of the BOLD signal participants repeat the task several times in the scanner. The type of task used alongside the timings for which the participant is instructed to perform the task inside the scanner are together known as the task paradigm or experimental design. Many different task conditions can be investigated within one task paradigm, however, it is fundamental that at least two conditions are included. This is because BOLD data is not quantitative, insofar that we are unable to interpret the level of neuronal activity from the absolute magnitude of the BOLD response alone. Instead, neuronal activity is inferred by using contrasts to measure the difference in the MR signal between two conditions. Commonly, the BOLD response to a task condition is contrasted with a baseline condition, where the participant is at rest within the scanner. However, it is equally acceptable to contrast two separate task conditions depending on the aims of the investigation.

An experimental design where the task condition is carried out for an extended period of time is said to have a block design (or boxcar design). One example of this could be a task paradigm where the participant is instructed to look at an animal photo for five seconds in each task repetition. Alternatively, in an event-related design the task or stimulus takes the form of a discrete, rapid event, such as a study where the participant experiences a mild electric shock. A graphical representation of each of these task paradigms is displayed in Figure 2.4, showing the onset timing function along with the anticipated response for both types of stimuli. While block designs have greater statistical power with a relatively larger BOLD response, the researcher has more control with respect to how the stimuli are delivered in an event-related design.

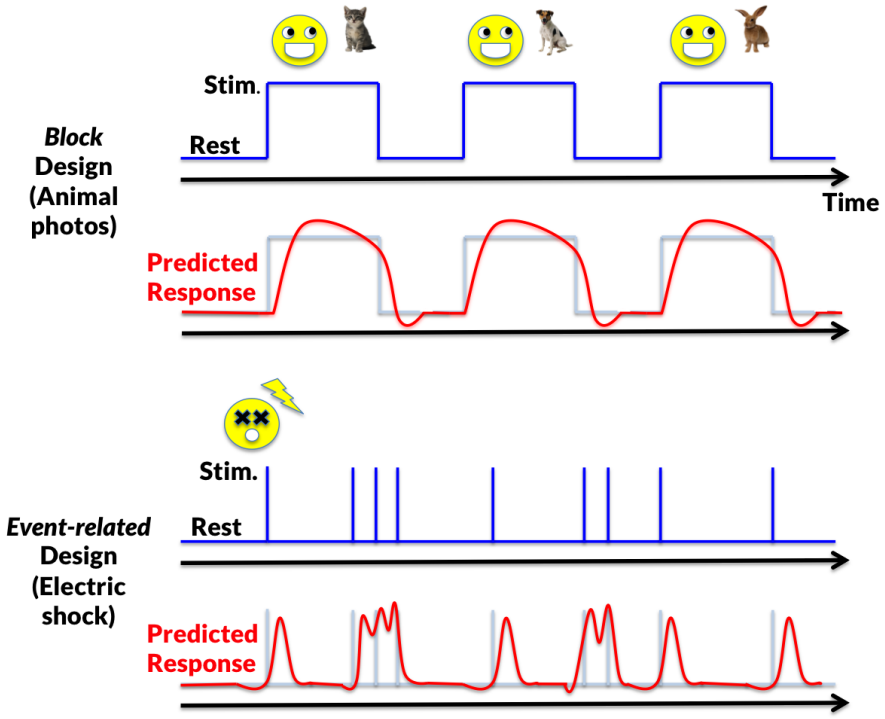


Figure 2.4: The stimulus onset timings (blue) and expected response (red) for a task paradigm using a block design where participants look at animal photos (top half), and a task paradigm using an event-related design where participants are given a mild electric shock (bottom half).

2.4 Overview of Analysis Pipeline

To analyze voxelwise t-fMRI BOLD data, a series of analysis steps are performed on the data in succession. Together, the complete chain of analysis procedures carried out is known as the analysis pipeline. Researchers have great flexibility as to how the analysis pipeline is comprised, with various options and adjustable parameters for each individual analysis step, and choices for the order in which certain procedures are conducted. Nevertheless, a standard analysis pipeline of t-fMRI data can be partitioned into three main stages: preprocessing, modelling, and statistical inference. In the upcoming sections we will describe the individual processing steps that are usually carried out in each of these analysis stages, while here we provide a brief overview.

The main goals of preprocessing are to reduce the severity of noise artefacts present in the raw BOLD fMRI data and to prepare the data for statistical analysis. At the modelling stage, a mass-univariate approach is adopted, whereby each voxel's functional time series data is considered independently as an instance of the general linear model (GLM) framework. Within the GLM, the contrasts discussed in the previous section are formulated to statistically test the main hypotheses investigated within the study. At the inference stage, a statistical parametric map is generated containing statistic values at each voxel for each contrast of interest. For subject-level inference, a participant's statistical parametric map is thresholded to display only voxels showing statistically significant results. This type of inference may be of interest in a clinical setting, particularly to aid in the diagnosis of a patient. However, for a research study there is usually a greater emphasis placed on finding results that generalize across the larger population. In this case, each participant's statistic map is entered into a second-level model for group-level inference, and a thresholded map is computed to localize effects that were consistent across all individuals in the study.

In practice, the analysis pipeline is usually conducted within a neuroimaging software package. Various software packages are available, many of which are freely distributed on the internet. The three most popular packages are AFNI (Cox, 1996), FSL (Jenkinson et al., 2012), and SPM (Penny et al., 2011). While there are several differences as to how each software package operates, most packages follow the same fundamental principles to implement the three main stages of the analysis pipeline.

2.5 Preprocessing

In this section, we present each of the analysis steps that are typically conducted within a t-fMRI preprocessing pipeline: brain extraction, distortion correction, slice timing correction, realignment, coregistration, spatial normalization, spatial smooth-

ing, temporal filtering and intensity normalization. These procedures are carried out to compensate for artefacts present in the data, and to ensure that the data satisfy the assumptions used for modelling and inference.

2.5.1 Brain Extraction

Brain extraction is commonly the first procedure carried out in the analysis pipeline, with the purpose of removing the skull and any other non-brain tissue from a participant's anatomical image. Since the purpose of the analysis is to infer areas of activation *within* the brain, it is sensible to remove any external structures that are not of interest. However, brain extraction also has a more important role in improving the outcome of subsequent steps in the preprocessing pipeline. In the upcoming sections we describe coregistration, where the subject's functional data is spatially realigned to the anatomical image, as well as spatial normalization, where each participant's data is registered to a standard space. Brain extraction helps to increase the robustness of both of these registration methods, since differences in non-brain structures can sidetrack the registration algorithms causing an inaccurate alignment of the respective images.

In SPM, brain extraction is carried out by first applying a segmentation to the anatomical image in order to generate probability maps of the gray and white matter tissue in the structural scan. The grey and white matter probability maps are summed and thresholded, creating a binary map containing the brain regions to be included in the analysis. Finally, the anatomical scan is masked with the binary map to remove any non-brain structures. AFNI and FSL both use variants of the Brain Extraction Tool (Smith, 2002) algorithm, implementing an adaptive model that evolves to fit the brain's surface in order to segment brain and non-brain tissue types.

2.5.2 Distortion Correction

Distortion correction is applied to account for signal loss and geometric distortions in the functional data that can manifest due to spatial inhomogeneities in the main static magnetic field during the acquisition. These inhomogeneities arise due to the different magnetic susceptibility properties of each tissue type in the brain, and the most severely affected regions are those close to air-filled synuses, such as the temporal or frontal lobe. If not corrected, signal dropout and distortion can cause failure in the registration of the functional data to the non-distorted anatomical image.

While it is not possible to recover regions of signal loss, field distortions can be rectified with the use of a field map. A field map is obtained as part of the acquisition to estimate the intensity of the static magnetic field. The analysis software uses the field map to calculate the magnitude of the geometric distortions, and then applies spatial transformations to unwarp the functional data. The field map is also used to de-weight areas of substantial signal dropout during the registration, and if the signal loss is particularly severe, ignore these locations in the analysis of the data.

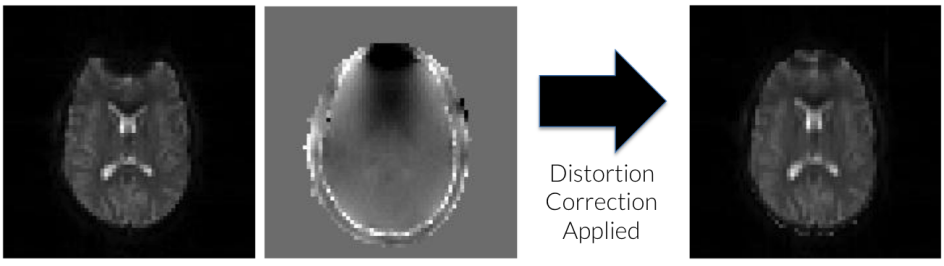


Figure 2.5: Distortion correction applied to functional data with the use of a field map. While lost signal can not be recovered, the correction has vastly improved distorted regions in the frontal lobe. Functional data and field map images reprinted from the *fMRIB Graduate Programme* lecture notes[1], with the kind permission of Mark Jenkinson.

[1] https://fsl.fmrib.ox.ac.uk/fslcourse/

2.5.3 Slice Timing Correction

While the statistical modelling of fMRI data assumes that the signal is measured over the entire brain simultaneously, in reality functional imaging is usually carried out on a slice-by-slice basis, creating a single 3D volume as a combination of multiple 2D slices. Slices are acquired sequentially from top-to-bottom or bottom-to-top, or by using an interleaved sequence where all odd-numbered slices are collected first, followed by the even slices. Because of this, the BOLD signal is sampled at different points of the HRF. This can create the illusion that the signal peaks earlier for slices that are collected later in the acquisition, even though the underlying response is identical.

Slice timing correction uses temporal interpolation methods to artificially obtain an intensity estimate for each brain voxel at a single time point, shifting the data to recreate the image as if all measurements were obtained collectively. The reference time point is commonly chosen to be halfway through the scanning procedure, and in this case the timings from all slices are corrected to match-up with the timings of the volume collected midway in the acquisition, which acts as the reference slice. The time series data from each voxel is temporally shifted to line-up with the signal response from the reference slice, and the voxel's data between the acquisition time points of the reference slice are re-estimated with interpolation. Commonly, this is done using either sinc or spline interpolation.

It is debatable whether slice timing correction should be conducted before or after realignment, and some practitioners have suggested that slice timing correction should be excluded from the analysis pipeline altogether. One alternative to slice timing correction is to account for timing differences at the modelling stage of the analysis. FSL recommend that temporal derivatives are incorporated as extra regressors into the GLM, effectively making the model flexible to temporal shifts in the signal response.

2.5.4 Realignment

While a participant is told to remain as still as possible in the scanner, over the course of the acquisition some head movement is inevitable. This is particularly problematic in fMRI, as it can corrupt the functional data in numerous ways. If left uncorrected, head movement may cause a voxel's time series data to contain signal from two different tissue types, and if the voxel is located at the edge of the brain, movement can cause a loss of signal altogether. Additionally, the change in signal intensity induced by head motion can be many orders of magnitude greater than the BOLD effect. Therefore, if head motion is elicited by the task the participant performs in the scanner this can lead to false activations in the statistical results that invalidate the analysis.

Realignment (or motion correction) of the functional data is performed to remove any substantial movement throughout the time series. To do this, each volume in the time series is spatially transformed to match a reference volume, usually chosen as the first volume of the data or an average image of all the scans. Specifically, a rigid-body transformation of translations and rotations is applied to superimpose each volume onto the reference image. The transformation is determined to optimize a cost function that quantifies the goodness of alignment between the images, e.g. a least squares (used by default in SPM) or normalized correlation (used by default in FSL) cost function. Finally, the transformed data are spatially interpolated to obtain estimates of the signal response on the same voxel grid as the reference image, usually with spline interpolation.

2.5.5 Coregistration

In order to carry out group analyses, corresponding voxels between each participant's functional data should contain information from the same anatomical location. However, prior to normalizing data *between* subjects, coregistration is conducted to

align a participant's functional time series data with their own anatomical image. Similar to realignment, coregistration is achieved via a rigid-body transformation chosen to minimize an appropriate cost function. However, to account for differences between the blurry, distorted functional data and the high-resolution structural image, scalings are also included as parameters of the rigid-body transformation and a mutual information cost function is commonly used.

2.5.6 Spatial Normalization

The goal of spatial normalization (or intersubject registration) is to warp all participants' functional time series data into a universal coordinate space, integrating the data between subjects to facilitate for group analyses. To remove structural variability between subjects, each participant's data is spatially transformed onto a standard template brain image. The most commonly used templates are the MNI152 images, created by the Montreal Neurological Institute by combining structural data from 152 healthy adults. The transformation is computed on a participant's structural image; the anatomy is registered to the template with a series of linear and non-linear transformations, permitting for local deformations to change the size and shape of the subject's structural image for a better alignment with the brain standard. Finally, the functional data are warped to standard space by concatenating the transformation from functional to structural space computed during coregistration with this transformation from structural to standard space.

2.5.7 Spatial Smoothing

Prior to statistical analyses, spatial smoothing is conducted on the functional data. Although this may seem unsound, as any smoothing will effectively reduce the spatial resolution of the fMRI data, the reasons for smoothing are twofold. The main reason for smoothing is to improve the SNR of the data by filtering out high-frequency regions. Intuitively, this works because averaging should reduce the intensity of noisy

areas, while leaving the underlying functional signal of interest relatively unaffected. The second reason is as a prerequisite for statistical analysis. Specifically, the methods used for parametric inference of fMRI data are adaptive in correcting for the multiple comparison problem dependent on the smoothness of the data (further details about this are provided in Section 2.7.1). However, a minimum amount of smoothing is required to obtain accurate control over the false discovery rate of activations in the thresholded statistical results.

In practice, the functional data are convolved with a three-dimensional Gaussian filter, and the amount of smoothing applied is proportional to the full width at half maximum (FWHM) of the kernel function. A suitable degree of smoothing is conditional on many factors, such as the quality of the data, the statistical power required, and the expected size of the final activation clusters. A typical smoothing kernel FWHM for fMRI data is between 6 and 10mm^3, although the preprocessing pipelines for recent high-quality, large-sample fMRI datasets have used a lesser degree of smoothing (e.g. 5mm FWHM for the UK Biobank, 4mm FWHM for the Human Connectome Project).

2.5.8 Temporal Filtering

Temporal filtering is another processing step that aims to increase the SNR of the functional data, by taking advantage of the fact that the BOLD signals fMRI sets out to measure generally have a consistent frequency range. Temporal filtering suppresses or removes frequencies outside of this range, implicitly eliminating any artefactual signals present in the data while leaving the neuronal signals of interest untouched.

A well-known source of noise is slow drifts that occur due to imperfections in the scanning hardware. As components of the scanner heat up, this can induce a gradual change in the MR signal resulting in low-frequency trends of less than 0.01 Hz in the data. The expected frequency of the BOLD signal response to a task stimulus is around 0.2Hz. Therefore, a high-pass filter can be applied, removing all frequen-

cies below a set threshold to attenuate scanner-related drifts. Other forms of noise are caused by physiological effects such as respiratory and cardiac cycles. These artefacts have a frequency range higher than the expected BOLD response (respiratory frequencies are ∼0.3hz, cardiac frequencies ∼1.0hz), although they may also manifest in the data as lower frequencies due to the effects of aliasing. A low-pass filter may be used to cut off higher frequencies and subdue artefacts such as physiological noise. While a high-pass filter is commonly included as part of the preprocessing pipeline, low-pass filters are more controversial as they can cultivate autocorrelation in the signal, violating the assumption of temporal independence made for statistical inference.

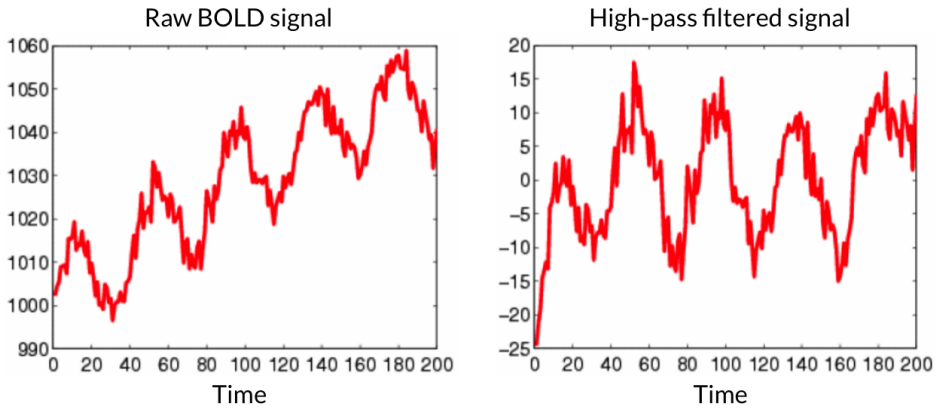


Figure 2.6: Showing the effect of high-pass filtering on one voxel's BOLD time series data. The high-pass filter has removed the slow drift seen in the raw BOLD signal on the left. Figure adapted from the *fMRIB Graduate Programme* lecture notes, with the kind permission of Mark Jenkinson.

2.5.9 Grand Mean Scaling

As touched on in Section 2.3, BOLD t-fMRI data are not quantitative. Because of this, the fMRI scanner assigns arbitrary units to the signal intensities during the acquisition, and data can be scaled differently across scanning sessions. Grand mean scaling (or intensity normalization) is applied to rescale each individual's functional time series to increase the interpretability of the data across the group of participants. This

is done by multiplying the functional time series (across all voxels and time points) by a constant so that the mean intensity takes a fixed value of, for example, 100. While grand mean scaling will not affect the statistical inference results, normalizing the data facilitates for comparability of the regression coefficient maps (i.e. beta maps) obtained for each task condition at the modelling stage of analysis.

2.6 Modelling of t-fMRI data with the General Linear Model

The general linear model (GLM) is the most widely used approach to modelling BOLD t-fMRI time series data, and a crucial part of any neuroimaging analysis. The GLM generalises a broad class of models that estimate the observed response as a linear combination of experimental and confounding variables. A key strength of this framework is its flexibility, allowing for analyses of data both within and between individuals, and providing a foundation for which experimental hypotheses can be assessed with a variety of statistical tests, using either parametric or nonparametric statistics. In this section we provide an overview of the GLM in the context of brain imaging, before describing some of the most commonly used statistical tests performed within the GLM for analysing fMRI data.

2.6.1 The GLM Set-up

To analyze voxelwise t-fMRI data, each voxel's time series is independently modelled within the GLM. This is commonly referred to as a mass-univariate analysis – the 'mass' term specifies that the same analysis is performed many times, and 'univariate' indicates that each analysis is performed separately at every brain voxel (as opposed to multivariate, which considers many locations as part of one analysis).

Mathematically, for a compact domain $S \subset \mathcal{R}^D$ (in fMRI, D = 3 and S is the brain mask), the GLM at location (or brain voxel) $\boldsymbol{s} \in S$ is expressed as

$$\boldsymbol{Y}(\boldsymbol{s}) = \boldsymbol{X}\boldsymbol{\beta}(\boldsymbol{s}) + \boldsymbol{\epsilon}(\boldsymbol{s}), \tag{2.1}$$

where $\boldsymbol{Y}(\boldsymbol{s})$ is a $N \times 1$ vector of observations at $\boldsymbol{s}$, $\boldsymbol{X}$ is a $N \times p$ design matrix containing explanatory variables linking the observations in $\boldsymbol{Y}(\boldsymbol{s})$ to the effect sizes in $\boldsymbol{\beta}(\boldsymbol{s})$, $\boldsymbol{\beta}(\boldsymbol{s})$ is a $p \times 1$ vector of the unknown parameters, and $\boldsymbol{\epsilon}(\boldsymbol{s})$ is a $N \times 1$ vector of error terms. It is assumed that the errors are independently distributed conditional on $\boldsymbol{X}$ by a Gaussian distribution with mean zero.

The aim of the regression is to find parameter estimates $\hat{\boldsymbol{\beta}}(\boldsymbol{s})$ that best fit the model to the data. The goodness of fit is determined by a method of least squares depending on additional constraints added to the model. The parameter estimates are then used at the inference stage to test hypotheses about the data expressed in terms of the unknown parameters contained in $\boldsymbol{\beta}(\boldsymbol{s})$.

2.6.2 Estimating the Parameters with Ordinary Least Squares (OLS)

OLS is used to solve (2.1) with the assumption that the errors are spherical, which means that there is no autocorrelation and that each error term has constant variance. Combined with the normality assumption stated in the previous section, this means

$$\boldsymbol{\epsilon}(\boldsymbol{s}) \mid \boldsymbol{X} \sim \mathcal{N}(0, \sigma^2(\boldsymbol{s})\boldsymbol{I}_N), \tag{2.2}$$

where $\boldsymbol{I}_N$ is the $N \times N$ identity matrix. OLS solves the GLM by minimizing the sum of squares cost function S given by

$$S(\boldsymbol{\beta}(s)) = \|\boldsymbol{Y}(\boldsymbol{s}) - \boldsymbol{X}\boldsymbol{\beta}(\boldsymbol{s})\|^2 = \sum_{i=1}^{N} |Y_i(\boldsymbol{s}) - \sum_{j=1}^{p} X_{ij}\beta_j(\boldsymbol{s})|^2. \tag{2.3}$$

This gives the OLS estimates

$$\hat{\boldsymbol{\beta}}(\boldsymbol{s}) = \arg\min_{\boldsymbol{\beta}} S(\boldsymbol{\beta}(\boldsymbol{s})) = (\boldsymbol{X}^\intercal \boldsymbol{X})^{-1}\boldsymbol{X}^\intercal \boldsymbol{Y}(\boldsymbol{s}). \tag{2.4}$$

By the Gauss-Markov Theorem, it can be shown that the OLS estimates are the Best Linear Unbiased Estimates (BLUE) of $\boldsymbol{\beta}(\boldsymbol{s})$ providing all the assumptions are satisfied.

2.6.3 Prewhitening

The key assumption of OLS is that the errors are spherical, however, this is often violated for fMRI data. As discussed in Section 2.5.8, functional data are characterized by slow drifts which induce temporal autocorrelation in the MR signal. While high-pass filtering can be applied in an attempt to remove the majority of low frequency components, another strategy is to estimate the autocorrelation directly and then remove it by prewhitening the data. This can be more efficient than filtering for event-related designs (Woolrich et al., 2001).

If the data are correlated, the error terms have marginal distribution

$$\boldsymbol{\epsilon}(\boldsymbol{s}) \mid \boldsymbol{X} \sim \mathcal{N}(0, \sigma^2(\boldsymbol{s})\boldsymbol{V}(\boldsymbol{s})), \tag{2.5}$$

where $\boldsymbol{V}(\boldsymbol{s})$ is the correlation matrix. Since $\boldsymbol{V}(\boldsymbol{s})$ is symmetric and positive-definite, $\boldsymbol{V}(\boldsymbol{s})$ satisfies the assumptions for the Cholesky decomposition, which means there exists a lower triangular matrix $\boldsymbol{K}(\boldsymbol{s})$ such that $\boldsymbol{V}^{-1}(\boldsymbol{s}) = \boldsymbol{K}^{\intercal}(\boldsymbol{s})\boldsymbol{K}(\boldsymbol{s})$. Providing that $\boldsymbol{K}(\boldsymbol{s})$ can be accurately determined, the idea is to update the model by multiplying both sides of the GLM by $\boldsymbol{K}(\boldsymbol{s})$ so that the error terms are spherical. Denoting $\boldsymbol{Y}^*(\boldsymbol{s}) = \boldsymbol{K}(\boldsymbol{s})\boldsymbol{Y}(\boldsymbol{s})$, and defining $\boldsymbol{X}^*(\boldsymbol{s})$ and $\boldsymbol{\epsilon}^*(\boldsymbol{s})$ similarly, then for the updated GLM

$$\boldsymbol{Y}^*(\boldsymbol{s}) = \boldsymbol{X}^*(\boldsymbol{s}) + \boldsymbol{\epsilon}^*(\boldsymbol{s}), \tag{2.6}$$

the conditional covariance of $\boldsymbol{\epsilon}^*(\boldsymbol{s})$ is

$$\mathrm{Cov}(\boldsymbol{\epsilon}^*(\boldsymbol{s}) \mid \boldsymbol{X}) = \boldsymbol{K}(\boldsymbol{s})\mathrm{Cov}(\boldsymbol{\epsilon}(\boldsymbol{s}) \mid \boldsymbol{X})\boldsymbol{K}^{\intercal}(\boldsymbol{s}) = \sigma^2(\boldsymbol{s})\boldsymbol{I}_N. \tag{2.7}$$

Therefore, the sphericity assumption is satisfied for the updated model, and OLS can be applied to obtain the BLUE of $\boldsymbol{\beta}(\boldsymbol{s})$.

2.6.4 Estimating the Variance

For statistical inference, the variance of the errors $\sigma^2(\boldsymbol{s})$ needs to be estimated. This can be done using the OLS estimates. The fitted values given by the OLS estimates are $\hat{\boldsymbol{Y}}(\boldsymbol{s}) = \boldsymbol{X}\hat{\boldsymbol{\beta}}(\boldsymbol{s})$. The differences between the observed data points and the fitted values are known as the residuals, denoted by $\hat{\boldsymbol{\epsilon}}(\boldsymbol{s}) = \boldsymbol{Y}(\boldsymbol{s}) - \hat{\boldsymbol{Y}}(\boldsymbol{s})$. The variance of the errors is estimated as the sum of squares of the residuals divided by the degrees of freedom of the model

$$\hat{\sigma}^2(\boldsymbol{s}) = \frac{\boldsymbol{\epsilon}^\intercal(\boldsymbol{s})\boldsymbol{\epsilon}(\boldsymbol{s})}{N-p}. \tag{2.8}$$

The degrees of freedom here are $N-p$, since there are N observations $Y_1(\boldsymbol{s}), ..., Y_N(\boldsymbol{s})$ and p parameters $\beta_1(\boldsymbol{s}), ..., \beta_p(\boldsymbol{s})$ to estimate.

2.6.5 Inference with Null-Hypothesis Significance Testing

The pay-off from obtaining the parameter estimates using OLS is that it enables us to statistically test hypotheses about the unknown effect sizes. This form of inference is known as null-hypothesis significance testing (NHST). Hypotheses are expressed using a contrast vector $\boldsymbol{c}$ to define a linear combination of the parameters. The null-hypothesis is always expressed in the form $H_0 : \boldsymbol{c}^\intercal\boldsymbol{\beta}(\boldsymbol{s}) = 0$, although this covers a wide variety of tests. For example, in a GLM with two parameters, $\boldsymbol{\beta}(\boldsymbol{s}) = (\beta_1(\boldsymbol{s}), \beta_2(\boldsymbol{s}))$, the contrast $\boldsymbol{c} = (1, 0)$ would lead to the null-hypothesis $H_0 : \beta_1(\boldsymbol{s}) = 0$, establishing a test to determine if the first parameter $\beta_1(\boldsymbol{s})$ is significantly different from zero. However, one could also test for significant differences between the two parameters by choosing the contrast vector $\boldsymbol{c} = (1, -1)$ to form the null-hypothesis $H_0 : \beta_1(\boldsymbol{s}) = \beta_2(\boldsymbol{s})$.

If the sphericity assumption is satisfied, then the distribution of $\boldsymbol{c}^\intercal\hat{\boldsymbol{\beta}}(\boldsymbol{s})$ is

$$\boldsymbol{c}^\intercal\hat{\boldsymbol{\beta}}(\boldsymbol{s}) \sim \mathcal{N}(\boldsymbol{c}^\intercal\boldsymbol{\beta}(\boldsymbol{s}), \sigma^2(\boldsymbol{s})\boldsymbol{c}^\intercal(\boldsymbol{X}^\intercal\boldsymbol{X})^{-1}\boldsymbol{c}). \tag{2.9}$$

Therefore, hypotheses about a contrast of the model parameters $\boldsymbol{c}^\intercal\hat{\boldsymbol{\beta}}(\boldsymbol{s})$ can be assessed with a t-test, using the knowledge that

$$\frac{\boldsymbol{c}^\intercal\hat{\boldsymbol{\beta}}(\boldsymbol{s}) - \boldsymbol{c}^\intercal\boldsymbol{\beta}(\boldsymbol{s})}{\sqrt{\hat{\sigma}^2(\boldsymbol{s})\boldsymbol{c}^\intercal(\boldsymbol{X}^\intercal\boldsymbol{X})^{-1}\boldsymbol{c}}} \sim t_{N-p}, \tag{2.10}$$

where t_{N-p} is a Student's t-distribution with $N - p$ degrees of freedom. In practice, the hypothesis $H_0 : \boldsymbol{c}^\intercal\boldsymbol{\beta}(\boldsymbol{s}) = 0$ is tested by computing the t-statistic

$$T(\boldsymbol{s}) = \frac{\boldsymbol{c}^\intercal\hat{\boldsymbol{\beta}}(\boldsymbol{s})}{\sqrt{\hat{\sigma}^2(\boldsymbol{s})\boldsymbol{c}^\intercal(\boldsymbol{X}^\intercal\boldsymbol{X})^{-1}\boldsymbol{c}}} \tag{2.11}$$

and then obtaining a p-value by comparing $T(\boldsymbol{s})$ to a t-distribution with $N - p$ degrees of freedom. For a one-sided hypothesis test where the alternative hypothesis is given by $H_A : \boldsymbol{c}^\intercal\boldsymbol{\beta}(\boldsymbol{s}) > 0$, the p-value is computed as $p = P(t_{N-p} \geq t)$. In fMRI, the statistical parametric map (or unthresholded statistic map) is an image containing the p-values computed at every voxel. A p-value is said to be statistically significant when $p < \alpha$, where α is a predetermined significance level set according to inference standards appropriate for the study (typically for fMRI, α is set at 5% *before* correction for multiple comparisons). In this case, the conclusion of the test is that there is sufficient evidence to reject the null-hypothesis in favour of the alternative. Thus, for the null $H_0 : \boldsymbol{c}^\intercal\boldsymbol{\beta}(\boldsymbol{s}) = 0$, a statistically significant p-value would suggest a non-zero effect size at location $\boldsymbol{s}$. A thresholded statistic map is obtained by masking the statistical parametric map to show only voxels with a statistically significant p-value.

In addition to testing a single contrast, one may also wish to test multiple contrasts at once. For example, in the two-parameter GLM described above, the null-hypothesis $H_0 : \beta_1(\boldsymbol{s}) = \beta_2(\boldsymbol{s}) = 0$ could be chosen to test for a significant effect size in either of the parameters. In this case, the contrast $\boldsymbol{c}$ is given as a matrix

$$\boldsymbol{c} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \tag{2.12}$$

where each row corresponds to each of the hypotheses being tested (for this example, $\beta_1(\boldsymbol{s}) = 0$ and $\beta_2(\boldsymbol{s}) = 0$ respectively). This time, inference is carried out using an F-test. The F-statistic is computed as

$$F(\boldsymbol{s}) = (\boldsymbol{c}^\intercal \hat{\boldsymbol{\beta}}(\boldsymbol{s}))^\intercal [r \boldsymbol{c}^\intercal \widehat{\mathrm{Cov}}(\hat{\boldsymbol{\beta}}(\boldsymbol{s}))\boldsymbol{c}]^{-1} (\boldsymbol{c}^\intercal \hat{\boldsymbol{\beta}}(\boldsymbol{s})), \tag{2.13}$$

where r is the rank of $\boldsymbol{c}$, and a p-value is obtained by comparing $F(\boldsymbol{s})$ to an F-distribution with r numerator and $N - p$ denominator degrees of freedom.

2.6.6 First-Level (Subject-Level) Analysis

In a first-level (or subject-level) analysis, the GLM set-up in Section 2.6.1 is used to analyze and test hypotheses related to the t-fMRI data obtained from an individual in a single scanning session.

At each brain voxel $\boldsymbol{s}$, the $N \times 1$ observations vector $\boldsymbol{Y}(\boldsymbol{s})$ contains the BOLD signal response data recorded by the scanner across all N sampled time-points during the session. The columns of the design matrix $\boldsymbol{X}(\boldsymbol{s})$ comprise of task-related and nuisance regressors to model the response in $\boldsymbol{Y}(\boldsymbol{s})$. The number of task-related regressors is dependent on the task paradigm and the statistical hypotheses the researcher wishes to test. For example, in the animal photo paradigm described at the end of Section 2.3, to test for activations when the participant was looking at any of the animal photos, a single regressor could be used to model the change in BOLD signal attributable to a photo being displayed. However, if instead the researcher wanted to test for changes in activation when the participant looked at photos of dogs compared to photos of cats, then multiple regressors would need to be used for pictures of each animal type.

The predicted response of each task-related regressor is estimated by convolving the onset timing function for the stimulus with the HRF. Figure 2.7 shows how the predicted response is obtained for the block design used in the animal photo

task. There are a variety of ways to model the HRF; by default FSL and SPM use a single canonical HRF, however alternative methods include use of a basis set of smooth functions (Friston et al., 1998) or a more flexible finite impulse response basis set (Goutte et al., 2000). For each task-related condition, temporal derivatives may be included as an additional regressor to account for differences between the actual and modelled HRF. See Lindquist et al. (2009) for a comparison of different HRF models.

The time series of motion-related parameters (e.g. translations and rotations) used for realignment are commonly added to the GLM as nuisance regressors to compensate for any leftover motion artefacts in the signal after preprocessing. Further nuisance regressors may include cardiac and respiratory recordings to account for fluctuations in the BOLD signal caused by changes in heart rate and breathing patterns during the scan.

To remove temporal autocorrelation in the BOLD signal the data are whitened (as in Section 2.6.3), after which unbiased estimators of the parameters in $\boldsymbol{\beta}(\boldsymbol{s})$ can be computed via OLS (as in Section 2.6.2). Finally, changes in neuronal activity between the task conditions is tested by contrasting the task-related regressors with an appropriate contrast vector $\boldsymbol{c}$ under the null-hypothesis $H_0 : \boldsymbol{c}^\intercal \boldsymbol{\beta}(\boldsymbol{s}) = 0$, using the NHST procedure described in Section 2.6.5.

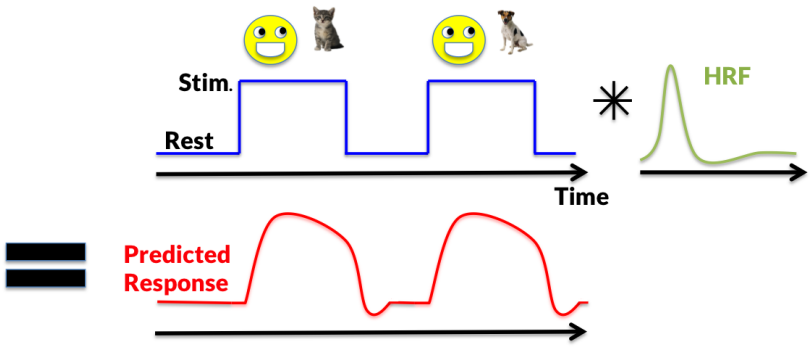


Figure 2.7: Showing how the predicted response for the animal photo task in Figure 2.4 is created by convolution of the onset timing function with the HRF.

2.6.7 Second-Level (Group-Level) Analysis

The contrast parameter estimates obtained for each individual during the first-level analysis are combined in a second-level (or group-level) analysis to test whether results can be generalized to the larger population. The second-level model aims to account for variation in quantity estimates within each individual's data, as well as variability *between* the participants stemming from biological differences between each subject. This is fundamental from a research perspective as it means that conclusions drawn from a second-level analysis can relate to the whole population from which the participants came from, rather than just the the specific cohort of individuals involved in the study. In this section we will describe a simplified second-level model for a one-sample t-test used to test for a consistent response in a first-level contrast effect across all individuals. However, within the GLM framework it is also possible to conduct a *between*-groups analysis with a two-sample t-test to assess activation differences in two distinct cohorts (e.g. a group of patients and a group of controls) and to incorporate other design types such as a paired t-test or one-way analysis of variance.

For the GLM set-up in Section 2.6.1, in a group-level analysis N represents the total number of participants in the study cohort. For a pre-specified contrast, the $N \times 1$ observations vector $\boldsymbol{Y}(\boldsymbol{s})$ contains each subject's first-level contrast effect, i.e. $\boldsymbol{Y}(\boldsymbol{s}) = \boldsymbol{c}^\intercal \boldsymbol{\beta}(\boldsymbol{s}) = [\boldsymbol{c}^\intercal \beta_1(\boldsymbol{s}), ..., \boldsymbol{c}^\intercal \beta_N(\boldsymbol{s})]^\intercal$. This gives the instance of the GLM

$$\boldsymbol{c}^\intercal \boldsymbol{\beta}(\boldsymbol{s}) = \boldsymbol{X}\boldsymbol{\beta}_G(\boldsymbol{s}) + \boldsymbol{\epsilon}(\boldsymbol{s}), \tag{2.14}$$

where the group-level parameters are notated with $\boldsymbol{\beta}_G(\boldsymbol{s})$ to distinguish from the subject-level parameters $\boldsymbol{\beta}(\boldsymbol{s})$. For a one-sample t-test the first column of the design matrix $\boldsymbol{X}$ is an intercept where all elements are set equal to 1. Other columns of the design are covariates to be considered for the analysis, such as the age of each participant or a score to describe each individual's performance in the scanning task.

For simplicity we will assume no covariates are included, so $\boldsymbol{X}(\boldsymbol{s})$ is a column vector of 1s. $\boldsymbol{\epsilon}(\boldsymbol{s})$ is the vector of group-level errors terms, assumed to have distribution

$$\boldsymbol{\epsilon}(\boldsymbol{s}) \sim \mathcal{N}(0, \sigma_G^2(\boldsymbol{s})\boldsymbol{I}_N), \tag{2.15}$$

where $\sigma_G^2(\boldsymbol{s})$ is the between-subject variance for the group.

Of course, in any practical analysis scenario the true subject-level contrast effects $\boldsymbol{c}^\intercal\boldsymbol{\beta}(\boldsymbol{s})$ are unknown, and therefore the contrast estimates obtained in the first-level analyses $\boldsymbol{c}^\intercal\hat{\boldsymbol{\beta}}(\boldsymbol{s})$ must be used instead. This leads to the GLM formulation

$$\boldsymbol{c}^\intercal\hat{\boldsymbol{\beta}}(\boldsymbol{s}) = \boldsymbol{X}\boldsymbol{\beta}_G(\boldsymbol{s}) + \boldsymbol{\epsilon}^*(\boldsymbol{s}), \tag{2.16}$$

where the errors $\boldsymbol{\epsilon}^*(\boldsymbol{s})$ must now account for within-subject variation caused by estimating each participant's parameter effect size as well as the between-subject variance. Rearranging (2.14) and (2.16),

$$\boldsymbol{\epsilon}^*(\boldsymbol{s}) = \boldsymbol{c}^\intercal\boldsymbol{\beta}(\boldsymbol{s}) - \boldsymbol{c}^\intercal\hat{\boldsymbol{\beta}}(\boldsymbol{s}) + \boldsymbol{\epsilon}(\boldsymbol{s}), \tag{2.17}$$

and therefore

$$\begin{aligned}\mathrm{Cov}(\boldsymbol{\epsilon}^*(\boldsymbol{s})) &= \mathrm{Cov}(\boldsymbol{c}^\intercal\hat{\boldsymbol{\beta}}(\boldsymbol{s})) + \sigma_G^2(\boldsymbol{s})\boldsymbol{I}_N \\ &= \underbrace{\begin{pmatrix} \sigma_1^2(\boldsymbol{s}) & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & \sigma_N^2(\boldsymbol{s}) \end{pmatrix}}_{\text{within-subject variance}} + \underbrace{\sigma_G^2(\boldsymbol{s})\boldsymbol{I}_N}_{\text{between-subject variance}},\end{aligned} \tag{2.18}$$

where $\sigma_i^2(\boldsymbol{s})$ is the within-subject variance for the ith subject.

From this point on, two approaches are frequently used in the fMRI literature. The first approach assumes that the within-subject variances are equal across indi-

viduals so that the errors are homoscedastic, meaning that each error term has the same variance. The second approach relaxes this assumption, allowing the within-subject variances to differ between individuals. In this case, the variance of the error terms also differ, meaning that the error terms are hetroscedastic.

2.6.8 Solving the Second-Level GLM with Homoscedastic Errors

Assuming that the within-subject variance terms are equal, $\sigma_1^2(\boldsymbol{s}) = \ldots = \sigma_N^2(\boldsymbol{s})$, (2.18) reduces to

$$\mathrm{Cov}(\boldsymbol{\epsilon}^*(\boldsymbol{s})) = (\sigma_S^2(\boldsymbol{s}) + \sigma_G^2(\boldsymbol{s}))\boldsymbol{I}_N, \tag{2.19}$$

where $\sigma_S^2(\boldsymbol{s})$ is the common within-subject variance. Since this means the errors are spherical, the group-level effect estimates can be obtained using OLS as described in Section 2.6.2. For the one-sample t-test where $\boldsymbol{X}$ is a column vector with all elements equal to one, this leads to the parameter estimates

$$\hat{\boldsymbol{\beta}}_G(\boldsymbol{s}) = \frac{1}{N}\sum_{i=1}^{N}\hat{\beta}_i(\boldsymbol{s}). \tag{2.20}$$

2.6.9 Solving the Second-Level GLM with Hetroscedastic Errors

Assuming that the within-subject variance terms are not equal, then the errors are hetroscedastic and therefore the sphericity assumption required for OLS is violated. In this case, a weighted least square (WLS) approach is used to solve the GLM instead. Conceptually, the idea of WLS is to de-weight the most variable subject-level parameter estimates contained in $\boldsymbol{c}^\intercal\hat{\boldsymbol{\beta}}(\boldsymbol{s})$. In practice, this is done by constructing a weight matrix $\boldsymbol{W}(\boldsymbol{s})$ equal to the inverse of the covariance matrix of the observations. In the

context of the second-level analysis, (2.18) leads to the weight matrix

$$\boldsymbol{W}(\boldsymbol{s}) = \begin{pmatrix} (\sigma_1^2(\boldsymbol{s}) + \sigma_G^2(\boldsymbol{s}))^{-1} & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & (\sigma_N^2(\boldsymbol{s}) + \sigma_G^2(\boldsymbol{s}))^{-1} \end{pmatrix}. \tag{2.21}$$

The WLS parameter estimates are then computed as

$$\hat{\boldsymbol{\beta}}_{WLS}(\boldsymbol{s}) = (\boldsymbol{X}^\intercal \boldsymbol{W}(\boldsymbol{s}) \boldsymbol{X})^{-1} \boldsymbol{X}^\intercal \boldsymbol{W}(\boldsymbol{s}) \boldsymbol{c}^\intercal \hat{\boldsymbol{\beta}}(\boldsymbol{s}), \tag{2.22}$$

which for the one-sample t-test example reduces to

$$\hat{\boldsymbol{\beta}}_{WLS}(\boldsymbol{s}) = \left(\sum_{i=1}^{N} \frac{1}{\sigma_i^2(\boldsymbol{s}) + \sigma_G^2(\boldsymbol{s})} \right)^{-1} \sum_{j=1}^{N} \frac{\boldsymbol{c}^\intercal \hat{\boldsymbol{\beta}}_j(\boldsymbol{s})}{\sigma_j^2(\boldsymbol{s}) + \sigma_G^2(\boldsymbol{s})}. \tag{2.23}$$

In the WLS approach to solving the GLM, the weight matrix $\boldsymbol{W}(\boldsymbol{s})$ effectively applies a whitening transformation to the data. In fact, it can be advantageous to simplify WLS by using a similar procedure as described in Section 2.6.3 for prewhitening. Using the same notation as Section 2.6.3, the error terms in the group-level GLM can be whitened using the diagonal matrix $\boldsymbol{K}(\boldsymbol{s})$, where $K_{ii}(\boldsymbol{s}) = \sqrt{W_{ii}(\boldsymbol{s})}$. The updated model

$$\boldsymbol{K}(\boldsymbol{s}) \boldsymbol{c}^\intercal \hat{\boldsymbol{\beta}}(\boldsymbol{s}) = \boldsymbol{K}(\boldsymbol{s}) \boldsymbol{X} \boldsymbol{\beta}_G(\boldsymbol{s}) + \boldsymbol{K}(\boldsymbol{s}) \boldsymbol{\epsilon}^*(\boldsymbol{s}) \tag{2.24}$$

satisfies the sphericity assumption and can therefore be solved using OLS. The parameter estimates obtained with this method are equivalent to solving the original model directly with WLS.

In this presentation, we have assumed that the within- and between-subject variance components are known, when in practice they must be estimated. Usually, the within-subject variance estimates obtained from the first-level analyses are also used in the group-level model. There are many proposed methods for estimating the

between-subject variance, several of which use an iterative procedure based on OLS or residual maximum likelihood estimates. For more, see (Searle et al., 2009; Woolrich et al., 2004; Worsley et al., 2000).

2.7 The Multiple Comparisons Problem

A fundamental issue with the mass-univariate approach for fMRI inference is the multiple comparisons problem. Because the GLM is applied at each voxel independently, across the entire brain mask this means thousands of statistical tests are performed simultaneously. The significance level for each test can be described as the probability of wrongly determining a 'discovery' when the null-hypothesis of no activation is in fact true. In statistical terms, this is the probability of making a type I error. The problem is, while a significance level of 5% may be appropriate for one test, as more inferences are carried out the probability of an erroneous inference also increases. To highlight this point, if 100 tests are performed independently, each with a significance level of 5%, then the probability of at least one false discovery is 99.4%. In the context of fMRI, where a typical brain mask contains over 100,000 voxels, this problem is especially severe – if the voxelwise significance level of 5% is not corrected, we can expect over 5,000 brain voxels to falsely be determined as active in the thresholded statistical results.

In the task-fMRI literature, two different approaches are commonly used to correct for the multiple comparisons problem. The false discovery rate (FDR) procedure corrects the significance level to control the expected proportion of type I errors across all detected voxels in the thresholded statistical results. For example, if a 0.05 FDR procedure is applied for subject-level inference on 20 subjects' data, in any individual's thresholded statistic image we expect 5% of activations to be false-positives. The second approach is to employ a familywise error (FWE) correction, used to control the expected frequency that *any* type I errors are made across the

whole brain. This is a more stringent form of correction – in the 20-subject example, using a 0.05 FWE procedure we expect only one individual's thresholded image to contain any false-positives.

In making a choice between FDR or FWE correction there is a trade-off. While FDR is more statistically powerful than FWE, the drawback is a greater risk of false activations. This has led to criticism concerning the spatial specificity of FDR-corrected results; particularly, since the location of false activations in an FDR-thresholded map are unknown, it is not possible to say with certainty that any given voxel is activated. On the other hand, FWE correction has also come under fire for being too conservative.

Multiple comparison correction procedures are carried out in fMRI at either the voxelwise or clusterwise level. Voxelwise inference is intuitive, once a corrected threshold has been determined via the FDR or FWE procedure, the thresholded statistical results are computed as all voxels whose t-statistic value exceeds the corrected threshold. Clusterwise inference involves a two-step procedure. First, a primary voxelwise threshold c is chosen, usually in correspondence with an uncorrected significance level such as $\alpha = 0.005$. Thresholding the statistic map at c creates groups of contiguous voxels above c, or 'clusters'. For this reason, c is commonly referred to as the cluster-forming threshold. Subsequently, a cluster-extent threshold k is determined based on the distribution of cluster sizes obtained under the null-hypothesis of no activation. The final thresholded results are computed as all suprathreshold clusters with a spatial extent larger than k.

FDR and FWE correction procedures can be applied for both voxelwise and clusterwise inference, with the key difference that voxelwise corrections are based on the sampling distribution of voxel intensities under the null-hypothesis of no signal while clusterwise corrections use the sampling distribution of cluster size. In the remainder of this section, we consider two methods for obtaining FWE-corrected results at the voxelwise level. First, we present how the body of mathematics called

Random Field Theory (RFT) is used to treat FWE correction with a parametric approach. Then, we show how FWE correction may also be obtained with permutation methods that make weaker assumptions about the data. While we will not consider FDR here, further information can be found in Benjamini and Hochberg (1995) where the method was originally proposed, or Genovese, Lazar, and Nichols (2002) for a presentation of FDR in the context of functional neuroimaging.

2.7.1 Random Field Theory for Voxelwise FWE Correction

Voxelwise FWE control is established using an intrinsic relationship linking the probability of making a false discovery with the distribution of the maximum t-statistic over the brain. For a threshold c,

$$P(\text{Reject } H_0 | H_0 \text{ true}) = P\Big(\max_{\boldsymbol{s} \in S} T(\boldsymbol{s}) > c \Big), \tag{2.25}$$

where $T(\boldsymbol{s})$ is the t-statistic map given in (2.11) and H_0 is the null-hypothesis of no activation.

The intuition of RFT is that under the global null of no signal anywhere, $T(\boldsymbol{s})$ can be modelled by a stationary continuous Gaussian random field $Z(\boldsymbol{s})$ of mean zero and unit variance over the same domain S. Note that this imposes assumptions that the data are stationary and that the discretely sampled statistic image $T(\boldsymbol{s})$ is sufficiently smooth enough to be approximated by a continuous random field. It is because of the latter that spatial smoothing of the data must be carried out during preprocessing. Supposing that the model is valid, then the right-hand side of (2.25) is approximated by $P\Big(\max_{\boldsymbol{s} \in S} Z(\boldsymbol{s}) > c \Big)$. Remarkably, this probability can be obtained with use of the Euler characteristic (EC), a measure originating from algebraic topology that provides information about a shape's fundamental structure irrespective of how the shape is distorted or deformed.

For a threshold c, defining the excursion set A_c as the set of voxels where the

random field exceeds c, i.e. $A_c = \Big\{ \boldsymbol{s} \in S : Z(\boldsymbol{s}) > c \Big\}$, the EC χ_c can be characterized as counting the number of clusters, minus the number of 'handles', plus the number of 'holes' in A_c. If c is chosen large enough then we expect the handles and holes to disappear, so the EC provides an approximation of the number of clusters. This relates back to the maximum distribution, because if $\max_{\boldsymbol{s} \in S} Z(\boldsymbol{s}) > c$, then clearly there must be at least one suprathreshold cluster in the excursion. Putting all this information together,

$$\begin{aligned} P\Big(\max_{\boldsymbol{s} \in S} T(\boldsymbol{s}) > c \Big) &\approx P\Big(\max_{\boldsymbol{s} \in S} Z(\boldsymbol{s}) > c \Big) \\ &\approx P\Big(\chi_c > 0 \Big). \end{aligned} \tag{2.26}$$

Finally, further increasing of the threshold c will result in fewer voxels contained in the excursion set until χ_c will almost certainly take the value of 1 (if the excursion set is made up of one suprathreshold cluster) or 0 (if the excursion set is empty). In this case,

$$P\Big(\chi_c > 0 \Big) = \mathbb{E}\Big(\chi_c \Big). \tag{2.27}$$

In practice, with a parametric approach the FWE-corrected p-value is always approximated using the expectation of the EC. When $S \subset \mathcal{R}^3$, e.g. S is the brain mask in a neuroimaging application, the expected EC has the closed-form solution

$$\mathbb{E}\Big(\chi_c \Big) \approx \lambda(S) |\Lambda|^{\frac{1}{2}} (c^2 - 1) \exp\left(-c^2/2\right) / (2\pi)^2, \tag{2.28}$$

where $\lambda(S)$ is the Lebesgue measure (i.e. the volume) of the brain mask and $|\Lambda|$ is the determinant of the covariance matrix of the gradient of $Z(\boldsymbol{s})$,

$$|\Lambda| = \left| \mathrm{Cov}\left(\left[\frac{\partial}{\partial x} Z(\boldsymbol{s}), \frac{\partial}{\partial y} Z(\boldsymbol{s}), \frac{\partial}{\partial z} Z(\boldsymbol{s}) \right] \right) \right|. \tag{2.29}$$

Essentially, $|\Lambda|$ provides a measure of the smoothness of the random field $Z(\boldsymbol{s})$; for a less smooth process $Z(\boldsymbol{s})$, the determinant is larger.

2.7.2 Permutation Testing for Voxelwise FWE Correction

In the previous section, we showed that the crux of estimating the FWE-corrected p-value is to approximate the maximum distribution of $T(\boldsymbol{s})$. We demonstrated how this was carried out with a parametric approach by assuming that under the global null, $T(\boldsymbol{s})$ can be modelled by a Gaussian random field. However, this imposed strong assumptions on the data which in practice are seldom fulfilled. In particular, it has been shown that RFT estimates of the FWE p-value are conservative unless the data are extremely smooth with high degrees of freedom. To remediate these problems, nonparametric methods have been proposed as an alternative.

The principle idea of permutation testing for FWE correction is that if the global null-hypothesis is true, i.e. there is really no signal anywhere, then the labels of each observation are arbitrary and the data is exchangeable. Therefore, the maximum distribution of $T(\boldsymbol{s})$ under the global null can be constructed empirically by creating a large number of surrogate realizations of the data, where on each realization the data labels are permuted randomly and the maximum value is obtained from the corresponding statistic map.

In the first-level analysis model described in Section 2.6.6, where the observations are fMRI time series data from one individual, permutation testing is usually inappropriate; due to temporal correlation in the data, the exchangeability assumption is violated. However, permutation testing is viable for the group-level model in Section 2.6.7, where the observations are first-level contrast of parameter estimates (cope) maps $\boldsymbol{c}^\intercal\hat{\beta}_1(\boldsymbol{s}), ..., \boldsymbol{c}^\intercal\hat{\beta}_N(\boldsymbol{s})$ obtained from each individual in the cohort.

In a two-sample t-test where cope maps are obtained from two groups, e.g. a group of patients and a group of controls, permutation testing is conducted by exchanging the labels between the groups. For a one-sample t-test, where there are no

group labels to swap, the principle of exchangability is replaced by an assumption of symmetry; if the null-hypothesis $H_0 : \boldsymbol{c}^\intercal \boldsymbol{\beta}(\boldsymbol{s}) = 0$ is true, then it should not matter if a change of sign is applied to all readings in any individual contrast image. In this case, a permutation test is carried out where each surrogate realization is established by randomly multiplying each individual contrast image $\boldsymbol{c}^\intercal \hat{\beta}_i(\boldsymbol{s})$ by 1 or -1. Clearly, the assumption of symmetry for the noise distribution is much weaker than Gaussianity required for RFT with the parametric approach.

In full, the voxelwise FWE-corrected p-value for a second-level one-sample t-test is obtained via permutation test using the following algorithm:

1. Let P be a large number of permutations that will be carried out for the permutation test. A larger P will provide a more precise approximation of the empirical null distribution, commonly $P = 10,000$ is used.

2. Create surrogate observations of the data $\boldsymbol{c}^\intercal \hat{\beta}_1^*(\boldsymbol{s}), ..., \boldsymbol{c}^\intercal \hat{\beta}_N^*(\boldsymbol{s})$, where each $\boldsymbol{c}^\intercal \hat{\beta}_j^*(\boldsymbol{s})$ is obtained by multiplying $\boldsymbol{c}^\intercal \hat{\beta}_j(\boldsymbol{s})$ randomly by either 1 or -1. Specifically, $\boldsymbol{c}^\intercal \hat{\beta}_j^*(\boldsymbol{s}) = r_j \boldsymbol{c}^\intercal \hat{\beta}_j(\boldsymbol{s})$, where the r_j are independent and identically distributed Rademacher variables (that is, $r_j = 1$ or -1, each with probability $1/2$).

3. Obtain the group-level statistic map $T^*(\boldsymbol{s})$ for the surrogate data, and compute the maximum statistic value across the image $t^* = max_{\boldsymbol{s} \in S} T^*(\boldsymbol{s})$.

4. Repeat steps 2 and 3 P times to create an empirical distribution $t_1^*, ..., t_P^*$ of the maximum statistic.

5. Assuming that $t_1^*, ..., t_P^*$ are ordered from smallest to largest (otherwise, rearrange the labelling so this is true), for a desired FWE rate of α, choose $c = t^*_{\lceil (1-\alpha)P \rceil}$, where $\lceil (1-\alpha)P \rceil$ is the smallest integer greater than $(1-\alpha)P$.

6. With this construction, it can be shown that c is the corrected p-value for a voxelwise FWE rate of α.

2.8 Conclusion

In this section, we provided the background that will form the basis of our research. We started with a general discussion of what it means to study brain function, before providing an overview of the biological phenomena behind the BOLD effect measured in a functional MRI study. In the rest of this section we gave an in-depth overview of task fMRI, which will be the main field of investigation in this thesis. We described the fundamental preprocessing steps carried out in a t-fMRI analysis, as well as the most common methods used to model subject- and group-level fMRI data within the general linear model. Finally, we described the multiple comparison problem for task fMRI inference, and then demonstrated how a parametic and nonparametric approach can be used to control the voxelwise familywise error rate in the thresholded statistic result maps computed at the end of a t-fMRI analysis.

CHAPTER 3

Exploring the Impact of Analysis Software on Task-fMRI Results

Functional magnetic resonance imaging (fMRI) for human brain mapping has given researchers remarkable power to probe the underpinnings of human cognition, behaviour, and emotion. As the field has evolved over the last 25 years, so to have the number of tools and techniques available to researchers, expanding our potential to gain insight into how the human brain works. However, this plurality of analysis methods has also lead to its own complications. The problem is, with numerous valid methodological strategies for analyzing a dataset, this also increases the chance of yielding distorted results with inflated levels of false activations (Hong et al., 2019; Ioannidis, 2005; Wager et al., 2009). Combined with selective reporting issues such as publication bias, where there has been evidence to suggest that studies with significant findings are disproportionately represented in the fMRI literature (David et al., 2013; Ioannidis et al., 2014), these conditions have created the perfect storm. In recent years, many attempts to replicate the results of published fMRI studies have been unsuccessful, in what has been deemed an ongoing reproducibility crisis within the field (Poldrack et al., 2017; Gorgolewski and Poldrack, 2016; Open Science Collaboration, 2015).

The choice of procedures, operating system, and software version implemented to conduct an analysis have all been shown to induce variability in the final research

results of an investigation. In a study examining the use of FreeSurfer to measure the cortical thickness and volume of structural brain images (Gronenschild et al., 2012), a change in software version was shown to have led to increases of over 10% in the observed anatomical measurement; a switch in workstation from which the software was run also manifested significant deviations in the final result. In related work (Glatard et al., 2015), changes in operating system led to differences in the results of an independent component analysis of resting state fMRI data carried out using FSL. Here, disparities in both the number of components determined as well as information between matched components were found when the analysis was conducted on two separate computing clusters.

In Chapter 1, we referenced a handful of studies where different choices in individual analysis procedures were found to influence the final activation maps for a task-fMRI investigation. However, in each of these studies the software package through which the analyses were carried out remained fixed. This is despite the fact that a variety of analysis packages are used throughout the fMRI literature, the most popular of which are AFNI (RRID:SCR_005927; (Cox, 1996)), FSL (RRID:SCR_002823; (Jenkinson et al., 2012)), and SPM (RRID:SCR_007037; (Penny et al., 2011)). While SPM is the oldest, FSL has grown in popularity and together the three packages have been estimated to account for 80% of published functional neuroimaging results (Carp, 2012b). Although there are differences in how each software package models and processes data, the analysis framework for task-fMRI – now a mature research area – is expected to be similar across software, and hence the results each package yields should be comparable. We therefore seek to answer the question: How much of the variability in neuroimaging results is attributable to the choice of analysis software package?

In this chapter we reanalyze data from three published neuroimaging studies using each of the three main software packages and quantify differences in the results. We choose three publications with data that have been made publicly available

on the OpenfMRI database (RRID:SCR_005031, http://openfmri.org; (Poldrack et al., 2013)), recently relaunched as OpenNeuro (http://openneuro.org), and attempt to recreate the main figure from each publication by replicating the original analysis within each package. These particular studies were selected on the basis that they reported clearly defined regions of brain activation and utilized analysis procedures feasible across the three software packages. We then make a number of comparisons to assess the similarity of our results. While a similar study from our group explored the results produced by each of these packages after implementing analysis pipelines using the default settings in each software (Pauli et al., 2016), here we attempt to make the analysis pipelines as similar as possible while still maintaining comparability across the three packages. While our primary focus is comparing standard results across software, we also aim to address recent concerns about the multiple-testing-corrected parametric inferences that each of these studies used (Eklund et al., 2016). For each study, we conduct equivalent inference procedures (when possible) using nonparametric statistics within each package, and then make intra-software comparisons between the parametric and nonparametric results.

All aspects of this work were carried out under the supervision and mentorship of Prof. Thomas Nichols and Dr. Camille Maumet. Dr. Maumet made additional contributions to the data analysis code and the Python Jupyter Notebooks used to create the figures for this work.

3.1 Data and Analysis Methods

3.1.1 Study Description and Data Source

We selected three t-fMRI studies for reanalysis from the publicly accessible OpenfMRI data repository: ds000001 (Revision: 2.0.4; (Schonberg et al., 2012)), ds000109 (Revision 2.0.2; (Moran et al., 2012)), and ds000120 (Revision 1.0.0; (Padmanabhan et al., 2011)). Each of the datasets had been organized in compliance with the Brain Imaging Data Structure (BIDS, RRID:SCR_016124; (Gorgolewski et al., 2016)). These datasets were chosen following an extensive selection procedure (carried out between May 2016 and November 2016), whereby we vetted the associated publication for each dataset stored in the repository. We sought to find studies with simple analysis pipelines and clearly reported regions of brain activation that would be easily comparable to our own results. Exclusion criteria included the use of custom software, activations defined using small volume correction, and application of more intricate methods such as region of interest and robust regression analysis, which we believed could be impractical to implement across all analysis software. A full description of the paradigm for each of our chosen studies is included in the respective publication, here we give a brief overview.

For the ds000001 study, 16 healthy adult subjects participated in a balloon analog risk task over three scanning sessions. On each trial, subjects were presented with a simulated balloon, and offered a monetary reward to 'pump' the balloon. With each successive pump the money would accumulate, and at each stage of the trial subjects had a choice of whether they wished to pump again or cash-out. After a certain number of pumps, which varied between trials, the balloon exploded. If subjects had cashed-out before this point they were rewarded with all the money they had earned during the trial, however if the balloon exploded all money accumulated was lost. Three different coloured 'reward' balloons were used between trials, each having a different explosion probability, as well as a gray 'control' balloon, which had no

monetary value and would disappear from the screen after a predetermined number of pumps. Here we reproduce the result contrasting the parametrically modulated activations of pumps of the reward balloons versus pumps of the control balloon, corresponding to Figure 3 and Table 2 in the original paper.

The ds000109 study investigated the ability of people from different age-groups to understand the mental state of others. A total of 48 subjects were scanned, although 43 had acceptable data for the false belief task: 29 younger adults and 14 older adults. In this task participants listened to either a 'false belief' or 'false photo' story. A false belief story would entail an object being moved from one place to another, with certain characters witnessing the change in location while others were unaware. False photo stories were similar except that they involved some physical representation of the missing object, such as a photo of an object in a location from which it had been subsequently removed. The task had a block design where stories were represented for ten seconds, after which participants had to answer a question about one of the characters' perceptions of the location of the object. We reproduce the contrast map of false belief versus false photo activations for the younger adults, corresponding to Figure 5a and Table 3 from the original publication.

Finally, the ds000120 study explored reward processing across different age groups. fMRI results were reported on 30 subjects, with 10 participants belonging to each of the three age groups (children, adolescents and adults). Participants took part in an antisaccade task where a visual stimulus was presented in each trial and subjects were instructed to quickly fixate their gaze on the side of the screen opposite to the stimulus. Prior to a trial, subjects were given a visual cue to signal whether or not they had the potential to win a monetary reward based on their upcoming performance (a 'reward' or 'neutral' trial). We reproduce the main effect of time activation map, an F-statistic for any non-zero coefficients in the sine HRF basis, corresponding to Figure 3 and Table 1 in the original publication.

3.1.2 Data Analyses

All data analyses were conducted using AFNI (version AFNI_18.1.09), FSL (version 5.0.10), and SPM (version SPM12, v6906). Computation was performed on a cluster comprised of 12 Dell PowerEdge servers (6 R410, 12 core 2.40GHz processors, 6 R420, 12 core 2.80GHz processors) running CentOS 7.3.

Pipeline

A full decomposition of the pipelines implemented within the three packages for each study is presented in Table 3.1. Here, we give a brief description of the procedures.

In AFNI, preprocessing and subject-level analyses were conducted using the @SSwarper program and afni_proc.py. For ds000001 and ds000109, we used the 3dMEMA program to perform a one-sample t-test, while for ds000120 we used the 3dMVM program at the second level to conduct a mixed-effects analysis, generating an F-statistic for the main effect of time.

In FSL, analyses were carried out using the FMRI Expert Analysis Tool (FEAT, v6.00). For each analysis, at the first level a separate .fsf file was created for each scanning session. Runs were then combined as part of a second-level fixed-effects model, yielding results which were subsequently inputted into a group analysis.

In SPM, preprocessing, subject- and group-level analyses were conducted by selecting the relevant modules within SPM's Batch Editor. In particular, subject-level and group-level analyses were conducted using the Specify 1st-level and Specify 2nd-level modules respectively.

Once analyses were complete, the results for each software package were exported as NIDM-Results packs (FSL and SPM only, (Maumet et al., 2016)) and uploaded to a public collection on the NeuroVault online data repository (RRID:SCR_003806, (Gorgolewski et al., 2015)).

Table 3.1: Software Processing Steps. Implementation of each of the processing steps (ds000001, ds000109, ds000120) within AFNI, FSL and SPM.

	Processing Step	AFNI	FSL	SPM
Preprocessing	**Script**	@SSWarper[2] afni_proc.py	FEAT First-level analysis	Batch (multiple modules)
	Slice-timing[1]	-tshift_opts_ts -tpattern	Pre-stats: Slice timing correction	Slice Timing
	Realignment/Motion Correction	-volreg_align_e2a	Pre-stats: Motion correction: MCFLIRT	Realign: Estimate and Reslice
	Segmentation	*Not applied*	*Not applied*	Segment
	Brain Extraction (Anatomical)	-copy_anat [@SSWarper result] -anat_has_skull no	bet (*command line*)	Image Calculator[3]
	Brain Extraction (Functional)	*Not applied*	Pre-stats: BET brain extraction	*Not applied*
	Intrasubject Coregistration	-align_opts_aea -giant_move -check_flip	Registration: Normal search, BBR	Coregister: Estimate
	Intersubject Registration	-tlrc_base -volreg_tlrc_warp -tlrc_NL_warp -tlrc_NL_warped _dsets [@SSWarper result]	Registration: Nonlinear, Warp resolution 10mm	Normalise: Write
	Analysis Voxel Size	-volreg_warp_dxyz (*overriding default determined from functional images*)	*determined by anatomical template voxel sizes*	Normalise: Write: Writing Options: Voxel Sizes
	Smoothing	-blur_size	Pre-stats: Spatial smoothing FWHM (mm)	Smooth
First-level	**Script**	afni_proc.py	FEAT First-level analysis	Specify 1st-level
	Model Specification	-regress_stim_times -regress_stim_labels -regress_basis_multi -regress_stim_types	Stats: Full model setup: EVs	fMRI model specification

Table 3.1 (continued)

	Processing Step	AFNI	FSL	SPM
	Inclusion of 6 Motion Parameters	*Implicitly added with 'regress' block*	Stats: Standard Motion Parameters	fMRI model specification: Data & Design: Multiple regressors: Realignment Param file
	Model Estimation	*Nothing to specify*	*Nothing to specify*	Model estimation
	Contrasts	-regress_3dD_stop -regress_reml_exec -regress_opts_3dD -gltsym	Stats: Full model setup: Contrasts	Contrast Manager
Second-level	**Script**	3dMEMA 3dMVM[1]	Feat Higher-level analysis	Specify 2nd-level
	Model Specification	3dMEMA -set -missing_data 0 3dMVM[1] -dataTable	Stats: Full model setup: EVs	Factorial design Specification: One-sample T-test Full factorial
	Model Estimation	*Nothing to specify*	*Nothing to specify*	Model estimation
	Contrasts	*Nothing to specify*	Stats: Full model setup: Contrasts	Contrast Manager
	Second-level Inference	3dMask_Tool (*Obtain group-mask*) 3dClustSim 3dClust 3dCalc (*Binarizing cluster masks and masking t-stat*) 3dTcat (*Obtaining one image in a 4D volume*)	Post-stats	Results Report
Results sharing	**NIDM-Results export**	*Not available*	nidmfsl	Results Report
	NeuroVault upload	*Upload all statistic images*	*Upload of 'group.gfeat.nidm.zip'*	*Upload of 'spm_****.nidm.zip'*

[1] ds000120 only.

[2] The @SSWarper program was run on each subject prior to afni_proc.py for brain extraction of the anatomical image, and to apply the nonlinear warp of the anatomy to MNI space.

[3] Image calculator was used to create the brain mask from grey matter, white matter and CSF images; see text.

Common Processing Steps

A number of processing steps for each package were included in all of our analyses, regardless of whether they had been implemented in the original study. While this meant deviating from an exact replication of the original pipeline, these processing steps were either fundamental to ensure that the results from each software package could be compared objectively, or steps that are widely accepted as best practices within the community. In this section we describe these steps.

Successful coregistration of the functional data to the structural brain images, and subsequently, registration to the MNI template, was of paramount importance to us for fair comparability of the results. During our first attempt at analyzing the ds000001 dataset we discovered that seven subjects had essential orientation information missing from the NIfTI header fields of their functional and structural data. As the source DICOM files were no longer available, the original position matrices for this dataset were unable to be retrieved. This caused coregistration to fail for several subjects across all three software packages in our initial analysis of this data. We rectified the issue by manually setting the origins of the functional and structural data. OpenfMRI released a revision (Revision: 2.0.4) of our amended dataset which we used for the analysis. Further to this, we also set a number of common preprocessing steps within each package to be applied in all of our analyses.

Firstly, brain extraction was conducted on the structural image in all software. We did this to improve registration and segmentation. In AFNI, brain extraction was carried out using 3dskullstrip. This was called implicitly from within the @SSwarper program. The skull-stripped anatomical volume obtained here was inputted into our afni_proc.py scripts where further preprocessing and first-level analyses were carried out. In FSL, brain extraction was performed on both the functional and structural data. The Brain Extraction Tool (BET; (Smith, 2002)) was applied to each structural image from the command line before preprocessing, and to the functional data with

the BET option within the Pre-stats module of FEAT. In SPM, brain extraction was implemented via the segmented structural images. Grey matter, white matter and CSF images were summed and binarised at 0.5 to create a brain mask, which was applied to the bias corrected structural image using the Image Calculator module.

Coregistration of the functional data to the anatomy was carried out for the most part using the default settings in each software. In AFNI, alignment of the data was conducted using the align_epi_anat.py program called implicitly from the align block within the afni_proc.py scripts. We included the -volreg_align_e2a option within our scripts to specify alignment of the functional data onto the anatomy, as by default AFNI conducts the inverse transformation of anatomy onto functional. Further to this, we also added the -align_opts_aea program to all of our scripts with the -giant_move and -check_flip options to allow for larger transformations between the images. In FSL, coregistration was carried out within FEAT using the default linear registration methods with a boundary-based registration (BBR) cost function. The default methods were also applied within SPM's Coregister: Estimate module, using a normalised mutual information cost function.

Registration of the structural and functional data to the anatomical template was executed using each package's nonlinear settings. In AFNI, nonlinear registration of the anatomical data to the MNI template was conducted as part of the @SSwarper program run prior to the afni_proc.py script. The transformations computed by @SSwarper were passed to afni_proc.py using the -tlrc_NL_warped_dsets option, and applied to the functional data within the tlrc block using the -volreg_tlrc_warp option. By default, the resampled functional data in MNI space has voxel size determined from the raw 4D data; we forced 2mm cubic voxels with the -volreg_warp_dxyz option for compatibility with FSL and SPM's 2mm default. In FSL, registration to the MNI template was conducted using FMRIB's Nonlinear Image Registration Tool (FNIRT; (Andersson et al., 2007)), controlling the degrees of freedom of the transformation with a warp resolution of 10mm. In SPM, the nonlinear deformations to MNI

space were obtained as part of the Segment module and then applied to the structural and functional data within the Normalise: Write module.

As a form of quality control, we created mean and standard deviation images of the subject-level MNI-transformed anatomical and mean functional images. Alongside the subject-level data, these images were assessed to check that registration to MNI space had been successful. When intersubject registration failed remedial steps were taken within each software; these are described in the software implementation parts of the following study-specific analysis sections.

Across all software packages six motion regressors were included in the analysis design matrix to regress out motion-related fluctuations in the BOLD signal. Use of six or more derived motion regressors is commonly recommended as good practice, and we chose to use just six regressors as this could be easily implemented across software.

Finally, we note that each software package uses a different default connectivity criterion for determining significant clusters: 6-connectivity for AFNI, 18-connectivity for SPM, and 28-connectivity for FSL. Since these settings are not typically modified we have kept these defaults in all of our analyses to reflect standard practices carried out within each software.

We now describe the task-specific analysis procedures for each of the three studies as carried out in the original publications, and how these methods were implemented within each package. While we decided to keep the above steps of the analysis pipelines fixed, for all remaining procedures we attempted to remain true to the original study. Any further deviations necessitated are discussed in the software implementation sections. Notably, apart from the addition of six motion regressors, all of our common steps relate to preprocessing. Therefore, for first- and group-level analysis we attempt to exactly replicate the original study.

ds000001 Analyses

In the publication associated with the ds000001 study all preprocessing and analysis was conducted within FSL (version 4.1.6). Data on all 16 subjects were available to us on OpenfMRI. In the original preprocessing, the first two volumes of the functional data were discarded and the highpass-filter was set to a sigma of 50.0s. Motion correction was conducted using MCFLIRT and brain extraction of the functional data was applied with BET, after which FSL's standard three-step registration procedure was carried out to align the functional images to the structural scan. Spatial normalization was implemented with FMRIB's Linear Image Registration Tool (FLIRT; (Jenkinson et al., 2002)), and data were smoothed using a 5mm full-width-half-maximum (FWHM) Gaussian kernel. At the run level, each of the events were convolved using a canonical double-gamma HRF; FEAT's (then newly available) outlier de-weighting was used. Subject-level analysis of the functional data were conducted using a GLM within FEAT, where a selection of the regressors were orthogonalized. The three scanning sessions for each participant were carried out separately and then combined together at the second level. A pair of one-sided t-tests were conducted at the group-level to test for positive and negative effects separately. For each test, clusterwise inference was performed using an uncorrected cluster-forming threshold of $p < 0.01$, FWE-corrected clusterwise threshold of $p < 0.05$ using Gaussian Random Field Theory.

We opted to not use outlier de-weighting on the basis that such methods were impractical to implement across all software packages.

AFNI Implementation

Using our default procedure for the AFNI analysis, we found that coregistration of the functional scans onto the anatomy failed for four subjects. To remedy this issue, for this study we modified our afni_proc.py scripts: within the -align_opts_aea mod-

ule, the '-ginormous move' option was added to align the centres of the functional and anatomical volumes, and the '-cost lpc+ZZ' option was used to apply a weighted combination of cost functionals. Both of these changes are recommended for data with little structural detail. Following these modifications all coregistrations were successful.

To replicate the orthogonalization methods from the original study, a separate orthogonalization script was run for each subject prior to the first-level analyses. Within this script, the (un-orthogonalized) regressors were generated by passing the event timing files to 3dDeconvolve, after which the 3dTproject command was used to obtain the desired projections. The orthogonalized regressor files outputted from this script were then entered into afni_proc.py to replicate the original subject-level analysis model.

Trials were convolved with a single gamma HRF using either the BLOCK or dmBLOCK option within the -regress_basis_multi module, determined by whether the event file had fixed or variable duration times respectively. The -regress_stim_types option was added to our afni_proc.py script to specify event files for regressors which had been parametrically modulated in the original study, and identify the orthogonalized regressors.

At the group level, we performed a mixed-effects analysis using 3dMEMA. The critical cluster size threshold was determined by Monte Carlo simulation with the 3dClustSim program.

FSL Implementation

Implementation in FSL closely followed the original procedure described above, with the exception that nonlinear registration was used to transform the data to standard space.

SPM Implementation

Implementation in SPM closely followed the pipeline outlined in Table 3.1.

ds000109 Analyses

The original preprocessing and statistical analysis for the ds000109 study was carried out using SPM8. Data were shared on 36 of the 40 subjects, 21 of which were young adults that had fMRI data compatible for our reanalysis. First, functional data were realigned and unwarped to correct for head motion and geometric distortions. After transforming the data into a standardized space, the normalized data were smoothed with an 8mm FWHM Gaussian kernel. Further to this, custom software was applied to exclude functional volumes where head motion had exceeded a certain limit, however this process was omitted from our pipelines since this feature was not available in any of the software packages. The preprocessed data were entered into a GLM for first-level analyses where trials were modeled using a block design and convolved using SPM's canonical HRF. Each participant's contrast images were then entered into a one-sample group analysis using clusterwise inference, cluster-forming threshold of $p < 0.005$, FWE-corrected clusterwise threshold of $p < 0.05$ using Random Field Theory; in their analysis, this amounted to a critical cluster size threshold of 56 voxels.

AFNI Implementation

Intersubject registration to the MNI atlas failed for one subject, causing a loss of the functional lobe in this subject's registered functional map. We addressed this by revising this study's AFNI pipeline to use the -pad_base 60 option within the -tlrc_opts_at module included in afni_proc.py. This gave extra padding to the MNI template so that no part of the functional image was lost during the alignment.

The HRF was modelled with SPM's canonical HRF using the SPMG1 option

for each event within the -regress_basis_multi option and passing the duration of the regressor as an argument to the function.

At the group level, we performed a mixed-effects analysis using 3dMEMA. p-values were determined by Monte Carlo simulation within 3dClustSim.

FSL Implementation

To recreate the original HRF model in FSL, we chose the Double-Gamma HRF from the convolution options within FEAT.

SPM Implementation

Implementation in SPM closely followed the original procedure described above.

ds000120 Analyses

A multi-software analysis procedure was used for the ds000120 study, where data were preprocessed with FSL and then analyzed using AFNI. fMRI data were shared on OpenfMRI for 26 of the original 30 subjects, and 17 had data available for the task of interest. This was the only study that applied slice-timing correction, adjusting the functional data for an interleaved slice acquisition. Functional scans were realigned to the middle volume, and following brain extraction with BET, registered to the structural scan in Talairach space using FLIRT and FNIRT. Data were high-pass filtered with a sigma value of 30.0s and smoothed with a 5mm FWHM Gaussian kernel. Like the previous study, further methods were used to remove functional volumes with excessive motion. These methods were left out from our analyses due to discordance across software. Subject-level analyses were conducted within AFNI. To allow for flexible modelling of the response to the saccade task, this study used an HRF basis consisting of eight sine functions with a post-stimulus window length of 24.0s. At the group level, subjects were entered into a mixed-effect model, with subjects as a random factor, trial type (reward, neutral) and time as within-group fac-

tors, and age group (child, adolescent, adult) as a between-group factor. Clusterwise inference was used on the main effect of time activation map ($F_{8,142}$ statistic), with a cluster-forming threshold of $p < 0.001$, FWE-corrected clusterwise threshold of $p < 0.05$ using Monte Carlo simulation methods. The computed critical cluster size threshold was 23 voxels.

For our replication exercise we only consider the main effect of time. This analysis is based on the corresponding time effect contrasts for each subject and requires a simpler model, with one random effect (subject) and one fixed effect (time).

AFNI Implementation

Slice timing was conducted using the -tshift_opts_ts program within afni_proc.py with the -tpattern option applied to specify an interleaved slice acquisition.

The sine basis set used for the HRF was modelled using the -regress_basis_multi module with the SIN option.

At the group level, a mixed-effect analysis was carried out with the 3dMVM program. Following this, 3dClustSim was used to obtain the cluster extent corresponding to the original study threshold. In our analysis we found the cluster size threshold to be 48 voxels.

FSL Implementation

The repeated-measures design used in the group-level analysis of the original study was not feasible to implement for parametric inference in FSL, and as such, we did not attempt an FSL reanalysis for this study. (The FEAT manual does describes "Repeated Measure" examples, but these are based on a restrictive assumption of compound symmetry; here this would entail assuming that all 8*7/2=28 correlations among the basis regression coefficients were equal.)

SPM Implementation

Slice timing was conducted using the Slice Timing module within the Batch Editor of SPM.

Although an exact equivalent of the original HRF model was not possible in SPM, we chose the closest equivalent using the Fourier basis set with an order of 4, leading to a total of 9 basis functions fit to each of the reward and neutral conditions for each of the three runs. A set of 9 first level contrasts computed the average Fourier coefficients over conditions and runs.

To reproduce the group-level analysis in SPM, a full factorial design was chosen within the 'Factorial design specification' module of the Batch Editor with a time factor (9 levels), and adding age-group to the model using two covariates (adolescent vs child, adult vs child); the main effect of time was tested with an F-contrast.

3.1.3 Comparison Methods

We applied three separate quantitative methods to measure the similarity between the group results obtained within each software package for each of the three studies.

Firstly, Bland-Altman plots comparing the unthresholded group statistic maps were created for each pairwise combination of software packages. These plotted the difference between the statistic values (y-axis) against the mean statistic value (x-axis) for all voxels lying inside the intersection of the two software's analysis masks. The plots provide an assessment of the level of agreement between two software packages about the magnitude of the statistic value observed at each voxel. If two software packages were in perfect agreement, all points on the bland-altman plot would lie on the x-axis, since the difference between the statistic values at each voxel would be zero. The degree of disagreement is therefore evaluated by the perpendicular distance of points from the x-axis; for example, for an "AFNI-FSL" Bland-Altman

plot, points lie above the x-axis for locations where AFNI's statistic value is larger than FSL's. With the difference plotted against the average, general patterns of disagreement can be discerned.

In addition to this, we created Bland-Altman plots to compare percentage BOLD change maps (for ds000120, partial R^2 maps) between software. For each package, an appropriate normalization of the group-level beta maps was conducted to convert to percentage BOLD change units. Due to differences in how each package scales the data, a different normalization was required for each of the three packages. For ds000120, the partial R^2 maps were computed via a transformation of the group-level F-statistic images. We provide full details on how each of these procedures were carried out in the appendix (Appendix A.1 for percent BOLD change, Appendix A.2 for partial R^2). In all of our Bland-Altman comparisons, we excluded white matter and cerebral spinal fluid voxels according to the MNI tissue probability maps thresholded at 0.5.

We also computed the Dice similarity coefficient for each pairwise combination of the group-level thresholded statistic maps. The coefficient is calculated as the cardinality of the intersection of the thresholded maps divided by the average of the cardinality of each thresholded map. While Bland-Altman is interested in the similarity between statistic values, Dice measures the overlap of voxels as a means to assess the spatial similarity of activated clusters. The coefficient takes a value between zero and one, where one indicates complete congruence between the size and location of clusters in both thresholded maps, while zero indicates no agreement. Dice coefficients were computed over the intersection of the pair of analysis masks, to assess only regions where activation could occur in both packages. We also calculated the percentage of 'spill over' activation, i.e. the percentage of activation in one software's thresholded statistic map that fell outside of the analysis mask of the other software.

A particular concern we had was that a pair of statistic images could in essence be very similar, but differ by a scale factor over all voxels. Another possibility was

that one software could have greater sensitivity for voxels where signal was present, causing differences between images only for relatively higher statistical values. Both of these features would not be identifiable using our previous comparison methods. To address this, we computed the Euler characteristic (EC) for each software's group-level t-statistic map (F-statistic for ds000120), thresholded using t-values between -6 to 6 (0 to 6 for ds000120; increasing with an increment of 0.2). Alongside the EC, we also computed the number of clusters in the statistic images using the same thresholds. As discussed in Section 2.7.1, for a given threshold t, the EC calculates the number of clusters, minus the numbers of 'handles', plus the number of 'holes' in a thresholded image. For large t, we expect the handles and holes to disappear, and therefore the EC provides an approximation of the number of clusters in the image. For smaller t, we expect our thresholded image to be one connected cluster with many holes and handles (like Swiss cheese) – it is in this situation where the EC is clearly more informative about differences between images than the cluster count alone. Over all t, the EC curve provides a signature of an entire statistic image, and provides a means to assess whether only superficial scaling differences are responsible for disparities between a pair of images.

For a qualitative assessment of whether similar activation patterns were displayed between packages, a NeuroSynth (RRID:SCR_006798, http://neurosynth.org) association analysis was conducted on each software's unthresholded statistic map. These analyses performed a cognitive decoding of the unthresholded statistic image with images in the NeuroSynth database, to find the words or phrases most strongly associated with the activation patterns found in the statistic map.

Finally, we visually compared the corresponding slices of each software's thresholded statistic map to those presented in the publication figure we had attempted to recreate. Ensuring we had found activation in approximately the same regions as the original publication gave us an indication that we had successfully replicated the study's analysis pipeline.

3.1.4 Permutation Test Methods

For ds000001 and ds000109, in parallel to our replication analyses we computed an additional set of group-level results applying nonparametric permutation test inference procedures available within each software package (a one-sample repeated measures permutation test needed for ds000120 was not available in AFNI). The first-level contrast maps obtained from our initial replications for each subject were entered into a group-level one-sample t-test where clusterwise inference was conducted using the same cluster-forming thresholds, and then 5% level FWE corrected thresholds were computed by permutation methods, with 10,000 permutations.

AFNI Implementation

In AFNI, permutation inference was carried out using the 3dttest++ module with the -ClustSim option. By applying this option, permutations generated multiple realizations of noise which 3dClustSim used to generate cluster-threshold tables. Significant clusters in the group-activation map were found with 3dclust, using a critical cluster size threshold extracted from the 3dClustSim output.

FSL Implementation

Permutation test inference was conducted in FSL using randomise (version 2.9; (Winkler et al., 2016)). This outputted a 'corrp' image which was then used to mask the raw t-statistic image to show significant voxels for the appropriate thresholds.

SPM Implementation

The Statistical nonParametric Mapping (SnPM, version SnPM13; RRID:SCR_002092; (Nichols and Holmes, 2002)) toolbox was used to carry out permutation tests in SPM. The "MultiSub: One Sample T test on diffs/contrasts", Compute and Inference modules within SnPM were applied to obtain the final group-level activation maps.

Each of the comparison methods described in the previous section were also applied to our permutation results to assess cross-software differences for nonparametric inference methods. In addition, we also generated intra-software Bland-Altman plots and Dice coefficients to understand differences between the parametric and nonparametric methods applied within each package.

These methods were excluded for ds000120, since it was not possible to conduct permutation inference for an F-test within AFNI, and parametric inference was unfeasible in FSL as discussed in the previous section.

3.1.5 Scripting of Analyses and Figures

AFNI and FSL scripts were written in Python 2.7.14 and SPM scripts were written in Matlab R2016b. Scripts were made generalizable, such that the only study-specific differences for each of the analyses in a software package were the raw data and working directory inputs, the subject- and group-level analysis templates (as well as a run-level template for FSL), and a unique conditions structure necessary for creating the onset files for the specified study. For each analysis package, a script was written to extract the stimulus timings from the raw data to create event files that were compatible within the software. Subject-level analysis templates were batch scripts created for each study containing all processing steps of the subject analysis pipeline for the respective software, with holding variables used where subject- or run-specific inputs were required. The main script would take the template as an input and, cycling through each of the subjects, replace the holding variables with appropriate pathnames to create distinct batch scripts for each participant. These were then executed to obtain subject-level results for all participants in the study.

A Python Jupyter Notebook (Kluyver et al., 2016) was created for each of the three studies. Each notebook harvests our results data from NeuroVault and applies the variety of methods discussed in the previous section using NiBabel 2.2.0 (Brett et al., 2017), NumPy 1.13.3 (Walt et al., 2011) and Pandas 0.20.3 (McKinney and Oth-

ers, 2010) packages. Figures were created using Matplotlib 2.1.0 (Hunter, 2007) and Nilearn 0.4.0 (Abraham et al., 2014).

3.2 Results

All scripts and results are available through our Open Science Framework (OSF; (Erin D. Foster, 2017)) Project at https://osf.io/U2Q4Y/ (Bowring et al., 2018a), and group-level statistic maps used to create the figures in this section are available on NeuroVault: https://neurovault.org/collections/4110/, https://neurovault.org/collections/4099/, https://neurovault.org/collections/4100/, for ds0000001, ds000109 and ds000120 respectively. All analysis scripts, results reports, and notebooks for each study are available through Zenodo (Nielsen and Smith, 2014) at https://doi.org/10.5281/zenodo.1203654 (Bowring et al., 2018b).

Registration of each subject's functional data onto the anatomy was visually assessed. The mean and standard deviation images of the MNI structural and (mean) functional data substantiate that registration was successful in all packages across the three studies (Fig. A.1; note that axial slices are slightly different beteen software due to different bounding boxes of the images).

3.2.1 Cross-Software Variability for Parametric Inference

While qualitatively similar, variability in the t-statistic values and locations of significant activation was substantial between software packages across all three studies.

Comparisons of the thresholded results with the published findings are shown in Figure 3.1, with further multi-slice comparisons across software in Figure 3.2 (also in Figs. A.2, A.4, and A.6). The ds000001 study described positive activation in the bilateral anterior insula, dorsal anterior cingulate cortex (ACC), and right dorsolateral prefrontal cortex, and negative activation in the ventromedial prefrontal cortex and bilateral medial temporal lobe. In our reanalyses (Fig. 3.1, left) all three software

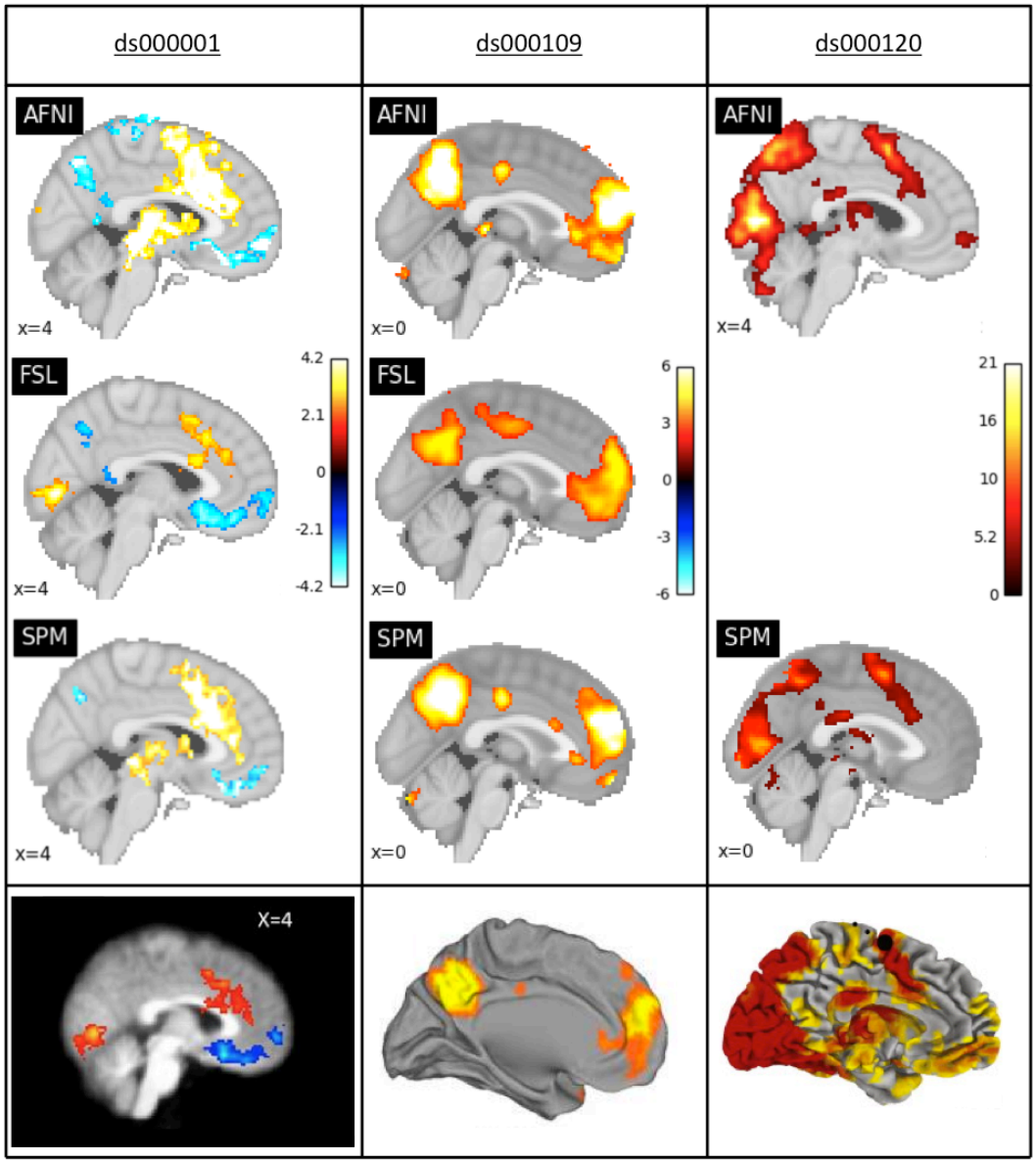


Figure 3.1: Comparison of the thresholded statistic maps from our reanalysis with the main figures from each of the three publications. Left: For ds000001 data, thresholded t-statistic images contrasting the parametric modulation of pumps of reward balloons versus the parametric modulation of the control balloon; beneath, a sagittal slice taken from Fig. 3 in Schonberg et al. (2012). Middle: For ds000109, thresholded t-statistic maps of the false belief versus false photo contrast; beneath, a midsagittal render from Moran et al. (2012). Right: For ds000120, thresholded F-statistic images of the main effect of time contrast; beneath, a midsagittal render from Fig. 3 in Padmanabhan et al. (2011). Note that for ds000109 and ds000120 the publication's figures are renderings onto the cortical surface while our results are slice views. While each major activation area found in the original study exists in the reanalyses, there is substantial variation between each reanalysis.

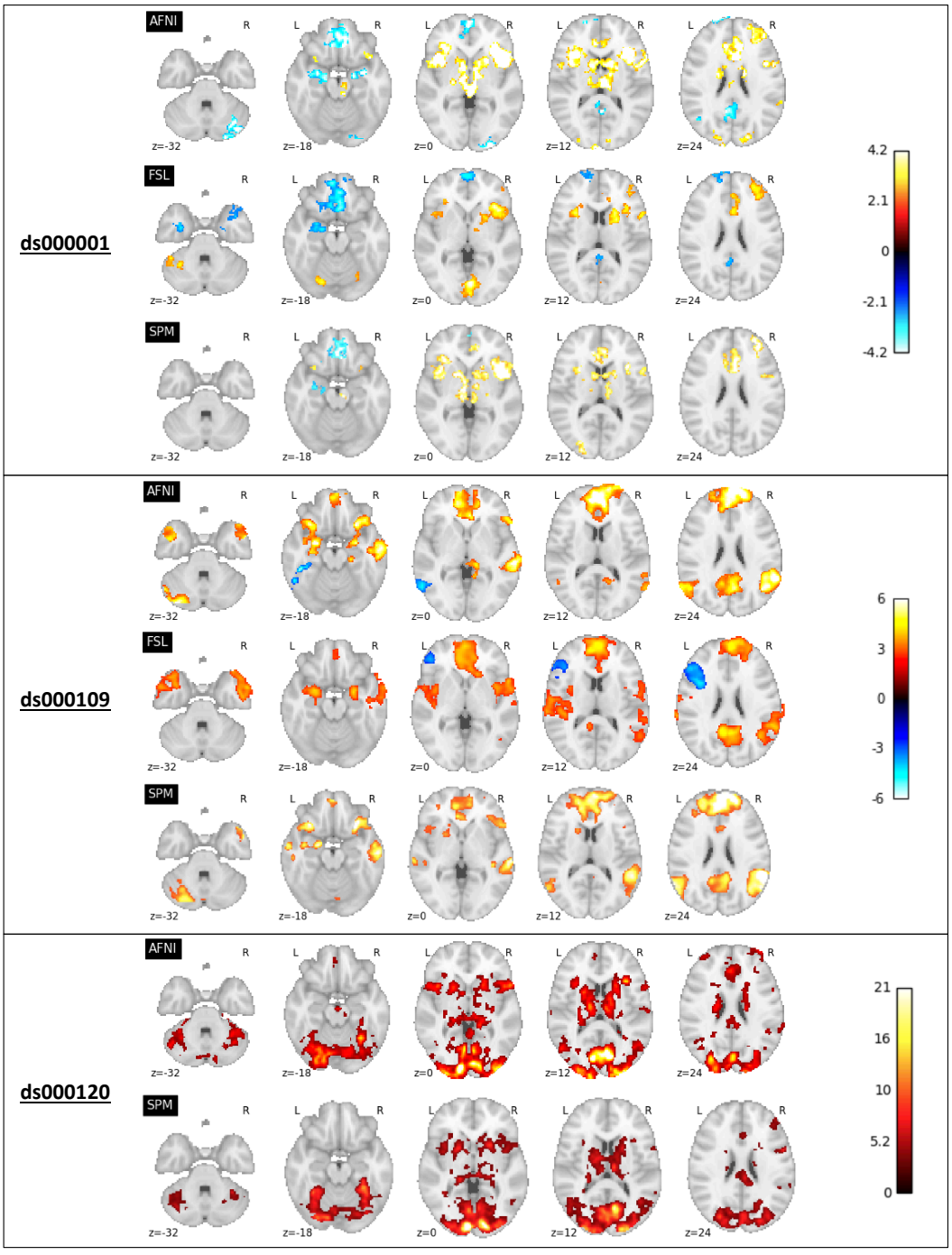


Figure 3.2: Comparison of the thresholded statistic maps from our reanalysis displayed as a series of axial slices. Top: ds000001's thresholded t-statistic maps contrasting parametric modulations of the reward balloons versus pumps of the control balloons. Middle: ds000109's thresholded t-statistic maps of the false belief versus false photo contrast. Bottom: ds000120's thresholded F-statistic maps of the main effect of time contrast. This figure complements the single slice views shown in Fig. 3.1.

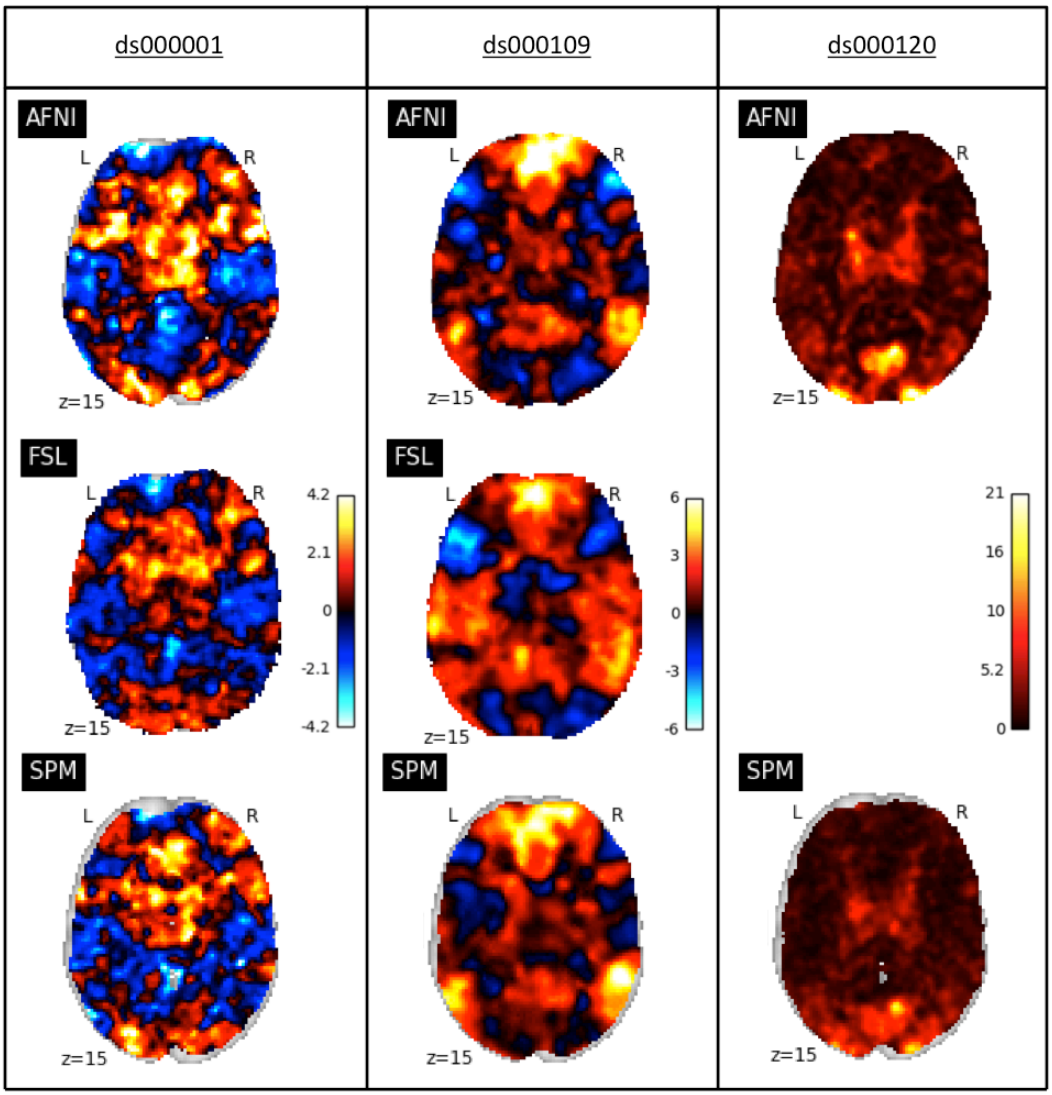


Figure 3.3: Comparison of the unthresholded statistic maps from our reanalysis of the three studies within each software package. Left: ds000001's unthresholded t-statistic maps of the parametric modulation of pumps of reward balloons versus the parametric modulation of the control balloon contrast. Middle: ds000109's unthresholded t-statistic maps of the false belief versus false photo contrast. Right: ds000120's unthresholded F-statistic maps of the main effect of time contrast. While areas of strong activation are somewhat consistent across all three sets of reanalyses, there is substantial variation in non-extreme values.

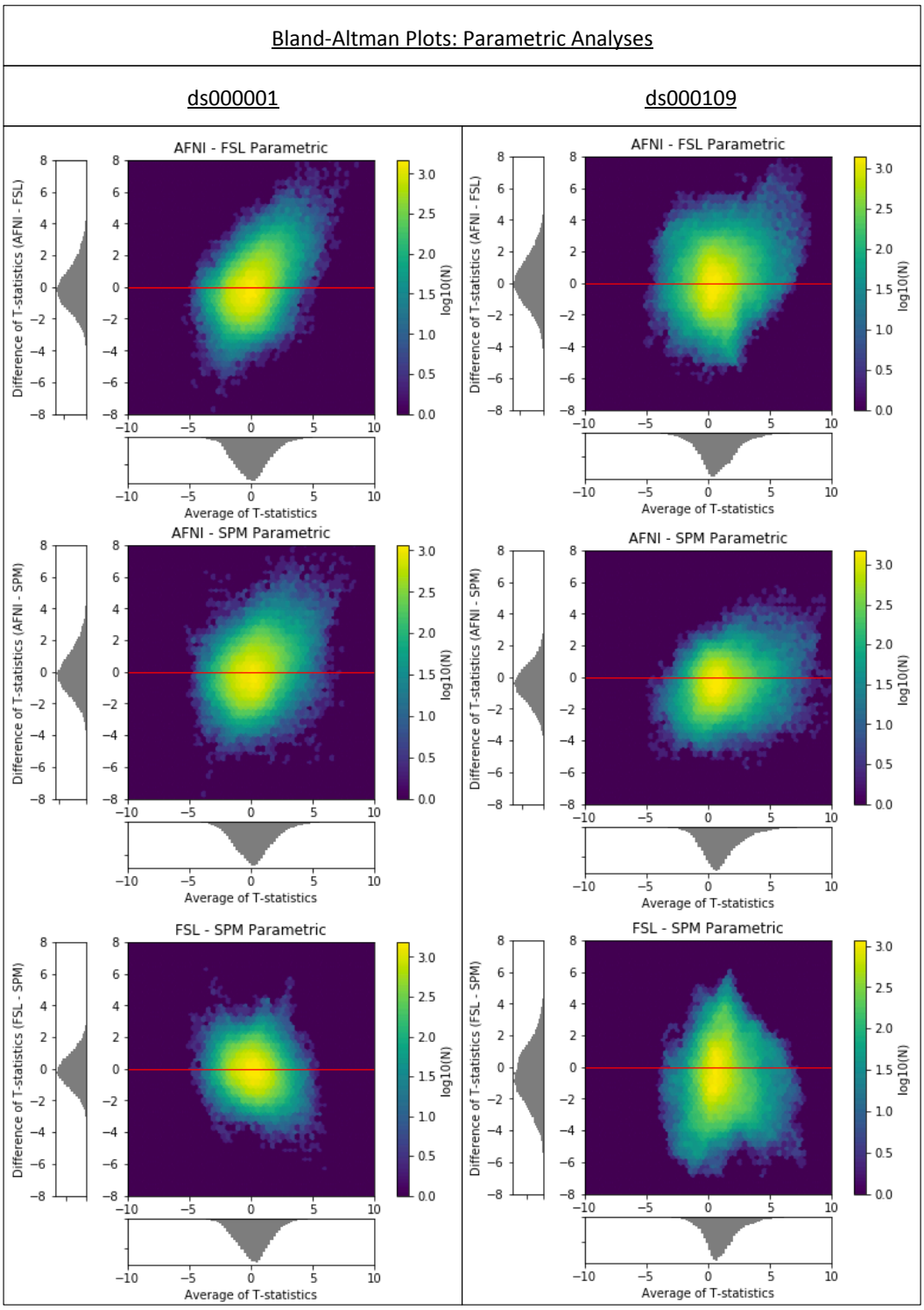


Figure 3.4: Cross-software Bland-Altman 2D histograms comparing the unthresholded group-level t-statistic maps computed as part our reanalyses of the ds000001 and ds000109 studies within AFNI, FSL, and SPM. Left; Comparisons for ds000001's balloon analog risk task, t-statistic images contrasting the parametric modulation of pumps of the reward balloons versus parametric modulation of pumps of the control balloon. Right; Comparisons for ds000109's false belief task, t-statistic images contrasting the false belief versus false photo conditions. Density images show the relationship between the average t-statistic value (abscissa) and difference of t-statistic values (ordinate) at corresponding voxels in the unthresholded t-statistic images for each pairwise combination of software packages. While there is no particular pattern of bias, as the t-statistic differences are centered about zero, there is remarkable range, with differences exceeding $\pm$ 4 in all comparisons.

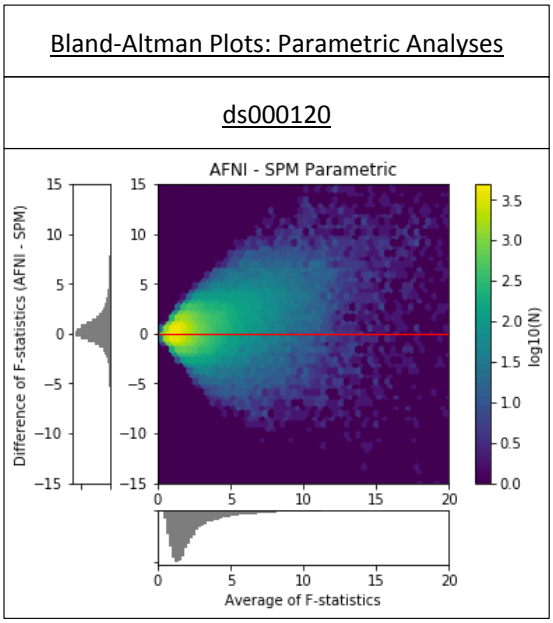


Figure 3.5: Cross-software Bland-Altman 2D histogram comparing the unthresholded main effect of time F-statistic maps computed in AFNI and SPM for reanalyses of the ds000120 study. The differences are generally centered about zero, with a trend of large F-statistics for AFNI. The funnel-like pattern is a consequence of the F-statistic taking on only positive values.

found activation in these set of regions, with the exception that decreases in the medial temporal lobe were unilateral in FSL and SPM (left only). FSL also localized a visual response that was not present in the corresponding AFNI and SPM thresholded maps (Fig. 3.1, left, and Fig. 3.2, top, z = 0 slice).

The ds000109 study reported activations in the bilateral temporoparietal junction (TPJ), precuneus, anterior superior temporal sulcus (aSTS), and dorsal medial prefrontal cortex (dmPFC). Similar activations from our reanalyses are seen in Figure 3.1, middle, although FSL only found activation in the right TPJ and aSTS. Further comparisons shown in Figure 3.2, middle, highlight disagreement in the results: AFNI and FSL detected significant deactivations in distinct brain regions (inferior temporal gyrus for AFNI, inferior frontal gyrus for FSL), while SPM did not determine any significant deactivation. FSL also found a positive response in the superior temporal gyrus (STG) where AFNI and SPM did not (Fig. 3.2, middle, z = 0 and z = 12 slices).

The original ds000120 study localized activation for the main effect of time in multiple regions of the brain – the frontal, supplementary, posterior parietal cortex, basal ganglia, prefrontal cortex, ventral striatum and orbitofrontal cortex all showed significant activation. Our reanalyses (Fig. 3.1, right) are consistent with these findings, with the exception that neither AFNI or SPM exhibited orbitofrontal (OFC) activation (however, the SPM analysis mask had poor OFC coverage). AFNI's F-statistic values look to be generally larger than SPM here (Fig. 3.2, bottom, z = 0 and z = 12 slices). The unthresholded statistic maps from our reanalyses (Figs. 3.3, A.7, A.9 and A.11) also show that while extreme values display moderate agreement, there are considerable differences across the brain in each given study.

NeuroSynth association analyses conducted on the unthresholded t-statistic maps (Table 3.2) show that the most strongly related term to the activation patterns displayed in all three sets of results was the same: 'anterior insula' for each software's ds000001 map, 'medial prefrontal' for ds000109, and 'visual' for ds000120. Phrases related to the task paradigm used in each study ('goal' for ds000001, 'theory mind' for ds000109, 'visual' for ds000120) were found across all software's activation patterns, alongside a range of common anatomical terms.

Figure 3.4 compares statistic values across packages using Bland-Altman plots (rendered as 2D histograms) for ds000001 and ds000109. The distribution of the pairwise differences in t-statistics (y-axes) is generally centered about zero, indicating no particular bias, however there is substantial variation here, with t-statistic differences exceeding 4.0 in magnitude. Pairwise correlations ranged from 0.429 to 0.747 for inter-software comparisons (Fig. 3.3). The Bland-Altman plots comparing percentage BOLD change maps (Fig. A.13) are more conclusive, showing a clear trend for SPM to report larger effect estimates than the other two packages. Figure 3.5 presents the Bland-Altman plot comparing the unthresholded F-statistic images for ds000120, which has a very different appearance since F-statistics are non-negative. The corresponding Bland-Altman plot comparing partial R^2 values (Fig.

Table 3.2: Neurosynth Analyses. The Neurosynth analysis terms most strongly associated (via Pearson correlation) to each software's group-level statistic map across the three studies. Non-anatomical terms are shown in bold.

	AFNI		FSL		SPM	
	Neurosynth Analysis	**Pearson Corr.**	**Neurosynth Analysis**	**Pearson Corr.**	**Neurosynth Analysis**	**Pearson Corr.**
ds000001	Anterior insula	0.359	Anterior insula	0.240	Anterior insula	0.322
	Insula	0.276	**Task**	0.233	Anterior	0.245
	Anterior	0.243	**Tasks**	0.203	Insula	0.240
	Insula anterior	0.233	Parietal	0.190	**Goal**	0.229
	Thalamus	0.221	**Goal**	0.188	**Task**	0.225
	Goal	0.211	**Working memory**	0.184	Insula anterior	0.214
	Pain	0.198	**Working**	0.181	Thalamus	0.201
	Supplementary	0.197	Basal ganglia	0.173	Acc	0.199
	Premotor	0.196	Ganglia	0.172	Anterior cingulate	0.196
	Anterior cingulate	0.192	Basal	0.169	Ganglia	0.188
ds000109	Medial prefrontal	0.422	Medial prefrontal	0.355	Medial prefrontal	0.361
	Medial	0.381	Medial	0.309	**Theory mind**	0.331
	Default	0.366	**Default**	0.301	**Default**	0.329
	Theory mind	0.348	Posterior cingulate	0.299	Precuneus	0.314
	Default mode	0.341	**Default mode**	0.290	**Default mode**	0.310
	Precuneus	0.334	**Social**	0.282	Medial	0.301
	Posterior cingulate	0.327	Cingulate	0.275	**Mind**	0.296
	Social	0.322	**Theory mind**	0.270	Prefrontal	0.294
	Mind	0.311	**Resting**	0.261	**Mind tom**	0.289
	Mind tom	0.287	Precuneus	0.259	Posterior cingulate	0.287
ds000120	**Visual**	0.377			**Visual**	0.481
	v1	0.317			Occipital	0.367
	Occipital	0.293			v1	0.340
	Eye	0.261			Visual cortex	0.267
	Eye movements	0.252			**Spatial**	0.248
	Visual cortex	0.243			Spl	0.245
	Early visual	0.241			**Eye**	0.242
	Spatial	0.232			Early visual	0.238
	Task	0.229			Lingual	0.238
	Parietal	0.222			Intraparietal	0.237

Table 3.3: Summary of the mean differences and correlations between the three softwares' test statistic images. Mean differences correspond to the y-axes of the Bland-Altman plots displayed in Figures 3.4, 3.4, 3.9 and 3.10. Each mean difference is the first item minus the second; for example, the AFNI vs. FSL mean difference is given by AFNI minus FSL. Correlation is the Pearson's r between the test statistic values for the given pair of software packages. Intersoftware differences are greater than intrasoftware.

		ds000001		ds000109		ds000120	
Software	**Inference Type**	**Mean Difference**	**Pearson Corr.**	**Mean Difference**	**Pearson Corr.**	**Mean Difference**	**Pearson Corr.**
AFNI vs. FSL	Parametric	0.009	0.616	0.035	0.585		
	Nonparametric	0.271	0.577	0.006	0.573		
AFNI vs. SPM	Parametric	0.061	0.614	-0.490	0.747	0.415	0.748
	Nonparametric	-0.096	0.628	-0.445	0.787	*n\a*	*n\a*
FSL vs. SPM	Parametric	-0.047	0.684	-0.529	0.429		
	Nonparametric	-0.479	0.720	-0.439	0.438		
AFNI	Para. vs Nonp.	0.155	0.984	-0.048	0.981		
FSL	Para. vs Nonp.	0.382	0.844	-0.064	0.946		
SPM	Para. vs Nonp.	0.000	1.000	0.000	1.000		

A.14) for this study is similar in shape. Broadly speaking, while there are no gross differences in sensitivity, there is a slight tendency for AFNI's extreme statistics to exceed the other two packages, and SPM's to exceed FSL's, most evident in ds000109.

Spatial localization of significant activation in the thresholded t-statistic images also varied across software packages. Figure 3.6 shows the Dice coefficients for all pairs of analyses (parametric results are presented in the first 3 rows of the larger triangles). For ds000001, the average value of Dice coefficients comparing locations of activations across reanalyses is 0.379. These values improve for ds000109, where the mean Dice coefficient for positive activations is 0.512. Here, AFNI and FSL were the only software packages to report significant negative clusters for the ds000109 study. Strikingly, these activations were found in completely different anatomical regions for each package, witnessed by the negative activation AFNI/FSL Dice coefficient of 0. Finally, the AFNI/SPM Dice coefficient for the thresholded F-statistic images obtained for ds000120 is 0.684; it is notable that across all studies, the AFNI/SPM Dice coefficients are consistently the largest.

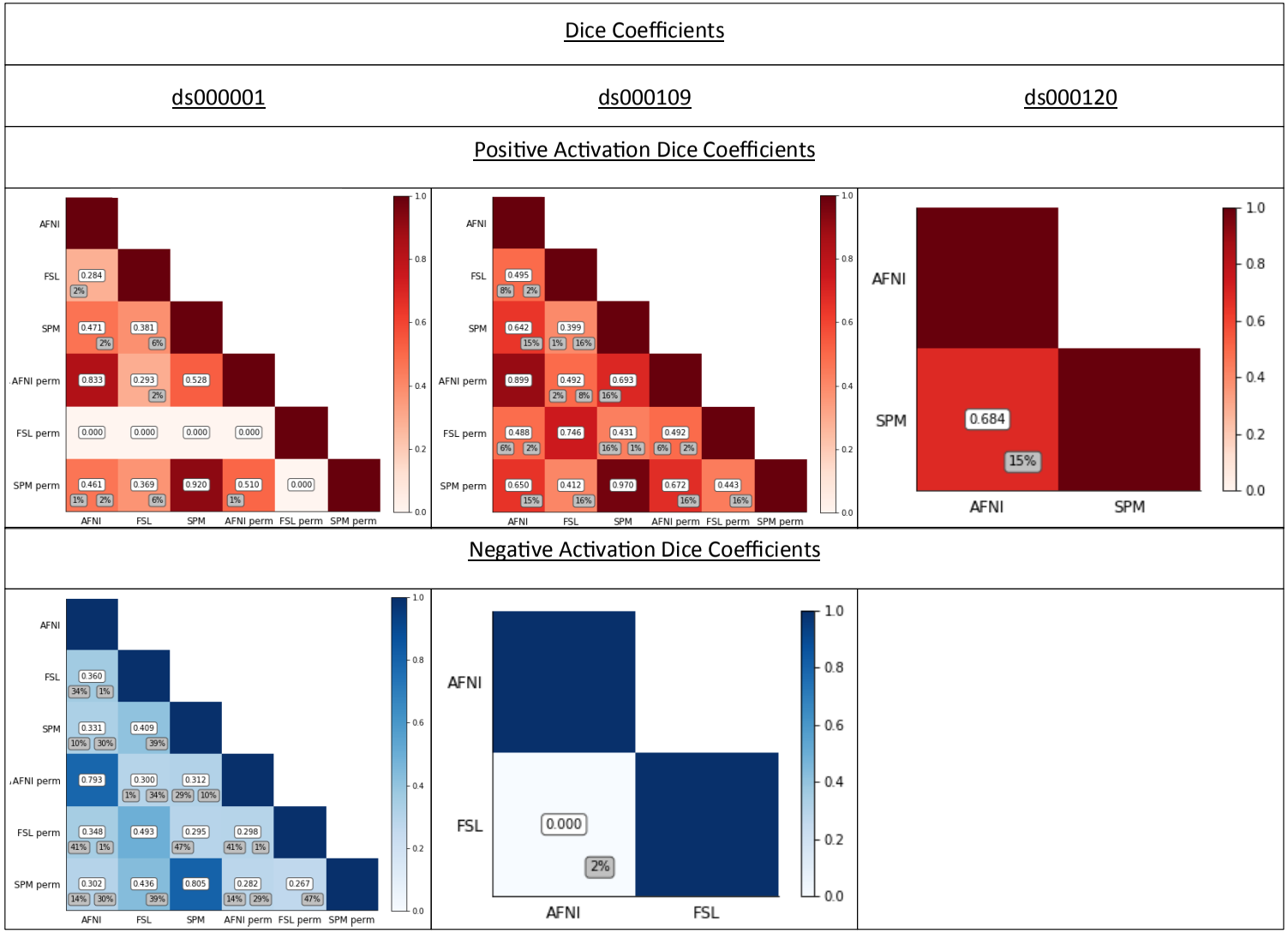


Figure 3.6: Dice coefficients comparing the thresholded positive and negative t-statistic maps computed for each pair of software package and inference method for each of the three reproduced studies. Dice coefficients were computed over the intersection of the pair of analysis masks, to assess only regions where activation could occur in both packages. Percentage of "spill over" activation, that is, the percentage of activation in one software's thresholded statistic map that fell outside of the analysis mask of the other software is displayed in grey; left value for row software, right value for column software. For ds000001 increases, FSL permutation obtained no significant results, thus generating Dice coefficients of zero; for ds000109 decreases, only AFNI and FSL parametric obtained a result and hence only one coefficient is displayed. Dice coefficients are mostly below 0.5, parametric-nonparametric intrasoftware results are generally higher; ds000120's F-statistic results are notably high, at 0.684, perhaps because it is testing a main effect with ample power

Spill over values, given by the grey values beneath the Dice coefficients in Figure 3.6, are generally largest for SPM comparisons. They are particularly prominent in the negative activation plot for ds000001, where there is at least 30% spill over for all parametric pairwise comparisons, the largest being 39% for SPM/FSL. Recalling that these values are the percentage of activation which occurred within one package that was outside the other package's analysis mask, this is likely due to the fact SPM consistently had the smallest analysis mask out of the three packages, while FSL had the largest. In our ds000001 reanalyses, SPM's group-level analysis mask was made up of 175269 voxels, while AFNI's had 198295 voxels and 251517 for FSL. For ds000109, SPM's group-level mask contained 178461 voxels compared to AFNI's 212721 and FSL's 236889. Finally, for ds000120, SPM had 174059 voxels to AFNI's 208340. (Note, FSL consistently had slightly more non-zero voxels in it's mask image than in it's corresponding statistical results images).

Further evidence of spatial variability is also exhibited by the Euler characteristic (EC) plots for the parametric analyses presented in Figure 3.7, top (and supplementary Fig. A.12 for ds000120), complemented by the cluster count plots in Figure 3.8, top. We note that because the EC plots were created by thresholding each software's statistic map at a fixed range of t-values (without the computation of p-values), differences between the parametric and nonparametric EC curves are due to differences between the first-level models used in each case (mixed-effects for parametric, OLS for nonparametric), rather than the actual parametric and nonparametric inference procedures used to obtain p-values.

As the EC counts the number of clusters, minus the number of 'handles', plus the numbers of 'holes' in an image, for large thresholds we expect the EC to closely approximate the number of clusters of significant activation present in the equivalent thresholded map. This is confirmed by Figures 3.7 and 3.8. Both figures show that across the two studies FSL had a smaller number of activated clusters at larger thresholds. For ds000001, the peak cluster count value (Fig. 3.8, top left) occurs at

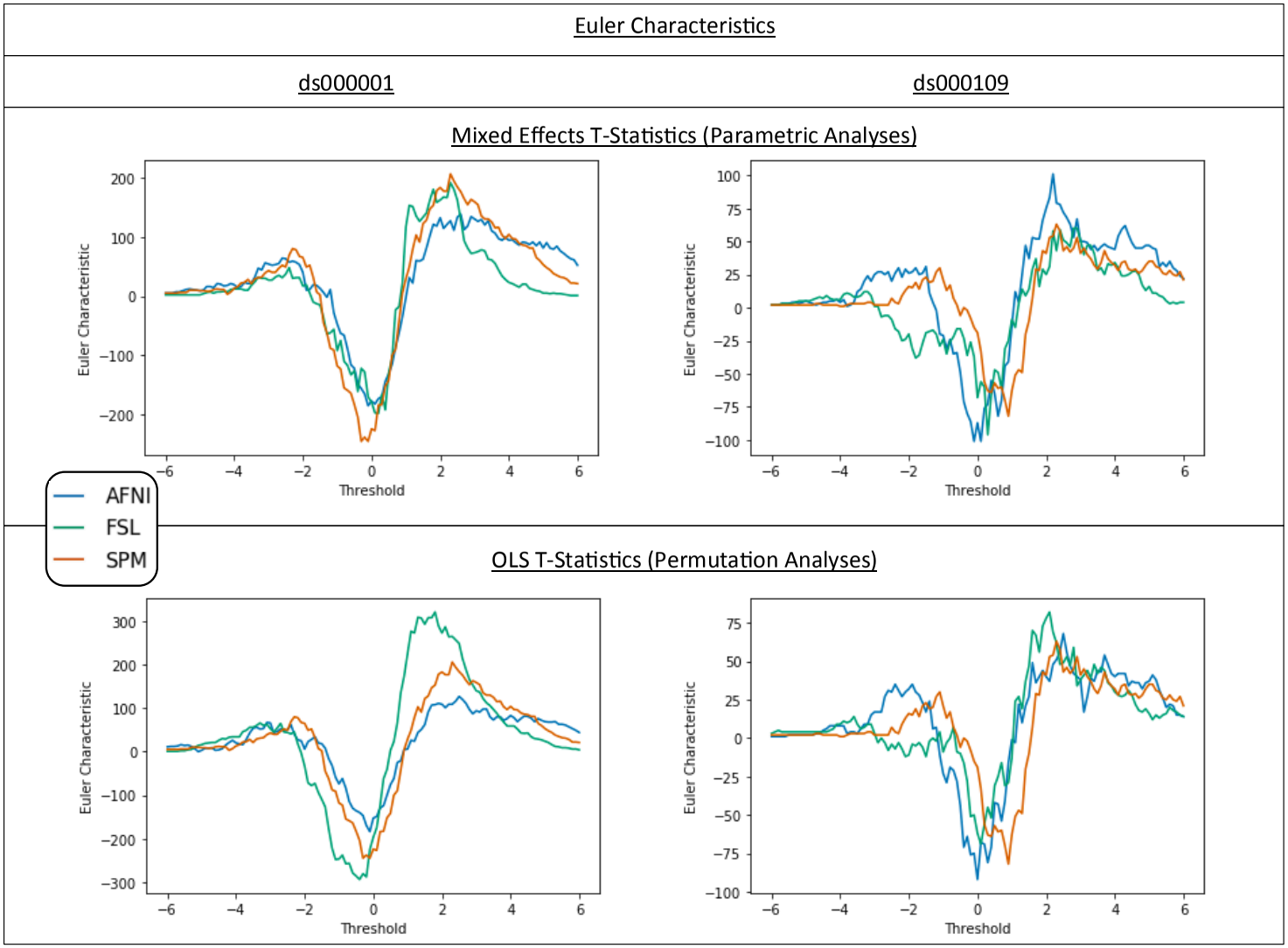


Figure 3.7: Euler characteristic (EC) plots for ds000001 and ds000109. On top, comparisons of the Euler characteristic computed for each software's t-statistic map from our reanalyses using a range of t-value thresholds between -6 and 6. Below, comparisons of the ECs calculated using the same thresholds on the corresponding t-statistic images for permutation inference within each package. For each t-value the EC summarises the topology of the thresholded image, and the curves provide a signature of the structure of the entire image. For extreme thresholds the EC approximates the number of clusters, allowing a simple interpretation of the curves: For example, for ds000001 parametric analyses, FSL clearly has the fewest clusters for positive thresholds.

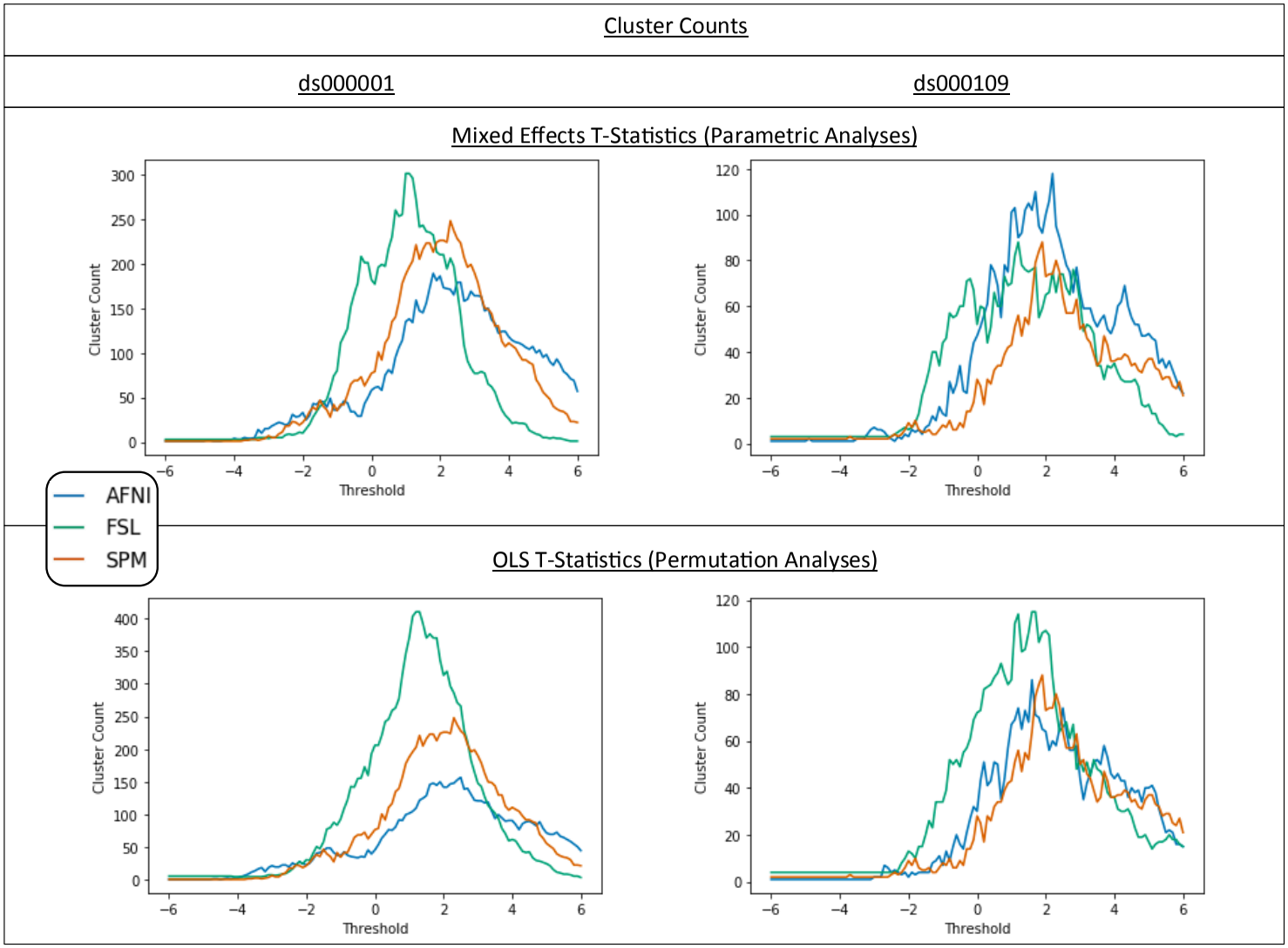


Figure 3.8: Cluster count plots for ds000001 and ds000109. On top, comparisons of the number of cluster found in each software's t-statistic map from our reanalyses using a range of t-value thresholds between -6 and 6. Below, comparisons of the cluster counts calculated using the same thresholds on the corresponding t-statistic images for permutation inference within each package.

a lower threshold for FSL. This plot suggests that in general FSL's t-statistic values were more liberal here – the initial rise of the FSL curve signifies the t-statistic image breaking up into clusters at lower thresholds than AFNI and SPM, and then as the clusters begin to get 'thresholded out' this causes the FSL curve to dip below the other two packages. The EC plots highlight overall topological differences in the statistic maps: if the images were the same up to an image-wide monotonic transformation, this would be revealed by the EC curves having the same general shape but with some portions shifted or compressed. In this sense, the distinct shapes seen in portions of the curves (e.g. for ds000109, negative thresholds) suggest differences in the topologies of each software's activation pattern.

3.2.2 Cross-Software Variability for Nonparametric Inference

Consistent with the parametric inference results, activation localization and statistic values varied greatly between packages for the permutation test results computed for ds000001 and ds000109.

Before reviewing statistic map comparisons, we stress that the goal of these nonparametric analyses was to obtain FWE-corrected cluster p-values with weaker assumptions. Therefore, the permutation test unthresholded statistic maps are not "nonparametric" maps, but rather usual one-sample t-test maps that form the basis of permutation analyses. While SPM's parametric analysis uses the same one-sample t-test as it's nonparametric counterpart, AFNI's and FSL's parametric models use a mixed-effects model and weighted least squares. In this sense, (in contrast to the thresholded maps) all comparisons of the nonparametric test statistic values do not convey information about nonparametric inference per se, but compare differences in the preprocessing and first-level modelling within the three packages while holding the second-level model constant.

Quantitative assessment with Dice coefficients are shown in Figure 3.6 ("perm" vs "perm" cells) and,in accordance with the parametric results, are generally poor.

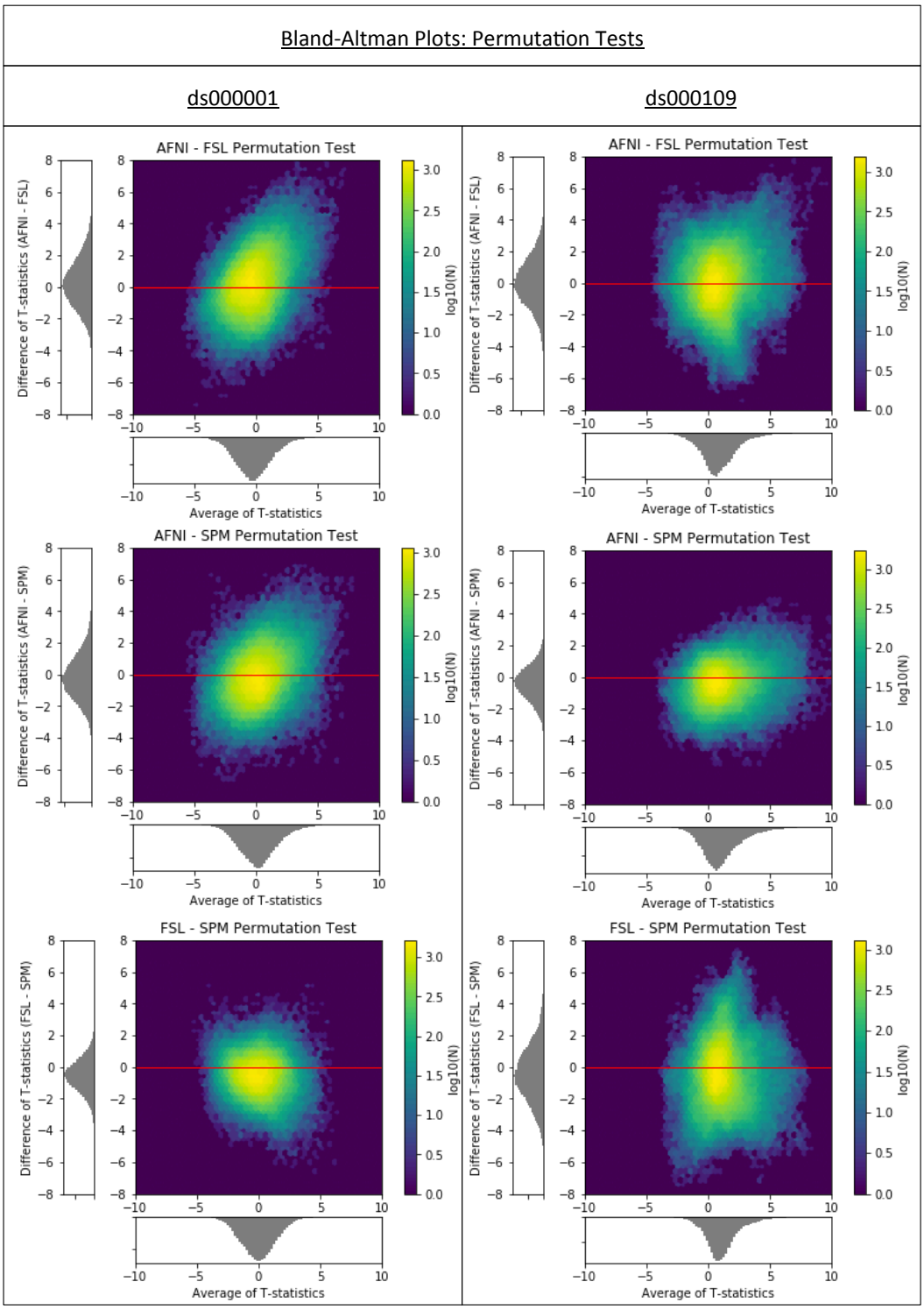


Figure 3.9: Cross-software Bland-Altman 2D histograms for the ds000001 and ds000109 studies comparing the unthresholded group-level t-statistic maps computed using permutation inference methods within AFNI, FSL, and SPM. Similar to the results obtained using parametric inferences in Figure 3.4, all of the densities indicate large differences in the size of activations determined within each package.

Like the parametric analyses, AFNI/SPM Dice values are altogether better than the other comparisons. For ds000001, FSL's nonparametric method found no significant clusters, and thus all Dice coefficients connected to this analysis are zero. However, note that the significant regions found in the other parametric and nonparametric results for this study mostly comprise of a single activation cluster spanning the lateral and medial frontal cortex, insular cortex, basal ganglia, and brainstem – an extensive and irregularly-shaped cluster that could easily become disconnected and thus lose significance. As before, ds000109 Dice values are generally better than ds000001.

The nonparametric Bland-Altman plots (Fig. 3.9) show substantial spread qualitatively similar to the parametric ones (Fig. 3.4), and correlations between the nonparametric statistics maps follow a similar trend to the parametric comparisons (Table 3.3). EC curves (Fig. 3.7, bottom) again exhibit considerable topological variation between software packages. Notably, while AFNI and SPM's EC curves are relatively similar across choice of inference method, FSL permutation inference determined substantially more clusters than parametric inference for small positive thresholds in both studies (Figs. 3.7 and 3.8, bottom).

3.2.3 Intra-Software Variability, Parametric vs Nonparametric

Comparisons of parametric and permutation test inference results within each package hold all preprocessing and first level modelling constant, only varying the second level model and inference procedure. The level of agreement between the two inferences *within* each package varied greatly across software. Before making comparisons, we note that since SPM's parametric and nonparametric inference share the same group-level model, the unthresholded statistic images produced using each inference model are identical here.

The thresholded statistic maps are generally similar within each of the software packages (ds000001: Fig. A.2 vs Fig. A.3; ds000109: Fig. A.4 vs A.5), with the exception of FSL's nonparametric inference 'decreases-only' finding for ds000001.

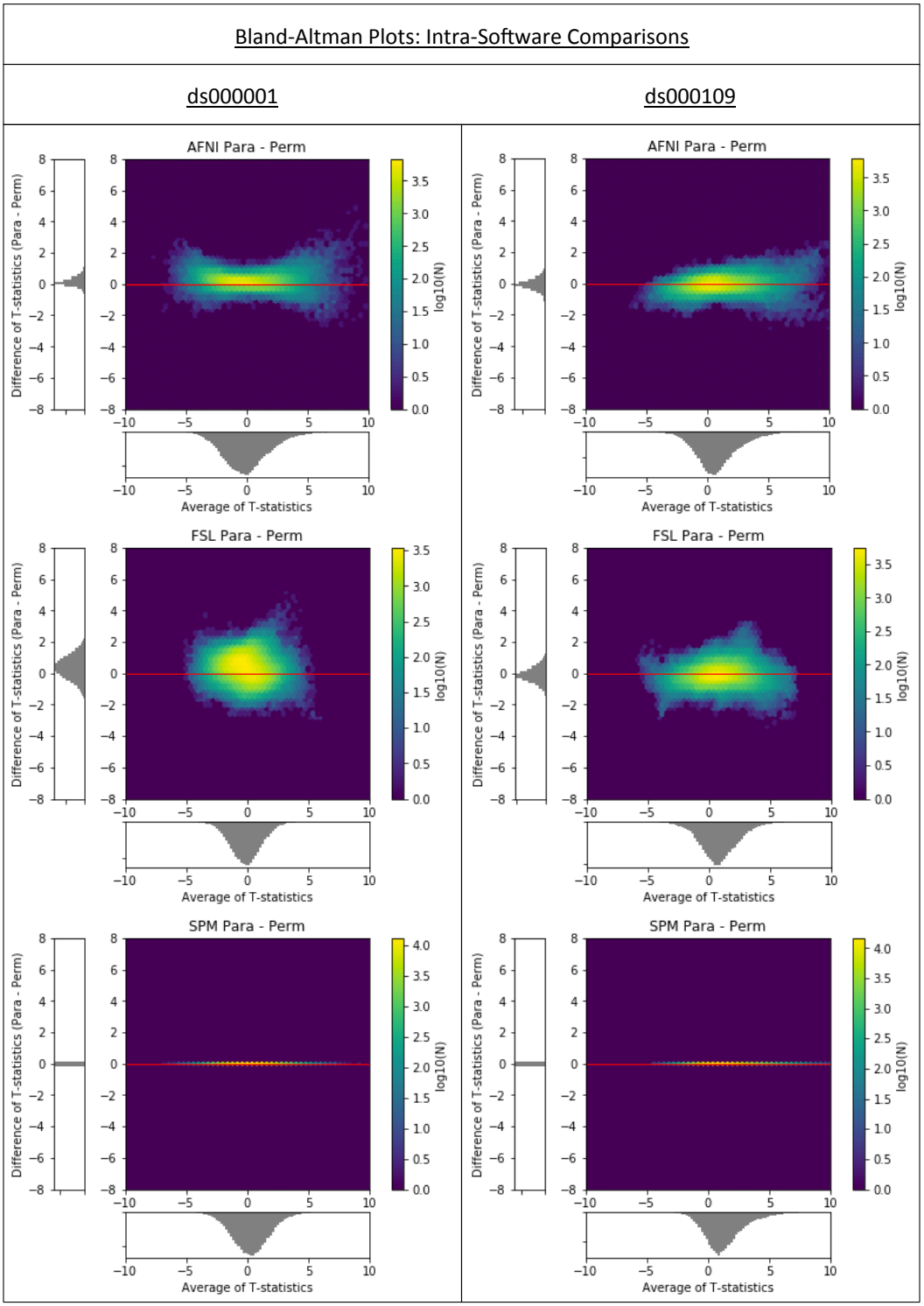


Figure 3.10: Intrasoftware Bland-Altman 2D histograms for the ds000001 and ds000109 studies comparing the unthresholded group-level t-statistic maps computed for parametric and nonparametric inference methods in AFNI, FSL and SPM. Each comparison here uses the same preprocessed data, varying only the second-level statistical model. SPM's parametric and nonparametric both use the same (unweighted) one-sample t-test, and thus show no differences. AFNI and FSL's parametric models use iterative estimation of between subject variance and weighted least squares and show some differences, but smaller than the between-software comparisons.

Unthresholded maps are notably more similar for ds000109 (Fig. A.9 vs A.10) than for ds000001 (Fig. A.7 vs Fig. A.10), again noting that SPM's pairs of maps are identical here.

Bland-Altman plots (Fig. 3.10) reveal much greater levels of parametric-nonparametric agreement, with AFNI displaying greater agreement than FSL. For FSL, we selectively investigated voxels that differed by the greatest amount and often found individual subjects responsible: a single subject with a large observation can drive a conventional one-sample t-test, but when the same subject also has large intrasubject variance FSL's mixed-effect model downweights that subject, leading to a substantially different t-test. For ds000109, the increased difference in AFNI's values for larger statistic values could also reflect a similar downweighting procedure within the package.

The Dice coefficients comparing the thresholded permutation test and parametric inferences are generally the best of any (Fig. 3.9, 3-element lower diagonal). In general, the origin of parametric-nonparametric differences are parametric inference determining a slightly larger number of significant clusters.

3.3 Discussion

Our results have displayed extensive differences in the results between the analysis pipelines of the three software packages. The low Dice values and differences in Euler characteristics (Figs. 3.6 and 3.7) are particularly salient, showing heterogeneity in the sizes and shapes of clusters determined across packages, indicating a strong dependence on software in terms of the anatomical regions covered by the activation. While some authors have pointed out the limitations of interpreting differences in the set of voxels that make up a significant activation when using clusterwise inference (which makes assessments based on topological properties of statistical maps) (Chumbley and Friston, 2009), we see merit in the use of quantitative measurements

such as Dice due to the ultimate application of statistical maps to infer the precise areas of the brain active during a task. While a deeper analysis on the differences between topological aspects of images across software would be valuable, there are inherent difficulties in this approach, such as identifying corresponding features between maps when the number and size of activations reported are variable across software.

It is notable that the level of variation in our analysis results also fluctuated across the datasets we analyzed. This is highlighted in our Dice comparisons, where the ds000001 Dice coefficients are considerably smaller than ds000109 for both the inter and intra-software comparisons. The relatively poor performance of ds000001 may be due to the smaller sample size for this study (16 vs 21 for d000109), as well as the particular inference method used in the study. For ds000001, group-level inference was conducted using a cluster-forming threshold of $p < 0.01$ uncorrected. A recent study (Eklund et al., 2016) found that parametric inference for a one-sample t-test at this threshold in AFNI, FSL and SPM resulted in false-positive rates far exceeding the nominal level – severely for cluster-forming threshold $p < 0.01$, modestly for $p < 0.001$ – while nonparametric permutation inference performed closer to the expected 5% FWE level. The results obtained here for ds000001 are consistent with these findings: across all three software packages, the thresholded images produced from permutation test inference display fewer significant clusters than the corresponding parametric maps. While the cluster-defining threshold $p < 0.005$ applied in the ds000109 study was not analyzed in Eklund et al., consistency between packages using parametric and nonparametric inference was greater for this study.

Notably, while all packages are purportedly using the same MNI atlas space, an appreciable amount of activation detected by AFNI and FSL fell outside of SPM's analysis mask (shown by the 'spill over' values displayed in grey, Fig. 3.6). Considerable differences in mask sizes are likely to have been a major factor for the disparities in activation and low Dice coefficients seen across packages. For effects close

to the edge of the brain, a larger analysis mask allows for a larger cluster volume, which can ultimately be the difference as to whether a cluster is determined as significantly active or not. This may explain why only FSL, which had the largest analysis mask, found an auditory response in the ds000109 study, or why Dice coefficients are generally worse for negative activations than positive in our ds000001 renanalyses, where positive clusters were on the whole reported in more central anatomical regions. Another possible reason for poor Dice values here is that the size and number of clusters determined for negative activations was smaller than that of positive activations. As Dice and spill-over values are proportional measurements, this means they will have been more susceptible to differences in cluster and mask size for the negative activations relative to the positive. Disagreement in atlas space may have contributed to the lack of structure in the Bland-Altman plots, however no gross misalignment between packages was evident (Fig. A.1). While far from perfect, the ds000120 AFNI and SPM thresholded results have the best Dice similarity score, likely due to the use of a very strong main effect as an outcome of interest.

Qualitative comparison of the results provide some optimism, with certain patterns of activation found across all packages. For example, the ds000001 parametric analyses were unanimous in determining significant activation in the anterior insula. While there is greater discordance over the precise location of activation within the anterior insular region, as well as the precise statistic values here, altogether our results align. This may substantiate that the strongest effects are robust across packages, supported by our own comparisons of the unthresholded maps that showed moderate agreement between software packages in areas with strong signal (many of these anatomical regions were identified across all three packages' NeuroSynth association analyses) but greater disagreement elsewhere, with ds000109 and ds000120 displaying more consistency than ds0000001. However, in making these qualitative comparisons, what has become most apparent is the importance for researchers to – at the very least – share their final statistical maps. The reasons

for this are exhibited most clearly by our ds000109 analyses; the visual slice comparisons of our replications of the main figure from the original study in Figure 3.1, shown alongside the publication figure itself, look remarkably similar and could lead to the conclusion that each package's results highly agree. It is only when analyzing these results over the whole brain that we discover broad differences in these activation patterns, e.g. positive activation identified in the auditory cortex in FSL that was not reported by AFNI and SPM, and significant deactivation determined only by AFNI and FSL.

At the start of our investigations, we selected a common set of preprocessing steps to be applied within each software package across all studies regardless of whether they had been used in the original analysis. This was to maximise the comparability of the results while maintaining consistency with best practices within the community. However, several complications arose during our analyses. For ds000001, orientation information was missing from seven of the subjects' structural and functional scans. Because the source DICOM files were no longer available, it was not possible to retrieve the original position matrices. As a consequence of this, the structural and functional images were misaligned, resulting in suboptimal coregistration during our analyses. Additionally, a bug in the event-files induced during data conversion to the BIDS standard had resulted in a loss of the some of the event timings. Thanks to the cooperation of BIDS and OpenfMRI these problems were solved; a revised dataset (Revision: 2.0.4) was uploaded to OpenfMRI and used in our analysis.

Future efforts would be strengthened by additional sharing of analysis scripts and statistic maps, enabling confirmation of analyses that follow original procedures and permitting more quantitative comparison of statistic maps. We have made all of our analysis scripts and statistic maps available, and we hope more researchers join this trend to advance openness in neuroimaging science.

3.3.1 Limitations

This study has mainly focused on comparing statistic maps, since these are the images studied to make judgements about localisation and determine the neuroscientific interpretation of results. However, by comparing the statistic maps obtained at the end of the pipeline, we have only assessed the net accumulation of differences across the entire analysis procedure. To illuminate the specific steps that contribute most to this variation, further in-depth assessment of software differences at each stage of the analysis pipeline will be required. One recent example of this was a study that investigated differences in the prewhitening procedures conducted in AFNI, FSL and SPM, by employing an analysis pipeline that used a single software package to carry out all other stages of processing (Olszowy et al., 2019). Further work could consider the factorial expansion of all possible combinations of preprocessing, first-level modelling, and second-level modelling, akin to previous efforts in assessing reproducibility over a number of pipelines (Strother et al., 2002).

Due to the restrictive requirements of this investigation – the necessity for published task-based fMRI data using analysis methods compatible in AFNI, FSL, and SPM – the three studies analyzed here were found to be the only datasets hosted on OpenfMRI suited to the aims of our investigation. Of the datasets that were not used, the most common reasons for exclusion were that no publication was associated to the data, that the sample size of the study was too small, or that custom software or region of interest analysis had been used as part of the analysis pipeline which was not feasible across the three software packages. Nevertheless, a greater sample of studies will need to be replicated across the packages to gain a more comprehensive understanding of the variability between software and validate the results found here. With increasing access to population neuroimaging studies, where thousands of fMRI subject data are available, a future study could test for non-zero software-related variation by splitting a large dataset (e.g. UK Biobank (Alfaro-Almagro et al.,

2018), N > 10,000) into smaller subsets to generate an extensive collection of replication analyses across the three packages. This may allow for the creation of a null-distribution from which differences between software results could be assessed in terms of statistical significance and confidence intervals, expanding on the raw concrete differences between t-statistic maps highlighted in this effort. As simulation techniques become more advanced, there is also the potential for the creation of synthetic subject-level fMRI data as a ground truth to which each software package's results could be compared (Ellis et al., 2019).

Of the datasets we did use, subject data were missing from both the ds000109 and ds000120 datasets. For ds000109, while 29 young adults were scanned for the false belief task, only 21 were present in the dataset; for ds000120, we analyzed 17 subjects instead of the 30 used in the original study. These analyses therefore should not be compared like-for-like with the published results, and have substantially less statistical power than the original studies. Overall, our sample sizes for the three datasets analyzed (ds000001, ds000109, ds000120) are 16, 21, and 17 respectively. While small, these sample sizes are fairly representative of a typical functional neuroimaging study over the past two decades – between 1995 and 2015, the median sample size of an fMRI study increased steadily from 8 to 22 (Poldrack et al., 2017). This increase has continued, and a review of 2017 publications found a median sample size of 33 (Yeung, 2018). Therefore, while our datasets are important for judging previous work, a future comparison exercise with larger datasets would be a valuable addition to the literature.

We have kept many parameters fixed in our analyses, such as the use of non-linear registration for all software packages, and the addition of motion regressors in all our design matrices. Further investigation is warranted into how changes in these variables influence the analysis; for example, while we decided to fix a 2mm cubic voxel size in all packages (since this is the default in FSL and SPM), a recent study found that alterations in this parameter can significantly impact statistical inference

(Mueller et al., 2017). There are also many areas of the parameter space we have not explored, such as the inclusion of analyses that use small volume corrections, more stringent cluster-forming thresholds (Eklund et al., 2016; Woo et al., 2014), and two-tailed testing (Chen et al., 2018).

3.4 Conclusion

Across all three of the studies reanalyzed here we have discovered considerable differences between the AFNI, FSL, and SPM results. The scale of these differences has been highlighted by each of the quantitative metrics applied to compare the group-level statistic maps: Dice coefficients were commonly less than 50% for cross-software comparisons, Bland-Altman plots showed that differences between reported t-statistic values were as large as 4 for a considerable quantity of voxels, and Euler characteristic curves displayed a divergence in the number of clusters being reported in each software, even at large thresholds.

In reporting these comparisons, we are not making any statements as to which software package is better or worse. Without a gold standard to compare against no such claims can be made, and we believe further development of well-validated pipelines by multiple groups can encourage innovation and ultimately benefit the field. Rather, we feel that the key contribution of our work is the quantitative measurement of inter-software differences on common datasets. Our finding that exceedingly weak effects may not generalize across packages, evidenced across all three of our analyses, is the primary take-home message of this work. While larger effects were found to be more robust – demonstrated by the similar Neurosynth association analysis results that suggest some alignment in the final qualitative conclusions that can be drawn from all three softwares' statistical maps – we stress that our analyses have been conducted under particularly favourable conditions; in this effort, we have sought out studies with a strong, primary effect and then made extensive ef-

forts to harmonize our analyses. Because of this, at best our results present an optimistic view of inter-software disparities. To better understand the underlying differences between software, further work on quantification of pipeline-related variation is needed, which in the long-term will hopefully lead to harmonisation in software implementation to reduce these differences. Another line of work would be the creation of integrative intra-study, ensemble learning techniques to integrate inconsistent findings. An additional contribution with this effort has been to provide generalizable measures and metrics in order to enable software validation, which we hope may benefit any further comprehensive comparison of software packages.

CHAPTER 4

Spatial Confidence Sets for Raw Effect Size Images

The mass-univariate approach for functional magnetic resonance imaging (fMRI) analysis remains a widely used statistical tool within neuroimaging. However, as discussed in Chapter 1, this method suffers from at least two fundamental limitations: First, with sufficient sample sizes there is high enough statistical power to reject the null hypothesis *everywhere*, making it difficult if not impossible to localize effects of interest. Second, with any sample size, when cluster-size inference is used a significant p-value only indicates that a cluster is larger than chance. Therefore, no notion of confidence is available to express the size or location of a cluster that could be expected with repeated sampling from the population.

For the reasons discussed above, alongside further concerns about misconceptions and the misuse of p-values in statistical testing (Nuzzo, 2014; Wasserstein et al., 2016), there has been a growing consensus among sections of the neuroimaging community that the statistical results commonly reported in the literature should be supplemented by effect estimates (Chen et al., 2017; Nichols et al., 2017). The main argument put forward supporting raw effect sizes is that they increase the interpretability of statistical results, highlighting the magnitude of statistically significant differences and providing another layer of evidence to support the overall scientific conclusions inferred from an fMRI study. This may also help tackle reproducibil-

ity concerns that have become prominent within the field due to failed attempts in replicating published neuroimaging results (Poldrack et al., 2017), a problem aggravated by the ubiquity of underpowered studies in the fMRI literature where traditional statistical inference methods are unlikely to detect the majority of meaningful effects (Cremers et al., 2017; Turner et al., 2018).

In this chapter, we address these issues by extending on a method proposed by Sommerfeld, Sain, and Schwartzman (2018) (*SSS*) to develop spatial Confidence Sets (CSs) on clusters found in thresholded raw effect size maps. Unlike hypothesis testing, the CSs allow for inference on *non-zero* raw effect sizes. While the method can be applied to any parameter in a mass-univariate general linear model, here we focus inference on the mean percentage blood-oxygen-level-dependent (BOLD) change raw effect. For a cluster-forming threshold c, and a predetermined confidence level $1-\alpha$, the CSs comprise of two sets: the upper CS (denoted $\hat{\mathcal{A}}_c^+$, red voxels in Fig. 4.1), giving all voxels we can assert have a percentage BOLD raw effect size truly *greater* than c; and the lower CS ($\hat{\mathcal{A}}_c^-$, blue voxels overlapped by yellow and red in Fig. 4.1), for which all voxels *outside* this set we can assert have a percentage BOLD raw effect size truly *less* than c. The upper CS is smaller and nested inside the lower CS, and the assertion is made with $(1-\alpha)100\%$ confidence holding simultaneously for both regions. Figure 4.1 provides an illustration of the schematic we will use to display the CSs, also showing the point estimate set ($\hat{\mathcal{A}}_c$, yellow voxels overlapped by red) obtained by thresholding the data at c.

This chapter is organized as follows: First of all we summarize the key theory of the CSs, before detailing our proposed advancements to the methods used in *SSS*. We then describe the settings for our simulations, and provide background information about the Human Connectome Project (HCP) dataset analyzed in this work. Finally we report the results of our simulations, before presenting the CSs obtained for the HCP data.

All aspects of this work were carried out under the supervision and mentor-

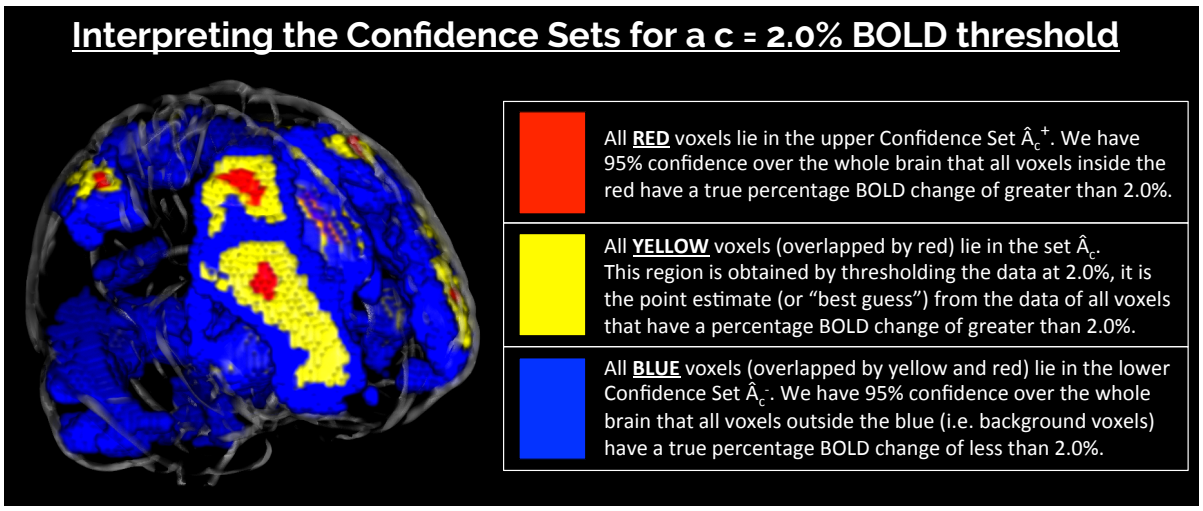


Figure 4.1: Schematic of the colour-coded regions used to visually represent the Confidence Sets (CSs) and point estimate set. CSs displayed in the glass brain were obtained by applying the method to 80 participants contrast data from the Human Connectome Project working memory task, using a a $c = 2.0\%$ BOLD change threshold at a confidence level of $1-\alpha = 95\%$.

ship of Prof. Thomas Nichols and Dr. Armin Schwartzman. Theoretical aspects of this work were developed in collaboration with Dr. Fabian Telschow. Dr. Telschow also made contributions to the simulation and simulations figures code.

4.1 Theory

4.1.1 Overview

A comprehensive treatment of the original method, including proofs, can be found in *SSS*. Here we develop the method specifically for the general linear model (GLM) and describe our own enhancements to the method. While the method can be performed for subject-level inference, we will motivate the method in the context of a group-level analysis, describing how the method can be applied to subject-level %BOLD estimate maps in order to obtain group-level CSs making confidence statements about %BOLD effect sizes relating to the entire population from which the participants were drawn.

For a compact domain $S \subset \mathbb{R}^D$, e.g. $D = 3$, consider the GLM at location

$\boldsymbol{s} \in S$,

$$\boldsymbol{Y}(\boldsymbol{s}) = \boldsymbol{X}\boldsymbol{\beta}(\boldsymbol{s}) + \boldsymbol{\epsilon}(\boldsymbol{s}) \tag{4.1}$$

where $\boldsymbol{Y}(\boldsymbol{s})$ is an $N \times 1$ vector of observations at $\boldsymbol{s}$, $\boldsymbol{X}$ is an $N \times p$ design matrix, $\boldsymbol{\beta}(\boldsymbol{s})$ is an $p \times 1$ vector of unknown coefficients, and $\boldsymbol{\epsilon}(\boldsymbol{s})$ an $N \times 1$ vector of mean-zero errors, independent over observations, and with each $\epsilon_i(\boldsymbol{s})$ having common variance $\sigma^2(\boldsymbol{s})$ and some unspecified spatial correlation. (Throughout we use boldface to indicate a vector- or matrix-valued variable.) In the context of a task-fMRI analysis, $\boldsymbol{Y}(\boldsymbol{s})$ is a vector of subject-level %BOLD response estimate maps obtained by applying a first-level GLM to each of the N participants functional data.

For a $p \times 1$ contrast vector $\boldsymbol{w}$, we seek to infer regions of the brain where a contrast of interest $\boldsymbol{w}^T\boldsymbol{\beta}(\boldsymbol{s})$ has exceeded a fixed threshold c. Particularly, we are interested in the noise-free, population cluster defined as:

$$\mathcal{A}_{\mathrm{c}} = \{\boldsymbol{s} \in S : \boldsymbol{w}^T\boldsymbol{\beta}(\boldsymbol{s}) \geq c\}. \tag{4.2}$$

Since we are unable to determine this excursion set in practice, our solution is to find spatial CSs: an upper set $\hat{\mathcal{A}}_{\mathrm{c}}^{+}$ and lower set $\hat{\mathcal{A}}_{\mathrm{c}}^{-}$ that surround $\mathcal{A}_{\mathrm{c}}$ for a desired confidence level of, for example, 95%. We emphasize that these clusters regard the raw units of the signal. Going forward, we assume that the design matrix $\boldsymbol{X}$ and contrast $\boldsymbol{w}$ have been carefully chosen so that $\boldsymbol{w}^T\hat{\boldsymbol{\beta}}(\boldsymbol{s})$ has the interpretation of mean %BOLD change across the group. For example, in a one-sample group fMRI model where data $\boldsymbol{Y}(\boldsymbol{s})$ have %BOLD units, choosing $\boldsymbol{X}$ as a column of 1's and $\boldsymbol{w} = (1)$ would ensure that $\boldsymbol{w}^T\hat{\boldsymbol{\beta}}$ has units of %BOLD change[1]. In this setting, we wish to obtain an upper CS, $\hat{\mathcal{A}}_{\mathrm{c}}^{+}$, such that we have 95% confidence all voxels *contained* in this set have a population raw effect size *greater* than, for example, $c = 2.0\%$ BOLD change, and a lower CS, $\hat{\mathcal{A}}_{\mathrm{c}}^{-}$, such that we have 95% confidence all voxels *outside* of this set

[1] For examples of how to set up more complex designs and contrasts, see Figure A.2. in the Appendix A section of (Poldrack et al., 2011).

have a population raw effect size *less* than 2.0% BOLD change. Moreover, we desire that the 95% confidence statement holds simultaneously across both CSs at once. *SSS* show that a construction of such CSs is possible within the general linear model framework using the following key result.

Result 1. *Consider the general linear model setup described in (4.1). Let $\hat{\boldsymbol{\beta}}(\boldsymbol{s})$ denote the ordinary least squares estimator of $\boldsymbol{\beta}(\boldsymbol{s})$, $\hat{\boldsymbol{\beta}}(\boldsymbol{s}) = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}(\boldsymbol{s})$, and define $v_w^2 = \boldsymbol{w}^T(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{w}$ to be the normalised variance of the contrast estimate.*

Then for a constant k, and for upper and lower CSs defined as

$$\hat{\mathcal{A}}_{\mathrm{c}}^{+} := \left\{ \boldsymbol{s} : \boldsymbol{w}^T\hat{\boldsymbol{\beta}}(\boldsymbol{s}) \geq c + k\,\hat{\sigma}(\boldsymbol{s})v_w \right\}, \quad \hat{\mathcal{A}}_{\mathrm{c}}^{-} := \left\{ \boldsymbol{s} : \boldsymbol{w}^T\hat{\boldsymbol{\beta}}(\boldsymbol{s}) \geq c - k\,\hat{\sigma}(\boldsymbol{s})v_w \right\},$$

the limiting coverage of the CSs is

$$\lim_{n\to\infty} P\left[\hat{\mathcal{A}}_{\mathrm{c}}^{+} \subset \mathcal{A}_{\mathrm{c}} \subset \hat{\mathcal{A}}_{\mathrm{c}}^{-} \right] = P\left[\sup_{\boldsymbol{s}\in\partial\mathcal{A}_{\mathrm{c}}} |G(\boldsymbol{s})| \leq k \right],$$

where $\partial\mathcal{A}_{\mathrm{c}}$ denotes the boundary of $\mathcal{A}_{\mathrm{c}}$, and $G(\boldsymbol{s})$ is a smooth Gaussian field on S with mean zero, unit variance, and with the same spatial correlation as each $\epsilon_i(\boldsymbol{s})$.

Result 1 is subject to continuity of the relevant fields and some basic conditions on the increments and moments of the error field $\boldsymbol{\epsilon}$. A list of these assumptions, as well as a proof of Result 1, are itemized in *SSS*.

For a predetermined confidence level $1-\alpha$ (e.g. $1-\alpha = 95\%$), by choosing k such that

$$P\left[\sup_{\boldsymbol{s}\in\partial\mathcal{A}_{\mathrm{c}}} |G(\boldsymbol{s})| \leq k \right] \geq 1-\alpha, \tag{4.3}$$

Result 1 ensures with asymptotic probability of (at least) $1-\alpha$ that $\hat{\mathcal{A}}_{\mathrm{c}}^{-}$ contains the true $\mathcal{A}_{\mathrm{c}}$, and $\hat{\mathcal{A}}_{\mathrm{c}}^{+}$ is contained within $\mathcal{A}_{\mathrm{c}}$. In practice, k is determined as the $(1-\alpha)100$ percentile of the maximum distribution of the asymptotic absolute error process $|G(\boldsymbol{s})|$ over the true boundary set $\partial\mathcal{A}_{\mathrm{c}} = \{\boldsymbol{s} : \boldsymbol{w}^T\boldsymbol{\beta}(\boldsymbol{s}) = c\}$. The upper CS taken away from the lower CS $\left(\hat{\mathcal{A}}_{\mathrm{c}}^{-} \cap \left(\hat{\mathcal{A}}_{\mathrm{c}}^{+} \right)^{\mathrm{c}} \right)$ can be interpreted analogously to a stan-

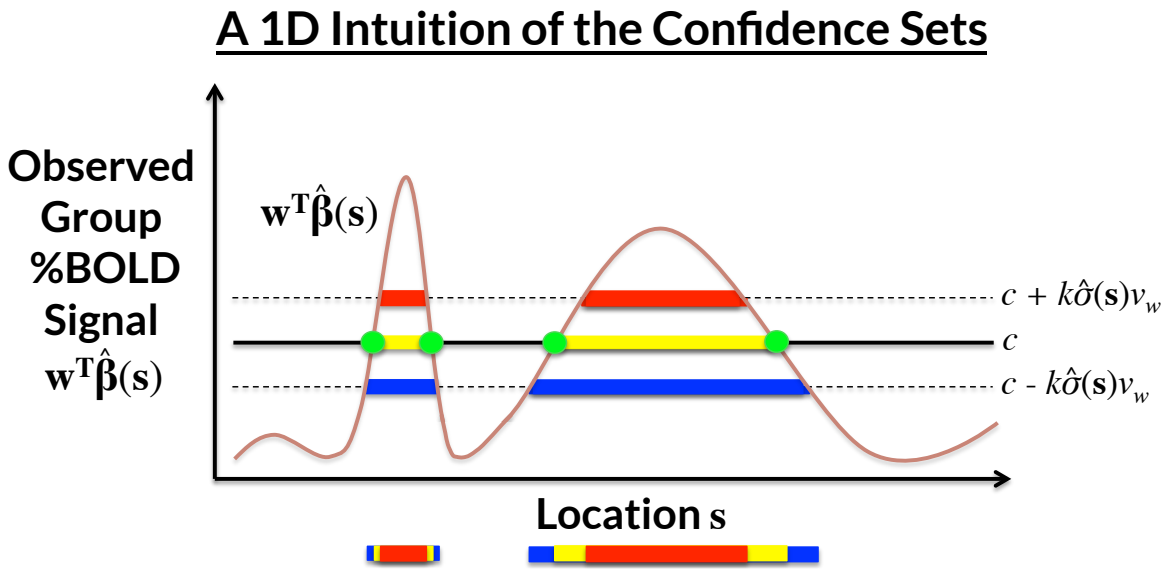


Figure 4.2: A demonstration of how the CSs are computed for a realization of the GLM $\boldsymbol{Y}(\boldsymbol{s}) = \boldsymbol{X}\boldsymbol{\beta}(\boldsymbol{s}) + \boldsymbol{\epsilon}(\boldsymbol{s})$ in 1-dimension, for each location $\boldsymbol{s}$. The yellow voxels $\hat{\mathcal{A}}_c$ are obtained by thresholding the observed group contrast map at a threshold c; this is the best guess of $\mathcal{A}_c$, the set of voxels whose true, noise-free raw effect surpasses c. The red upper CS $\hat{\mathcal{A}}_c^+$ and blue lower CS $\hat{\mathcal{A}}_c^-$ are computed by thresholding the signal at $c + k\,\hat{\sigma}(\boldsymbol{s})v_w$ and $c - k\,\hat{\sigma}(\boldsymbol{s})v_w$ respectively. We have $(1-\alpha)100\%$ confidence that $\hat{\mathcal{A}}_c^+ \subset \mathcal{A}_c \subset \hat{\mathcal{A}}_c^-$, i.e. that $\hat{\mathcal{A}}_c^+$ (red) is completely within the true $\mathcal{A}_c$, and $\mathcal{A}_c$ is completely within $\hat{\mathcal{A}}_c^-$ (blue). We find the critical value k from the $(1-\alpha)100$ percentile of the maximum distribution of the absolute error process over the estimated boundary $\partial\hat{\mathcal{A}}_c$ (green ●'s) using the Wild t-Bootstrap; $\hat{\sigma}(\boldsymbol{s})$ is the estimated standard deviation and v_w is the normalised contrast variance.

dard confidence interval: with a confidence of $1-\alpha$, we can assert the true boundary $\partial\mathcal{A}_c$ lies within this region. Here, we allude to the classical frequentist interpretation of confidence, where stated precisely, there is a probability of $1-\alpha$ that the region $\left(\hat{\mathcal{A}}_c^- \cap (\hat{\mathcal{A}}_c^+)^c\right)$ computed from a future experiment fully encompasses the true set boundary $\partial\mathcal{A}_c$.

Application of Result 1 presents us with two challenges: that the boundary set $\partial\mathcal{A}_c$ and the critical value k are both unknown. To solve the first problem, *SSS* propose using $\partial\hat{\mathcal{A}}_c$ as a plug-in estimate of $\partial\mathcal{A}_c$. However, there remain technicalities at to how the boundary is determined in any non-abstract setting, and in particular in a 3D image. In Section 4.1.3 we propose our own novel method for boundary estimation. Before that, we address the second problem, finding the critical value k via a Wild Bootstrap resampling scheme.

4.1.2 The Wild *t*-Bootstrap Method for Computation of *k*

To apply Result 1, we require knowledge of the tail distribution of the limiting Gaussian field $G(\boldsymbol{s})$ along the boundary $\partial\mathcal{A}_c$. However, the distribution of this field is unknown, because it is dependent on the unknown spatial correlation of the errors $\epsilon_i(\boldsymbol{s})$. We can approximate the maximum distribution of $G(\boldsymbol{s})$ using the Gaussian Wild Bootstrap (Chernozhukov et al., 2013), also known as the Gaussian Multiplier Bootstrap, which multiplies residuals by random values to create surrogate instances of the random errors.

SSS construct $G(\boldsymbol{s})$ as follows: The standardized residuals,

$$\tilde{\boldsymbol{\epsilon}}(\boldsymbol{s}) = \frac{\boldsymbol{Y}(\boldsymbol{s}) - \boldsymbol{X}\hat{\boldsymbol{\beta}}(\boldsymbol{s})}{\sigma(\boldsymbol{s})}, \tag{4.4}$$

are multiplied by i.i.d. Gaussian random variables, $r_1^*, \ldots, r_N^*$, summed and scaled, producing a bootstrap field

$$G^*(\boldsymbol{s}) = \frac{1}{\sqrt{N}} \sum_{i=1}^{N} r_i^* \tilde{\epsilon}_i(\boldsymbol{s}) \tag{4.5}$$

with approximately the same covariance as each error $\epsilon_i(\boldsymbol{s})$, where the superscript asterisk (*) indicates these are just one of many bootstrap realizations. With B bootstrap samples $G^*(\boldsymbol{s})$, we choose k as the $(1-\alpha)100$ percentile of the B suprema $\sup_{\boldsymbol{s}\in\partial\hat{\mathcal{A}}_c}|G^*(\boldsymbol{s})|$ to approximate the LHS of (4.3) and apply Result 1 to obtain the CSs.

Up to this point, we have summarized the Gaussian Wild Bootstrap methodology as proposed in *SSS*. However, when applying this method to our own simulations, we consistently found that our coverage results fell below the nominal level. This was particularly severe for 3D simulations we conducted using a small sample size ($N = 60$), where our results in some cases suffered from under-coverage 40% or more below the nominal level (see Fig. 4.8). Hence we made two alterations: First, while *SSS* used Gaussian multipliers, we found improved performance using Rademacher variables, where each r_i takes on 1 or -1 with probability 1/2; others have also reported improved performance with Rademacher variables as well (Davidson and Flachaire, 2008). Second, we implemented a Wild t-Bootstrap (Telschow and Schwartzman, 2019) method, normalizing the bootstrapped residuals $\tilde{\epsilon}_i(\boldsymbol{s})$ by their standard deviation $\hat{\sigma}^*(\boldsymbol{s})$. This detail was omitted in the proof of Result 1 provided in *SSS*, where the true standard deviation was assumed to be known. By taking into account the estimation of the standard deviation via the Wild t-Bootstrap, we found improved performance for moderate sample sizes. The Wild t-Bootstrap version of G is

$$\tilde{G}^*(\boldsymbol{s}) = \frac{1}{\sqrt{N}}\sum_{i=1}^{N} r_i^* \frac{\tilde{\epsilon}_i(\boldsymbol{s})}{\hat{\sigma}^*(\boldsymbol{s})}, \tag{4.6}$$

where $\hat{\sigma}^*(\boldsymbol{s})$ is the standard deviation of the present realization of the bootstrapped residuals $r_i^*\tilde{\epsilon}_i(\boldsymbol{s})$. We then determine k as described above but using $\tilde{G}^*(\boldsymbol{s})$ instead of $G^*(\boldsymbol{s})$. Going forward, we refer to this method as the "Wild t-Bootstrap", to be contrasted with the original "Gaussian Wild Bootstrap" method proposed in *SSS*.

With these two alterations we found a dramatic increase in performance for small sample sizes in 3D simulations. Notably, in contrast to the Gaussian Wild Boot-

strap, our simulation results presented in Section 4.3 suggest that empirical coverage rates for this modified procedure remain valid, i.e. stay *above* the nominal level.

4.1.3 Approximating the Boundary on a Discrete Lattice

In the previous section, we described the ideal bootstrap procedure used to obtain the maximum distribution of $G(\boldsymbol{s})$ along the boundary $\partial\mathcal{A}_{\mathrm{c}}$ in order to apply Result 1. However, in any practical application, data will be observed on a discrete grid of lattice points at a fixed resolution. Therefore, a key challenge is how to appropriately approximate the true continuous boundary $\partial\mathcal{A}_{\mathrm{c}}$ from the lattice representation of the data.

In *SSS*, spline-interpolation was used to estimate a 1D boundary at a resolution greater than their 2D sampled field (*SSS*, Section 4.1). However, to apply the method to fMRI data we will work with 3D images, and estimating a 2D spline boundary for a 3D field is more involved, requiring careful tuning of the spline basis to accommodate the structure of the 3D signal. Instead, we choose to use a first-order weighted linear interpolation method to approximate the signal values at estimated locations along the true, continuous boundary $\partial\mathcal{A}_{\mathrm{c}}$, providing a method of boundary estimation that is less computationally intensive than spline interpolation.

Consider two adjacent points on the lattice, $\boldsymbol{s_O}$ and $\boldsymbol{s_I}$, such that $\boldsymbol{s_O}$ lies outside of $\mathcal{A}_{\mathrm{c}}$, while $\boldsymbol{s_I}$ lies inside $\mathcal{A}_{\mathrm{c}}$. By the definition of $\mathcal{A}_{\mathrm{c}}$, $\boldsymbol{w}^T\boldsymbol{\beta}(\boldsymbol{s_O}) < c$, and $\boldsymbol{w}^T\boldsymbol{\beta}(\boldsymbol{s_I}) \geq c$. Under the assumption that the component of the signal between $\boldsymbol{s_O}$ and $\boldsymbol{s_I}$ increases linearly, we can find the location $\boldsymbol{s}^*$ between $\boldsymbol{s_O}$ and $\boldsymbol{s_I}$ such that $\boldsymbol{w}^T\boldsymbol{\beta}(\boldsymbol{s}^*) = c$, our estimate of where the true continuous boundary $\partial\mathcal{A}_{\mathrm{c}}$ crosses between $\boldsymbol{s_O}$ and $\boldsymbol{s_I}$. We can then construct a linear interpolant for the location $\boldsymbol{s}^*$, using weights

$$m_1 = \frac{\boldsymbol{w}^T\boldsymbol{\beta}(\boldsymbol{s_I}) - c}{\boldsymbol{w}^T\boldsymbol{\beta}(\boldsymbol{s_I}) - \boldsymbol{w}^T\boldsymbol{\beta}(\boldsymbol{s_O})}, \quad m_2 = \frac{c - \boldsymbol{w}^T\boldsymbol{\beta}(\boldsymbol{s_O})}{\boldsymbol{w}^T\boldsymbol{\beta}(\boldsymbol{s_I}) - \boldsymbol{w}^T\boldsymbol{\beta}(\boldsymbol{s_O})}, \tag{4.7}$$

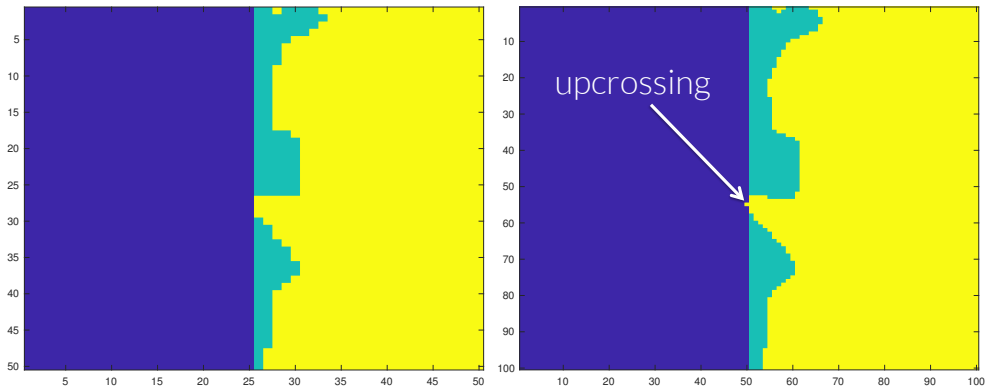
for the locations $\boldsymbol{s_O}$ and $\boldsymbol{s_I}$ respectively. By construction, applying m_1 and m_2 to the

contrast image returns the threshold: $m_1 \boldsymbol{w}^T \boldsymbol{\beta}(\boldsymbol{s_O}) + m_2 \boldsymbol{w}^T \boldsymbol{\beta}(\boldsymbol{s_I}) = \boldsymbol{w}^T \boldsymbol{\beta}(\boldsymbol{s}^*) = c$. Applied to the standardized residuals $\tilde{\boldsymbol{\epsilon}}(\boldsymbol{s_O})$ and $\tilde{\boldsymbol{\epsilon}}(\boldsymbol{s_I})$, we can likewise obtain the residuals at the estimated continuous boundary point $\tilde{\boldsymbol{\epsilon}}(\boldsymbol{s}^*) = m_1 \tilde{\boldsymbol{\epsilon}}(\boldsymbol{s_O}) + m_2 \tilde{\boldsymbol{\epsilon}}(\boldsymbol{s_I})$.

By repeating this procedure for all adjacent points $\boldsymbol{s_O}$ and $\boldsymbol{s_I}$ that lie on the lattice either side of $\partial\mathcal{A}_c$, we are able to estimate the standardized residual values at locations that should approximately sample the true continuous boundary $\partial\mathcal{A}_c$, and therefore we can apply the ideal bootstrap procedure outlined in Section 4.1.2. Of course, in practice we apply this interpolation method on the observed, noisy data, using the plug-in estimated boundary $\partial\hat{\mathcal{A}}_c$.

In the simulation results in Section 4.3, we assess performance of the method when the bootstrap procedure is carried out over the true boundary $\partial\mathcal{A}_c$, and the plug-in estimated boundary $\partial\hat{\mathcal{A}}_c$ that must be used in practice.

4.1.4 Assessment of Continuous Coverage on a Discrete Lattice



(a) Low Resolution Simulation (b) High Resolution Simulation

Figure 4.3: Demonstrating the resolution issue for testing the subset condition $\hat{\mathcal{A}}_c^+ \subset \mathcal{A}_c \subset \hat{\mathcal{A}}_c^-$.
Figure 4.3(a): Here $\mathcal{A}_c$ is comprised of the right half of the image (all green and yellow pixels), and $\hat{\mathcal{A}}_c^+$ is shown as yellow pixels. It appears that $\hat{\mathcal{A}}_c^+ \subset \mathcal{A}_c$.
Figure 4.3(b): The same configuration as **Fig. 4.3(a)** at double the resolution. Here, we have enough detail to see that $\hat{\mathcal{A}}_c^+$ has crossed the boundary $\partial\mathcal{A}_c$ (yellow seeping into blue), and the subset condition $\hat{\mathcal{A}}_c^+ \subset \mathcal{A}_c$ has been violated.

In testing the finite-sample validity of our method through simulation, it is im-

perative that we are able to accurately measure when violations of the subset condition $\hat{\mathcal{A}}_c^+ \subset \mathcal{A}_c \subset \hat{\mathcal{A}}_c^-$ occur. While this may seem a trivial task, as touched on in the previous section, the boundaries of each of these three sets can become ambiguous when data are collected on a discrete lattice.

To illustrate this point, consider the configuration of sets displayed in Figure 4.3(a). In this instance, suppose the right half of the image corresponds to $\mathcal{A}_c$ (green pixels overlapped by yellow), and yellow pixels belong to $\hat{\mathcal{A}}_c^+$. We wish to determine whether the condition $\hat{\mathcal{A}}_c^+ \subset \mathcal{A}_c$ has been violated or not. One may argue that at the resolution for which the data have been acquired, all pixels that belong to $\hat{\mathcal{A}}_c^+$ also belong to $\mathcal{A}_c$, and therefore no violation has occurred. However, the example presented in Fig. 4.3(a) has in fact been derived from a 2D simulation conducted at a higher resolution: this 50×50 simulation was obtained by down-sampling a 100×100 grid by dropping every other pixel. Fig. 4.3(a) displays the sets $\mathcal{A}_c$ and $\hat{\mathcal{A}}_c^+$ from the down-sampled, low resolution simulation, while Figure 4.3(b) shows the same set of results at the original resolution. In Fig. 4.3(b) we see that there *has* been an upcrossing of the yellow pixels belonging to $\hat{\mathcal{A}}_c^+$ over the boundary of the green, and therefore the subset condition $\hat{\mathcal{A}}_c^+ \subset \mathcal{A}_c$ *has* been violated. From this simulation, it is clear that if we conclude no violation has occurred in situations like Fig. 4.3(a), our empirical coverage will miss violations and be positively biased. By an analogous argument the same issue occurs when testing violations of $\mathcal{A}_c \subset \hat{\mathcal{A}}_c^-$.

This direct comparison of the lattice representation of the three sets was used to assess coverage for the simulations carried out in *SSS*. While they observed the phenomenon of missed violations leading to over-coverage, their proposed solution was to sequentially increase the resolution of the data. We instead tackle the issue by again making use of interpolation.

Since in a simulation we know the true continuous mean image and $\mathcal{A}_c$, following the method described in Section 4.1.3 we can obtain weights m_1 and m_2 to interpolate between points $\boldsymbol{s_O}$ and $\boldsymbol{s_I}$ either side of the true, continuous boundary

$\partial\mathcal{A}_c$, in order to find a location $\boldsymbol{s}^*$ that approximately lies on the boundary (if the true mean is linear, $\boldsymbol{s}^*$ lies exactly on the boundary). To determine if $\boldsymbol{s}^* \in \hat{\mathcal{A}}_c^+$, we then reapply the weights m_1 and m_2 and assess whether

$$\boldsymbol{w}^T\hat{\boldsymbol{\beta}}(\boldsymbol{s}^*) - k\,\hat{\sigma}(\boldsymbol{s}^*)v_w = m_1\left(\boldsymbol{w}^T\hat{\boldsymbol{\beta}}(\boldsymbol{s_O}) - k\,\hat{\sigma}(\boldsymbol{s_O})v_w\right) + m_2\left(\boldsymbol{w}^T\hat{\boldsymbol{\beta}}(\boldsymbol{s_I}) - k\,\hat{\sigma}(\boldsymbol{s_I})v_w\right) \geq c. \tag{4.8}$$

If the inequality holds, then by definition $\boldsymbol{s}^* \in \hat{\mathcal{A}}_c^+$. Otherwise, $\boldsymbol{s}^* \notin \hat{\mathcal{A}}_c^+$, and therefore we can conclude that the subset condition $\hat{\mathcal{A}}_c^+ \subset \mathcal{A}_c$ has been violated. By checking whether $\boldsymbol{w}^T\hat{\boldsymbol{\beta}}(\boldsymbol{s}^*) + k\,\hat{\sigma}(\boldsymbol{s}^*)v_w \geq c$, we can similarly test for a violation of $\mathcal{A}_c \subset \hat{\mathcal{A}}_c^-$.

By applying this interpolation scheme to all pairs of lattice points with one point inside, one outside, the lattice representation of the boundary, we have devised a method to more accurately assess violations of the subset condition $\hat{\mathcal{A}}_c^+ \subset \mathcal{A}_c \subset \hat{\mathcal{A}}_c^-$ for configurations similar to 4.3(a). We applied this method for testing the subset condition in our simulations alongside a direct comparison of the lattice representations of the three sets of interest as was done in *SSS*. The addition of the weighted interpolation method caused a considerable decrease in the empirical coverage results towards the nominal level in all of our 3D simulations. Using the direct comparison of the three sets on its own here essentially determined total empirical coverage ($\hat{\mathcal{A}}_c^+ \subset \mathcal{A}_c \subset \hat{\mathcal{A}}_c^-$ for all simulation runs), even when using small sample sizes and a low nominal coverage level (see Fig. 4.8). This is likely to be because the discrete lattice of observed data points is relatively less dense inside the true continuous process for larger, 3D settings, and therefore more violations of the subset condition are missed if only a direct comparison of the lattice representation of the CSs is carried out.

4.2 Method

4.2.1 Simulations

In this section we describe the settings used in order to evaluate the CSs obtained for synthetic data. As a simplified instance of the general linear model setup described in Section 4.1.1, we simulate 3000 independent samples of the signal-plus-noise model

$$Y_i(\boldsymbol{s}) = \mu(\boldsymbol{s}) + \epsilon_i(\boldsymbol{s}), \qquad i = 1, ..., N$$

using a range of signals $\mu(\boldsymbol{s})$, Gaussian noise structures $\epsilon_i(\boldsymbol{s})$ with stationary and non-stationary variance, in two- and three-dimensional regions S. We compute the critical value k, applying the Wild t-Bootstrap method outlined in Section 4.1.2 with $B = 5000$ bootstrap samples to both the true boundary $\partial\mathcal{A}_{\mathrm{c}}$ and the plug-in boundary $\partial\hat{\mathcal{A}}_{\mathrm{c}}$ that would be used in practice. The boundaries are obtained using the interpolation method outlined in Section 4.1.3. We then compare the empirical coverage, i.e./ the percentage of trials that the true thresholded signal is completely contained between the upper and lowers CSs (i.e. the number of times for which $\hat{\mathcal{A}}_{\mathrm{c}}^{+} \subset \mathcal{A}_{\mathrm{c}} \subset \hat{\mathcal{A}}_{\mathrm{c}}^{-}$), across the two sets of results, using the assessment method outlined in Section 4.1.4. In each simulation, we apply the method for sample sizes of $N = 60, 120, 240$ and 480, and using three nominal coverage probability levels $1 - \alpha = 0.80, 0.90$ and 0.95.

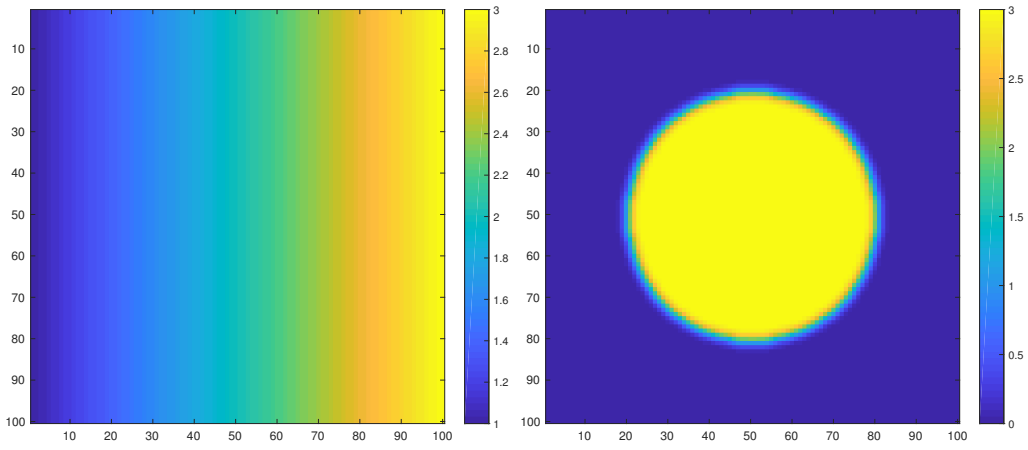
4.2.2 2D Simulations

We analyzed the performance of the CSs on a square region of size 100×100. For the true underlying signal $\mu(\boldsymbol{s})$ we considered two different raw effects: first, a linear ramp that increased from a magnitude of 1 to 3 in the x-direction while remaining constant in the y-direction (Figure 4.4(a)). Second, a circular effect, created by placing a circular phantom of magnitude 3 and radius 30 in the centre of the search region, which was then smoothed using a 3 voxel FWHM Gaussian kernel (Figure

4.4(b)). If we were to assume that each voxel had a size of 2mm^3, this would amount to applying smoothing with a 6mm FWHM kernel, a fairly typical setting used in fMRI analyses.

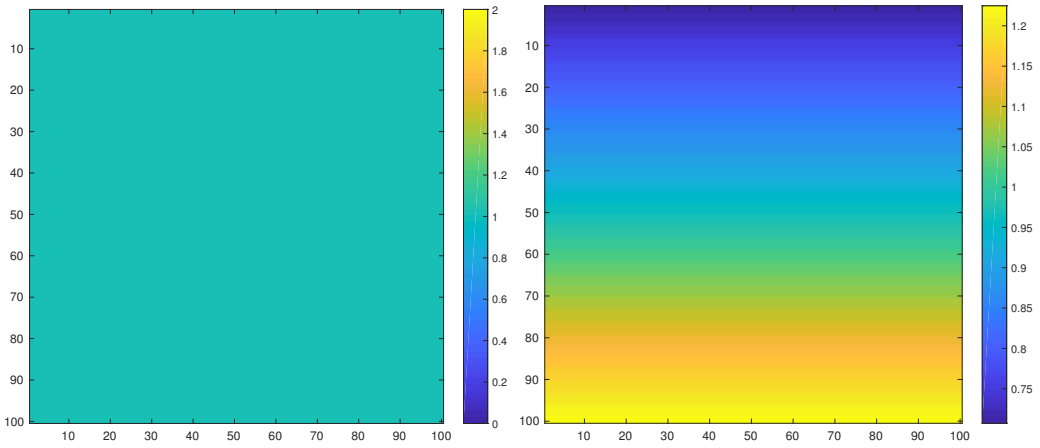
To each of these signals we added subject-specific Gaussian noise $\epsilon_i(\boldsymbol{s})$, also smoothed using a 3 voxel FWHM Gaussian kernel, with homogeneous and non-homogeneous variance structures: the first noise field had a spatially constant standard deviation of 1 (Figure 4.5(a)), the second field had a linearly increasing standard deviation structure in the y-direction from $\sqrt{0.5}$ to $\sqrt{1.5}$ while remaining constant in the x-direction (Figure 4.5(b)). Thus, the variance of this noise field spatially increased in the y-direction from 0.5 to 1.5 in a non-linear fashion.

Altogether, the two underlying signals and two noise sources gave us four separate trials; across all of the simulations, we obtained Confidence Sets for the noise-free cluster $\mathcal{A}_c$ at a cluster-forming threshold of $c = 2$.



(a) **Signal 1.** (b) **Signal 2.**

Figure 4.4: Linear ramp and circular signals $\mu(\boldsymbol{s})$.
Figure 4.4(a): Signal 1. A linear ramp signal that increases from magnitude of 1 to 3 in the x-direction.
Figure 4.4(b): Signal 2. A circular signal with magnitude of 3 and radius of 30, centred within the region and convolved with a 3 voxel FWHM Gaussian kernel.



(a) **Signal 1.** (b) **Signal 2.**

Figure 4.5: Stationary and non-stationary standard deviation fields of the noise $\epsilon_i(\boldsymbol{s})$.
Figure 4.5(a): Standard Deviation 1. Stationary variance of 1 across the region.
Figure 4.5(b): Standard Deviation 2. Non-stationary (linear ramp) standard deviation field increasing from $\sqrt{0.5}$ to $\sqrt{1.5}$ in the y-direction.

4.2.3 3D Simulations

Four signal types $\mu(\boldsymbol{s})$ were considered to analyze performance of the method in three dimensions. The first three of these signals were generated synthetically on a cubic region of size $100 \times 100 \times 100$: Firstly, a small spherical effect, created by placing a spherical phantom of magnitude 3 and radius 5 in the centre of the search region, which was then smoothed using a 3 voxel FWHM Gaussian kernel (Figure 4.6(a)). Secondly, a larger spherical effect, generated identically to the first effect with the exception that the spherical phantom had a radius of 30 (Figure 4.6(b)). Lastly, we created an effect by placing four spherical phantoms of magnitude 3 in the region of varying radii and then smoothing the entire image using a 3 voxel FWHM Gaussian (Figure 4.6(c)). For each of these signals, the final image was rescaled to have a maximum intensity of 3.

Similar to the two-dimensional simulations, for the three signals described above we added 3-voxel smoothed Gaussian noise of homogeneous and heterogeneous variance structures. The first noise field had a spatially constant standard de-

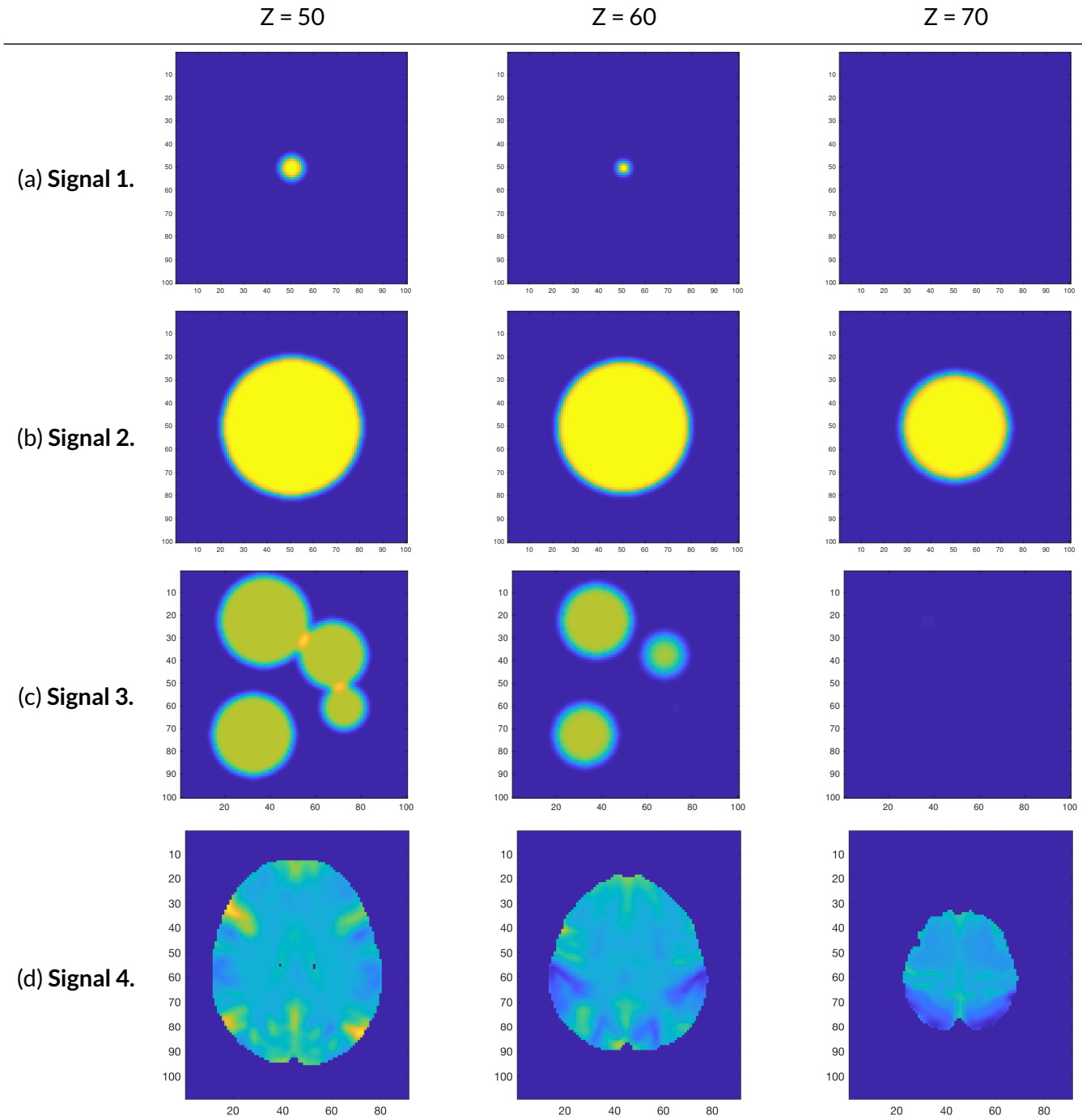


Figure 4.6: The four 3D signal types $\mu(\boldsymbol{s})$, from top-to-bottom: small sphere, large sphere, multiple spheres, and the UK Biobank full mean image. Note that the colormap limits for the first three signal types are from 0 to 3, while the colormap limits for the UK Biobank mean image is from -0.4 to 0.5.

viation of 1, while the second field had a linearly increasing standard deviation in the z-direction from $\sqrt{0.5}$ to $\sqrt{1.5}$, while remaining constant in both the x- and y- directions. For all three effects, we obtained Confidence Sets for the threshold $c = 2$.

For the final signal type, we took advantage of big data that has been made available through the UK Biobank in an attempt to generate an effect that replicated the true %BOLD change induced during an fMRI task. We randomly selected 4000 subject-level contrast of parameter estimate result maps from the Hariri Faces/Shapes task-fMRI data collected as part of the UK Biobank brain imaging study. Full details on how the data were acquired and processed is given in Miller et al. (2016), Alfaro-Almagro et al. (2018) and the UK Biobank Showcase; information on the task paradigm is given in Hariri et al. (2002). From these contrast maps, we computed a group-level full mean (Figure 4.6(d)) and full standard deviation image. In the final simulation, we used the group-level Biobank mean image as the true underlying signal $\mu(\boldsymbol{s})$ for each subject, and the full standard deviation image was used for the standard deviation of each simulated subject-specific Gaussian noise field $\epsilon_i(\boldsymbol{s})$ added to the true signal. Because of the considerably large sample size of high-quality data from which these maps have been obtained, we anticipate that both of these images are highly representative of the true underlying fields that they approximate. Both images were masked using an intersection of all 4000 of the subject-level brain masks.

Once again, we smoothed the noise field using a 3 voxel FWHM Gaussian kernel; since the Biobank maps were written with voxel sizes of 2mm^3, this is analogous to applying 6mm FWHM smoothing to the noise field of the original data. We obtained Confidence Sets for a threshold of $c = 0.25\%$ BOLD change.

4.2.4 Application to Human Connectome Project Data

For a real-data demonstration of the method proposed here, we computed CSs on 80 participants data from the Unrelated 80 package released as part of the Human

Connectome Project (HCP, S1200 Release). We applied the method to subject-level contrast maps obtained for the 2-back vs 0-back contrast from the working memory task results included with the dataset. To compare the CSs with results obtained from standard fMRI inference procedures, we also performed a traditional statistical group-level analysis on the data. A one-sample t-test was carried out in SPM, using a voxelwise FWE-corrected threshold of $p < 0.05$ obtained via permutation test with SPM's SnPM toolbox. We chose to use the HCP for its high-quality task-fMRI data, the working memory task specifically picked for its association with cognitive activations in subcortical networks that can not be distinguished by the anatomy. Full details of the task paradigm, scanning protocol and analysis pipeline are given in Barch et al. (2013) and Glasser et al. (2013), here we provide a brief overview.

For the working memory task participants were presented with pictures of places, tools, faces and body parts in a block design. The task consisted of two runs, where on each run a separate block was designated for each of the image categories, making four blocks in total. Within each run, for half of the blocks participants undertook a 2-back memory task, while for the other half a 0-back memory task was used. Eight EVs were included in the GLM for each combination of picture category and memory task (e.g. 2-back Place); we compute CSs on the subject-level contrast images for the 2-back vs 0-back contrast results that contrasted the four 2-back related EVs to the four 0-back EVs.

Imaging was conducted on a 3T Siemans Skyra scanner using a gradient-echo EPI sequence; TR = 720ms, TE = 33.1 ms, 208×180 mm FOV, 2.0 mm slice thickness, 72 slices, 2.0 mm isotropic voxels, and a multi-band acceleration factor of 8. Preprocessing of the subject-level data was carried out using tools from FSL and Freesurfer following the 'fMRIVolume' HCP Pipeline fully described in Glasser et al. (2013). To summarize, the fundamental steps carried out to each individual's functional 4D time-series data were gradient unwarping, motion correction, EPI distortion correction, registration of the functional data to the anatomy, non-linear reg-

istration to MNI space (using FSL's Non-linear Image Registration Tool, FNIRT), and global intensity normalization. Spatial smoothing was applied using a Gaussian kernel with a 4mm FWHM.

Modelling of the subject-level data was conducted with FSL's FMRIB's Improved Linear Model (FILM). The eight working task EVs were included in the GLM, with temporal derivatives terms added as confounds of no interest, and regressors were convolved using FSL's default double-gamma hemodynamic response function. The functional data and GLM were temporally filtered with a high pass frequency cutoff point of 200s, and time series were prewhitened to remove autocorrelations from the data.

In comparison to a typical fMRI study, the 4mm FWHM smoothing kernel size used in the HCP preprocessing pipeline is modest. Because of this, we applied additional smoothing to the final contrast images to emulate maps smoothed using a 6mm FWHM Gaussian kernel.

4.3 Results

4.3.1 Methodological Comparisons

In this work we have proposed two fundamental methodological changes to the procedures carried out in *SSS*: in Section 4.1.2 we suggested the Wild t-Bootstrap instead of the Gaussian Wild Bootstrap used for *SSS*, and in Section 4.1.4 we introduced the interpolation method for assessing empirical coverage alongside the direct comparison methods used for *SSS*. Here, we show the impact of these methodological innovations on the empirical coverage results from simulations carried out using two different synthetic signals, the 2D circular signal (**Signal 2.** in Fig. 4.4(b)) and the 3D large spherical signal (**Signal 2.** in Fig. 4.6(b)). The standard deviation of the subject-specific Gaussian noise fields $\epsilon_i(\boldsymbol{s})$ had a stationary variance of 1 across the region in both simulations (for the 2D case, this corresponds to **Standard Deviation 1.** in Fig.

4.5(a)).

Empirical coverage results for each of the three confidence levels $1-\alpha = 0.80, 0.90$ and 0.95 are presented for the 2D circular signal in Figure 4.7 and for the 3D large spherical signal in Figure 4.8. In both simulations, for all methods the bootstrap procedure was carried out over the estimated boundary $\partial\hat{\mathcal{A}}_{\mathrm{c}}$ (as must be done with real data). In each figure, the green curves highlight the results for the Gaussian Wild Bootstrap and coverage assessment method that were applied in *SSS*. The red curves highlight the results for the Wild t-Bootstrap and interpolation assessment method that we have proposed.

In Fig. 4.7 and Fig. 4.8, all simulations using the direct comparison assessment (SSS Simulation Assessment) produced results substantially above the nominal level, converging to almost 100% for both the Gaussian Wild Bootstrap (green curves) and Wild t-Bootstrap (blue curves) methods across all three confidence levels. We suspect this is due to the resolution issue described in Section 4.1.4, suggesting that this assessment method missed violations of the coverage condition $\hat{\mathcal{A}}_{\mathrm{c}}^{+} \subset \mathcal{A}_{\mathrm{c}} \subset \hat{\mathcal{A}}_{\mathrm{c}}^{-}$ causing a considerable positive bias in all of these results. Further evidence of this is suggested by the empirical coverage obtained for simulations using the interpolation assessment method (BTSN Simulation Assessment, pink and red curves), which appear to be converging much closer to the nominal level as is theoretically expected by Result 1.

Considering only the results using the interpolation assessment, in both figures empirical coverage for the Wild Bootstrap method (pink curves) came below the nominal level for small sample sizes. For the 2D circle simulation, the empirical coverage result for 60 subjects was 84.7% for the nominal target of $1-\alpha = 0.95$ (right plot in Fig. 4.7). For the 3D spherical simulation this under-coverage was even more severe, where the corresponding empirical coverage result was 54.9% (right plot in Fig. 4.8). In comparison, coverage performance for the Wild t-Bootstrap method (red curves) was much improved, staying close to the nominal level in both the 2D and 3D

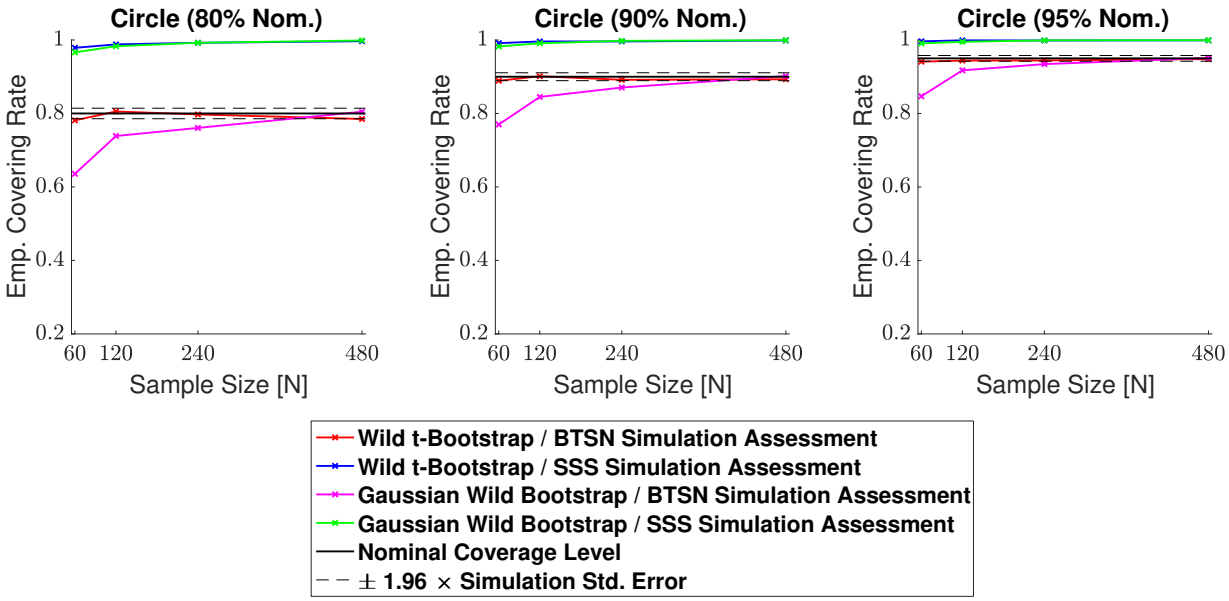


Figure 4.7: Coverage results for the 2D circular signal simulation with homogeneous Gaussian noise (**Signal 2.** (**Standard deviation 1.**) in Fig. 4.4(b) (Fig. 4.5(a)). Empirical coverage results are presented for implementations of the CS method with and without the Wild t-Bootstrap we propose in Section 4.1.2 and the interpolation schema for assessing simulations results we propose in Section 4.1.4. All empirical coverage results for simulations using the *SSS* assessment method are close to 100%, suggesting that this assessment substantially biases the results upwards. Using our proposed assessment method, while both the Wild t-Bootstrap and Gaussian Wild bootstrap converge to the nominal level, the Wild t-Bootstrap performed better for small sample sizes.

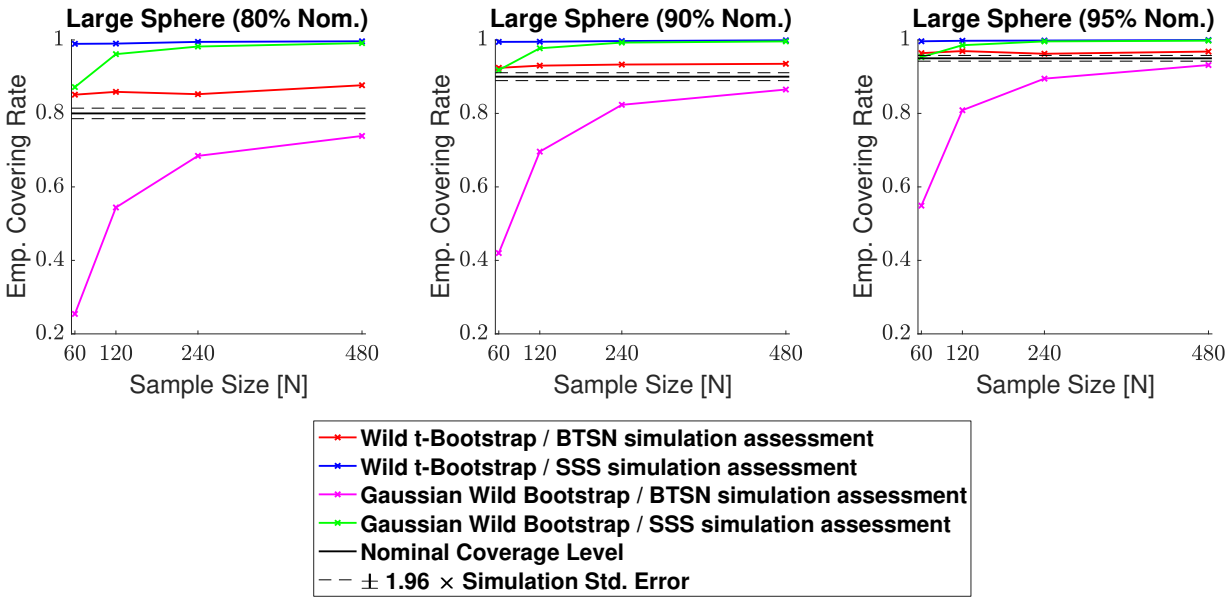


Figure 4.8: Coverage results for the 3D large spherical signal (**Signal 2.** in Fig. 4.6(b)) simulation with homogeneous Gaussian noise. Empirical coverage results are presented for implementations of the CS method with and without the Wild t-Bootstrap we propose in Section 4.1.2, and the interpolation schema for assessing simulations results we propose in Section 4.1.4. Once again, all simulations using the *SSS* assessment method quickly converge to close to 100%. Using our proposed assessment method, the Gaussian Wild bootstrap had severe under-coverage for small sample sizes, while the Wild t-Bootstrap results hover slightly above the nominal level for all sample sizes.

simulations across all sample sizes. While for the 3D spherical signal the empirical coverage remained slightly above the nominal target, for the circular signal almost all results lie within the 95% confidence interval of the nominal coverage level. For these reasons, in the remaining simulation results presented in this section we only consider the Wild t-Bootstrap method with our proposed interpolation assessment.

4.3.2 2D Simulations

Empirical coverage results for each of the three confidence levels $1-\alpha = 0.80, 0.90$ and 0.95, are presented for the linear ramp signal (**Signal 1.** in Fig. 4.4(a)) in Figure 4.9, and for the circular signal (**Signal 2.** in Fig. 4.4(b)) in Figure 4.10. Results are also presented in tabular format in Table B.1. In both plots, results obtained for simulations applying the bootstrap procedure over the estimated boundary $\partial\hat{\mathcal{A}}_{\mathrm{c}}$ are displayed with a solid line, while results for simulations using the true boundary $\partial\mathcal{A}_{\mathrm{c}}$ are displayed with a dashed line. We emphasize that when computing CSs for real data, only the estimated boundary can be used.

For the linear ramp, across all confidence levels we observed valid, over-coverage for the estimated boundary method, and under-coverage for the true boundary method. In both cases, the degree of agreement between our empirical results and the nominal coverage level improved for larger confidence levels, and as the sample size increased. For instance, while our estimated boundary empirical results were around 88% when the nominal target level was set at 80% (Fig. 4.9, left), corresponding empirical coverage results hovered around 97% for a nominal target of 95% (Fig. 4.9, right). Comparing the differences between the solid and dashed curves, there is also greater harmonization between the estimated and true boundary results for higher confidence levels. The method performed similarly regardless of whether homogeneous or heterogeneous noise was added to the model, evidenced by the minimal differences between the red and the blue curves for each of the two boundary methods seen in the plots.

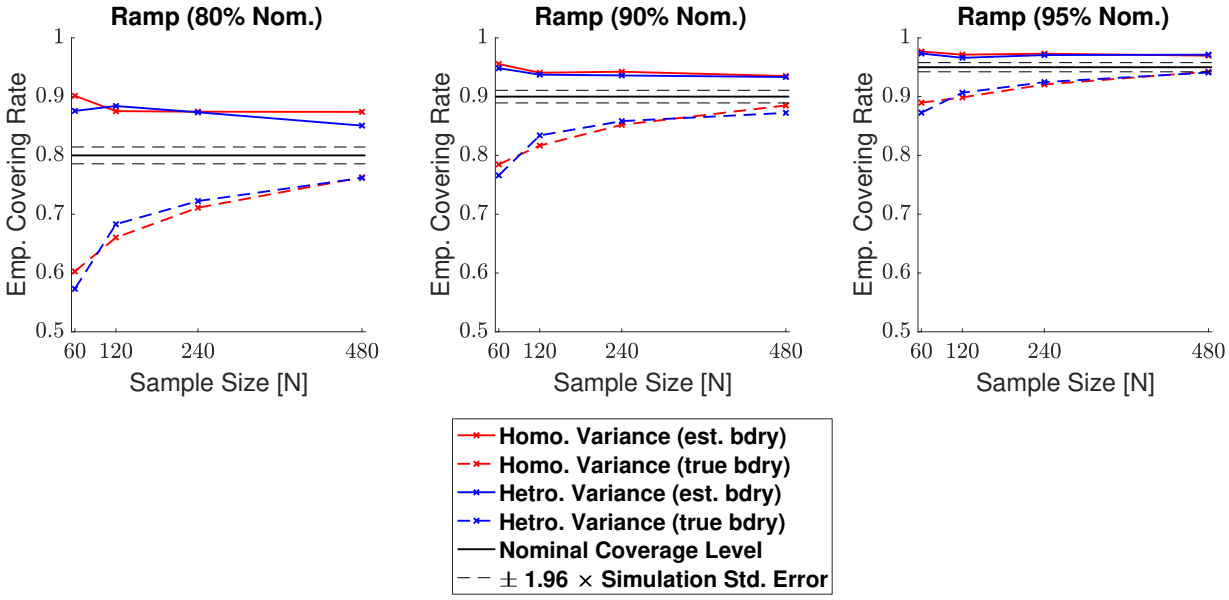


Figure 4.9: Coverage results for **Signal 1.**, the 2D linear ramp signal. While the true boundary coverage results (dashed curves) fall under the nominal level, results for the estimated boundary method (solid curves) that must be applied to real data remain above the nominal level. Performance of the method improved for larger confidence levels, and in particular, the estimated boundary results for a 95% confidence level seen in the right plot hover slightly above nominal coverage for all sample sizes.

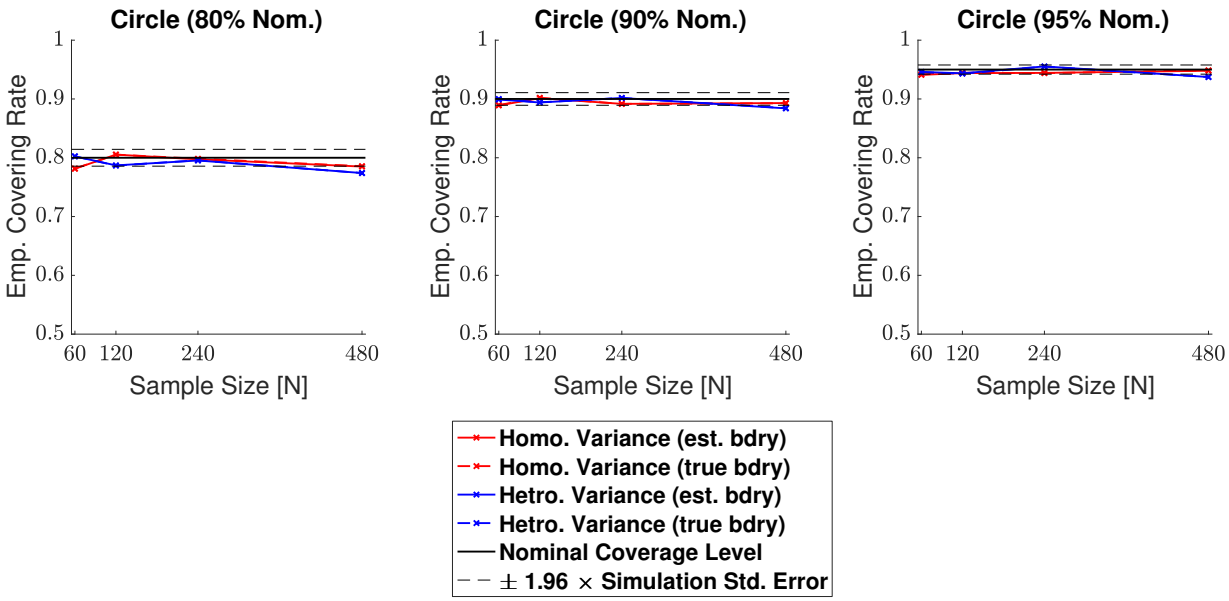


Figure 4.10: Coverage results for **Signal 2.**, the 2D circular signal. Coverage performance was close to nominal level in all simulations. The method was robust as to whether the subject-level noise had homogeneous (red curves) or heterogeneous variance (blue curves), or as to whether the estimated boundary (dashed curves) or true boundary (solid curves) method was used; in all plots, all of the curves lie practically on top of each other.

For the circular signal the method performed remarkably well, with almost all our empirical coverage results lying within the 95% confidence interval of the nominal coverage rate (red and blue curves sandwiched between black dashed lines for all three plots in Fig. 4.10). Once again, the use of homogeneous or heterogeneous noise in the model had minimal difference on the method's empirical coverage performance, and in this setting, our results were virtually identical whether the estimated boundary or true boundary was used for the bootstrap procedure. This has made the dashed curves hard to distinguish in the plots, as the solid curves lie practically on top of them.

4.3.3 3D Simulations

Empirical coverage results for each of the three confidence levels $1-\alpha = 0.80, 0.90$ and 0.95, are presented in Figures 4.11, 4.12, 4.13 and 4.14 respectively for each of the four signal types (small sphere, large sphere, multiple spheres, Biobank full mean) displayed in Fig. 4.6. Results are also presented in tabular format in Table B.2. Once again, results obtained for simulations applying the bootstrap procedure over the estimated boundary $\partial\hat{\mathcal{A}}_c$ are displayed with a solid line, and results for simulations using the true boundary $\partial\mathcal{A}_c$ are displayed with a dashed line.

Overall, the results for all four signal types were consistent: In general, empirical coverage always came above the nominal target level, and the extent of over-coverage diminished when a higher confidence level was used. Particularly, for a nominal target of $1-\alpha = 0.95$, all of our 3D empirical coverage results lie between 95% and 98%. The method was robust as to whether the bootstrap procedure was applied over the true or estimated boundary, or as to whether the variance of the noise field was homo- or heterogeneous. The similarity of the empirical coverage results, in spite of differences in these specific settings, is exhibited in all of the plots by the uniformity of the red and blues curves (indicating minimal differences in performance as to whether the noise had homogeneous or heterogeneous variance), and agreement between the solid and dashed curves (indicating minimal differences in performance as to whether the true boundary or estimated boundary was used). In the empirical coverage plots for the small and large spherical signals shown in Figs. 4.11 and 4.12, all of these curves lie virtually on top of each other.

While performance with the multiple spheres and Biobank signals presented in Figs. 4.13 and 4.14 was slightly better when using the true boundary, the true- and estimated boundary performance converged as the sample size increased.

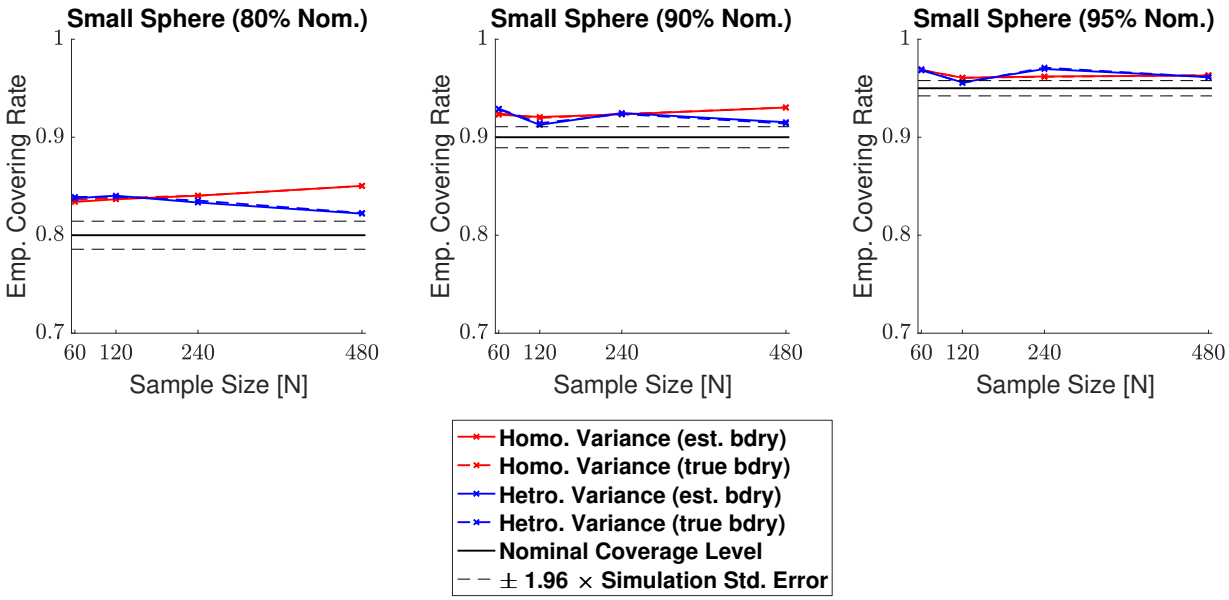


Figure 4.11: Coverage results for **Signal 1.**, the 3D small spherical signal. For all confidence levels, coverage remained above the nominal level in all simulations, and for a 95% confidence level (right plot), coverage hovered slightly above the nominal level for all sample sizes. The method was robust as to whether the subject-level noise had homogeneous (red curves) or heterogeneous variance (blue curves), or as to whether the estimated boundary (dashed curves) or true boundary (solid curves) method was used.

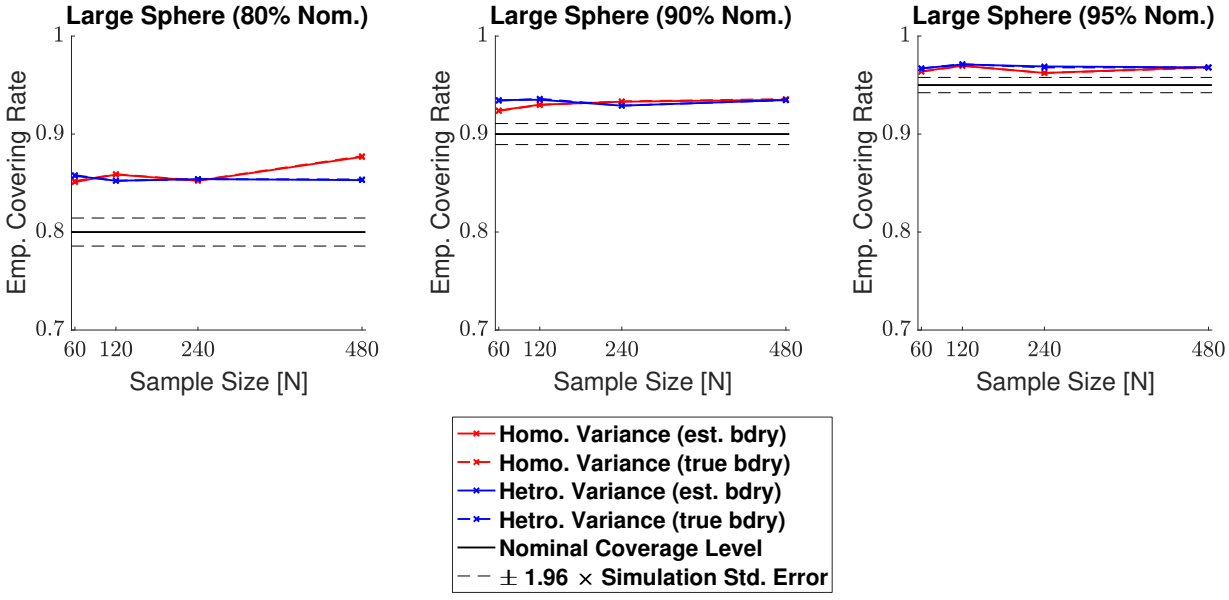


Figure 4.12: Coverage results for **Signal 2.**, the large 3D spherical signal. Coverage results here were very similar to the results for the small spherical signal shown in Fig. 4.11, suggesting that the method is robust to changes in boundary length.

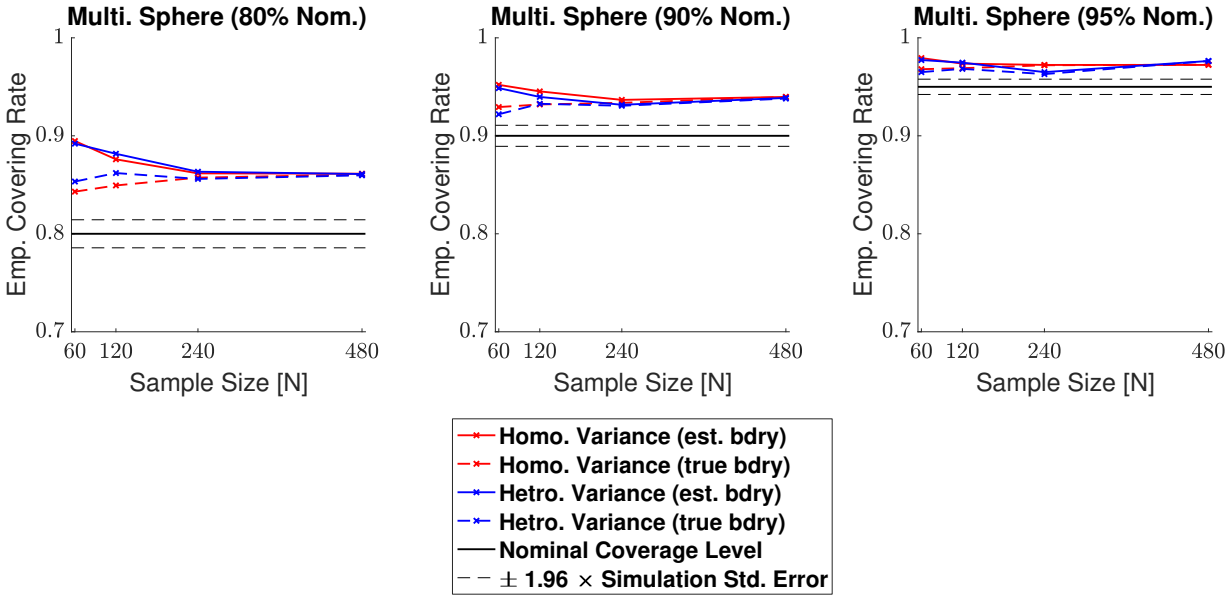


Figure 4.13: Coverage results for **Signal 3.**, the multiple spheres signal. Once again, for all confidence levels, coverage remained above the nominal level in all simulations. Here, the true boundary method (dashed curves) performed slightly better than the estimated boundary method (solid curves) in small sample sizes, although the choice of boundary made less of a difference for a higher confidence level. For a 95% confidence level (right plot), all results hover slightly above nominal coverage for all sample sizes.

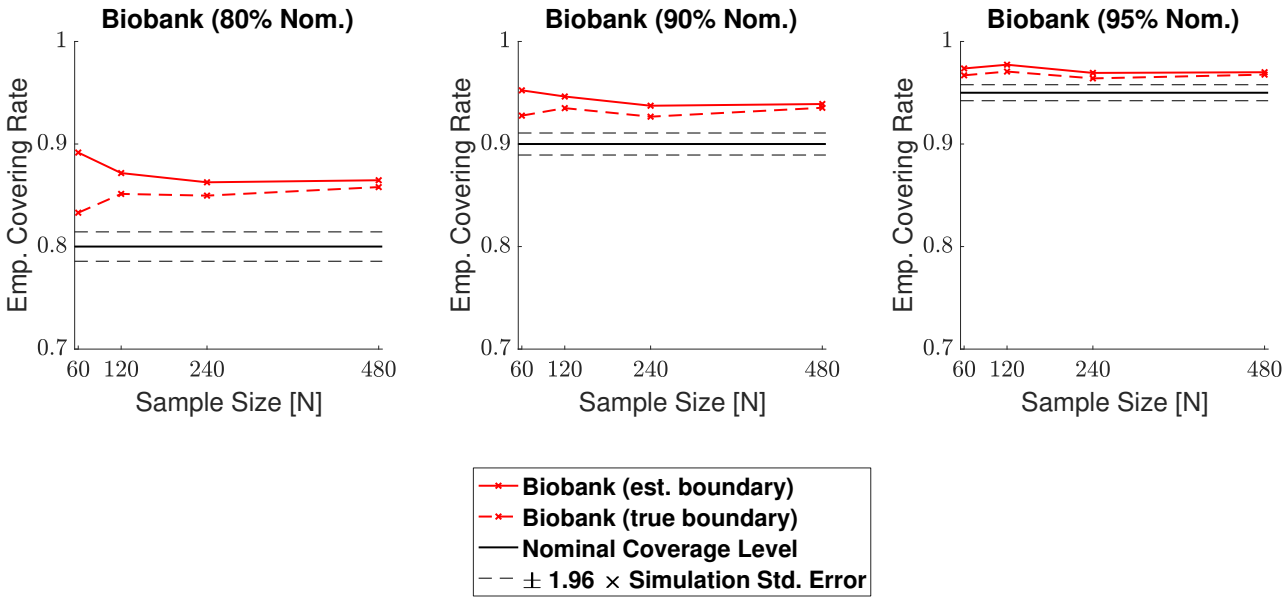


Figure 4.14: Coverage results for **Signal 4.**, the UK Biobank full mean signal, where the full standard deviation image was used as the standard deviation of the subject-level noise fields. Coverage results here were similar to the results for the multiple spheres signal shown in Fig. 4.13: In small sample sizes, coverage was slightly improved for the true boundary method (dashed curves) compared to the estimated boundary method (solid curves), however, for a 95% confidence level (right plots), all results hover slightly above nominal coverage for all sample sizes.

4.3.4 Human Connectome Project

Confidence Sets obtained from applying the method to 80 subjects contrast data from the Human Connectome Project working memory task are shown in Figure 4.15 and Figure 4.16.

In both Fig. 4.15 and Fig. 4.16, the red upper CS localized brain regions within the frontal cortex commonly associated to working memory. This included areas of the middle frontal gyrus (left and right; Fig. 4.15, sagittal and coronal slices), superior frontal gyrus (left and right; Fig. 4.16, coronal slice) anterior insula (left and right; Fig. 4.15, sagittal and axial slices), as well as the anterior cingulate (Fig. 4.16, all slices). In all of the above regions, the method identified clusters of voxels for which we can assert with 95% confidence there was a percentage BOLD change raw effect greater than 2.0% (Fig. 4.15 and Fig. 4.16, bottom plots).

Further brain areas localized by the upper CS were the frontal pole (left and right; Fig. 4.15, sagittal and axial slices), supramarginal gyrus (left and right; Fig. 4.15, sagittal slice and Fig. 4.16, coronal and axial slices), precuneous (Fig. 4.16, sagittal slice) and cerebellum (Fig. 4.15, sagittal slice). While for these areas we can assert with 95% confidence there was a percentage BOLD change raw effect greater than at least 1.0% (Fig. 4.15 and Fig. 4.16, top plots), on the whole the method only localized areas where there was a BOLD change of at least 2.0% in parts of the frontal cortex. This can be observed by the 'disappearance' of the red CSs in brain regions located in the ocipital lobe for the 2.0% BOLD change plots when compared with the corresponding 1.0% and 1.5% BOLD change plots in Fig. 4.15 and Fig. 4.16.

As the percentage BOLD change threshold increases between plots, there is a shrinking of both the blue lower CSs and red upper CSs; by using a larger threshold, there are less voxels we can confidently declare have surpassed this higher level of percentage BOLD change, and thus the volume of the red upper CSs decreases (in some cases, vanishing). At the same time, there are more voxels we expect to be able

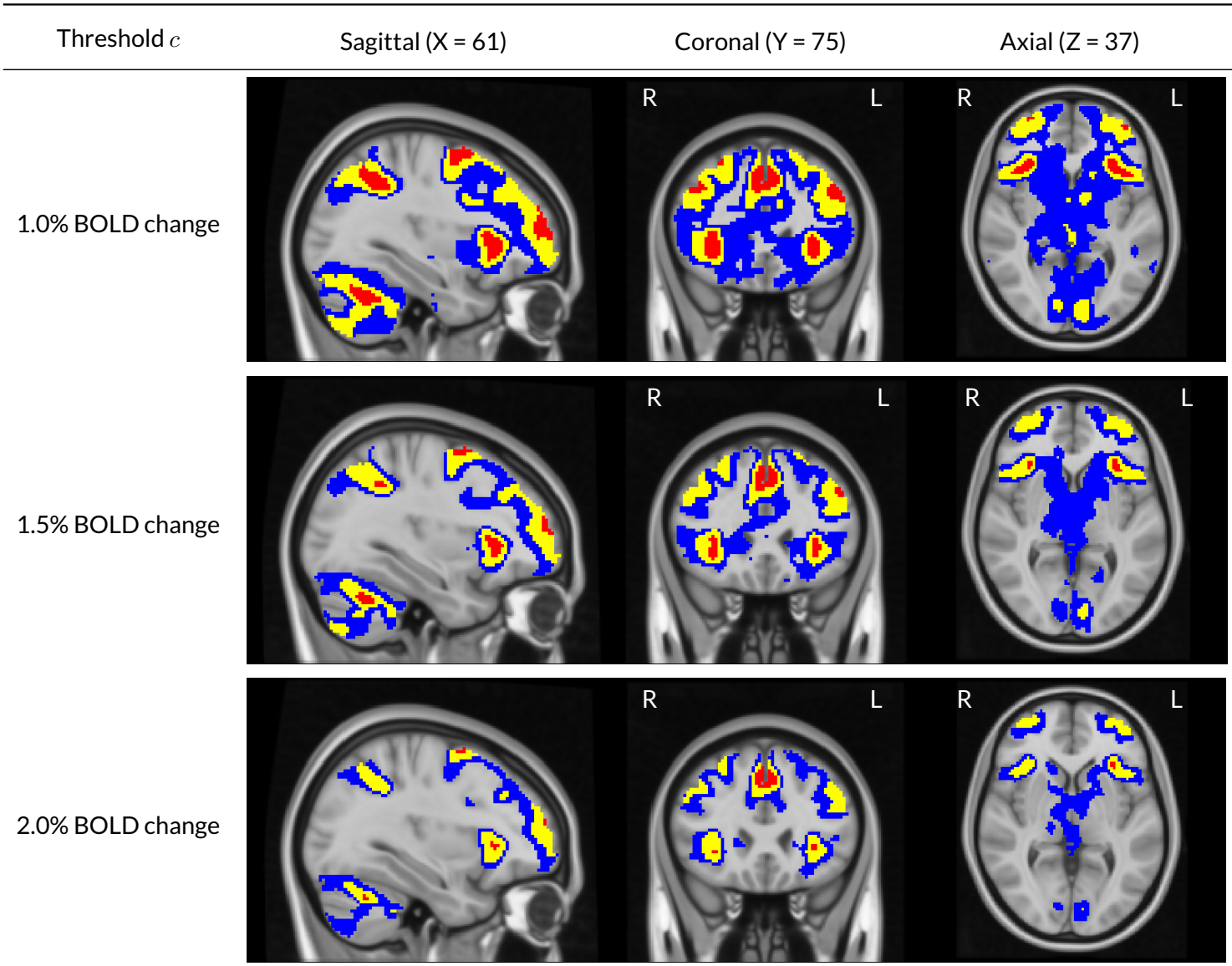


Figure 4.15: Slice views of the Confidence Sets for 80 subjects data from the HCP working memory task for $c = 1.0\%, 1.5\%$ and 2.0% BOLD change thresholds. The upper CS $\hat{\mathcal{A}}_c^+$ is displayed in red, and the lower CS $\hat{\mathcal{A}}_c^-$ displayed in blue. In yellow is the point estimate set $\hat{\mathcal{A}}_c$, the best guess from the data of voxels that surpassed the BOLD change threshold. The red upper CS has localized regions in the frontal gyrus, frontal pole, anterior insula, supramarginal gyrus and cerebellum for which we can assert with 95% confidence that there has been (at least) a 1.0% BOLD change raw effect.

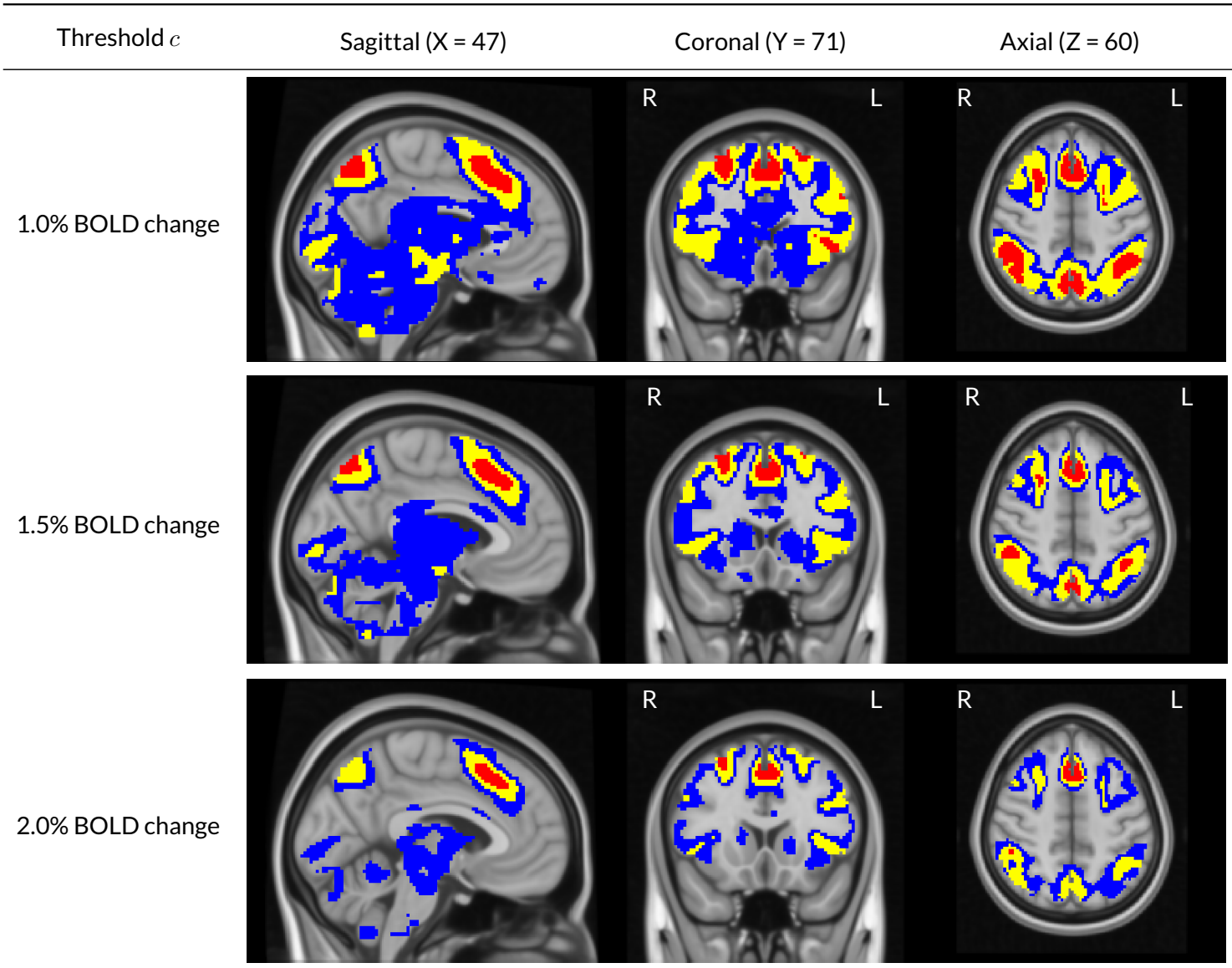


Figure 4.16: Further slice views of the Confidence Sets. Here, we see that the red upper CS has also localized regions in the anterior cingulate, superior front gyrus, supramarginal gyrus, and precuneous for which we can assert with 95% confidence that there has been (at least) a 1.0% BOLD change raw effect.

to confidently declare have fallen below the threshold. Since these are precisely the (grey background) voxels that lie outside of the lower blue CSs, the volume of the blue lower CSs also decreases.

Finally, in Figure B.1 and Figure B.2 the red upper CSs are compared with the thresholded t-statistic map (green-yellow voxels) obtained from applying a traditional one-sample t-test group-analysis to the 80 subjects working memory task contrast data, using a voxelwise FWE-corrected threshold of $p < 0.05$. Differences here highlight how statistical significance may not translate to practical significance; while over 28,000 voxels were declared as active in the thresholded t-statistic results, only 4,818 voxels were contained in the upper CS indicating a percentage BOLD change of at least 1.0%.

4.4 Discussion

4.4.1 Spatial Inference on %BOLD Raw Effect Size

Thorough interpretation of neuroimaging results requires an appreciation of the practical (as well as statistical) significance of differences through visualization of raw effect magnitude maps with meaningful units (Chen et al., 2017).

In this work, we have presented a method to create confidence sets for raw effect size maps, providing formal confidence statements on regions of the brain where the %BOLD response magnitude has exceeded a specified activation threshold, alongside regions where the %BOLD response has *not* surpassed this threshold. Both of these statements are made simultaneously, and across the entire brain. This not only enables researchers to infer brain areas that have responded to a task, but also allows for inference on areas that did not respond to the task. In this sense, the method goes beyond statistical hypothesis testing, where the null-hypothesis of no activation can 'fail to be rejected', but never accepted. By operating on percentage BOLD change units, instead of t-statistic values, the confidence set maps present a clear

and more direct interpretation of the biophysical changes that occur during a neuroimaging study, which can be distorted by the thresholded statistic maps commonly reported at the end of an investigation (Engel and Burton, 2013). In essence, the CSs synthesize information that is usually provided separately in a raw effect size and t-statistic map, determining practically significant effects in terms of effect magnitude, that are also reliable in terms of statistical significance traditionally given by p-values in a statistic image. While in this work we have focused on BOLD fMRI, the methods presented here are applicable to any neuroimaging measure that can be fit in a group-level GLM.

The choice of threshold c is ultimately up to the user, and may depend on the aims of the investigation. Researchers may choose a threshold based on prior knowledge of raw effect sizes observed in previous similar studies, and it is likely that localization of larger raw effects will be possible as sample sizes increase. Obtaining the CSs for the Human Connectome Project contrast data in this work was computationally quick, each analysis taking no longer than a couple of minutes. Therefore, one possible strategy is to evaluate a variety of different c's on pilot or historical data before fixing a value to use on a study of interest.

4.4.2 Analysis of HCP data and Simulation Results

In our analysis of the HCP emotional faces task-fMRI dataset, we have primarily focused on activated areas localized by the red upper CS. However, the confidence set maps in Fig. 4.15 and Fig. 4.16 also quantify the spatial precision of the point estimate 'best guess from the data' activation clusters. While so far we have described the confidence sets in terms of the red and blue upper and lower CSs, we now highlight that the set difference between the upper and lowers CSs acts as a confidence region itself; with 95% confidence, we can assert that the boundary of the point estimate set (raw effect size $>$ threshold; yellow voxels overlapped by red in Fig. 4.15 and Fig. 4.16) is completely contained within this region. The set difference region,

visualized by blue and yellow voxels (but not red) in Fig. 4.15 and Fig. 4.16, therefore anticipates how the point estimate clusters may fluctuate if the experiment was to be repeated again. From this perspective, the vast areas of the brain covered by blue in Fig. 4.15 and Fig. 4.16 demonstrate the high level of uncertainty in localizing a raw effect size of, for example, 1.0% BOLD change, despite the large sample size of $N = 80$ used for the HCP. The regions of greatest uncertainty were sub-cortical areas, covered by expansive clusters of blue as seen in the axial slices displayed in Fig. 4.15 and sagittal slices in Fig. 4.16. Large intersubject variability here may be explained by the high multi-band acceleration factor used in the HCP scanning protocol, which is generally more suited for scanning the cortex (Smith et al., 2013).

For the 2D simulations, the method achieved close to nominal coverage for the circular signal, but performed less well for the ramp signal, obtaining under-coverage for the true boundary method and over-coverage for the estimated boundary method. We believe differences in the circle and ramp results are not due to changes in the signal shape per se, but instead are caused by differences in the slope of each shape close to the true boundary $\partial\mathcal{A}_{\mathrm{c}}$. Since the linear ramp signal has a shallower gradient at the true boundary compared to the circle, local changes in the observed signal around the boundary are dominated by changes in the noise. Since the noise is more wavey than the signal, the linear interpolation method for obtaining the boundary is likely to be less accurate for the ramp, causing too many violations of the subset condition, which may explain the under-coverage for the true boundary results seen here.

For the 3D simulations, the method obtained over-coverage in all of our results. Here, the degree of over-coverage was consistently larger for the smaller confidence level of $1-\alpha = 0.80$ in comparison to the larger confidence level of $1-\alpha = 0.95$. Notably, the over-coverage was also more severe for signals with a longer boundary, such as the multiple spheres and Biobank signals, when compared to the Small Sphere signal that had a shorter boundary length. One possible reason for this is that

our proposed method for assessing coverage may still be missing instances where violations of the subset condition $\hat{\mathcal{A}}_{\mathrm{c}}^{+} \subset \mathcal{A}_{\mathrm{c}} \subset \hat{\mathcal{A}}_{\mathrm{c}}^{-}$ occur, causing the results to be slightly positively biased. While our assessment method reduces the influence of grid coarseness by sampling locations on the true continuous boundary ∂A_c, ultimately we can still only assess coverage at a discrete set of points on a continuous process. For signals with a longer boundary length, the set of sampled locations obtained with the interpolation method is relatively less dense within the true continuous boundary, and thus it is more likely violations of the subset condition are missed. Over-coverage for smaller confidence levels may also be explained by this, as theoretically more violations should occur here, but these may be missed due to inaccuracies caused by the discreteness of the lattice. This line of reasoning is consistent with Section 4.4 of *SSS*, where it was shown that coverage approached the nominal level as the resolution of the grid was increased.

4.4.3 Methodological Innovations

In this work, we have advanced on the original methods applied in *SSS*. From a theoretical standpoint, we have proposed a Wild t-Bootstrap method (dividing bootstrap residuals by bootstrap standard deviation) to compute the critical quantile value k. We have also introduced an interpolation scheme for obtaining the boundary and assessing the simulation coverage results to reduce the influence of grid coarseness. In Section 4.3.1, we demonstrated that applying the assessment method in *SSS* could lead to empirical coverage of close to 100%, suggesting that this method may considerably bias the simulation results upwards. When using our proposed assessment, the Wild Bootstrap method suffered from under-coverage, most severely for small sample sizes in the 3D setting of the large spherical signal presented in Fig. 4.8. This was greatly remedied by the Wild t-Bootstrap method, for which empirical results stayed close to the nominal target independent of sample size.

Our simulations using the original procedures may not seem consistent with

the simulation results published in Figure 5 of *SSS*, where empirical coverage stayed close to the nominal target. However, the signal-plus-noise models investigated to test the performance of the CSs in *SSS* were much smoother than the synthetic signals considered to emulate fMRI data with this effort. By applying a larger degree of smoothing, the signals used in *SSS* effectively had a much higher resolution. Because of this, it is likely the resolution issue presented in Fig. 4.3 was less critical, reducing the positive bias in empirical coverage induced from using the original simulation assessment procedure. Further evidence for this is provided in Figure 7 of *SSS*, where they observed an increase in coverage after repeating their simulations on a coarser lattice. In our simulation results in Section 4.3.1, the scale of under-coverage from using the Gaussian Wild bootstrap method was much more severe for the 3D simulation on the spherical signal in Fig. 4.8 compared to the 2D circular signal in Fig. 4.7. This may explain why the Gaussian Wild bootstrap method performed relatively well in *SSS*, as only 2D signals were considered there.

4.4.4 Limitations

The principal limitation of this work is one that is intentional and explicit: Our method is for spatial inference on maps of raw and not standardized effects, such as Cohen's d or partial R^2 (t- or F-statistics, which scale with sample size, do not estimate population quantities and are not suitable for making confidence statements). Even when scaled to percentage BOLD change, it has been shown that raw effects can modulate with acquisition parameters such as the scanner field strength or echo time (Uludag et al., 2009). Users should therefore be cautious when combining effect estimates from studies using heterogeneous acquisition setups, and clearly specify such differences when reporting the results of any meta-analysis on raw effects. It is also known that inhomogeneities in the vasculature of the brain is a cause of variation in the BOLD response. Therefore, we recommend that any interpretation of %BOLD change inferred from the CSs is referenced against a variance map or similar image

that indicates the most venous brain regions. We note that each of these points are general complications of raw effect sizes within fMRI, rather than issues with the method proposed in this effort per se. Nonetheless, the use of standardized effect estimates may help to remedy these problems.

The need for resampling to conduct inference is another limitation of this effort, especially given the big data motivation of this work. However, the bootstrap is only conducted on the estimated boundary, $\partial\hat{\mathcal{A}}_c$, not the whole 3D volume, which substantially reduces the computational burden. For very large datasets, techniques for approximating empirical distributions can be used to improve the accuracy of the estimation of k based on a smaller number (e.g. $B = 500$) of bootstrap samples (Winkler et al., 2016).

CHAPTER 5

Spatial Confidence Sets for Cohen's *d* Effect Size Images

In Chapter 4, we extended on a method initially proposed by *Sommerfeld, Sain, and Schwartzman (2018) (SSS)* to obtain Confidence Sets (CSs) for inference on %BOLD change effect size maps. Unlike traditional hypothesis testing methods, where inference is only provided in terms of the presence of an effect, we developed the CSs to make confidence statements about the precise brain regions where raw effect sizes had surpassed, and fallen short of, a *non-zero* %BOLD threshold.

In this chapter, we set out to adapt the CSs for application to fMRI Cohen's d effect size images. For a one-sample t-test, the Cohen's d estimate is computed as the sample mean effect size divided by the population standard deviation $\left(\text{i.e. } \hat{d} = \frac{\hat{\mu}}{\hat{\sigma}}\right)$. Therefore, in contrast to percentage BOLD, Cohen's d provides an interpretation of the effect magnitude in terms of $standardized$ units. Standardized effects are arguably more appropriate for evaluation of functional imaging data, since the parameter estimates obtained from an fMRI analysis are commonly reported in units that are essentially arbitrary. To acquire contrast estimates interpretable as %BOLD change requires that the analysis data, analysis design and contrast vector are appropriately normalized, and this rescaling is dependent on both the task paradigm and software package used for the analysis (see Appendix A.1 for further discussion). While this process can be cumbersome and susceptible to error, obtaining Cohen's d images is

relatively straightforward; due to the simple relationship connecting the Cohen's d estimator and the t-statistic, $\hat{d} = \frac{t}{\sqrt{N}}$, Cohen's d effect estimate maps can be easily obtained from the unthresholded t-statistic images reported within all the main neuroimaging software packages. At the end of the previous chapter we discussed further issues associated with raw effects, such as physiological factors and choices in acquisition parameters that can induce variance in the BOLD signal. For all of these reasons, Cohen's d may provide a more interpretable measure of effect size with improved comparability of functional results between studies.

The statistical characteristics of Cohen's d effect size maps are fundamentally different to the raw %BOLD images motivating the Confidence Sets in the previous chapter. In this work, we derive the properties of Cohen's d before proposing three separate methods for computing CSs on Cohen's d effect size images. Our ultimate goal here is to provide a method of obtaining Cohen's d CSs with adequate empirical coverage performance on images representative of fMRI Cohen's d effect size maps.

The remainder of this chapter is organized as follows: First, we contextualize the problem of obtaining Confidence Sets for Cohen's d images, exemplifying the key differences which distinguish Cohen's d from the %BOLD effect size. We then derive properties of the Cohen's d estimator, exploring how the methods developed in the previous chapter may be modified to give a procedure for Cohen's d images. From this, we propose three separate algorithms to compute Cohen's d CSs. We assess the empirical coverage performance of each of these methods on 2D and 3D Monte Carlo simulations, and finally, present the CSs obtained from applying each algorithm to Human Connectome Project working memory task-fMRI data.

All aspects of this work were carried out under the supervision and mentorship of Prof. Thomas Nichols and Dr. Armin Schwartzman. Theoretical aspects of this work were developed in collaboration with Dr. Fabian Telschow.

5.0.1 From %BOLD to Cohen's d

For a compact domain $S \subset \mathbb{R}^D$, e.g. $D = 3$, for $i = 1, ..., N$ consider the one-sample model at location $\boldsymbol{s} \in S$,

$$Y_i(\boldsymbol{s}) = \mu(\boldsymbol{s}) + \epsilon_i(\boldsymbol{s}) \tag{5.1}$$

where $Y_1(\boldsymbol{s}), ..., Y_N(\boldsymbol{s})$ are the observations at $\boldsymbol{s}$, $\mu(\boldsymbol{s})$ is the true underlying mean intensity across the observations, and $\epsilon_1(\boldsymbol{s}), ..., \epsilon_N(\boldsymbol{s})$ are i.i.d. mean-zero errors with common variance $\sigma^2(\boldsymbol{s})$ and some unspecified spatial correlation. In Chapter 4, we motivated the method to obtain CSs in the context of a one-sample group-level fMRI analysis for inference on the raw %BOLD change effect size. In this instance, $\mu(\boldsymbol{s})$ represented the true mean %BOLD change across the group, and the CSs localized brain regions to assert where $\mu(\boldsymbol{s})$ had exceeded, and fallen short of, a non-zero fixed threshold c. Here, we instead seek to obtain CSs for the standardized Cohen's d effect size defined as:

$$d(\boldsymbol{s}) = \frac{\mu(\boldsymbol{s})}{\sigma(\boldsymbol{s})}. \tag{5.2}$$

Therefore, in this case we are interested in the excursion set $\mathcal{A}_{\mathrm{c}}$ where the true population Cohen's d effect size surpasses the threshold c, defined as:

$$\mathcal{A}_{\mathrm{c}} = \{\boldsymbol{s} \in S : d(\boldsymbol{s}) \geq c\}. \tag{5.3}$$

Since $\mathcal{A}_{\mathrm{c}}$ is unknown, we pursue a method for constructing pairs of spatial CSs: an upper set $\hat{\mathcal{A}}_{\mathrm{c}}^{+}$ and a lower set $\hat{\mathcal{A}}_{\mathrm{c}}^{-}$ such that for a desired confidence level of, for example, $1 - \alpha = 95\%$, the CSs surround the true excursion set $\mathcal{A}_{\mathrm{c}}$ (i.e. $\hat{\mathcal{A}}_{\mathrm{c}}^{+} \subset \mathcal{A}_{\mathrm{c}} \subset \hat{\mathcal{A}}_{\mathrm{c}}^{-}$). Contingent on this, with 95% confidence we may assert that all voxels *contained* in the upper CS $\hat{\mathcal{A}}_{\mathrm{c}}^{+}$ have a Cohen's d effect size *greater* than, for example, $c = 0.8$, and simultaneously, we are 95% confident all voxels *outside* the lower CS $\hat{\mathcal{A}}_{\mathrm{c}}^{-}$ have a Cohen's d effect size *less* than 0.8. Here we emphasize the classical frequentist connotation of the term 'confidence'; letting $\partial\mathcal{A}_{\mathrm{c}}$ denote the boundary of $\mathcal{A}_{\mathrm{c}}$, then precisely, there is

a probability of $1-\alpha$ that the region $\left(\hat{\mathcal{A}}_{\mathrm{c}}^{-} \cap \left(\hat{\mathcal{A}}_{\mathrm{c}}^{+}\right)^{\mathrm{c}}\right)$ computed from a future experiment fully encompasses the true set boundary $\partial\mathcal{A}_{\mathrm{c}}$. In this sense, the upper CS taken away from the lower CS $\left(\hat{\mathcal{A}}_{\mathrm{c}}^{-} \cap \left(\hat{\mathcal{A}}_{\mathrm{c}}^{+}\right)^{\mathrm{c}}\right)$ is similar to a standard confidence interval.

In Chapter 4, we adapted the mathematical theory first proposed in *SSS* to obtain such a method for the mean %BOLD change $\mu(\boldsymbol{s})$. Let $\bar{X}(\boldsymbol{s}) = \frac{1}{N}\sum_{i=1}^{N} Y_i(\boldsymbol{s})$, the sample mean %BOLD change effect. Then subject to continuity of the relevant fields and some basic conditions on the error terms $\epsilon_i(\boldsymbol{s})$, for the excursion set $\mathcal{A}_{\mathrm{c},\mu}$ of voxels with a true %BOLD effect size greater than c:

$$\mathcal{A}_{\mathrm{c},\mu} = \{\boldsymbol{s} \in S : \mu(\boldsymbol{s}) \geq c\}, \tag{5.4}$$

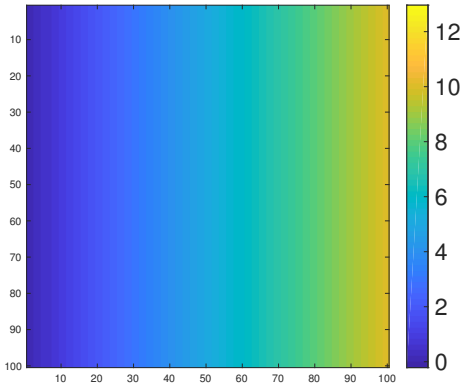
we showed that for a critical constant k, the upper and lower CSs constructed as

$$\hat{\mathcal{A}}_{\mathrm{c},\mu}^{+} := \left\{\boldsymbol{s} : \bar{X}(\boldsymbol{s}) \geq c + \frac{k}{\sqrt{N}}\,\hat{\sigma}(\boldsymbol{s})\right\}, \quad \hat{\mathcal{A}}_{\mathrm{c},\mu}^{-} := \left\{\boldsymbol{s} : \bar{X}(\boldsymbol{s}) \geq c - \frac{k}{\sqrt{N}}\,\hat{\sigma}(\boldsymbol{s})\right\} \tag{5.5}$$

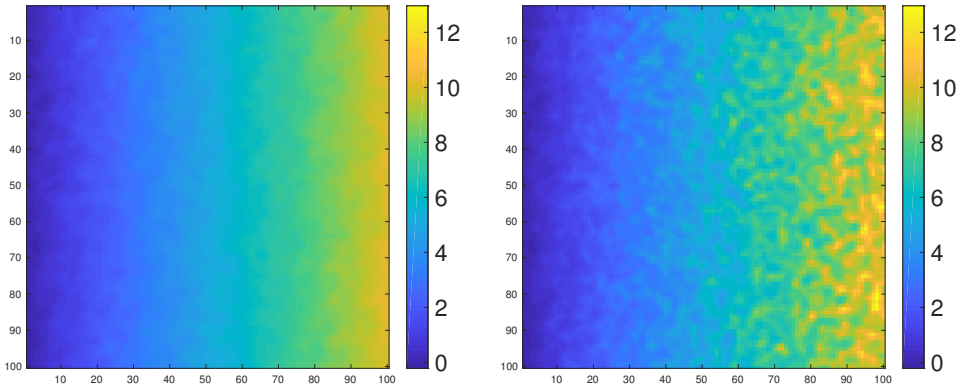
asymptotically satisfied the desired properties described above for the mean %BOLD change effect size. Further to this, we proposed a Wild t-bootstrap method for determining the critical value k, and demonstrated that on applying this method the CSs were also valid for data with smaller sample sizes.

We now seek to develop a similar methodology for the Cohen's d effect size. However, the statistical properties of the Cohen's d estimator $\hat{d}(\boldsymbol{s}) = \frac{\bar{X}(\boldsymbol{s})}{\hat{\sigma}(\boldsymbol{s})}$ are considerably different to the sample mean $\bar{X}(\boldsymbol{s})$. To provide a visual intuition of this in the case of Gaussian data, in Figure 5.1 we display images of both of these fields from a 2D simulation over a square region $S = 100 \times 100$. For $N = 60$ subjects, we simulated a toy run of the signal-plus-noise model in (5.1) where the true underlying signal $\mu(\boldsymbol{s})$ was a linear ramp effect increasing from a magnitude of 0 to 10 in the x-direction while remaining constant in the y-direction (shown in Fig. 5.1(a)). To the signal we added subject-specific Gaussian noise $\epsilon_i(\boldsymbol{s})$ with a spatially constant standard deviation of $\sigma(\boldsymbol{s})$ = 1, smoothed using a 3 voxel FWHM Gaussian kernel. Notably, in this

setup the true Cohen's d field $d(\boldsymbol{s})$ was identical to $\mu(\boldsymbol{s})$.



(a) The true underlying signal $\mu(\boldsymbol{s}) = d(\boldsymbol{s})$, a linear ramp increasing from 0 to 10 in the x-direction.



(b) The sample mean field $\bar{X}(\boldsymbol{s})$

(c) The sample Cohen's d field $\hat{d}(\boldsymbol{s}) = \frac{\bar{X}(\boldsymbol{s})}{\hat{\sigma}(\boldsymbol{s})}$

Figure 5.1: Visualizing the differences between the sample mean and sample Cohen's d field from the 2D simulation. While the sample mean image appears to be uniformly smooth across the region, the sample Cohen's d field becomes rougher from left-to-right.

In Fig. 5.1(b) and Fig. 5.1(c) we show the sample mean and sample Cohen's d fields from this simulation. While the sample mean image is uniformly smooth across the space, the Cohen's d field becomes more speckled from left-to-right. Comparing the sample fields with the true underlying signal, readings in each column of the sample mean image appear to be evenly centred around the true underlying effect size. This is not the case for the estimated Cohen's d image, where looking at the far-right

column, an appreciable quantity of values are close to 12 when the true underlying effect size was 10. Altogether, this suggests that the moments of the sample Cohen's d may be dependent on the true underlying effect size. In the following section we will derive the limiting properties of the Cohen's d estimator for Gaussian data, before proposing our theoretical adjustments to the method presented in Chapter 4 to obtain CSs for Cohen's d effect size images.

5.0.2 Limiting Properties of the Cohen's d Estimator

Motivated by the example in Fig. 5.1, we now consider the one-sample model

$$Y_i(\boldsymbol{s}) = \mu(\boldsymbol{s}) + \epsilon_i(\boldsymbol{s}), \qquad i = 1, ..., N \tag{5.6}$$

for i.i.d. Gaussian data $Y_1(\boldsymbol{s}), ..., Y_N(\boldsymbol{s}) \sim \mathcal{N}(\mu(\boldsymbol{s}), \sigma^2(\boldsymbol{s}))$, with sample mean and standard deviation:

$$\bar{X}(\boldsymbol{s}) = \frac{1}{N}\sum_{i=1}^{N} Y_i(\boldsymbol{s}), \quad \hat{\sigma}^2(\boldsymbol{s}) = \frac{1}{N}\sum_{i=1}^{N}\Big(Y_i(\boldsymbol{s}) - \bar{X}(\boldsymbol{s})\Big)^2. \tag{5.7}$$

We wish to understand the limiting structure of the Cohen's d estimator $\hat{d}(\boldsymbol{s}) = \frac{\bar{X}(\boldsymbol{s})}{\hat{\sigma}(\boldsymbol{s})}$. Applying the multivariate central limit theorem, $\bar{X}(\boldsymbol{s})$ and $\hat{\sigma}(\boldsymbol{s})$ have joint limiting distribution:

$$\sqrt{N}\Big(\Big(\bar{X}(\boldsymbol{s}), \hat{\sigma}(\boldsymbol{s})\Big) - \Big(\mu(\boldsymbol{s}), \sigma(\boldsymbol{s})\Big)\Big) \to \mathcal{N}\Big(0, \Sigma(\boldsymbol{s})\Big), \tag{5.8}$$

where

$$\Sigma(\boldsymbol{s}) = \begin{pmatrix} \sigma^2(\boldsymbol{s}) & 0 \\ 0 & \frac{\sigma^2(\boldsymbol{s})}{2} \end{pmatrix}. \tag{5.9}$$

Applying the multivariate delta method to (5.8) with the function $f(x, y) = \frac{x}{\sqrt{y}}$, this

yields:

$$\sqrt{N}\left(\frac{\bar{X}(\boldsymbol{s})}{\hat{\sigma}(\boldsymbol{s})} - \frac{\mu(\boldsymbol{s})}{\sigma(\boldsymbol{s})}\right) \to \mathcal{N}\left(0, 1 + \frac{\mu^2(\boldsymbol{s})}{2\sigma^2(\boldsymbol{s})}\right). \tag{5.10}$$

Therefore, the limiting field of the Cohen's d estimator $\hat{d}(\boldsymbol{s})$ is asymptotically normal with asymptotic variance $1 + \frac{d^2(\boldsymbol{s})}{2}$. As alluded to in the previous section, it is notable that unlike the sample mean, the asymptotic variance of the Cohen's d estimator is dependent on the underlying true effect size.

We will now assume that the errors are i.i.d. $\epsilon_1(\boldsymbol{s}), ..., \epsilon_N(\boldsymbol{s}) \sim \epsilon(\boldsymbol{s})$, where $\epsilon(\boldsymbol{s})$ is a mean zero Gaussian random field such that for all $\boldsymbol{s}, \boldsymbol{t} \in S$:

$$\mathrm{Cov}[\epsilon(\boldsymbol{s}), \epsilon(\boldsymbol{t})] = \sigma(\boldsymbol{s})\sigma(\boldsymbol{t})\rho(\boldsymbol{s}, \boldsymbol{t}), \tag{5.11}$$

where $\rho(\boldsymbol{s}, \boldsymbol{t})$ is the Pearson correlation coefficient. In this case, we can also derive the covariance of the limiting distribution. Consider the vector:

$$\begin{pmatrix} Y_i(\boldsymbol{s}) \\ (Y_i(\boldsymbol{s}) - \mu(\boldsymbol{s}))^2 \\ Y_i(\boldsymbol{t}) \\ (Y_i(\boldsymbol{t}) - \mu(\boldsymbol{t}))^2 \end{pmatrix} \tag{5.12}$$

which has mean $\Big(\mu(\boldsymbol{s}), \epsilon^2(\boldsymbol{s}), \mu(\boldsymbol{t}), \epsilon^2(\boldsymbol{t})\Big)$ and covariance matrix:

$$\Sigma(\boldsymbol{s}, \boldsymbol{t}) = \begin{pmatrix} \sigma^2(\boldsymbol{s}) & 0 & \sigma(\boldsymbol{s})\sigma(\boldsymbol{t})\rho(\boldsymbol{s}, \boldsymbol{t}) & 0 \\ 0 & 2\sigma^4(\boldsymbol{s}) & 0 & 2\sigma^2(\boldsymbol{s})\sigma(\boldsymbol{t})^2\rho^2(\boldsymbol{s}, \boldsymbol{t}) \\ \sigma(\boldsymbol{s})\sigma(\boldsymbol{t})\rho(\boldsymbol{s}, \boldsymbol{t}) & 0 & \sigma^2(\boldsymbol{t}) & 0 \\ 0 & 2\sigma^2(\boldsymbol{s})\sigma^2(\boldsymbol{t})\rho^2(\boldsymbol{s}, \boldsymbol{t}) & 0 & 2\sigma^4(\boldsymbol{t}) \end{pmatrix} \tag{5.13}$$

By implementing the multivariate central limit theorem once again, from above we

deduce the asymptotic joint distribution:

$$\sqrt{N}\Big(\Big(\bar{X}(\boldsymbol{s}),\hat{\sigma}^2(\boldsymbol{s}),\bar{X}(\boldsymbol{t}),\hat{\sigma}^2(\boldsymbol{t})\Big)-\Big(\mu(\boldsymbol{s}),\sigma^2(\boldsymbol{s}),\mu(\boldsymbol{t}),\sigma^2(\boldsymbol{t})\Big)\Big)\to\mathcal{N}\Big(0,\Sigma(\boldsymbol{s},\boldsymbol{t})\Big). \tag{5.14}$$

Applying the multivariate delta method with the function $g(x_1,y_1,x_2,y_2)=\left(\frac{x_1}{\sqrt{y_1}},\frac{x_2}{\sqrt{y_2}}\right)$ yields:

$$\sqrt{N}\left(\left(\frac{\bar{X}(\boldsymbol{s})}{\hat{\sigma}(\boldsymbol{s})},\frac{\bar{X}(\boldsymbol{t})}{\hat{\sigma}(\boldsymbol{t})}\right)-\left(\frac{\mu(\boldsymbol{s})}{\sigma(\boldsymbol{s})},\frac{\mu(\boldsymbol{t})}{\sigma(\boldsymbol{t})}\right)\right)\to\mathcal{N}\Big(0,\Sigma^*(\boldsymbol{s},\boldsymbol{t})\Big), \tag{5.15}$$

where:

$$\Sigma^*(\boldsymbol{s},\boldsymbol{t})=\begin{pmatrix}1+\frac{d^2(\boldsymbol{s})}{2} & \rho(\boldsymbol{s},\boldsymbol{t})+\rho^2(\boldsymbol{s},\boldsymbol{t})\frac{d(\boldsymbol{s})d(\boldsymbol{t})}{2}\\ \rho(\boldsymbol{s},\boldsymbol{t})+\rho^2(\boldsymbol{s},\boldsymbol{t})\frac{d(\boldsymbol{s})d(\boldsymbol{t})}{2} & 1+\frac{d^2(\boldsymbol{t})}{2}\end{pmatrix}. \tag{5.16}$$

5.0.3 Spatial Confidence Sets for Cohen's d Effect Size Images

Once again, consider the model outlined at the start of Section 5.0.1. For clarity, we reiterate that the spatial CSs for the raw %BOLD change field $\mu(\boldsymbol{s})$ of focus in our previous work took the form:

$$\hat{\mathcal{A}}^{+}_{\mathrm{c},\mu}:=\left\{\boldsymbol{s}:\bar{X}(\boldsymbol{s})\geq c+\frac{k}{\sqrt{N}}\hat{\sigma}(\boldsymbol{s})\right\},\quad \hat{\mathcal{A}}^{-}_{\mathrm{c},\mu}:=\left\{\boldsymbol{s}:\bar{X}(\boldsymbol{s})\geq c-\frac{k}{\sqrt{N}}\hat{\sigma}(\boldsymbol{s})\right\}, \tag{5.17}$$

where k was determined via a Wild t-bootstrap procedure. Such a construction of the CSs was motivated by the limiting properties of the field:

$$M(\boldsymbol{s})=\sqrt{N}\cdot\frac{\bar{X}(\boldsymbol{s})-\mu(\boldsymbol{s})}{\hat{\sigma}(\boldsymbol{s})}. \tag{5.18}$$

In particular, letting $\partial\mathcal{A}_{\mathrm{c},\mu}$ denote the boundary of $\mathcal{A}_{\mathrm{c},\mu}$ defined in (5.4), then on a neighbourhood U of $\partial\mathcal{A}_{\mathrm{c},\mu}$, it was shown in *SSS* that $M(\boldsymbol{s})$ converges weakly to a smooth Gaussian field $G(\boldsymbol{s})$ on U with mean zero, unit variance, and with the same (unknown) spatial correlation as each of the ϵ_i.

In the previous section, for the model given in (5.6) we derived the pointwise

convergence of the function:

$$N(\boldsymbol{s}) = \sqrt{N} \cdot \frac{\hat{d}(\boldsymbol{s}) - d(\boldsymbol{s})}{\sqrt{1 + \frac{d^2(\boldsymbol{s})}{2}}} \tag{5.19}$$

to a Gaussian field $\mathcal{G}(\boldsymbol{s})$ with mean zero, unit variance, and covariance structure:

$$\text{Cov}[\mathcal{G}(\boldsymbol{s}), \mathcal{G}(\boldsymbol{t})] = \frac{\rho(\boldsymbol{s}, \boldsymbol{t}) + \rho(\boldsymbol{s}, \boldsymbol{t})^2 \frac{d(\boldsymbol{s})d(\boldsymbol{t})}{2}}{\sqrt{\left(1 + \frac{d(\boldsymbol{s})^2}{2}\right)\left(1 + \frac{d(\boldsymbol{t})^2}{2}\right)}}. \tag{5.20}$$

This suggests a natural analog to the construction of CSs in (5.17) for the Cohen's d effect size given by:

$$\hat{\mathcal{A}}^{+}_{\text{c},d} := \left\{ \boldsymbol{s} : \hat{d}(\boldsymbol{s}) \geq c + \frac{k}{\sqrt{N}} \sqrt{1 + \frac{\hat{d}^2(\boldsymbol{s})}{2}} \right\}, \quad \hat{\mathcal{A}}^{-}_{\text{c},d} := \left\{ \boldsymbol{s} : \hat{d}(\boldsymbol{s}) \geq c - \frac{k}{\sqrt{N}} \sqrt{1 + \frac{\hat{d}^2(\boldsymbol{s})}{2}} \right\}. \tag{5.21}$$

Ideally, we wish to apply the same Wild t-bootstrap procedure described in Section 4.1.2 of Chapter 4 to approximate the limiting field $G(\boldsymbol{s})$ in order to determine k. However, we will now show that such an approach is not viable for Cohen's d, before proposing a modified procedure to solve the problem. Going forward our focus will mainly be on the Cohen's d effect size, and thus for brevity, we will drop the subscript from our notation and refer to the Cohen's d CSs above as $\hat{\mathcal{A}}^{+}_{\text{c}}$ and $\hat{\mathcal{A}}^{-}_{\text{c}}$ respectively.

5.0.4 Modified Residuals for the Cohen's d Wild t-bootstrap

In *SSS*, it was shown that the limiting coverage of the CSs for the %BOLD effect size $\mu(\boldsymbol{s})$ is governed by the maximum distribution of the limiting Gaussian field $G(\boldsymbol{s})$ on the boundary $\partial\mathcal{A}_{\text{c},\mu}$ such that:

$$\lim_{n\to\infty} P\left[\hat{\mathcal{A}}^{+}_{\text{c},\mu} \subset \mathcal{A}_{\text{c},\mu} \subset \hat{\mathcal{A}}^{-}_{\text{c},\mu}\right] = P\left[\sup_{\boldsymbol{s}\in\partial\mathcal{A}_{\text{c},\mu}} |G(\boldsymbol{s})| \leq k\right]. \tag{5.22}$$

Since the limiting Gaussian field $G(\boldsymbol{s})$ is unknown, in Chapter 4 we implemented a Wild t-bootstrap procedure to approximate $G(\boldsymbol{s})$ on the boundary $\partial\mathcal{A}_{c,\mu}$. Defining the standardized residuals:

$$\tilde{\epsilon}_i(\boldsymbol{s}) = \frac{Y_i(\boldsymbol{s}) - \bar{X}(\boldsymbol{s})}{\hat{\sigma}(\boldsymbol{s})}, \tag{5.23}$$

then the Wild t-bootstrap approximating field is given by:

$$\tilde{G}^*(\boldsymbol{s}) = \frac{1}{\sqrt{N}} \sum_{i=1}^{N} r_i^* \frac{\tilde{\epsilon}_i(\boldsymbol{s})}{\hat{\sigma}^*(\boldsymbol{s})}, \tag{5.24}$$

where the r_i^* are i.i.d. Rademacher variables, and $\hat{\sigma}^*(\boldsymbol{s})$ is the standard deviation of the current realization of bootstrapped residuals $r_i^*\tilde{\epsilon}_i(\boldsymbol{s})$. The asterisk $(*)$ indicates that $\tilde{G}^*(\boldsymbol{s})$ is one of many bootstrap samples; in practice, we would obtain a large number B of bootstrap samples $\tilde{G}^*(\boldsymbol{s})$, and approximate k as the $(1-\alpha)100$ percentile of the B suprema $\sup_{\boldsymbol{s}\in\partial\mathcal{A}_c} |\tilde{G}^*(\boldsymbol{s})|$.

While this method is valid in regards to %BOLD change, for Cohen's d we demonstrate that asymptotically the correlation structure of the approximating field $\tilde{G}^*(\boldsymbol{s})$ is incorrect. Consider again the Gaussian model in Section 5.0.2. In this instance, the covariance of the approximating field is:

$$\begin{aligned}
\mathrm{Cov}[\tilde{G}^*(\boldsymbol{s}), \tilde{G}^*(\boldsymbol{t})] &= \frac{1}{N} \sum_{i,j=1}^{N} \frac{\tilde{\epsilon}_i(\boldsymbol{s})\tilde{\epsilon}_j(\boldsymbol{t})}{\hat{\sigma}^*(\boldsymbol{s})\hat{\sigma}^*(\boldsymbol{t})} \mathrm{Cov}[r_i, r_j] \\
&= \frac{1}{N} \cdot \frac{1}{\hat{\sigma}^*(\boldsymbol{s})\hat{\sigma}^*(\boldsymbol{t})} \cdot \frac{1}{\hat{\sigma}(\boldsymbol{s})\hat{\sigma}(\boldsymbol{t})} \sum_{i=1}^{N} (Y_i(\boldsymbol{s}) - \bar{X}(\boldsymbol{s}))(Y_i(\boldsymbol{t}) - \bar{X}(\boldsymbol{t})) \\
&\to \rho(\boldsymbol{s}, \boldsymbol{t}),
\end{aligned} \tag{5.25}$$

where we note that since the standardized residuals $\tilde{\epsilon}_i(\boldsymbol{s})$ are asymptotically Gaussian with unit variance, the bootstrap estimate of the standard deviation of the standardized residuals $\hat{\sigma}^*(\boldsymbol{s})$ also converges pointwise to 1. Comparing this with the correlation structure of the true limiting field $\mathcal{G}(\boldsymbol{s})$ given in (5.20), letting $d(\boldsymbol{s}) = d(\boldsymbol{t})$ and

setting $q = \frac{d^2(\boldsymbol{s})}{2} \geq 0$, we observe that:

$$\mathrm{Cov}[\mathcal{G}(\boldsymbol{s}), \mathcal{G}(\boldsymbol{t})] = \frac{\rho(\boldsymbol{s},\boldsymbol{t}) + \rho^2(\boldsymbol{s},\boldsymbol{t})\frac{d(\boldsymbol{s})d(\boldsymbol{t})}{2}}{\sqrt{\left(1+\frac{d^2(\boldsymbol{s})}{2}\right)\left(1+\frac{d^2(\boldsymbol{t})}{2}\right)}} = \rho(\boldsymbol{s},\boldsymbol{t})\frac{1+\rho(\boldsymbol{s},\boldsymbol{t})q}{1+q} \leq \rho(\boldsymbol{s},\boldsymbol{t}). \quad (5.26)$$

Therefore, the limiting covariance of the approximating field *overestimates* the covariance function of $\mathcal{G}(\boldsymbol{s})$.

To solve this, we implement a Taylor expansion transformation recently proposed in *FABIAN's PAPER* to construct modified residuals with the desired limiting properties. Motivated by the delta method procedures used in Section 5.0.2, an estimation of the residual field for a single subject i is given by:

$$\mathcal{E}_i(\boldsymbol{s}) = \frac{Y_i(\boldsymbol{s})}{Y_i(\boldsymbol{s}) - \hat{\mu}(\boldsymbol{s})} - \frac{\hat{\mu}(\boldsymbol{s})}{\hat{\sigma}(\boldsymbol{s})} = f\Big(Y_i(\boldsymbol{s}), (Y_i(\boldsymbol{s}) - \hat{\mu}(\boldsymbol{s}))^2\Big) - f\Big(\hat{\mu}(\boldsymbol{s}), \hat{\sigma}^2(\boldsymbol{s})\Big), \quad (5.27)$$

where the function $f(x,y) = \frac{x}{\sqrt{y}}$. A first-order Taylor expansion of $f(x,y)$ about the point $(\hat{\mu}(\boldsymbol{s}), \hat{\sigma}^2(\boldsymbol{s}))$ yields the approximating Cohen's d residuals:

$$\begin{aligned} R_i(\boldsymbol{s}) &= \nabla f(\hat{\mu}(\boldsymbol{s}), \hat{\sigma}^2(\boldsymbol{s}))\left(\Big(Y_i(\boldsymbol{s}), (Y_i(\boldsymbol{s}) - \hat{\mu}(\boldsymbol{s}))^2\Big) - \Big(\hat{\mu}(\boldsymbol{s}), \hat{\sigma}^2(\boldsymbol{s})\Big)\right)^{\intercal} \\ &= \frac{Y_i(\boldsymbol{s}) - \hat{\mu}(\boldsymbol{s})}{\hat{\sigma}(\boldsymbol{s})} - \frac{\hat{\mu}(\boldsymbol{s})}{2\hat{\sigma}(\boldsymbol{s})}\left(\frac{\big(Y_i(\boldsymbol{s}) - \hat{\mu}(\boldsymbol{s})\big)^2}{\hat{\sigma}^2(\boldsymbol{s})} - 1\right). \end{aligned} \quad (5.28)$$

Normalizing by the estimated standard deviation of the limiting field $\mathcal{G}(\boldsymbol{s})$, we obtain the modified standardized residuals:

$$\tilde{R}_i(\boldsymbol{s}) = \frac{R_i(\boldsymbol{s})}{\sqrt{1+\frac{\hat{d}^2(\boldsymbol{s})}{2}}}. \quad (5.29)$$

In *FABIAN's PAPER*, they prove that the limiting covariance of $\tilde{R}_i(\boldsymbol{s})$ is equal to the covariance function of $\mathcal{G}(\boldsymbol{s})$. Therefore, a modification of (5.24) leads us to the Cohen's

d version of the Wild t-bootstrap approximating field:

$$\tilde{G}^*(\boldsymbol{s}) = \frac{1}{\sqrt{N}} \sum_{i=1}^{N} r_i^* \frac{\tilde{R}_i(\boldsymbol{s})}{\hat{\sigma}^*(\boldsymbol{s})}, \tag{5.30}$$

where now $\hat{\sigma}^*(\boldsymbol{s})$ is the standard deviation of the bootstrapped Cohen's d residuals $r_i^* \tilde{R}_i(\boldsymbol{s})$.

While normalization of the $R_i(\boldsymbol{s})$ by an estimator of the standard deviation of $\mathcal{G}(\boldsymbol{s})$ provides us with residuals that have the correct limiting properties, for application to fMRI data we are compelled to optimize the performance of the method for smaller sample sizes. In this regard, it may be preferable to standardize the $R_i(\boldsymbol{s})$ using an estimator tailored to the sample. Noting that the sample mean of the approximating residuals, $\bar{X}_R(\boldsymbol{s}) = \frac{1}{N}\sum_{i=1}^{N} R_i(\boldsymbol{s})$, is equal to zero for all N, letting:

$$\hat{\sigma}_R^2(\boldsymbol{s}) = \frac{1}{N} \sum_{i=1}^{N} R_i^2(\boldsymbol{s}), \tag{5.31}$$

then an alternative to (5.29) is to normalize the Cohen's d residuals by their sample standard deviation so that:

$$\tilde{R}_i(\boldsymbol{s}) = \frac{R_i(\boldsymbol{s})}{\hat{\sigma}_R(\boldsymbol{s})}. \tag{5.32}$$

These standardized residuals could then be used for the Wild t-bootstrap approximating field given in (5.30). In this case, the sample standard deviation should also be accounted for in the formation of the CSs, suggesting an alternate construction to (5.21) given by:

$$\hat{\mathcal{A}}_{\mathrm{c}}^{+} := \left\{ \boldsymbol{s} : \hat{d}(\boldsymbol{s}) \geq c + \frac{k}{\sqrt{N}} \hat{\sigma}_R(\boldsymbol{s}) \right\}, \quad \hat{\mathcal{A}}_{\mathrm{c}}^{-} := \left\{ \boldsymbol{s} : \hat{d}(\boldsymbol{s}) \geq c - \frac{k}{\sqrt{N}} \hat{\sigma}_R(\boldsymbol{s}) \right\}. \tag{5.33}$$

In Section **SIMULATIONS SECTION**, we assess the performance of the CSs on synthetic data when the residuals are standardized using either the estimated limiting variance $\sqrt{1 + \frac{\hat{d}^2(\boldsymbol{s})}{2}}$ or the sample standard deviation $\hat{\sigma}_R(\boldsymbol{s})$.

5.0.5 Finite Properties of the Cohen's d Estimator and a Variance-stabilizing Transformation

Up to now, we have used the limiting properties of the Cohen's d estimator to motivate two possible constructions for Cohen's d CSs ((5.21) and (5.33)). In this section, we will draw our attention to the distributional properties of $\hat{d}(\boldsymbol{s})$ for finite samples to make further improvements on these methods, as well as develop another novel procedure for obtaining CSs based on Gaussianizing the distibution of $\hat{d}(\boldsymbol{s})$.

Again, assuming the Gaussian model described in Section 5.0.2, we observe that the Cohen's d estimator can be expressed in the form:

$$\hat{d} = \frac{\bar{X}}{\hat{\sigma}} = \frac{1}{\sqrt{N}} \cdot \frac{\bar{X}}{\hat{\sigma}/\sqrt{N}} = \frac{1}{\sqrt{N}} \cdot \frac{\frac{\bar{X}-\mu}{\sigma/\sqrt{N}} + \frac{\mu}{\sigma/\sqrt{N}}}{\sqrt{\left(\frac{\hat{\sigma}^2}{\sigma^2/(N-1)}\right)/(N-1)}}. \tag{5.34}$$

Since $\frac{\bar{X}-\mu}{\sigma/\sqrt{N}}$ is normally distributed with unit variance and zero mean, and $\frac{\hat{\sigma}^2}{\sigma^2/(N-1)}$ follows a chi-square distribution with $N-1$ degrees of freedom, from the RHS of the equality we see that $\sqrt{N}\hat{d}$ is characterized by a noncentral t-distribution with noncentrality parameter $\sqrt{N}d$ and $N-1$ degrees of freedom. It is known that the noncentral t is asymmetric unless $\mu = 0$; in general, the size of the asymmetry scales with the magnitude of the noncentrality parameter and is inversely proportional to the degrees of freedom. Therefore, we expect the distribution of the Cohen's d estimator to be highly skew when the true effect size is large and the sample size is small. This conflicts with the symmetric construction of the upper and lower CSs $\hat{\mathcal{A}}_{\mathrm{c}}^{+}$ and $\hat{\mathcal{A}}_{\mathrm{c}}^{-}$ given in (5.21) and (5.33), suggesting that the coverage performance of these two methods may decline in such situations.

To account for skewness, we adapt a method originally proposed in *Laubscher (1960)* to stabilize the variance of the noncentral t, transforming to a distribution which is approximately Gaussian, and hence, symmetric.

Theorem 1. *Assume the one-sample Gaussian model described in Section 5.0.2. For fixed*

N, *let:*

$$a = \sqrt{\frac{N-1}{N-3}}, \qquad b = \sqrt{\frac{8N^2 - 17N + 11}{(N-3)(4N-5)^2}},$$

and define

$$\alpha = \frac{1}{b}, \qquad \beta = \frac{a}{b}.$$

Then for $b^* = \sqrt{N}b, \alpha^* = \frac{\alpha}{\sqrt{N}}$, *and* $\beta^* = \sqrt{N}\beta$, *define the transformation* $f : \mathbb{R} \to \mathbb{R}$ *as:*

$$f\left(\hat{d}\right) = \sqrt{N}\left[\alpha^* \operatorname{arcsinh}\left(\beta^* \hat{d}\right) - \alpha^* \operatorname{arcsinh}\left(\beta^* d\left(1 - \frac{3}{4N-5}\right)^{-1}\right) + \frac{1}{2N} b^{*2}\left(d\left(1 - \frac{3}{4N-5}\right)^{-1}\right)\left(\frac{N-1}{N-3} + Nd^2\left(\frac{8N^2 - 17N + 11}{16(N-3)(N-2)^2}\right)\right)^{-\frac{1}{2}}\right]. \tag{5.35}$$

Then the random variable $f\left(\hat{d}\right)$ *has, approximately, zero mean and unit variance.*

Proof. We closely follow the workings given in the '**2. Noncentral *t*.**' section of *Lauβcher (1960)*. We have shown that $\sqrt{N}\hat{d}$ is distributed by a noncentral t-distribution with noncentrality parameter $\sqrt{N}\hat{d}$ and $N-1$ degrees of freedom. Defining:

$$C_N = \sqrt{\frac{N-1}{2}} \frac{\Gamma\left(\frac{N-2}{2}\right)}{\Gamma\left(\frac{N-1}{2}\right)}, \tag{5.36}$$

where Γ is the gamma function, then the expectation and variance of $\sqrt{N}\hat{d}$ are:

$$\mathbb{E}\left[\sqrt{N}\hat{d}\right] = \sqrt{N} d C_N, \tag{5.37}$$

$$\operatorname{Var}\left[\sqrt{N}\hat{d}\right] = \frac{N-1}{N-3} + Nd^2\left(\frac{N-1}{N-3} - C_N^2\right). \tag{5.38}$$

It is known that C_N is well-approximated by the polynomial:

$$C_N \approx \left(1 - \frac{3}{4N-5}\right)^{-1}. \tag{5.39}$$

Substituting into equations (5.37) and (5.38), we deduce approximations of the expectation and variance of $\sqrt{N}\hat{d}$:

$$\mathbb{E}\left[\sqrt{N}\hat{d}\right] \approx \sqrt{N}d\left(1 - \frac{3}{4N-5}\right)^{-1} = m_1, \tag{5.40}$$

$$\mathrm{Var}\left[\sqrt{N}\hat{d}\right] \approx \frac{N-1}{N-3} + Nd^2\left(\frac{8N^2-17N+11}{16(N-3)(N-2)^2}\right) = m_2. \tag{5.41}$$

Now, noting that:

$$m_1^2 = Nd^2\frac{(4N-5)^2}{16(N-2)^2}, \tag{5.42}$$

then m_2 can be expressed in the form:

$$m_2 = a^2 + b^2 m_1^2, \tag{5.43}$$

where $a = \sqrt{\frac{N-1}{N-3}}$, $b = \sqrt{\frac{8N^2-17N+11}{(N-3)(4N-5)^2}}$. Using the variance expression in (5.43) and applying Corollary 1. in *Laubscher*, the approximate variance-stabilizing transformation of $\sqrt{N}\hat{d}$ is given by:

$$\begin{aligned}\psi\left(\sqrt{N}\hat{d}\right) &= \int_0^{\sqrt{N}\hat{d}} \left(a^2 + b^2x^2\right)^{-\frac{1}{2}} dx \\ &= \alpha \,\mathrm{arcsinh}\left(\beta\sqrt{N}\hat{d}\right),\end{aligned} \tag{5.44}$$

where $\alpha = \frac{1}{b}$ and $\beta = \frac{a}{b}$. The quadratic Taylor approximation of $\psi\left(\sqrt{N}\hat{d}\right)$ about the point $\sqrt{N}\hat{d} = \mathbb{E}\left[\sqrt{N}\hat{d}\right] \approx m_1$ is given by:

$$\psi\left(\sqrt{N}\hat{d}\right) \approx \psi(m_1) - \frac{1}{2}b^2 m_1 m_2^{-\frac{1}{2}}. \tag{5.45}$$

Therefore, the random variable:

$$g\left(\sqrt{N}\hat{d}\right) = \psi\left(\sqrt{N}\hat{d}\right) - \psi(m_1) + \frac{1}{2}b^2 m_1 m_2^{-\frac{1}{2}} \tag{5.46}$$

will have, approximately, mean zero and unit variance. Substituting the precise expressions for ψ, m_1, and m_2 into (5.46) yields:

$$g\left(\sqrt{N}\hat{d}\right) = \alpha \operatorname{arcsinh}\left(\beta\sqrt{N}\hat{d}\right) - \alpha \operatorname{arcsinh}\left(\beta\sqrt{N}d\left(1 - \frac{3}{4N-5}\right)^{-1}\right) + \frac{1}{2}b^2\left(\sqrt{N}d\left(1 - \frac{3}{4N-5}\right)^{-1}\right)\left(\frac{N-1}{N-3} + Nd^2\left(\frac{8N^2-17N+11}{16(N-3)(N-2)^2}\right)\right)^{-\frac{1}{2}}. \tag{5.47}$$

At this point, we have established a variance-stabilizing transformation in terms of $\sqrt{N}\hat{d}$, when in practice, we require a transformation in terms of the Cohen's d estimator $\hat{d}$. This is possible by applying a change of variables to b, α and β. Defining $b^* = \sqrt{N}b, \alpha^* = \frac{1}{b^*} = \frac{1}{\sqrt{N}}\alpha$, and $\beta^* = \frac{b^*}{a} = \sqrt{N}\beta$, substituting into (5.47) obtains the desired transformation:

$$g\left(\sqrt{N}\hat{d}\right) := f\left(\hat{d}\right) = \sqrt{N}\left[\alpha^* \operatorname{arcsinh}\left(\beta^*\hat{d}\right) - \alpha^* \operatorname{arcsinh}\left(\beta^* d\left(1 - \frac{3}{4N-5}\right)^{-1}\right) + \frac{1}{2N}b^{*2}\left(d\left(1 - \frac{3}{4N-5}\right)^{-1}\right)\left(\frac{N-1}{N-3} + Nd^2\left(\frac{8N^2-17N+11}{16(N-3)(N-2)^2}\right)\right)^{-\frac{1}{2}}\right]. \tag{5.48}$$

□

Numerical work presented in *Laubscher (1960)* shows that the 90th percentile value of $f\left(\hat{d}\right)$ in (5.48) closely estimates $\phi^{-1}(0.9)$ for a range of true effect sizes d when the sample size is larger than 40. This suggests that, for moderate sample sizes, the distribution of $f\left(\hat{d}\right)$ is approximately Gaussian.

An immediate observation from the proof of Theorem 1. is that the estimated expectation of $\hat{d}$ is:

$$\mathbb{E}\left[\hat{d}\right] \approx d\left(1 - \frac{3}{4N-5}\right)^{-1}, \tag{5.49}$$

and therefore, unlike the sample mean, the Cohen's d estimator is biased. To improve

the performance of the CSs for small sample sizes, this bias should be accounted for in the formulation of the CSs. This leads to a bias-corrected version of the CS construction in (5.21) given by:

$$\hat{\mathcal{A}}_{\mathrm{c}}^{\pm} := \left\{ \boldsymbol{s} : \hat{d}(\boldsymbol{s}) \geq c\left(1 - \frac{3}{4N-5}\right)^{-1} \pm \frac{k}{\sqrt{N}} \sqrt{1 + \frac{\hat{d}^2(\boldsymbol{s})}{2}} \right\}, \tag{5.50}$$

and similarly, a bias-corrected version of the alternate construction in (5.33) given by:

$$\hat{\mathcal{A}}_{\mathrm{c}}^{\pm} := \left\{ \boldsymbol{s} : \hat{d}(\boldsymbol{s}) \geq c\left(1 - \frac{3}{4N-5}\right)^{-1} \pm \frac{k}{\sqrt{N}} \hat{\sigma}_R(\boldsymbol{s}) \right\}. \tag{5.51}$$

Additionally, for any application to real data the Wild t-Bootstrap described in Section 5.0.4 must be applied over an approximation of the boundary $\partial\mathcal{A}_{\mathrm{c}} = \{\boldsymbol{s} \in S : d(\boldsymbol{s}) = c\}$. Taking into consideration the bias of the Cohen's d estimator, we will therefore use the plug-in boundary:

$$\partial\hat{\mathcal{A}}_{\mathrm{c}} = \left\{ \boldsymbol{s} \in S : \hat{d}(\boldsymbol{s}) = c\left(1 - \frac{3}{4N-5}\right)^{-1} \right\}. \tag{5.52}$$

Finally, by the monotonicity of the mapping $x \mapsto \alpha^* \operatorname{arcsinh}\left(\beta^* x\right)$, another possibility is to construct the CSs in the transformed domain $f(S)$. From Theorem 1, the transformed CSs are constructed as:

$$\hat{\mathcal{A}}_{\mathrm{c}}^{\pm} = \left\{ \boldsymbol{s} : \alpha^* \operatorname{arcsinh}\left(\beta^* \hat{d}(\boldsymbol{s})\right) \geq \alpha^* \operatorname{arcsinh}\left(\beta^* c\left(1 - \frac{3}{4N-5}\right)^{-1}\right) \right.$$
$$\left. - \frac{1}{2N} b^{*2} \left(c\left(1 - \frac{3}{4N-5}\right)^{-1}\right) \left(\frac{N-1}{N-3} + N c^2 \left(\frac{8N^2 - 17N + 11}{16(N-3)(N-2)^2}\right)\right)^{-\frac{1}{2}} \pm \frac{k}{\sqrt{N}} \right\}. \tag{5.53}$$

In this case, the Cohen's d residuals given in (5.28) for the Wild t-Bootstrap must also be modified. An estimation of the transformed residual field for a single subject i is

given by:

$$\begin{aligned}\mathcal{E}_i(\boldsymbol{s}) &= \alpha^* \operatorname{arcsinh}\left(\beta^* \frac{Y_i(\boldsymbol{s})}{Y_i(\boldsymbol{s}) - \hat{\mu}(\boldsymbol{s})}\right) - \alpha^* \operatorname{arcsinh}\left(\beta^* \frac{\hat{\mu}(\boldsymbol{s})}{\hat{\sigma}(\boldsymbol{s})}\right) \\ &= h\Big(Y_i(\boldsymbol{s}), (Y_i(\boldsymbol{s}) - \hat{\mu}(\boldsymbol{s}))^2\Big) - h\Big(\hat{\mu}(\boldsymbol{s}), \hat{\sigma}^2(\boldsymbol{s})\Big), \end{aligned} \tag{5.54}$$

where the function $h(x, y) = \alpha^* \operatorname{arcsinh}\left(\beta^* \frac{x}{\sqrt{y}}\right)$. Similarly to the methods applied in Section 5.0.4, a first-order Taylor expansion of $h(x, y)$ about the point $(\hat{\mu}(\boldsymbol{s}), \hat{\sigma}^2(\boldsymbol{s}))$ obtains the transformed Cohen's d residuals:

$$\begin{aligned}\tilde{R}_i(\boldsymbol{s}) &= \nabla h(\hat{\mu}(\boldsymbol{s}), \hat{\sigma}^2(\boldsymbol{s}))\left(\Big(Y_i(\boldsymbol{s}), (Y_i(\boldsymbol{s}) - \hat{\mu}(\boldsymbol{s}))^2\Big) - \Big(\hat{\mu}(\boldsymbol{s}), \hat{\sigma}^2(\boldsymbol{s})\Big)\right)^{\mathsf{T}} \\ &= \frac{\alpha^*\beta^*}{\sqrt{1 + \beta^{*2}\frac{\hat{\mu}^2(\boldsymbol{s})}{\hat{\sigma}^2(\boldsymbol{s})}}}\left(\frac{Y_i(\boldsymbol{s}) - \hat{\mu}(\boldsymbol{s})}{\hat{\sigma}(\boldsymbol{s})} - \frac{\hat{\mu}(\boldsymbol{s})}{2\hat{\sigma}(\boldsymbol{s})}\left(\frac{\big(Y_i(\boldsymbol{s}) - \hat{\mu}(\boldsymbol{s})\big)^2}{\hat{\sigma}^2(\boldsymbol{s})} - 1\right)\right). \end{aligned} \tag{5.55}$$

In practice, the critical value k in (5.53) is computed by applying the Wild t-Bootstrap over $\partial\hat{\mathcal{A}}_{\mathrm{c}}$ using the transformed Cohen's d residuals given above for the bootstrap approximating field in (5.30).

For coherence, we will now formalize the complete procedures to obtain CSs for each of our three proposed CS constructions in (5.50), (5.51), and (5.53).

5.0.6 Three Algorithms for Computing Cohen's d CSs

Based on our derivations up to this point, we give three algorithms to compute Cohen's d CSs for data modelled within the one-sample model in Section 5.0.1. While the first two algorithms are similar, the key difference separating these methods is the estimator of the variance used in the formation of the CSs and for standardizing the Cohen's d residuals. We first describe Algorithm 1., where we use $1 + \frac{\tilde{d}^2(\boldsymbol{s})}{2}$ as the estimator of the variance, motivated by the variance of the limiting field $\mathcal{G}(\boldsymbol{s})$ derived in Sections 5.0.2 and 5.0.3.

Algorithm 1. *For observations $Y_1(\boldsymbol{s}), \ldots, Y_N(\boldsymbol{s})$ over a spatial domain S following the one-*

sample linear model in (5.1), the following procedure yields CSs for the Cohen's d image $\hat{d}(\boldsymbol{s}) = \frac{\hat{\mu}(\boldsymbol{s})}{\hat{\sigma}(\boldsymbol{s})}$ corresponding to a fixed threshold c and confidence level $(1-\alpha)\%$.

1. *For each observation $Y_i(\boldsymbol{s})$ let $\hat{\epsilon}_i(\boldsymbol{s})$ denote the residual field, $\hat{\epsilon}_i(\boldsymbol{s}) = Y_i(\boldsymbol{s}) - \hat{\mu}(\boldsymbol{s})$. Then compute the Cohen's d residuals as:*

$$R_i(\boldsymbol{s}) = \frac{\hat{\epsilon}_i(\boldsymbol{s})}{\hat{\sigma}(\boldsymbol{s})} - \frac{\hat{\mu}(\boldsymbol{s})}{2\hat{\sigma}(\boldsymbol{s})}\left(\frac{\hat{\epsilon}_i^2(\boldsymbol{s})}{\hat{\sigma}^2(\boldsymbol{s})} - 1\right).$$

2. *Normalize the Cohen's d residuals by the estimated limiting standard deviation of the Cohen's d image to obtain the standardized residuals:*

$$\tilde{R}_i(\boldsymbol{s}) = \frac{R_i(\boldsymbol{s})}{\sqrt{1 + \frac{\hat{\mu}^2(\boldsymbol{s})}{2\hat{\sigma}^2(\boldsymbol{s})}}}.$$

3. *Draw N i.i.d. Rademacher variables $r_1^*, ..., r_N^*$, and compute the Wild t-Bootstrap approximating field:*

$$\tilde{G}^*(\boldsymbol{s}) = \frac{1}{\sqrt{N}} \sum_{i=1}^{N} r_i^* \frac{\tilde{R}_i(\boldsymbol{s})}{\hat{\sigma}^*(\boldsymbol{s})},$$

where $\hat{\sigma}^*(\boldsymbol{s})$ is the bootstrap standard deviation of the bootstrapped residuals $r_i\tilde{R}_i(\boldsymbol{s})$.

4. *Obtain the value $k^* = \sup_{\boldsymbol{s}\in\partial\hat{\mathcal{A}}_c} |G^*(\boldsymbol{s})|$, using the bias-corrected estimator of the boundary $\partial\hat{\mathcal{A}}_c = \left\{\boldsymbol{s} \in S : \hat{d}(\boldsymbol{s}) = c\left(1 - \frac{3}{4N-5}\right)^{-1}\right\}$.*

5. *For a large number of bootstrap replications B repeat steps 3. and 4., obtaining the empirical distribution of the absolute maximum $\mathcal{K}_B = \{k_1^*, ..., k_B^*\}$. Compute k as the $(1-\alpha)$ percentile of $\mathcal{K}_B$.*

6. *Obtain the Cohen's d CSs:*

$$\hat{\mathcal{A}}_c^{\pm} := \left\{\boldsymbol{s} : \hat{d}(\boldsymbol{s}) \geq c\left(1 - \frac{3}{4N-5}\right)^{-1} \pm \frac{k}{\sqrt{N}}\sqrt{1 + \frac{\hat{d}^2(\boldsymbol{s})}{2}}\right\}.$$

We now describe Algorithm 2. Here, we use the sample variance of the Co-

hen's d residuals $\hat{\sigma}_R^2(\boldsymbol{s})$ as the variance estimator, motivated by our workings in Section 5.0.4.

Algorithm 2. *For observations $Y_1(\boldsymbol{s}), ..., Y_N(\boldsymbol{s})$ over a spatial domain S following the one-sample linear model in (5.1), the following procedure yields CSs for the Cohen's d image $\hat{d}(\boldsymbol{s}) = \frac{\hat{\mu}(\boldsymbol{s})}{\hat{\sigma}(\boldsymbol{s})}$ corresponding to a fixed threshold c and confidence level $(1-\alpha)\%$.*

1. *For each observation $Y_i(\boldsymbol{s})$ let $\hat{\epsilon}_i(\boldsymbol{s})$ denote the residual field, $\hat{\epsilon}_i(\boldsymbol{s}) = Y_i(\boldsymbol{s}) - \hat{\mu}(\boldsymbol{s})$. Then compute the Cohen's d residuals as:*

$$R_i(\boldsymbol{s}) = \frac{\hat{\epsilon}_i(\boldsymbol{s})}{\hat{\sigma}(\boldsymbol{s})} - \frac{\hat{\mu}(\boldsymbol{s})}{2\hat{\sigma}(\boldsymbol{s})}\left(\frac{\hat{\epsilon}_i^2(\boldsymbol{s})}{\hat{\sigma}^2(\boldsymbol{s})} - 1\right).$$

2. *Normalize the Cohen's d residuals by their sample standard deviation to obtain the standardized residuals:*

$$\tilde{R}_i(\boldsymbol{s}) = \frac{R_i(\boldsymbol{s})}{\hat{\sigma}_R(\boldsymbol{s})}.$$

3. *Draw N i.i.d. Rademacher variables $r_1^*, ..., r_N^*$, and compute the Wild t-Bootstrap approximating field:*

$$\tilde{G}^*(\boldsymbol{s}) = \frac{1}{\sqrt{N}} \sum_{i=1}^{N} r_i^* \frac{\tilde{R}_i(\boldsymbol{s})}{\hat{\sigma}^*(\boldsymbol{s})},$$

where $\hat{\sigma}^*(\boldsymbol{s})$ is the bootstrap standard deviation of the bootstrapped residuals $r_i\tilde{R}_i(\boldsymbol{s})$.

4. *Obtain the value $k^* = \sup_{\boldsymbol{s}\in\partial\hat{\mathcal{A}}_c} |G^*(\boldsymbol{s})|$, using the bias-corrected estimator of the boundary $\partial\hat{\mathcal{A}}_c = \left\{\boldsymbol{s} \in S : \hat{d}(\boldsymbol{s}) = c\left(1 - \frac{3}{4N-5}\right)^{-1}\right\}$.*

5. *For a large number of bootstrap replications B repeat steps 3. and 4., obtaining the empirical distribution of the absolute maximum $\mathcal{K}_B = \{k_1^*, ..., k_B^*\}$. Compute k as the $(1-\alpha)$ percentile of $\mathcal{K}_B$.*

6. *Obtain the Cohen's d CSs:*

$$\hat{\mathcal{A}}_c^{\pm} := \left\{\boldsymbol{s} : \hat{d}(\boldsymbol{s}) \geq c\left(1 - \frac{3}{4N-5}\right)^{-1} \pm \frac{k}{\sqrt{N}}\,\hat{\sigma}_R(\boldsymbol{s})\right\}.$$

Finally, we describe Algorithm 3. Based on the derivations in Section 5.0.5, we transform the estimated Cohen's d image to a field which is approximately Gaussian. This is done to stabilize the variance and remove the skew of the Cohen's d estimator, which may adversely effect the performance of the CSs. By the monotonicity of the transformation, the CSs obtained using this method are valid for inference on the true (un-transformed) Cohen's d effect size.

Algorithm 3. *For observations $Y_1(\boldsymbol{s}), ..., Y_N(\boldsymbol{s})$ over a spatial domain S following the one-sample linear model in (5.1), the following procedure yields CSs for the Cohen's d image $\hat{d}(\boldsymbol{s}) = \frac{\hat{\mu}(\boldsymbol{s})}{\hat{\sigma}(\boldsymbol{s})}$ corresponding to a fixed threshold c and confidence level $(1-\alpha)\%$.*

1. *For each observation $Y_i(\boldsymbol{s})$ let $\hat{\epsilon}_i(\boldsymbol{s})$ denote the residual field, $\hat{\epsilon}_i(\boldsymbol{s}) = Y_i(\boldsymbol{s}) - \hat{\mu}(\boldsymbol{s})$. Then compute the transformed, variance-stabilized Cohen's d residuals as:*

$$\tilde{R}_i(\boldsymbol{s}) = \frac{\alpha^*\beta^*}{\sqrt{1+\beta^{*2}\frac{\hat{\mu}^2(\boldsymbol{s})}{\hat{\sigma}^2(\boldsymbol{s})}}}\left(\frac{Y_i(\boldsymbol{s})-\hat{\mu}(\boldsymbol{s})}{\hat{\sigma}(\boldsymbol{s})} - \frac{\hat{\mu}(\boldsymbol{s})}{2\hat{\sigma}(\boldsymbol{s})}\left(\frac{\left(Y_i(\boldsymbol{s})-\hat{\mu}(\boldsymbol{s})\right)^2}{\hat{\sigma}^2(\boldsymbol{s})} - 1\right)\right).$$

2. *Draw N i.i.d. Rademacher variables $r_1^*, ..., r_N^*$, and compute the Wild t-Bootstrap approximating field:*

$$\tilde{G}^*(\boldsymbol{s}) = \frac{1}{\sqrt{N}}\sum_{i=1}^{N} r_i^* \frac{\tilde{R}_i(\boldsymbol{s})}{\hat{\sigma}^*(\boldsymbol{s})},$$

where $\hat{\sigma}^*(\boldsymbol{s})$ is the bootstrap standard deviation of the bootstrapped residuals $r_i\tilde{R}_i(\boldsymbol{s})$.

3. *Obtain the value $k^* = \sup_{\boldsymbol{s}\in\partial\hat{\mathcal{A}}_c}|G^*(\boldsymbol{s})|$, using the bias-corrected estimator of the boundary $\partial\hat{\mathcal{A}}_c = \left\{\boldsymbol{s}\in S : \hat{d}(\boldsymbol{s}) = c\left(1-\frac{3}{4N-5}\right)^{-1}\right\}$.*

4. *For a large number of bootstrap replications B repeat steps 3. and 4., obtaining the empirical distribution of the absolute maximum $\mathcal{K}_B = \{k_1^*, ..., k_B^*\}$. Compute k as the $(1-\alpha)$ percentile of $\mathcal{K}_B$.*

5. *Obtain the Cohen's d CSs:*

$$\hat{\mathcal{A}}_{\mathrm{c}}^{\pm} = \left\{ \boldsymbol{s} : \alpha^* \operatorname{arcsinh}\left(\beta^* \hat{d}(\boldsymbol{s})\right) \geq \alpha^* \operatorname{arcsinh}\left(\beta^* c\left(1 - \frac{3}{4N-5}\right)^{-1}\right) \right.$$
$$\left. - \frac{1}{2N} b^{*2} \left(c\left(1 - \frac{3}{4N-5}\right)^{-1}\right)\left(\frac{N-1}{N-3} + Nc^2\left(\frac{8N^2-17N+11}{16(N-3)(N-2)^2}\right)\right)^{-\frac{1}{2}} \pm \frac{k}{\sqrt{N}} \right\}. \tag{5.56}$$

5.1 Methods

5.1.1 Simulations

In this section we describe the settings used to evaluate the performance of each of the three algorithms for obtaining Cohen's d CSs on synthetic data. We simulate 3000 independent samples of the Gaussian one-sample model

$$Y_i(\boldsymbol{s}) = \mu(\boldsymbol{s}) + \epsilon_i(\boldsymbol{s}), \qquad i = 1, ..., N$$

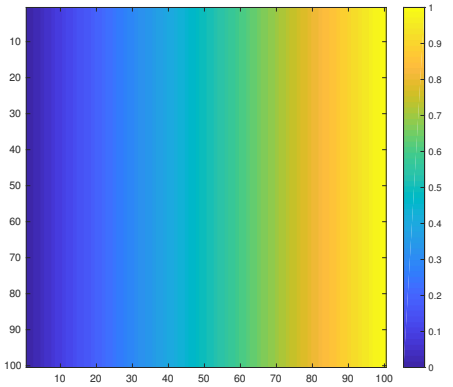
using a range of signals $\mu(\boldsymbol{s})$, Gaussian noise structures $\epsilon_i(\boldsymbol{s})$ with stationary and non-stationary variance $\sigma^2(\boldsymbol{s})$, in two- and three-dimensional regions S. To compute the critical value k, we apply the Wild t-Bootstrap method with $B = 5000$ bootstrap samples on the estimated boundary $\partial\hat{\mathcal{A}}_{\mathrm{c}}$ that must be used for applications of the method to real data. We obtain the boundary using the linear interpolation method described in Section 4.1.3 of Chapter 4. We then compare the empirical coverage – the percentage of trials that the true thresholded signal is completely contained between the upper and lowers CSs (i.e. the number of times for which $\hat{\mathcal{A}}_{\mathrm{c}}^{+} \subset \mathcal{A}_{\mathrm{c}} \subset \hat{\mathcal{A}}_{\mathrm{c}}^{-}$) – between the three algorithms, using the interpolation assessment method described in Section 4.1.4 of Chapter 4. In each simulation, we apply the method for sample sizes of $N = 60, 120, 240$ and 480, and using three nominal coverage probability levels $1 - \alpha = 0.80, 0.90$ and 0.95.

5.1.2 2D Simulations

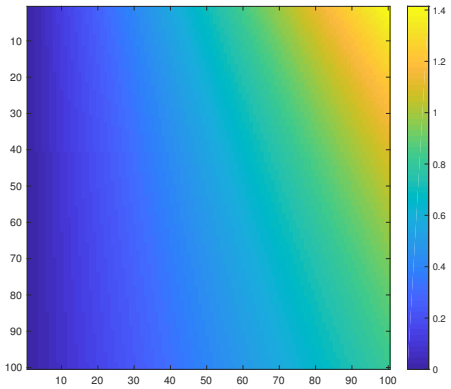
We analyzed the performance of the three algorithms to obtain Cohen's d CSs on a square region of size 100×100. For the true underlying signal $\mu(\boldsymbol{s})$ we considered two different raw effects: First, a linear ramp that increased from a magnitude of 0 to 1 in the x-direction while remaining constant in the y-direction. Second, a circular effect, created by placing a circular phantom of magnitude 1 and radius 30 in the centre of the search region, which was then smoothed using a 3 voxel FWHM Gaussian kernel. If we were to assume that each voxel had a size of 2mm^3, we note that this would amount to applying smoothing with a 6mm FWHM kernel, a fairly typical setting used in fMRI analyses.

To each of these signals we added subject-specific Gaussian noise $\epsilon_i(\boldsymbol{s})$, also smoothed using a 3 voxel FWHM Gaussian kernel, with homogeneous and non-homogeneous variance structures: The first noise field had a spatially constant standard deviation of 1, and therefore in this case the true Cohen's d effect was identical to the underlying signal field $\mu(\boldsymbol{s})$. The second field had a linearly increasing standard deviation structure in the y-direction from $\sqrt{0.5}$ to $\sqrt{1.5}$ while remaining constant in the x-direction. Thus, the variance of this noise field spatially increased in the y-direction from 0.5 to 1.5 in a non-linear fashion.

The true Cohen's d fields $d(\boldsymbol{s})$ for the linear ramp signal with homogeneous and heterogeneous noise are shown in Figure 5.2. The corresponding Cohen's d fields for the circular signal are shown in Figure 5.3. Altogether, for the three algorithms, the two underlying signals and two noise sources gave us 12 separate trials; across all of the simulations, we obtained Cohen's d Confidence Sets for the noise-free cluster $\mathcal{A}_\text{c}$ at a cluster-forming threshold of $c = 0.8$. In Chapter 2.2 of *Cohen (2013)*, $d = 0.8$ was classified as a 'large effect'; for group-level analyses of large-sample fMRI data with ample statistical power (such as the HCP or UK Biobank), effect sizes of this magnitude may be used to assess brain areas where practically significant activation

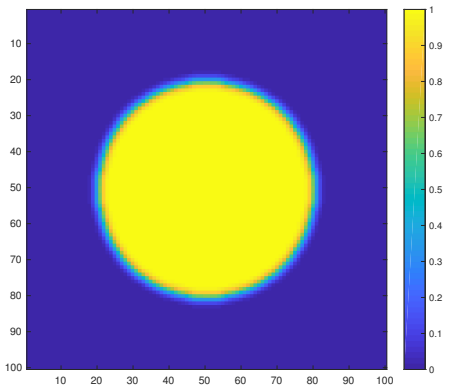


(a) The Cohen's d field $d(\boldsymbol{s})$ for the linear ramp signal $\mu(\boldsymbol{s})$ and homogeneous noise structure.

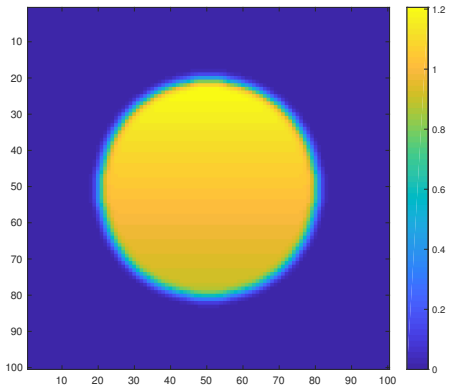


(b) The Cohen's d field $d(\boldsymbol{s})$ for the linear ramp signal $\mu(\boldsymbol{s})$ and heterogeneous noise structure.

Figure 5.2: The two Cohen's d effects corresponding to the linear ramp signal $\mu(\boldsymbol{s})$. On the left, the subject-specific Gaussian noise field $\epsilon_i(\boldsymbol{s})$ has a spatially constant standard deviation of 1, and therefore $d(\boldsymbol{s}) = \mu(\boldsymbol{s})$. On the right, $\epsilon_i(\boldsymbol{s})$ had a spatially increasing standard deviation structure in the y-direction (from top-to-bottom), while remaining constant in the x-direction.



(a) The Cohen's d field $d(\boldsymbol{s})$ for the circular signal $\mu(\boldsymbol{s})$ and homogeneous noise structure.



(b) The Cohen's d field $d(\boldsymbol{s})$ for the circular signal $\mu(\boldsymbol{s})$ and heterogeneous noise structure.

Figure 5.3: The two Cohen's d effects corresponding to the circular signal $\mu(\boldsymbol{s})$. On the left, the subject-specific Gaussian noise field $\epsilon_i(\boldsymbol{s})$ has a spatially constant standard deviation of 1, and therefore $d(\boldsymbol{s}) = \mu(\boldsymbol{s})$. On the right, $\epsilon_i(\boldsymbol{s})$ had a spatially increasing standard deviation structure in the y-direction (from top-to-bottom), while remaining constant in the x-direction.

has occurred.

5.1.3 3D Simulations

Four signal types $\mu(\boldsymbol{s})$ were considered to analyze the performance of the three algorithms in three dimensions. The first three of these signals were generated synthetically on a cubic region of size $100 \times 100 \times 100$: Firstly, a small spherical effect, created by placing a spherical phantom of magnitude 1 and radius 5 in the centre of the search region, which was then smoothed using a 3 voxel FWHM Gaussian kernel. Secondly, a larger spherical effect, generated identically to the first effect with the exception that the spherical phantom had a radius of 30. Lastly, we created an effect by placing four spherical phantoms of magnitude 1 in the region of varying radii and then smoothing the entire image using a 3 voxel FWHM Gaussian.

Each of the images were re-scaled after smoothing to have a maximum intensity of 1. For the small and large spherical effect an image-wise re-scaling was applied, where all locations in the smoothed map were divided through by the maximum intensity across the region. For the final effect, because parts of the four spherical phantoms overlapped after smoothing, the signal intensities in these regions summed to greater than 1. In this case, we re-scaled the smoothed map by reducing the intensities in these areas to have a magnitude of 1 while leaving the rest of the image untouched.

Similar to the two-dimensional simulations, for the three signals described above we added 3-voxel smoothed Gaussian noise with homogeneous and heterogeneous variance structures. The first noise field had a spatially constant standard deviation of 1, while the second field had a linearly increasing standard deviation in the z-direction from $\sqrt{0.5}$ to $\sqrt{1.5}$, while remaining constant in both the x- and y- directions. As demonstrated for the 2D simulations in Figures 5.2 and 5.3, this lead to two different true Cohen's d effect-size images $d(\boldsymbol{s})$ corresponding to the homogeneous and heterogeneous standard deviation fields $\sigma(\boldsymbol{s})$ used for the noise.

For the final signal type, we took advantage of big data that has been made available through the UK Biobank in an attempt to generate an effect that replicated the true %BOLD change induced during an fMRI task. We randomly selected 4000 subject-level contrast of parameter estimate result maps from the Hariri Faces/Shapes task-fMRI data collected as part of the UK Biobank brain imaging study. Full details on how the data were acquired and processed is given in *Miller et al. (2016)*, *Alfaro-Almagro et al. (2018)* and the UK Biobank Showcase; information on the task paradigm is given in *Hariri et al. (2002)*. From these contrast maps, we computed a group-level full mean and full standard deviation image. In the final simulation, we used the group-level Biobank mean image as the true underlying signal $\mu(\boldsymbol{s})$ for each subject, and the full standard deviation image was used for the standard deviation of each simulated subject-specific Gaussian noise field $\epsilon_i(\boldsymbol{s})$ added to the true signal. Because of the considerably large sample size of high-quality data from which these maps have been obtained, we anticipate that both of these images are highly representative of the true underlying fields that they approximate. Both images were masked using an intersection of all 4000 of the subject-level brain masks.

Once again, we smoothed the noise field using a 3 voxel FWHM Gaussian kernel; since the Biobank maps were written with voxel sizes of 2mm^3, this is analogous to applying 6mm FWHM smoothing to the noise field of the original data.

In Figure 5.4, we show the true underlying Cohen's d fields for the three synthetic 3D effects with homogeneous noise structure, and the Cohen's d field corresponding to the UK Biobank full mean and standard deviation. For all four signal types, we obtained Cohen's d Confidence Sets for the threshold $c = 0.8$.

5.1.4 Application to Human Connectome Project Data

To provide a real-data demonstration of the three methods proposed in this work, we computed Cohen's d CSs on 80 participants data from the Unrelated 80 package released as part of the Human Connectome Project (HCP, S1200 Release) us-

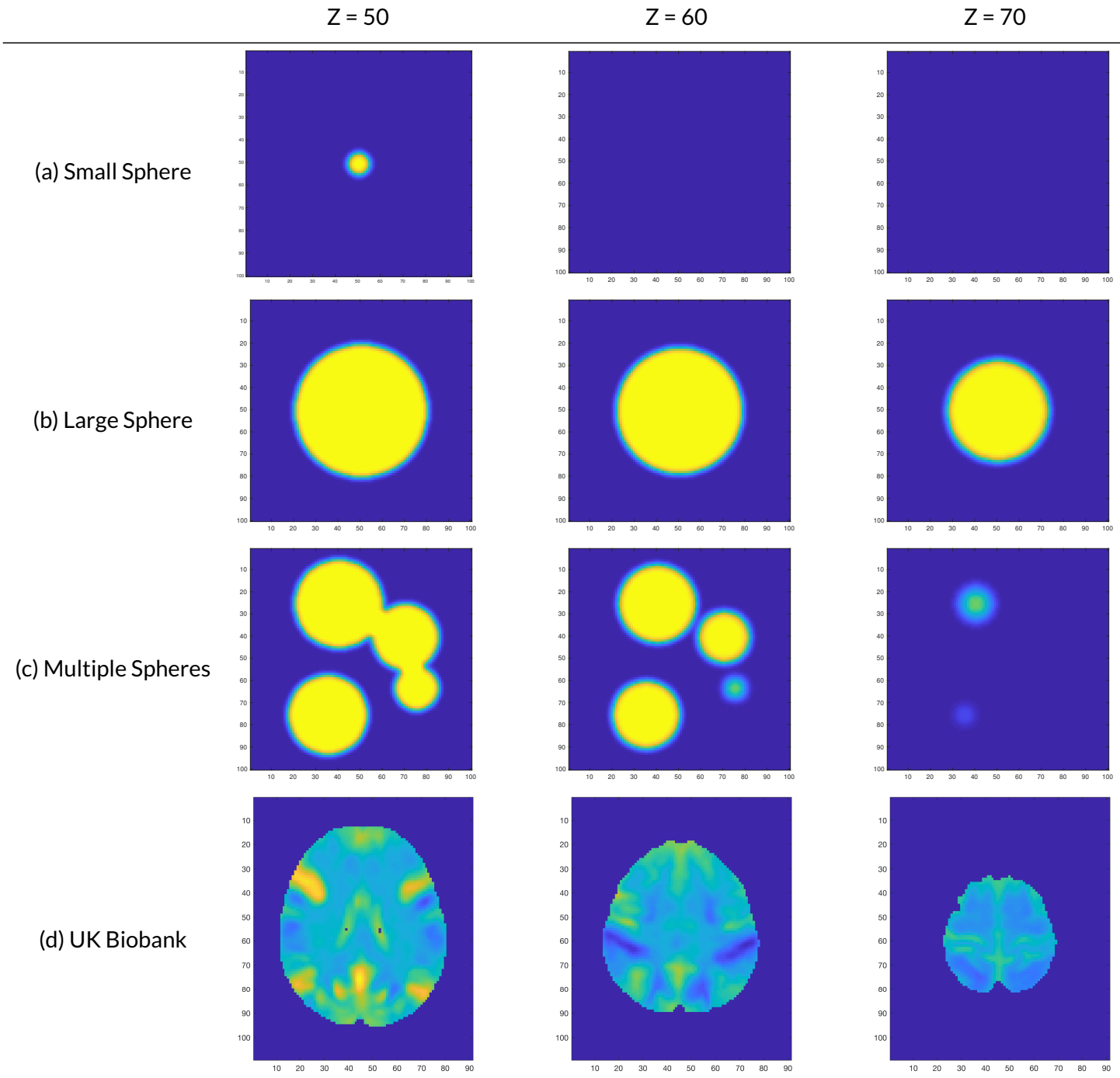


Figure 5.4: Four of the Cohen's d fields $d(\boldsymbol{s})$ used for the 3D simulations. In (a) to (c), we show the Cohen's d field for the three different spherical effects $\mu(\boldsymbol{s})$ when Gaussian noise with spatially homogeneous standard deviation was added to the signal. In (d) we show the Cohen's d field corresponding to the UK Biobank full mean and standard deviation images. Note that the colormap limits for the first three Cohen's d effect-size images are from 0 to 1, while the colormap limits for the UK Biobank image is from -0.8 to 0.8.

ing all three algorithms described in Section 5.0.6. Cohen's d CSs were obtained for the subject-level 2-back vs 0-back contrast maps from the working memory task results included with the HCP dataset. For a comparison with standard fMRI inference procedures, we also performed a traditional statistical group-level analysis on the data. A one-sample t-test was carried out in SPM, using a voxelwise FWE-corrected threshold of $p < 0.05$ obtained via permutation test with SPM's SnPM toolbox. Details about HCP's scanning protocol and data-analysis pipeline alongside a description of the working memory task are provided in Section 4.2.4 of Chapter 4. Mirroring the methods used in the previous chapter, we applied additional smoothing to the final contrast maps to mimic images smoothed using a 6mm FWHM Gaussian kernel. This is a more typical degree of smoothing applied to functional data than the 4mm kernel originally used in the HCP analysis pipeline.

5.2 Results

5.2.1 2D Simulations

Empirical coverage results for each of the three algorithms are presented for the linear ramp signal in Fig. 5.5 and for the circular signal in Fig. 5.6. In both figures, on the top row we display the coverage results obtained when the standard deviation field of the noise was homogeneous across the region (corresponding to Fig. 5.2(a) for the linear ramp, Fig. 5.3(a) for the circle), and on the bottom row we display the equivalent results when the standard deviation field was spatially heterogeneous (Fig. 5.2(b) and Fig. 5.3(b) for the linear ramp and circle respectively). Coverage results obtained with Algorithm 1. are displayed with a red curve, Algorithm 2. with a blue curve, and Algorithm 3. with a magenta curve.

For the linear ramp, across all confidence levels $1-\alpha = 0.80, 0.90,$ and 0.95 we generally observed valid, over-coverage for all three algorithms, particularly when larger sample sizes were used. In all plots, it appears that the coverage rates for the

three algorithms are converging to the same value, slightly above the nominal target. Specifically, for the nominal target level of 80%, in both the homogeneous and heterogeneous cases all empirical results seem to be converging to around 88% (Fig. 5.5, left-side plots). For the 95% target, the scale of disagreement between the empirical results and the nominal target is smaller; here, all coverage results hover close to 96% for $N = 240$ and 480 (Fig. 5.5, right-side plots).

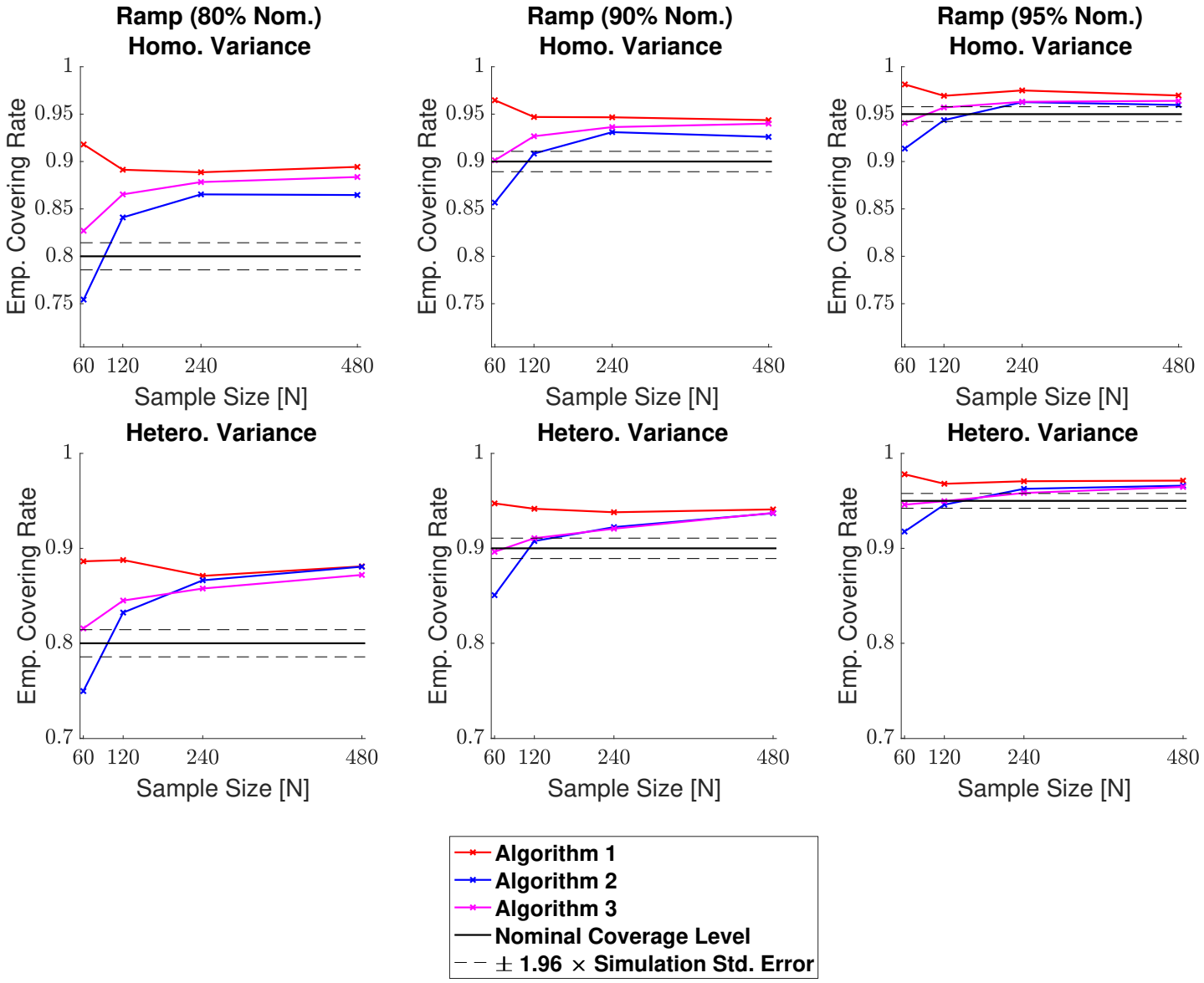


Figure 5.5: Coverage results for the linear ramp signal, with homogeneous (top row) and heterogeneous (bottom row) Gaussian noise structures. For large sample sizes the empirical coverage performance of all three algorithms was similar, hovering slightly above the nominal level in all simulations. For $N = 60$ the degree of over-coverage became larger for Algorithm 1., while empirical coverage for Algorithm 2. fell below the nominal target. Algorithm 3. performed best, with all results remaining particularly close to the nominal target level for simulations using a 95% confidence level (right plots).

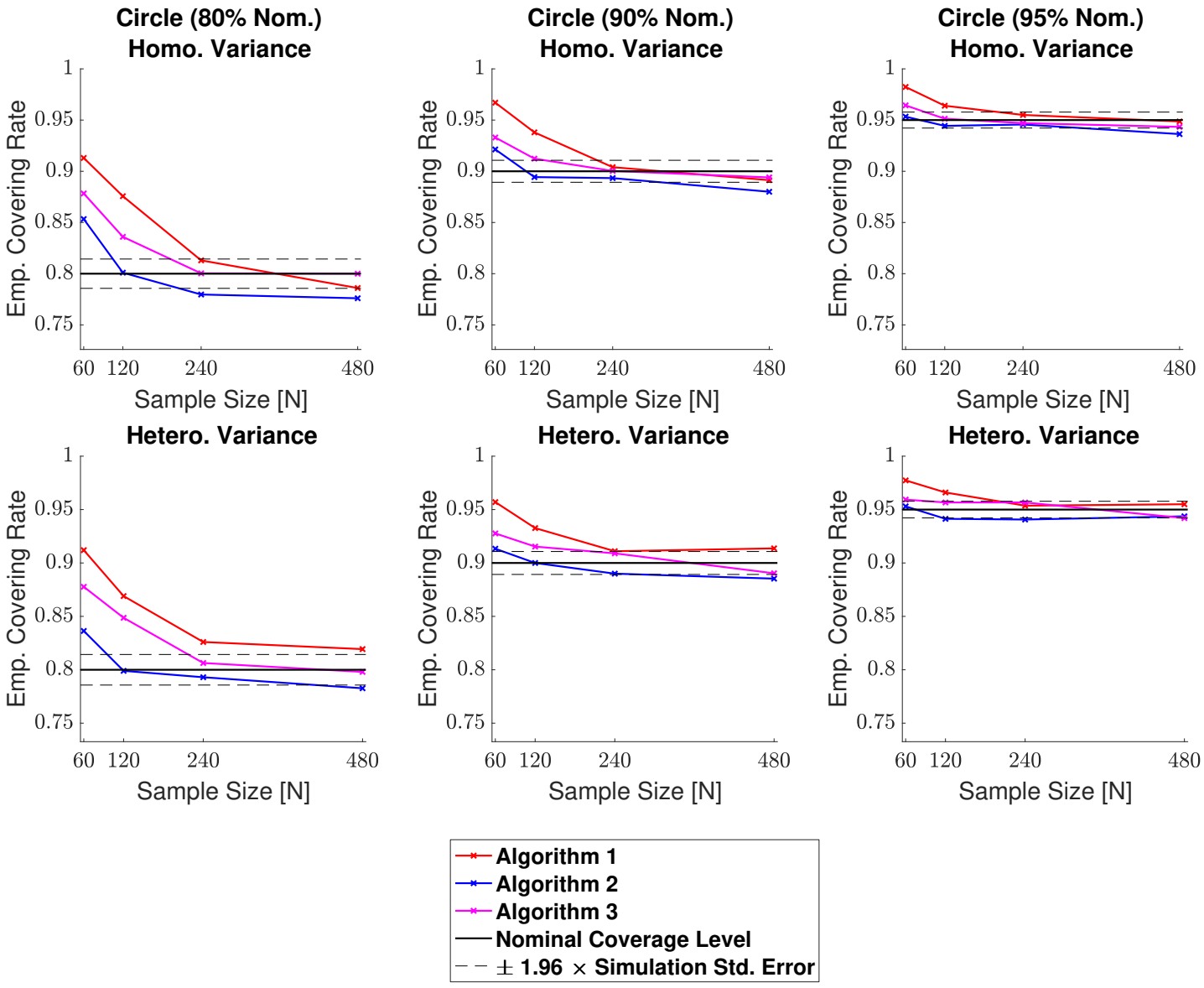


Figure 5.6: Coverage results for the circular signal, with homogeneous (top row) and heterogeneous (bottom row) Gaussian noise structures. All algorithms performed very well, and unlike the linear ramp, empirical coverage for all three methods converged towards the nominal level. For smaller sample sizes there was a slight degree of over-coverage, most noticeably for simulations using the 80% nominal target. Overall, Algorithm 2. performed marginally better than the other two methods, and Algorithm 1. performed the worst.

While for larger sample sizes the performance of all three algorithms was similar, there was greater disparity between the methods for simulations using the small sample size of $N = 60$. Here, the empirical coverage results for Algorithm 1. were consistently higher than the other two methods, and at the same time, results for Algorithm 2. were always the smallest. Notably, unlike the other two methods, the empirical coverage rate for Algorithm 2. fell below the nominal level in all simulations with a sample size of $N = 60$.

For the circular signal, on-the-whole all three methods performed very well. In this instance, almost all empirical results for Algorithm 2. and Algorithm 3. lied within the 95% confidence interval of the nominal coverage rate (blue and magenta curves sandwiched between black dashed lines for all plots in Fig. 5.6), with Algorithm 2. performing marginally better. While we observed slight over-coverage with the three methods for $N = 60$, most substantially in simulations using the 80% nominal target (Fig. 5.6, left-side plots), empirical coverage converged towards the nominal level for all three algorithms.

Finally, the use of homogeneous or heterogeneous noise in the model had minimal impact on any of the algorithm's empirical coverage performance for either of the signals. This is exemplified in Figs. 5.5 and 5.6, where in both cases the homogeneous coverage plots presented in the top row are almost identical to the corresponding heterogeneous plots shown below.

5.2.2 3D Simulations

Empirical coverage results for each of the three algorithms are presented in Figs. 5.7, 5.8, 5.9 and 5.10 respectively for each of the four 3D signal types displayed in Fig. 5.4 (small sphere, large sphere, multiple spheres, UK Biobank). For the spherical effects (Figs. 5.7, 5.8 and 5.9), on the top row we display the coverage results obtained when the standard deviation field of the noise was homogeneous across the region, and on the bottom row we display the equivalent results when the standard deviation field

was spatially heterogeneous. For the UK Biobank signal (Fig. 5.10), the full standard deviation image computed from the UK Biobank data was used for the standard deviation field of the noise, and hence in this case there is only one row of results. Once again, coverage results obtained with Algorithm 1. are displayed with a red curve, Algorithm 2. with a blue curve, and Algorithm 3. with a magenta curve.

Across all 3D simulations, we observed consistencies between the results obtained with each of the three algorithms: In general, empirical coverage for all methods came above the nominal target, and similar to the 2D simulations, the extent of over-coverage was smaller when a larger confidence level was used. Comparing the three methods, coverage results for Algorithm 1. were considerably higher than the other two methods, particularly when a small sample size and confidence level were used. For the 'large sphere' and 'multiple spheres' signal types, Algorithm 1. suffered with over-coverage of above 15% in simulations with a sample of size of $N = 60$ and a nominal target level of 80% (Figs. 5.8 and 5.9, left-side plots). For both of these signals, there was still a considerable amount of over-coverage when larger sample sizes of $N = 240$ and 480 were used. On the other hand, Algorithm 2. and Algorithm 3. performed similarly in large sample sizes across all simulations, with empirical coverage results coming slightly above the nominal target. Notably, both of these algorithms performed very well for simulations with a 95% nominal target level (all figures, right-side plots). Differences between these two methods were more distinguished for smaller sample sizes of $N = 60$ and 120, where coverage results for Algorithm 2. were moderately less than Algorithm 3. Consequentially, Algorithm 2.'s results came closer to the nominal target here, although for the 'multiple sphere' and 'UK Biobank' signal types, in some cases Algorithm 2.'s results fell *below* the nominal level (Figs. 5.9 and 5.10). Overall, empirical coverage for Algorithm 3. was the most uniform of the three methods with respect to changes in sample size.

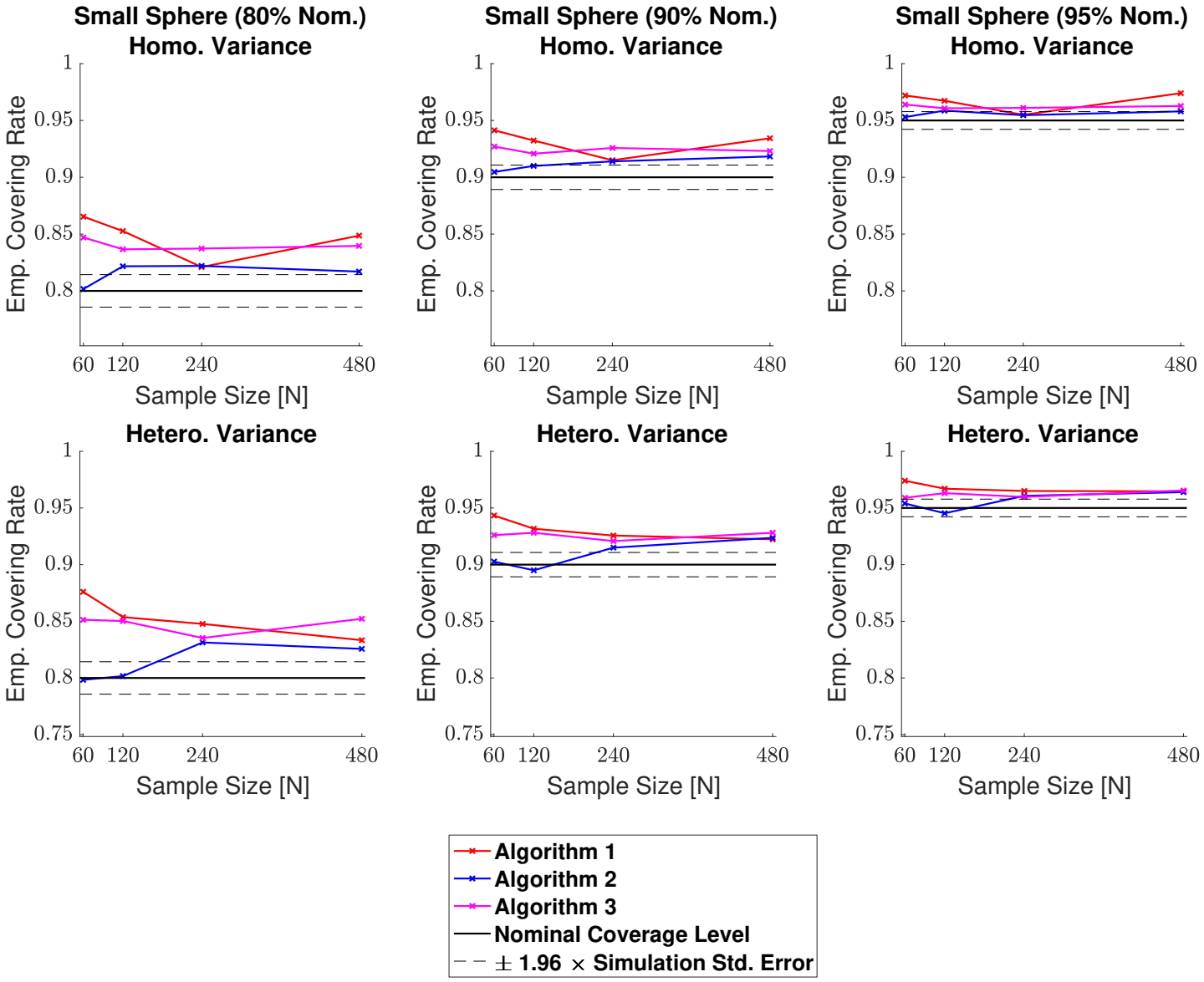


Figure 5.7: Coverage results for the small sphere signal type, with homogeneous (top row) and heterogeneous (bottom row) Gaussian noise structures. In general, empirical coverage remained above the nominal level across all simulations, and for the 95% confidence level (right plots), the results of all three methods fell close to the nominal target. All methods were robust as to whether the subject-level noise had homogeneous or heterogeneous variance structure. Because of this, there are minimal differences comparing the plots between both rows.

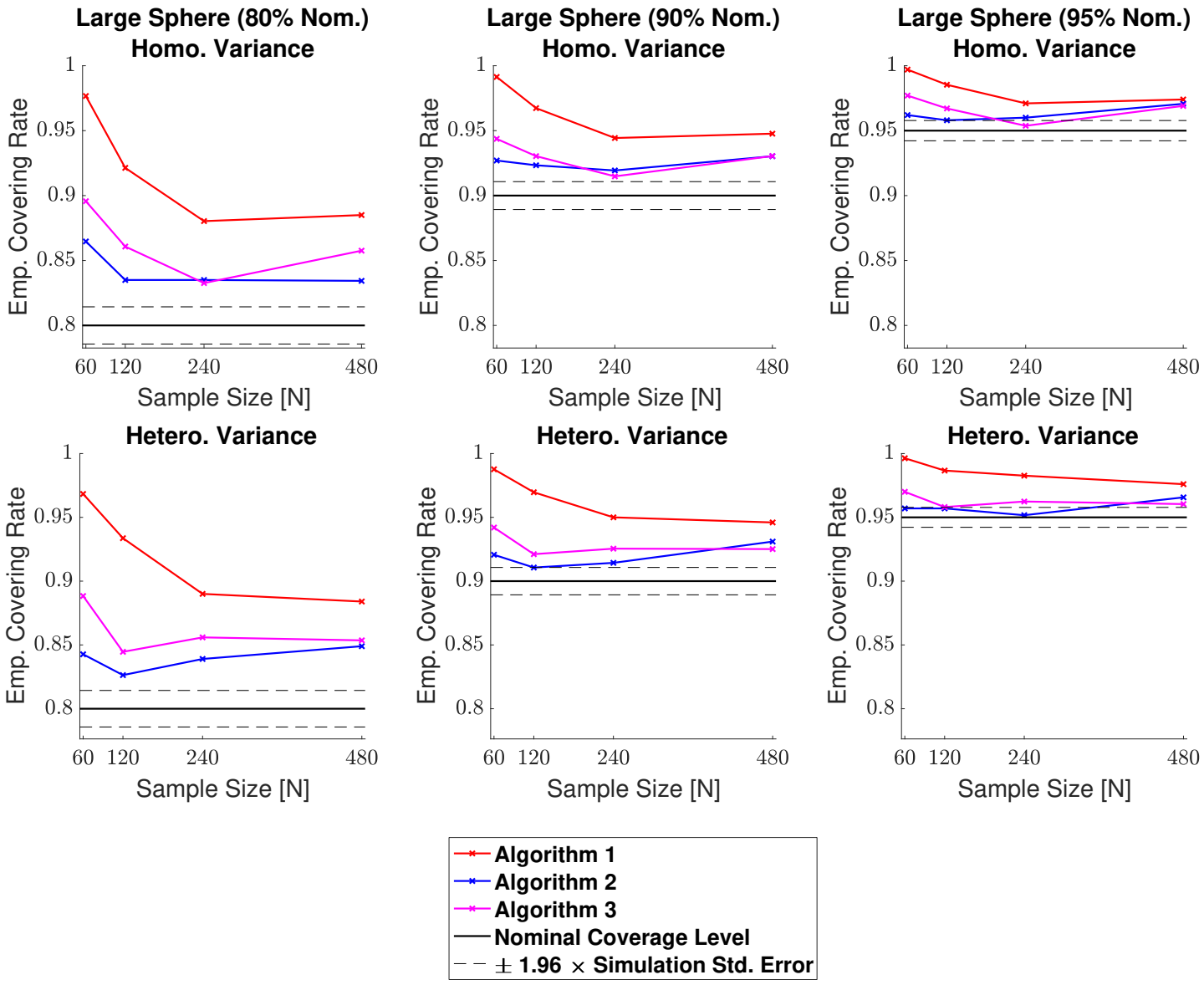


Figure 5.8: Coverage results for the large sphere signal type, with homogeneous (top row) and heterogeneous (bottom row) Gaussian noise structures. Compared with the small sphere results displayed in Fig. 5.8, empirical coverage results were higher for all three methods here. Algorithm 1. suffered from a particularly large degree of over-coverage for simulations with a small sample size. Coverage performance for Algorithm 2. and Algorithm 3. was closer in resemblance to the corresponding small sphere results, with Algorithm 2. performing slightly better. This suggests that both of these methods are fairly robust to changes in the boundary length.

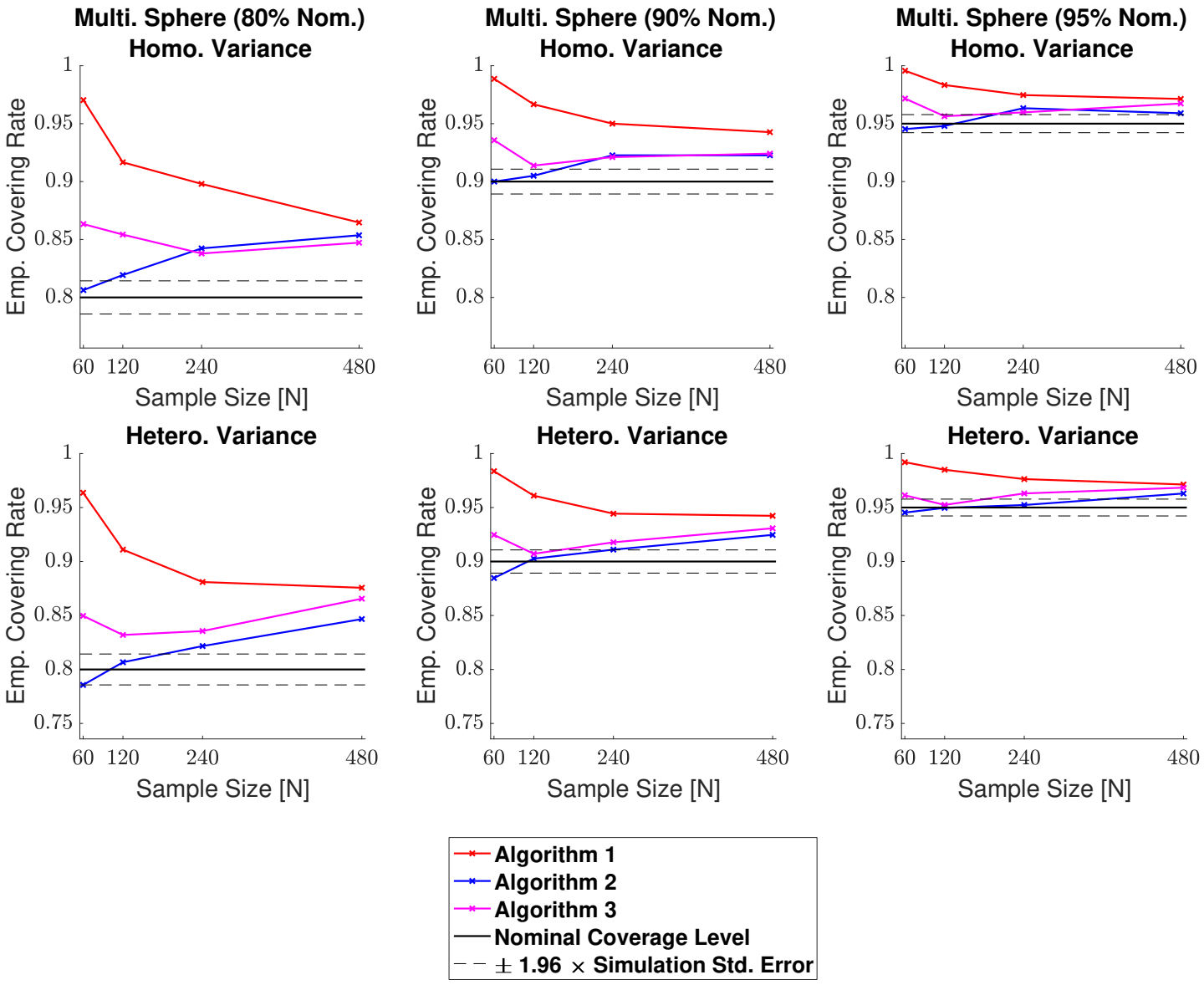


Figure 5.9: Coverage results for the multiple spheres signal type, with homogeneous (top row) and heterogeneous (bottom row) Gaussian noise structures. Algorithm 2. and Algorithm 3. both performed well, particularly for the 95% confidence level, where coverage levels remained in the vicinity of the 95% confidence interval of the nominal target. While Algorithm 2. was closer to the nominal target for $N = 60$, in some cases empirical coverage went slightly below the nominal level. In comparison to the other two methods, Algorithm 1. suffered from a large degree of over-coverage, which slightly improved as the sample size increased.

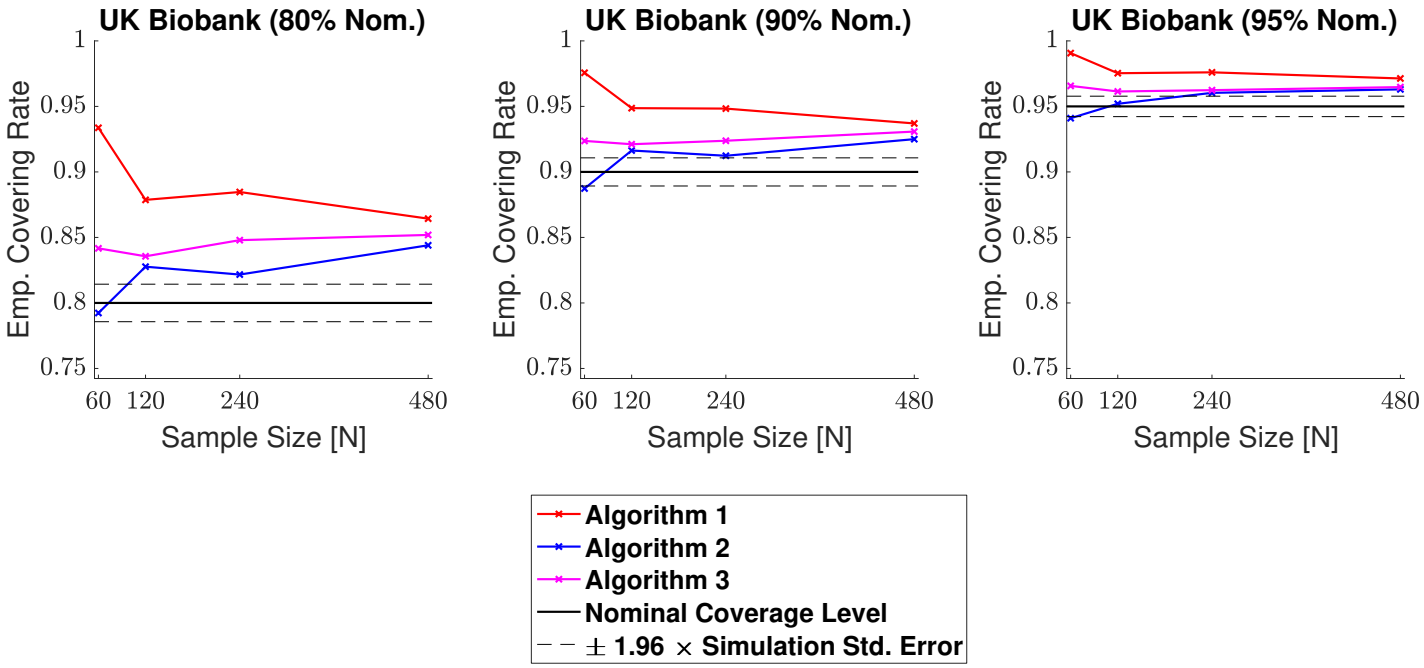


Figure 5.10: Coverage results for the UK Biobank signal type, where the full standard deviation image was used as the standard deviation of the subject-level noise fields. Covereage results here were similar to the results for the multiple spheres signal type shown in Fig. 5.9. Once again, both Algorithm 2. and Algorithm 3. performed well, with empirical coverage rates hovering above the nominal target for large sample sizes, while results for Algorithm 1. came further above the nominal level.

Comparing Figs. 5.7 and 5.8, we observed a slight deterioration in the performance of all three algorithms when moving from the small sphere signal type to the large sphere. In particular, results obtained from applying the three methods to the large sphere fell further above the nominal target relative to the small sphere. This was most severe for Algorithm 1., where differences between the two sets of results were larger than 10% for the 80% confidence level (Figs. 5.7 and 5.8, left-side plots). These differences were comparatively marginal for Algorithm 2. and Algorithm 3., where we observed only a slight increase in empirical coverage. This would suggest that both of these methods are fairly robust to changes in the boundary length.

Finally, the use of homogeneous or heterogeneous noise in the model once again had very little impact on the performance of all three algorithms. Nonetheless, for simulations with small sample sizes, a heterogeneous noise structure lead to a

slight decrease in the empirical coverage results for Algorithm 2. and Algorithm 3. (Figs. 5.7, 5.8 and 5.9, left-side plots).

5.2.3 Human Connectome Project

Cohen's d Confidence Sets obtained by applying Algorithm 3. to 80 subjects contrast data from the Human Connectome Project are shown in Fig. 5.11. CSs computed on the same data using Algorithm 1. and Algorithm 2. are displayed in Figs. C.1 and C.2 respectively. For each figure, we display the CSs obtained from applying the specified algorithm with three separate thresholds, $c = 0.5, 0.8,$ and 1.2. These three Cohen's d effect sizes were classified as medium, large, and very large in *Cohen (2013)*.

In the top plot of Fig. 5.11, the red upper CS localized brain regions within the frontal cortex commonly associated to working memory. This included areas of the superior frontal gyrus (left and right, all slices), middle frontal gyrus (left, coronal slice), paracingulate gyrus (left and right, axial slice) and insular cortex. Other brain areas encapsulated inside the upper CS were the angular gyrus (left and right, axial slice), cerebellum (left and right, sagittal slice) and precuneus (left and right, axial slice). For all these regions, the method identified clusters of voxels where we can assert with 95% confidence there was a Cohen's d effect size greater than 0.5.

By increasing the threshold to $c = 0.8$ (Fig. 5.11, middle plot), there was a shrinking of both the blue lower CSs and red upper CSs. Therefore, while we can confidently declare a medium effect size in all of the brain areas identified above, the quantity of voxels within each region that we can proclaim to have a large effect size is considerably smaller. In the case of the right cerebellar hemisphere (left, sagittal slice) and insular cortex, the upper CS vanished completely, indicating that the method did not locate any voxels in these regions where we can assert a Cohen's d effect size greater than 0.8.

For the largest threshold assessed, $c = 1.2$, the red upper CS was empty, and hence we can not assert any region of the brain to to have attained a very large effect

size. Notably, the yellow point estimate set contains a small but appreciable number of voxels, signifying that based on the data alone, these voxels were estimated to have a Cohen's d effect size greater than 1.2. Conversely, the large quantity of (grey backgroud) voxels lying outside the blue lower CS in Fig. 5.11 imply an effect size *less* than 1.2 across the vast majority of the brain.

In Fig. 5.12 the red upper CSs computed with Algorithm 3. are compared with the thresholded t-statistic map (green-yellow voxels) obtained from applying a one-sample t-test group-analysis to the 80 subjects contrast data, using a voxelwise FWE-corrected threshold of $p < 0.05$. This figure demonstrates the improved spatial specificity that can be provided with the Confidence Sets in comparison to the traditional approach. Specifically, while the thresholded statistic map contains one large cluster covering a sizeable portion of the parietal lobe across both brain hemispheres, the red upper CSs pinpoint precise areas in the precuneus and angular gyrus where a practically significant medium (or large) Cohen's d effect size can be inferred (Fig. 5.12, axial slices).

5.3 Discussion

5.3.1 Spatial Inference on Cohen's d Effect Size

To fully appreciate the outcomes of a neuroimaging study requires that information about the *magnitude* (as well as presence) of effects is reported at the end of an investigation. It is only with this knowledge that one can truly determine the practical relevance (and potential clinical importance) of any discoveries made during the analysis. In this work, we have presented three methods to create Confidence Sets for Cohen's d effect size maps, providing formal confidence statements on regions of the brain where the Cohen's d effect size has exceeded a specified activation threshold, alongside regions where the effect size has *not* surpassed this threshold. Both of these statements are made simultaneously across the entire brain, enabling researchers

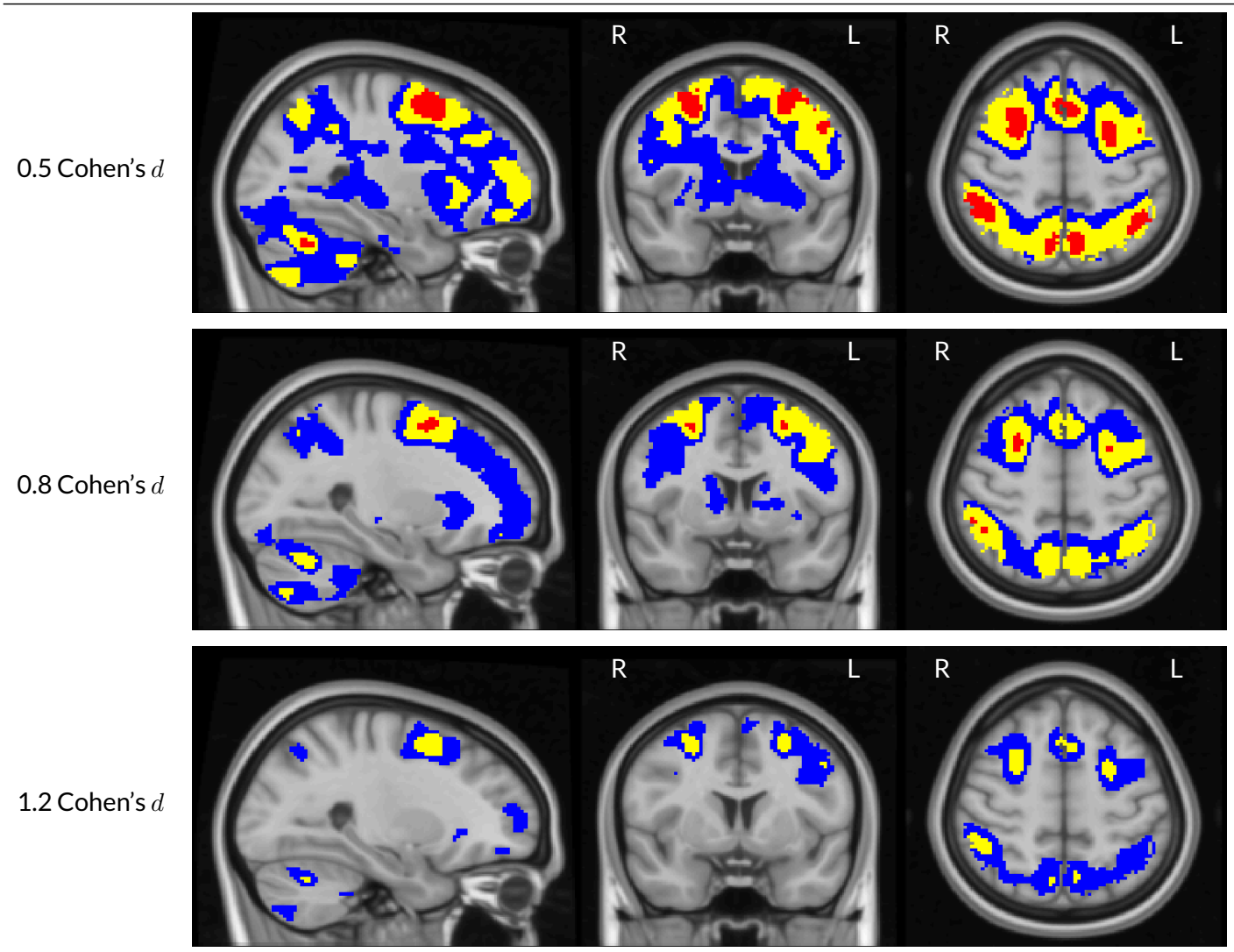


Figure 5.11: Slices views of the Cohen's d Confidence Sets obtained from applying Algorithm 3. to the HCP working memory task data, using three Cohen's d effect size thresholds, $c = 0.5, 0.8$ and 1.2. The upper CS $\hat{\mathcal{A}}_c^+$ is displayed in red, and the lower CS $\hat{\mathcal{A}}_c^-$ in blue. Yellow voxels represent the point estimate set $\hat{\mathcal{A}}_c$, the best guess from the data of voxels that have surpassed the Cohen's d threshold. The red upper CS has localized regions in the frontal gyrus, paracingulate gyrus, angular gyrus, cerebellum and precuneus which we can assert with 95% confidence have attained (at least) a 0.5 Cohen's d effect size.

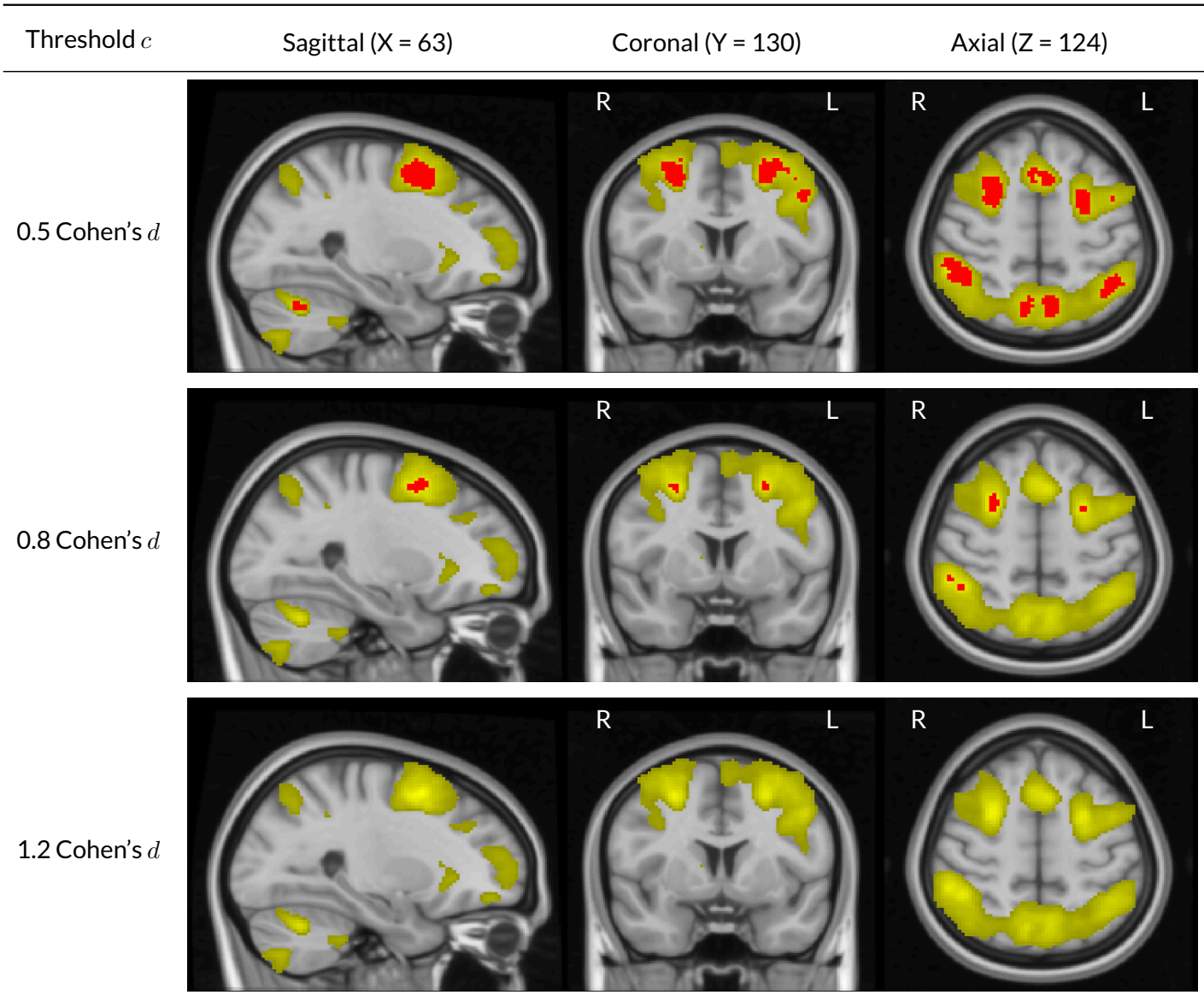


Figure 5.12: Comparing the upper Confidence Sets computed with Algorithm 3. on the HCP working memory task data (same slice views as Fig. 5.11) with the thresholded t-statistic results obtained by applying a traditional group-level one-sample t-test, voxelwise $p < 0.05$ FWE correction (green-yellow voxels). While the thresholded statistic map contains a single cluster covering a sizable portion of the parietal lobe across both hemispheres (axial slices), the upper CSs have localized the precise areas of the precuneus and anglur gyrus where we can confidently declare a Cohen's d effect size of at least 0.5. This demonstrates how the CSs can provide improved spatial specificity in determining regions with practically significant activation.

to pinpoint the precise regions where meaningful differences have occured as well as identifying areas that have not responded to the task. This is in contrast to the traditional statistical approach, where a p-value can only quantify the compatibility between the observed data and what would be expected under the null-hypothesis of no activation, and no information is provided about the effect size when a finding is deemed to be statistically significant. Since the p-value is confounded by the sample size of the study, it is not possible to implement a similar framework to obtain Confidence Sets for statistic images; as sample sizes increase p-values also become arbitrarily large, until ultimately there is enough statistical power to determine even the smallest of effects as statistically significant. On the other hand, as the Cohen's d CSs are computed using estimates of the effect strength, for larger samples we expect the red upper CS to closely approximate the population parameter $\mathcal{A}_c$ representing the set of voxels with a true population effect magnitude above a purposeful activation threshold c.

The use of Confidence Sets for inference on effect size may also help to alleviate issues associated with hypothesis testing for studies with lower statistical power. Specifically, for studies with small sample sizes it has been reported that applying traditional inference procedures can lead to spurious or irreproducible results with considerably inflated observed effect sizes (Poldrack et al., 2017; Cremers et al., 2017). In regards to the latter, this is often caused by a form of selection bias known as the 'winners curse' (Reddan et al., 2017; Button et al., 2013), whereby voxels whose observed effect size has exceeded their expected performance are intrinsically more likely to be determined as statistically significant. This becomes a problem as magnitudes are commonly reported *only* for significant voxels, a practice that leads to positively biased effect estimates. Our analysis results from the Human Connectome Project dataset exemplify how the CSs can help to negate this issue. In Fig. 5.11, the yellow point estimate 'best guess from the data' clusters identified a number of voxels with a Cohen's d effect size greater than 1.2 that were also included in the thresh-

olded statistic map obtained from applying a one-sample t-test, voxelwise $p < 0.05$ FWE correction (Fig. 5.12). However, by synthesizing information about the effect magnitude as well as the *reliability* of the estimate, the CSs presented in Fig. 5.11 affirm that there are in fact *no* voxels that can be confidently declared to have an effect size larger than 1.2. On the contrary, only a handful of brain regions were contained in the red upper CS asserting a Cohen's d effect size exceeding 0.5 for the HCP working memory task data.

In our previous effort we described a method to obtain CSs for unstandardized percentage BOLD change maps, rather than the Cohen's d images that have served as our main focus here. The use of Cohen's d instead of %BOLD is likely to be advantageous due to complications associated with the BOLD effect, and at a more rudimentary level, the difficulties involved in obtaining percentage BOLD change images. While all of the main fMRI software packages provide contrast of parameter estimate maps, each of the three most widely-used analysis packages (AFNI, FSL & SPM) scale the raw data differently; the parameter estimates are often given in arbitrary units which deviate between packages. Conversion to percentage BOLD change therefore requires a software-dependent normalization, where one must take into consideration how to appropriately scale the data, design matrix and analysis contrasts. While this can be cumbersome and prone to human error, conversion to Cohen's d is relatively simple. Due to the straightforward relationship between the Cohen's d effect size and the t-statistic $\left(\hat{d} = \frac{t}{\sqrt{N}}\right)$, users can easily generate Cohen's d images from the unthresholded t-statistic maps outputted by all the main neuroimaging packages. As discussed in our previous work, %BOLD effect sizes have also been shown to modulate according to acquisition parameters such as the scanner field strength or MRI pulse sequence, and inhomogenieties in vascularity between different brain regions can cause further variation in the BOLD response. For all of these reasons, the Cohen's d CS maps may also be more suitable for comparison between studies.

In this work, we have used classifications of the Cohen's d effect size as initially suggested in *Cohen (2013)*, describing 0.5 as a 'medium' effect, 0.8 as 'large', and 1.2 as 'very large'. While these benchmarks provide basic descriptors of effect size, in general we recommend that users take appropriate steps to contextualize what sort of magnitude constitutes as a meaningful finding in their own study. Users should factor in the aims of their investigation, the quality of the study and, if possible, the effect sizes reported in similar previous efforts before choosing a threshold. Obtaining the CSs for the Human Connectome Project contrast data in this work was computationally quick, with each analysis taking less than one minute for all three proposed algorithms. Therefore, one possible strategy is to evaluate a variety of different c's on pilot or historical data before fixing a value to use on a study of interest.

5.3.2 Three Algorithms for Cohen's d Confidence Sets

In this work, we have theoretically motivated three algorithms for obtaining Cohen's d CSs. Our simulation results in Sections 5.2.1 and 5.2.2 have demonstrated differences in the coverage performance for each of these algorithms. Across all sets of simulation results, empirical coverage for Algorithm 1. came above the nominal level, with particularly severe over-coverage for 3D simulations carried out on large synthetic signals when small sample sizes were used (Figs. 5.7, 5.8 and 5.9). The cause for such poor performance here is likely to be due to the variance term used to construct the CSs in Algorithm 1. Recalling the derivations in Section 5.0.2, the variance term $\sqrt{1 + \frac{\hat{d}^2(\boldsymbol{s})}{2}}$ used for Algorithm 1. was chosen as an estimator of the variance of the *limiting* Gaussian field $\mathcal{G}(\boldsymbol{s})$. Therefore, while we expect this term to correctly approximate the variance of the bootstrap approximating field asymptotically, our theory provides no indication about the accuracy of this term in small samples. The over-coverage seen in our simulation results suggests that this term overestimates the true variance of the approximating field when the sample size is low. While there was some improvement in Algorithm 1.'s results as N increased, even for the largest

sample size we analyzed, $N = 480$, empirical coverage for the other two methods was consistently closer to the nominal target level.

Algorithm 2. and Algorithm 3. performed well in all of our 2D and 3D simulations. For simulations using a 95% confidence level, the empirical coverage performance of these two methods was remarkably similar for $N \geq 120$ (in most cases, slightly above the nominal target). It is therefore difficult to conclude which method should be implemented in practice. For our 3D simulations (Figs. 5.7 to 5.10), Algorithm 2.'s empirical coverage results fell slightly closer to the nominal level in most cases. However, the results for Algorithm 3. were more robust to changes in sample size; for the UK Biobank and multiple spheres simulations (Figs. 5.9 and 5.10), the empirical coverage performance for Algorithm 3. was relatively uniform in terms of sample size, while Algorithm 2.'s coverage fell below the nominal level for $N = 60$. This may imply a further drop-off in performance for Algorithm 2. on data with sample sizes below 60, suggesting that Algorithm 3. could be favourable here.

From a theoretical standpoint, the variance-stabilizing transformation approach used in Algorithm 3. assumes that the observations are Gaussian, while this is somewhat relaxed for Algorithm 2., where the bootstrap is applied to estimate the standard deviation directly from the data. While this supports that Algorithm 2. may be preferable for non-Gaussian data, in our Human Connectome Project analyses the CSs maps obtained using both methods (Fig. 5.11 for Algorithm 3., Fig. C.2 for Algorithm 2.) were virtually identical, indicating that both methods could be equally effective for fMRI data with sample sizes on the order of the HCP.

CHAPTER 6

Conclusion and Future Work

In this thesis, we have focussed our attention on some of the key issues at the forefront of task-based functional magnetic resonance imaging (fMRI). Our work has been carried out at a pivotal moment in the field's history, during the emergence of population-size neuroimaging studies that are delivering functional data on an unprecedented scale. While fMRI has traditionally been a small data enterprise, where sample sizes of 20 to 30 subjects are common for a task-based study, datasets such as the UK Biobank and Human Connectome Project are now giving researchers the opportunity to analyze fMRI data acquired from tens of thousands of participants. Population neuroimaging projects promise to transform our understanding of brain function, and are already yielding rich results (Miller et al., 2016; David C. Van Essen, 2016). However, the arrival of these datasets have also made this a critical time to reassess the current analysis methods implemented for task-based fMRI, as well as presenting new questions about how to appropriately analyze data of such magnitude.

The challenges posed by the big data revolution have guided all aspects of this effort. With the plurality of tools and techniques being applied to analyze population-size datasets, there has been a growing apprehension within the field about the degree of analytic variability across neuroimaging results. In Chapter 3, we helped ad-

dress these concerns in regards to variation between fMRI analysis software. By reanalyzing three publicly available fMRI datasets in the three most popular analysis packages, we ultimately demonstrated the fragility of group-level fMRI results dependant on the software package chosen to analyze the data. Our main contribution here was the measurement of inter-software differences using a variety of quantitative methods; Dice coefficients, Bland-Altman plots and Euler Characteristic curves illuminated disparities between the final group-level statistic maps obtained with each analysis package in terms of the location, magnitude and topology of each software's activation profile.

Our findings that weak effects may not generalize across software have highlighted the need for further examination of pipeline-related variation. While here we focussed on the statistic maps obtained at the end of an analysis, a limitation of this approach was that it meant our comparisons reflected the net accumulation of differences across the *entire* pipeline, rather than diagnosing the precise procedures where the greatest variation between software transpired. In future work we plan to carry out additional analyses on the three datasets used in this effort, implementing a common preprocessing strategy before utilizing procedures from different packages in the remaining stages of the analysis. This will distinguish whether the largest sources of software-variability are during the preprocessing or statistical modelling of fMRI data, prompting a further assessment of individual procedures to find areas where the three packages can be harmonised.

Alternatively, on the basis that each analysis package represents a peer-reviewed and valid analytic strategy, another line of work may include the development of techniques to synthesize inconsistent findings. While meta-analysis methods are routinely used in fMRI to aggregate data from separate studies, our setting is unique since all results are obtained by analyzing a *single* source dataset within multiple software packages. Therefore, unlike a traditional meta-analysis, any inter-result variation is solely due to methodological differences between analysis software. As part

of future work, we look to expand on current approaches to develop a series of 'same-data meta-analysis' techniques. By accounting for the correlation between individual statistic maps obtained from numerous analyses of a single dataset, we hope to establish methods that can provide a calibrated, consensus result across software. These sort of techniques may benefit further efforts to combine data from replication studies, such as the *Neuroimaging Analysis Replication and Prediction Study* (Botvinik-Nezer et al., 2019), where a single fMRI dataset has been analyzed by multiple teams worldwide.

In the second part of this thesis, we shifted our attention to the methods carried out for fMRI inference. Once again, big data played a key role in motivating this work; the introduction of population fMRI datasets have shed light on the practical limitations of current statical inference methods, providing ample power such that traditional testing methods essentially declare whole-brain activation. To overcome this, we endeavoured on developing a method that bypassed statistical testing to make inference on the practical *effect sizes* of interest.

In Chapter 4, we extended on recent work by Sommerfeld, Sain, and Schwartzman (2018) (*SSS*) to compute Confidence Sets (CSs) on fMRI %BOLD change maps, providing confidence statements about brain regions where %BOLD effect sizes had exceeded, and fallen short of, a *non-zero* cluster-forming threshold. In this work, we discovered that the methods used to evaluate empirical coverage for the simulations carried out in *SSS* had biassed the results upwards. By developing a more accurate assessment procedure based on linear interpolation, we demonstrated that the approach for computing CSs described in *SSS* could suffer from under-coverage when applied to 3D synthetic data with moderate sample sizes. To remedy this, we adapted the method to incorporate a Wild t-Bootstrap and the use of Rademacher variables for multiplication of the bootstrapped residuals. Across a range of 2D and 3D simulations, we showed that empirical coverage performance for our modified procedure remained close to the nominal target level, suggesting that the method could be ef-

fectively applied to fMRI %BOLD data.

In Chapter 5, we made further advancements on the Confidence Sets for application to fMRI Cohen's d effect size images. By deriving the statistical characteristics of the Cohen's d estimator, we explored how the methods studied in Chapter 4 could be adapted to compute Cohen's d CSs. One of our main contributions here was the development of a transformation to normalize the distribution of the sample Cohen's d, motivating a procedure for obtaining Cohen's d CSs in the transformed domain. Altogether, we formulated three separate algorithms to compute Cohen's d CSs. By testing our methods on 2D and 3D Monte Carlo simulations, we found that two of these procedures performed particularly well. In the final part of this work, we demonstrated the Cohen's d CSs on Human Connectome Project working memory task-fMRI data, and by comparing the CSs with statistical results obtained using a traditional inference procedure, we showed how the CSs can provide an improved localization of meaningful effects.

While in this work we have focussed on CSs, recently an increased amount of attention has been given to developing methods that estimate simultaneous confidence *bands* for functional data (Degras, 2017; Telschow and Schwartzman, 2019). A future study may investigate how confidence bands could be applied to fMRI effect size maps, similar in vein to the work carried out here on Confidence Sets. Whereas the CSs localize brain regions to assert where effect sizes exceed (and fall short of) a threshold, the purpose of the confidence bands would be to envelope the activation across the *entire* brain. This would provide information about the maximum and minimum effect sizes in an image, that could then be used, for example, for determining an appropriate threshold to compute CSs. Finally, although Cohen's d is commonly applied in fMRI to estimate effect sizes following a t-test, partial R^2 is also used to assess group-level results when an F-test or ANOVA are carried out. A generalization of the Confidence Sets for application to fMRI partial R^2 maps may therefore also be desirable, to provide another layer for interpretable effect size inference.

APPENDIX A

Software Comparison Supplementary Material

A.1 Percentage BOLD change Maps

While each of the three software packages provide contrast of parameter estimate maps (going forth "contrast estimate"), the units of the analysis differs between software. We first review the issue of units for fMRI then how we addressed this for each software.

Raw fMRI data has arbitrary units, and a normalization step is required to produce effect estimates that are both comparable across subjects and give an interpretable magnitude of the BOLD effect. The final units of the contrast estimates depend on how the data, design matrix and contrast vectors are scaled.

Consider arbitrary first level data at a voxel $\boldsymbol{Y}$ (N-vector), fMRI design matrix $\boldsymbol{X}$ ($N \times P$ matrix), related as $\mathrm{E}(\boldsymbol{Y}) = \boldsymbol{X}\boldsymbol{\beta}$, where $\boldsymbol{\beta}$ (P-vector) are the regression coefficients and the effect of interest is $\boldsymbol{c}\hat{\boldsymbol{\beta}}$ where $\boldsymbol{c}$ (P-row-vector) is the contrast. The following scaling is needed to ensure interpretable contrasts of parameter estimates.

- **Data**: The data needs to be scaled so that 1 unit change corresponds to 1%,

$$\boldsymbol{Y}^* = \frac{100}{B}\boldsymbol{Y} \tag{A.1}$$

where B is an estimate of the mean or baseline.

- **Design**: The design needs to be scaled such that a unit change in a coefficient gives rise to a unit effect change in the fitted data,

$$\boldsymbol{X}^* = \frac{1}{h}\boldsymbol{X} \tag{A.2}$$

where h is the (assumed common) predictor baseline-to-peak (or baseline-to-plateau for long blocks).

- **Contrast**: The contrast needs to be scaled to preserve the units of the coefficients,

$$\boldsymbol{c}^* = \frac{1}{s}\boldsymbol{c}, \tag{A.3}$$

where s is the sum of the positive contrast elements (or, if all less than zero, minus the sum of the negative elements).

The ideally scaled data, design and contrast gives rise to the contrast estimate $\boldsymbol{c}^*\hat{\boldsymbol{\beta}}^*$ which we can relate to the arbitrarily scaled data:

$$\begin{aligned} \boldsymbol{c}^*\hat{\boldsymbol{\beta}}^* &= \boldsymbol{c}^*(\boldsymbol{X}^{*\top}\boldsymbol{X}^*)^{-1}\boldsymbol{X}^{*\top}\boldsymbol{Y}^* \\ &= \frac{100h}{Bp}\boldsymbol{c}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{Y} \\ &= \frac{100h}{Bp}\boldsymbol{c}\hat{\boldsymbol{\beta}}. \end{aligned} \tag{A.4}$$

This scaling is described in terms of first level models, but can be applied to a 2nd level contrast estimate if the values of h, B and p are the same for all subjects.

In AFNI, a localized (i.e. voxelwise) approach is taken to overcome this hurdle (Chen et al., 2017). By including the 'scale' block in our subject-level analysis scripts, each voxel's time series was normalized (in our notation, using a voxelwise B) to a voxelwise mean of 100 for all subjects. Scaling of the design matrix is conducted implicitly within AFNI, and as the sum of the contrast elements was 1 in all our analyses,

no contrast scaling was necessary. Because of this, the effect estimate maps obtained at the group-level in our AFNI analyses could be directly interpreted as percentage BOLD change maps.

In FSL and SPM a global (i.e. brain-wide) approach is taken for data scaling, so that the modelled data ($\boldsymbol{Y}$) has a typical mean value of $B = 100$ for SPM, and $B = 10,000$ for FSL. In practise, it is known that SPM can underestimate the global mean intensities of each subject, leading to a grand mean intensity *larger* than 100 (Nichols, 2012). To account for this, in SPM we calculated B empirically by computing the mean image of all subject-level functional maps and then finding the median value over subjects. We computed h directly for the block design (ds000109) and based on an isolated event for the event-related design (ds00001); in FSL this required creating dummy designs with high-pass filtering disabled. Contrast scaling was only required in SPM ($s = 3$ for ds000001's 3-run design, $s = 2$ for ds000109's 2-run design).

A.2 Partial R^2 Maps

Since ds000120 used a sine basis HRF, this study's general linear model contained multiple predictor variables which hampered the computation of a percent BOLD change measure. Instead we computed partial R^2 maps to assess the explained variance (relative to data variance not already explained by other terms) of the main effect of the saccade condition.

With ν_1 and ν_2 as the numerator and denominator degrees of freedom of the F-statistic, respectively, we used the relationship between R^2 and the F-statistic given by the identity:

$$F = \frac{R^2}{1 - R^2} \frac{\nu_2}{\nu_1}, \tag{A.5}$$

which solves for R^2 as

$$R^2 = 1 - \frac{1}{1 + \frac{\nu_1}{\nu_2} F}. \tag{A.6}$$

A.3 Supplementary Figures

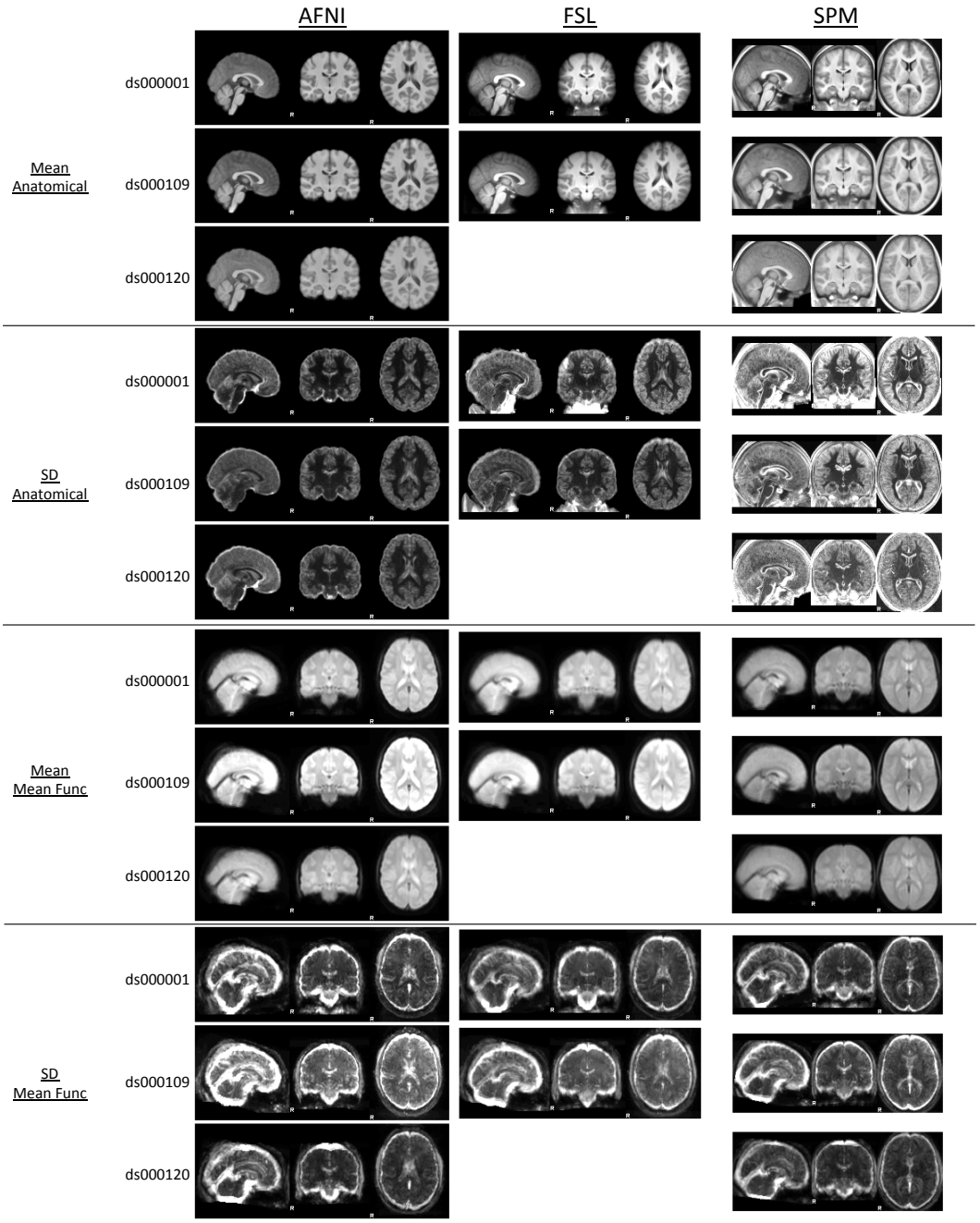


Figure A.1: Registration Quality Control: Mean and standard deviation of anatomical and mean functional images

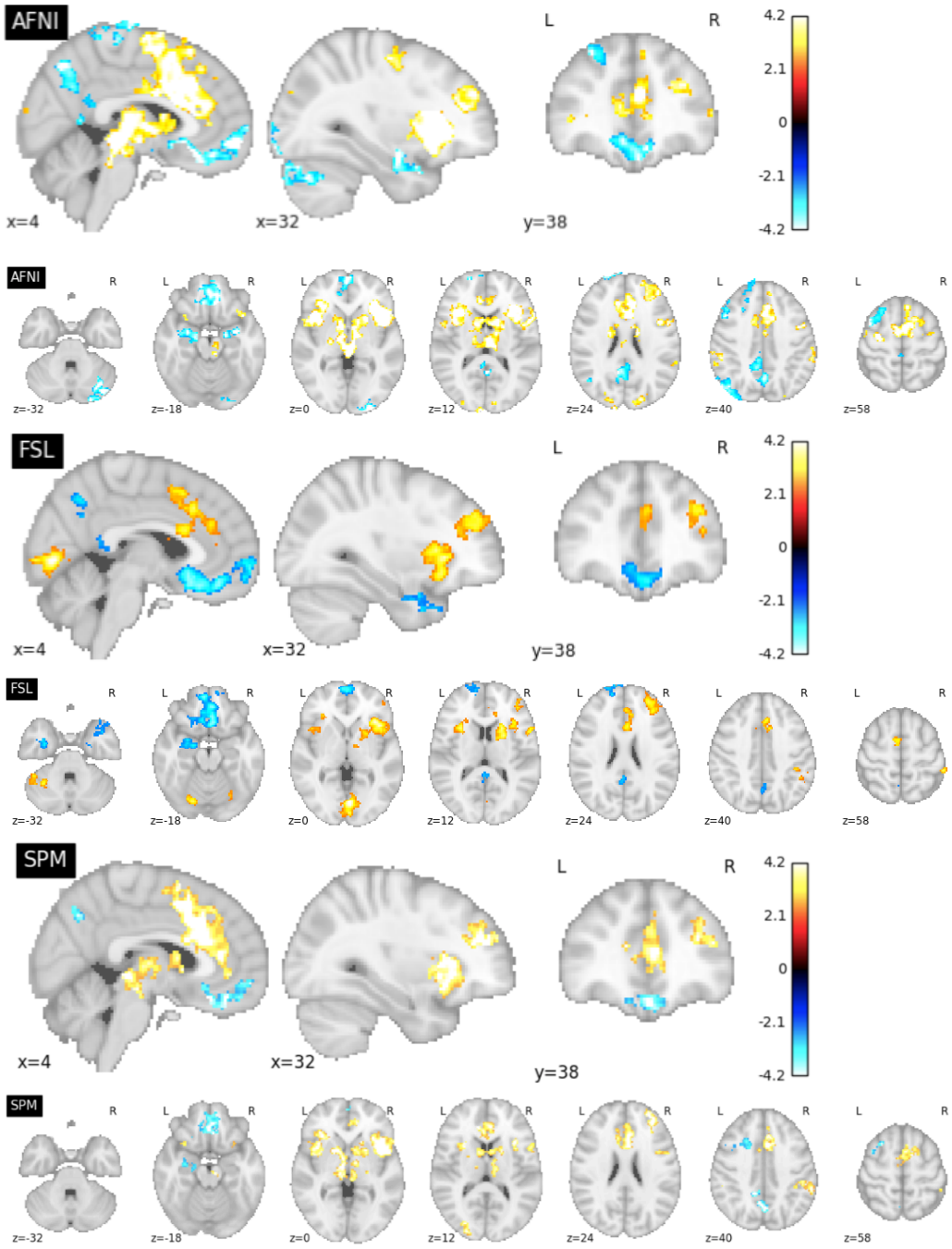


Figure A.2: ds000001 inter-software comparisons, 5% FWE clusterwise inference

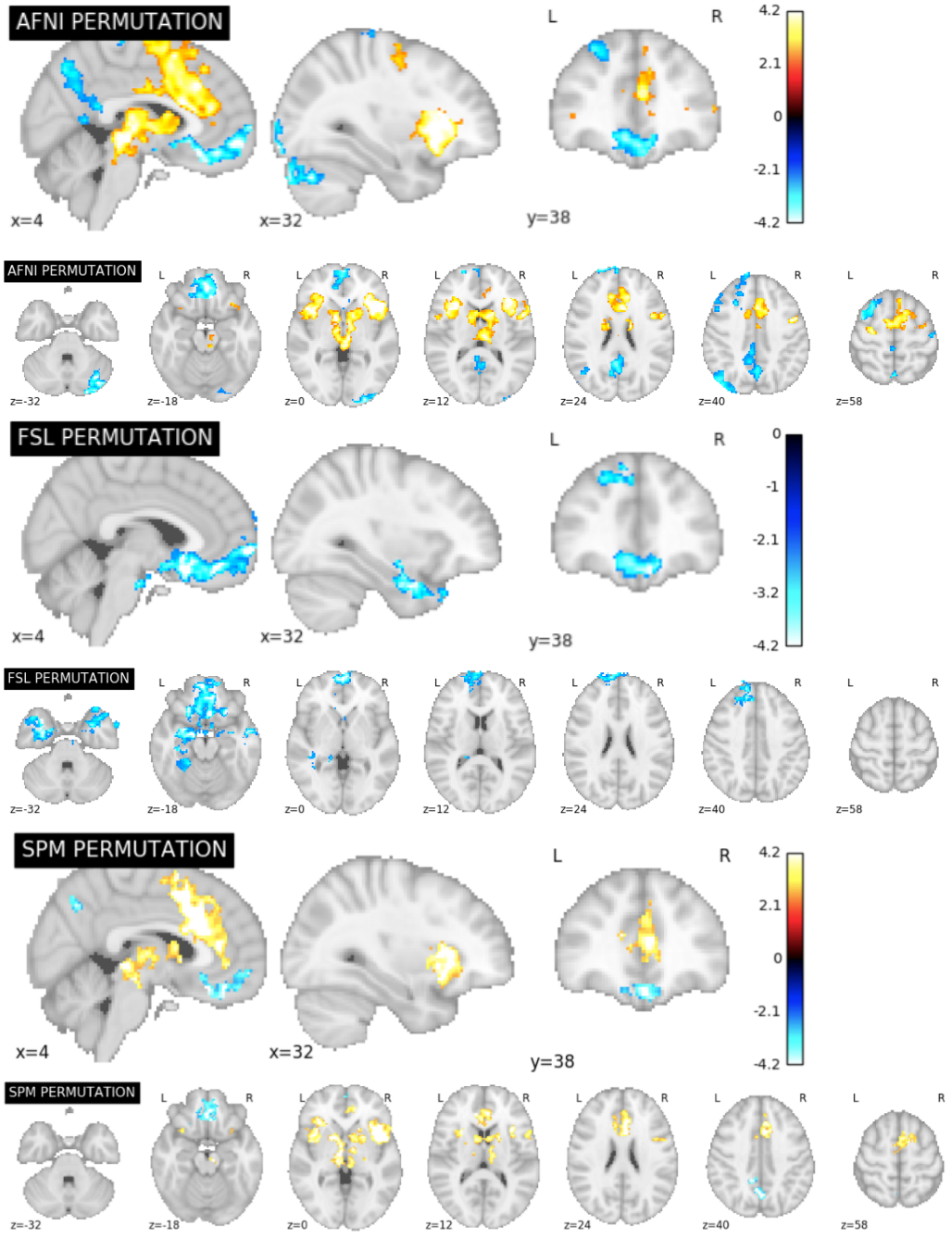


Figure A.3: ds000001 inter-software comparisons, 5% FWE clusterwise permutation inference

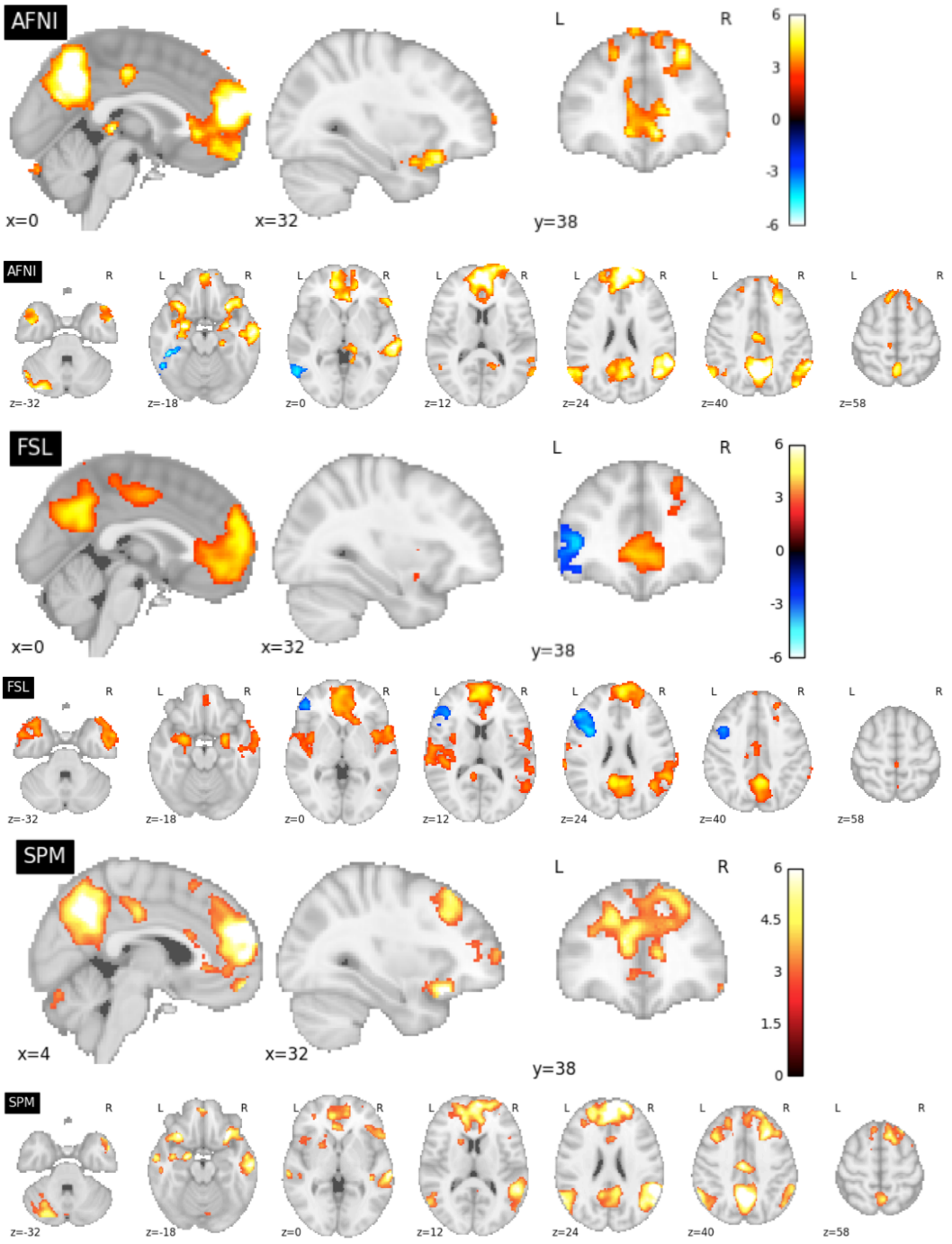


Figure A.4: ds000109 inter-software comparisons, 5% FWE clusterwise inference

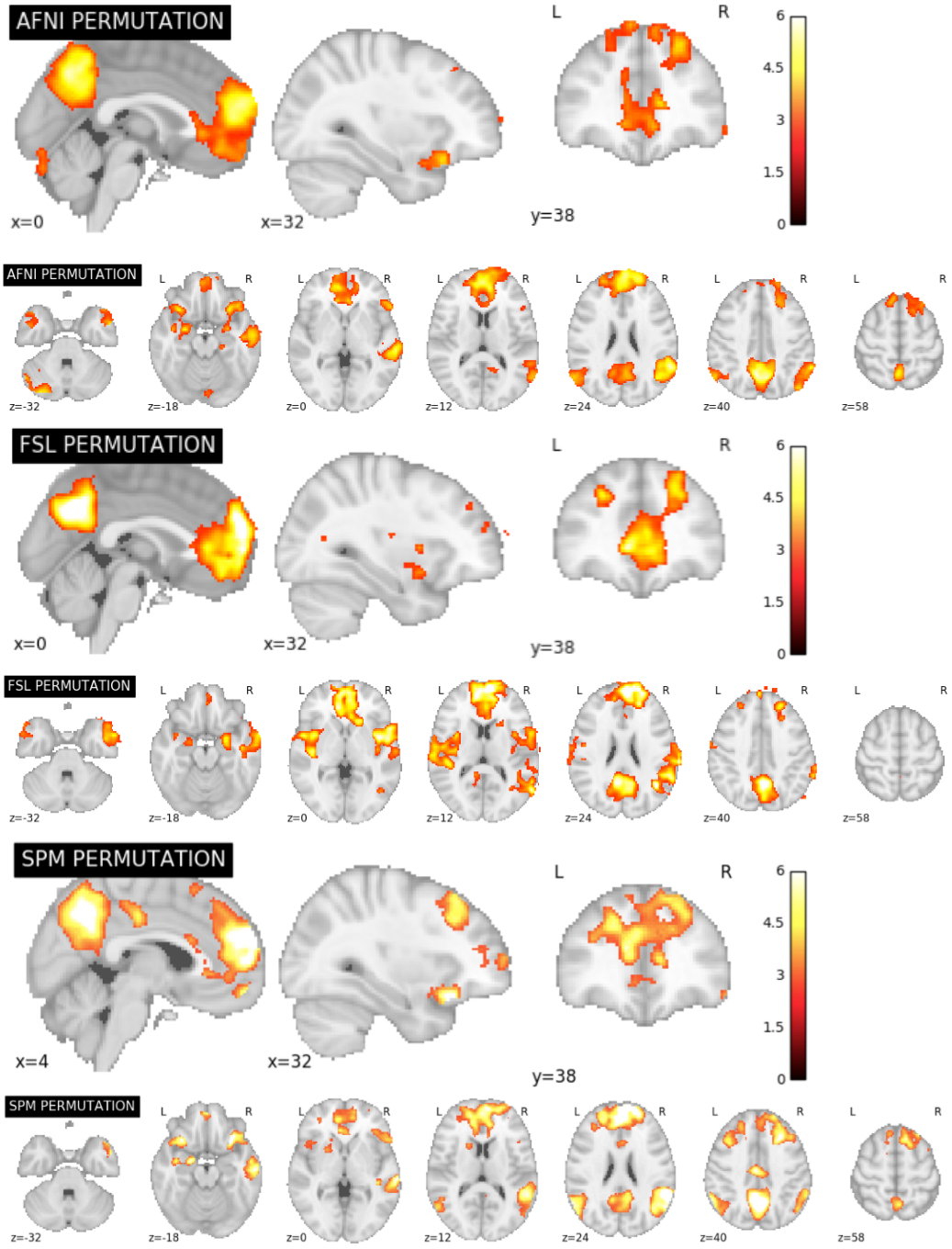


Figure A.5: ds000109 inter-software comparisons, 5% FWE clusterwise permutation inference

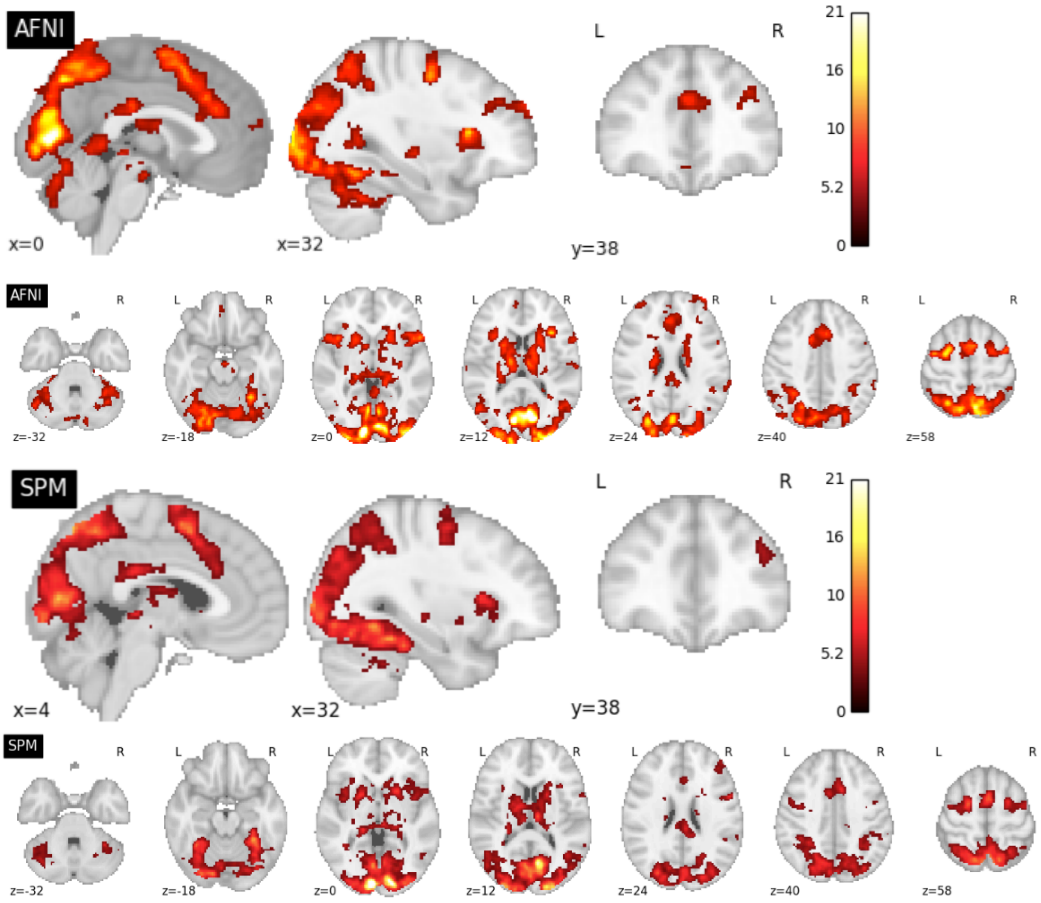


Figure A.6: ds000120 inter-software comparisons, 5% FWE clusterwise inference

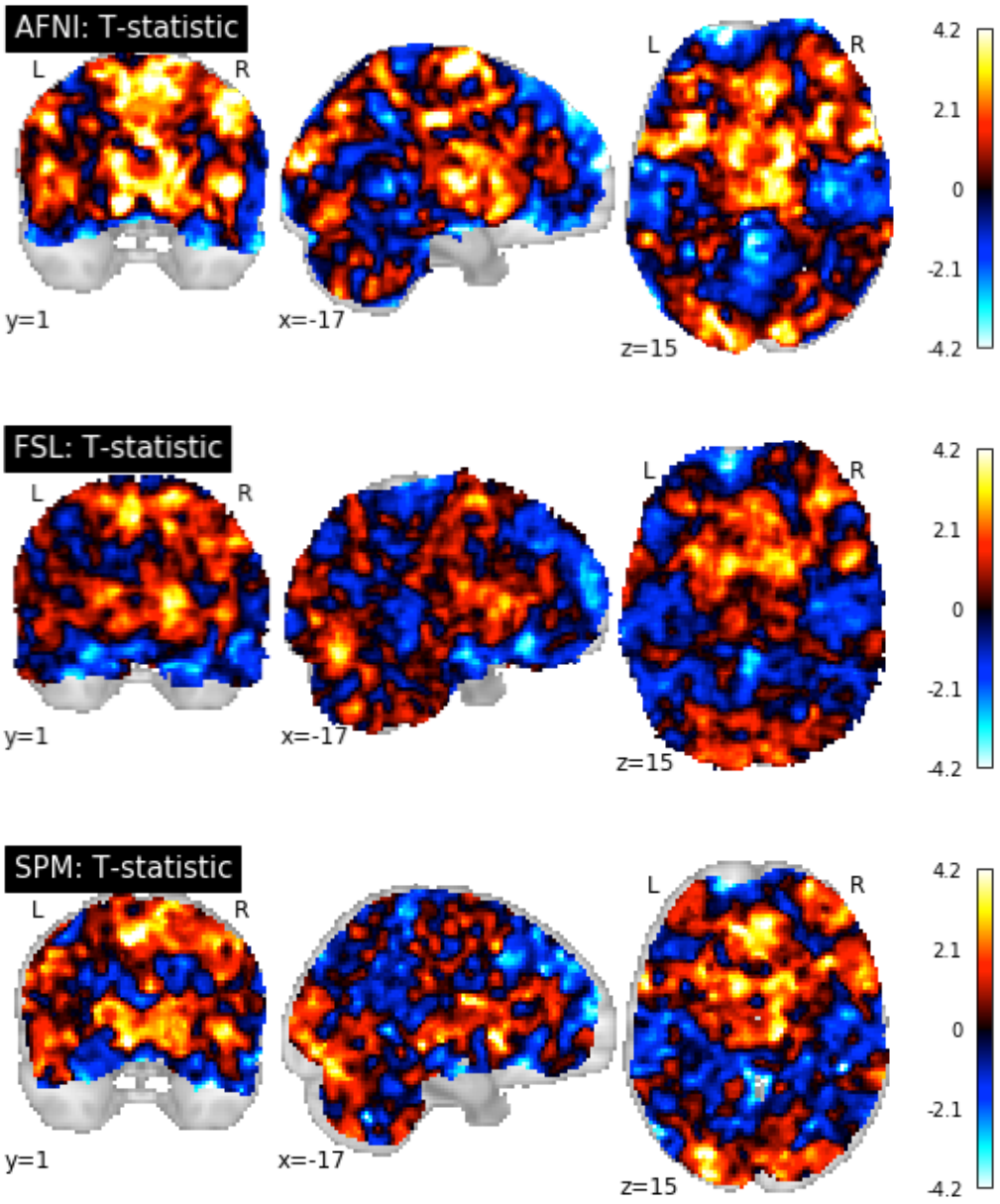


Figure A.7: ds000001 inter-software comparisons, t-statistic maps

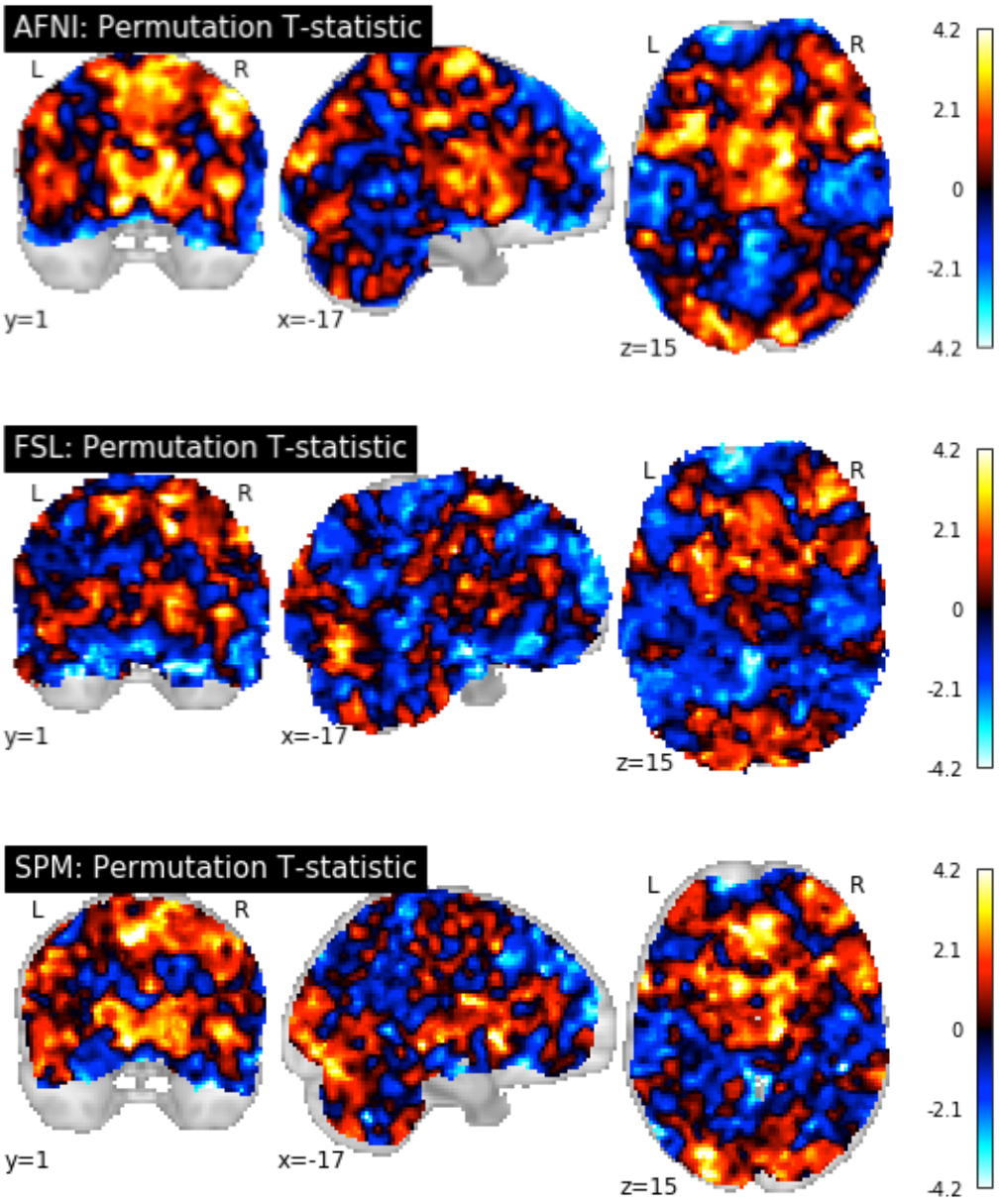


Figure A.8: ds000001 inter-software comparisons, t-statistic maps from permutation

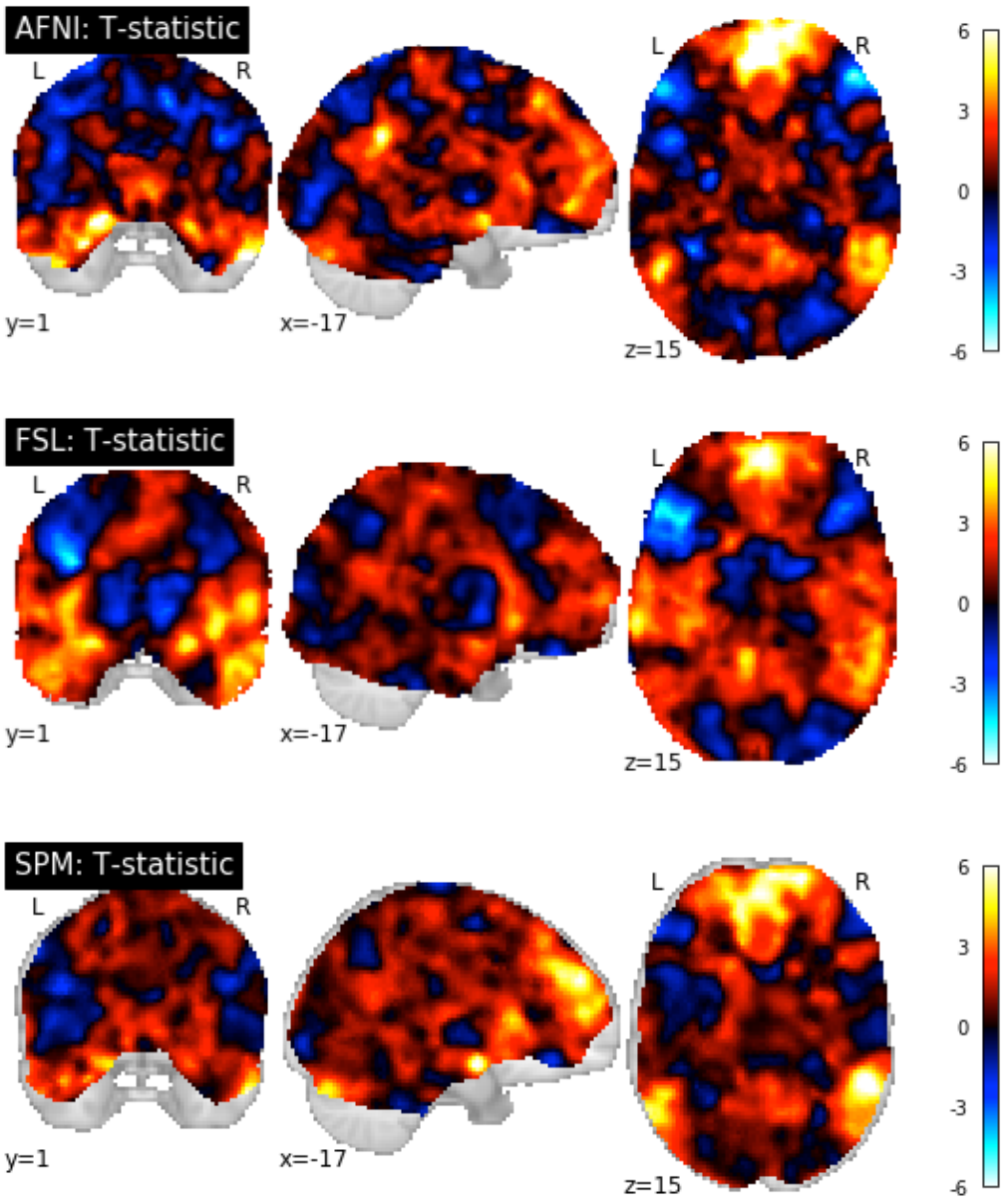


Figure A.9: ds000109 inter-software comparisons, t-statistic maps

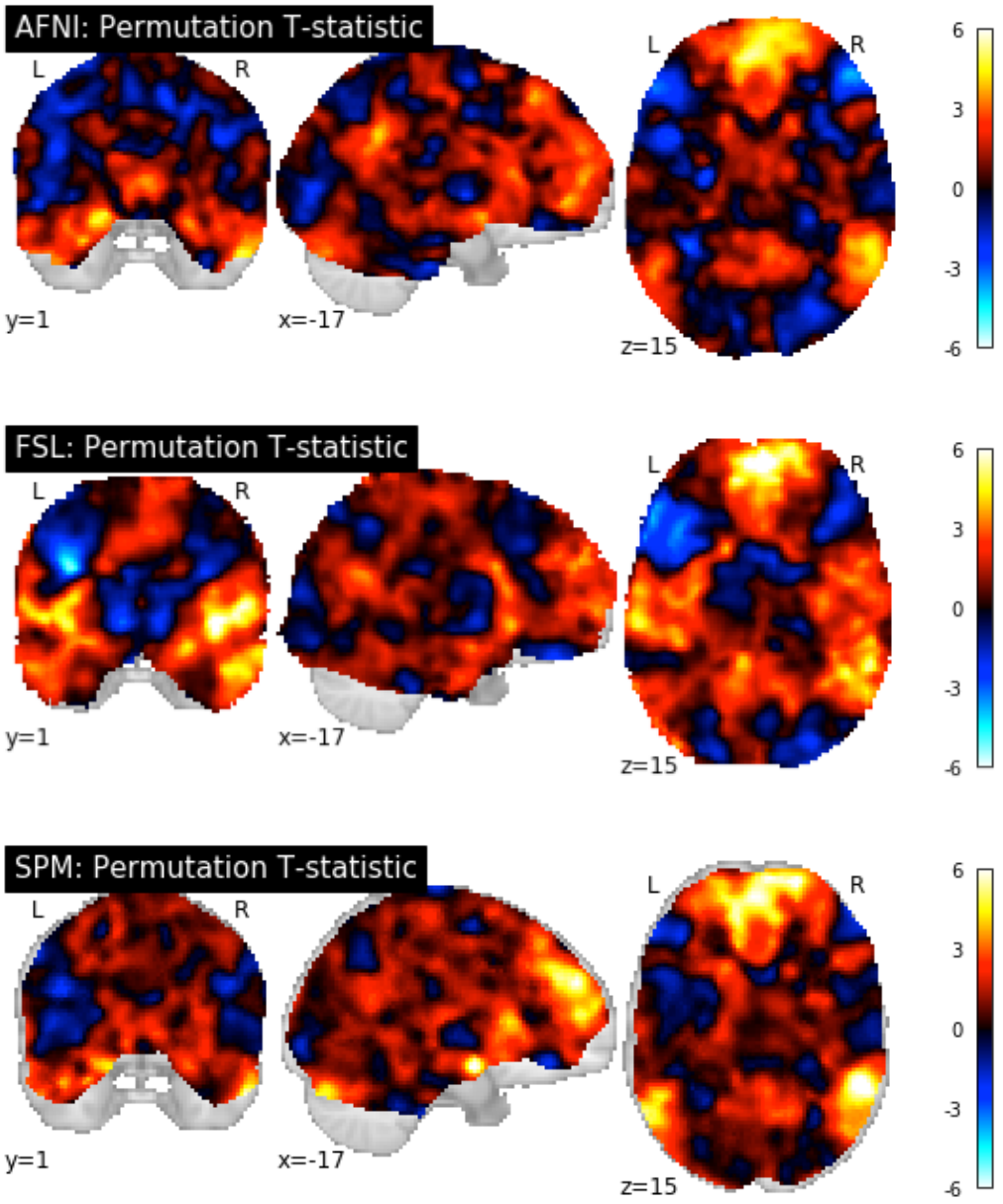


Figure A.10: ds000109 inter-software comparisons, t-statistic maps from permutation

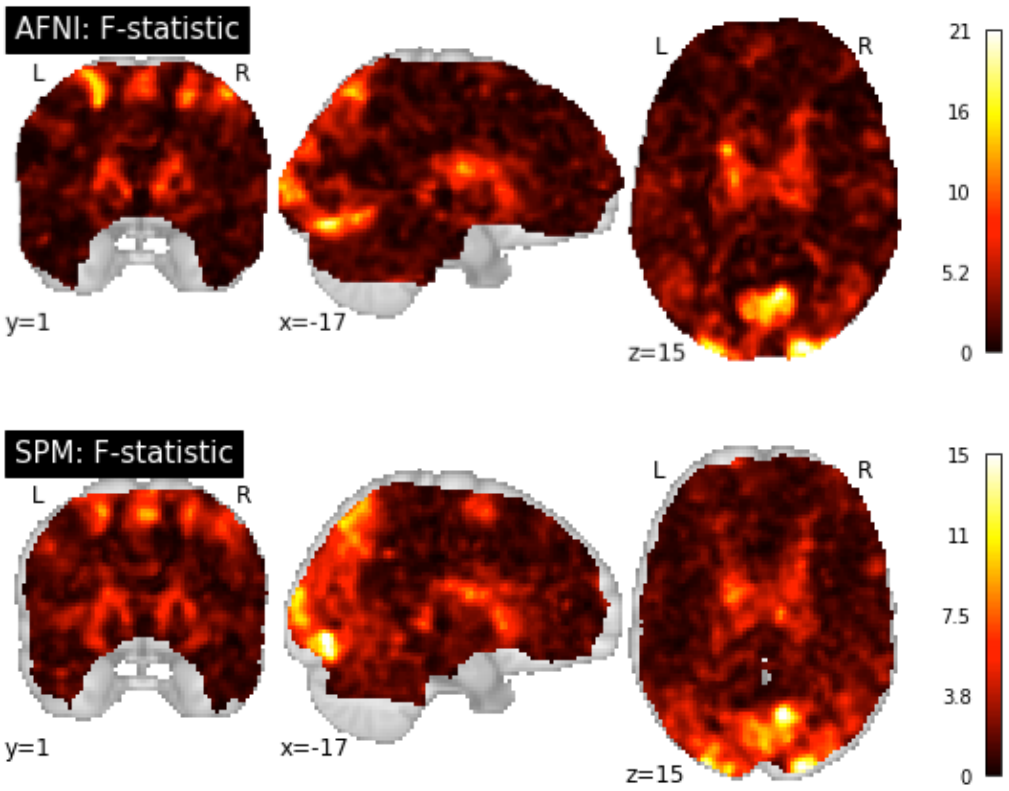


Figure A.11: ds0001020 inter-software comparisons, F-statistic maps

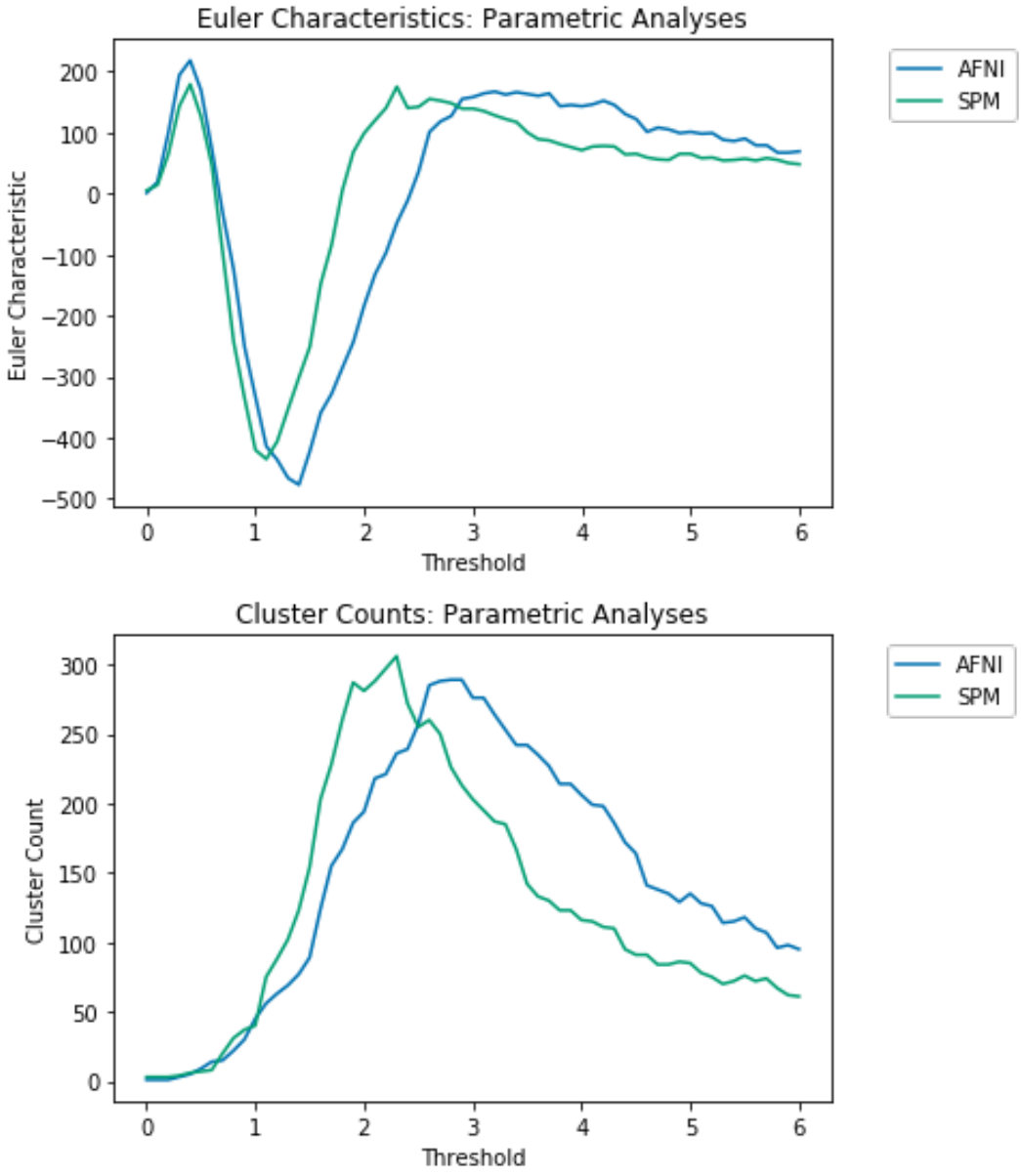


Figure A.12: ds000120 inter-software comparisons, Euler characeristic and cluster count curves for F-statistic maps

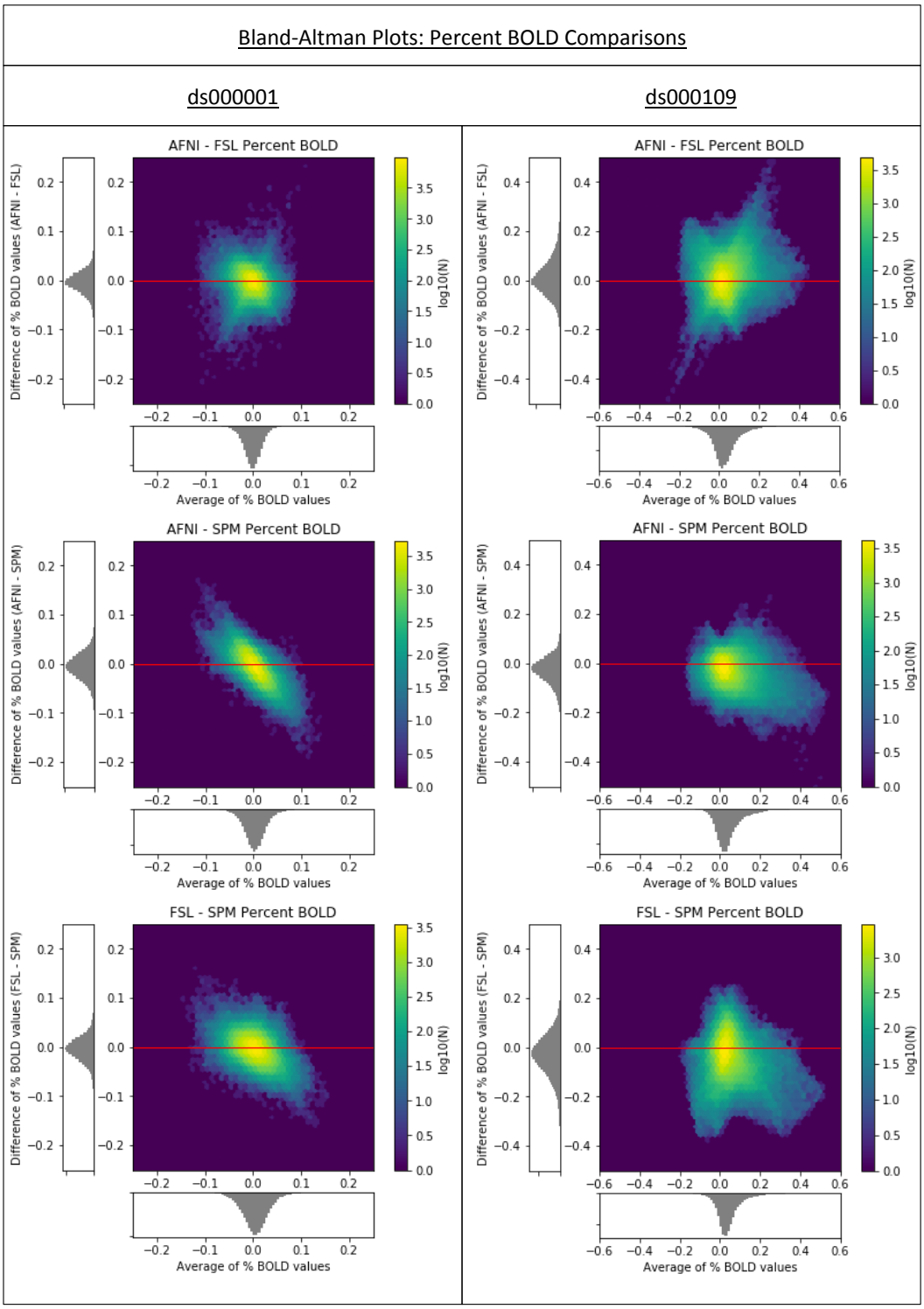


Figure A.13: Bland-Altman percent BOLD comparisons

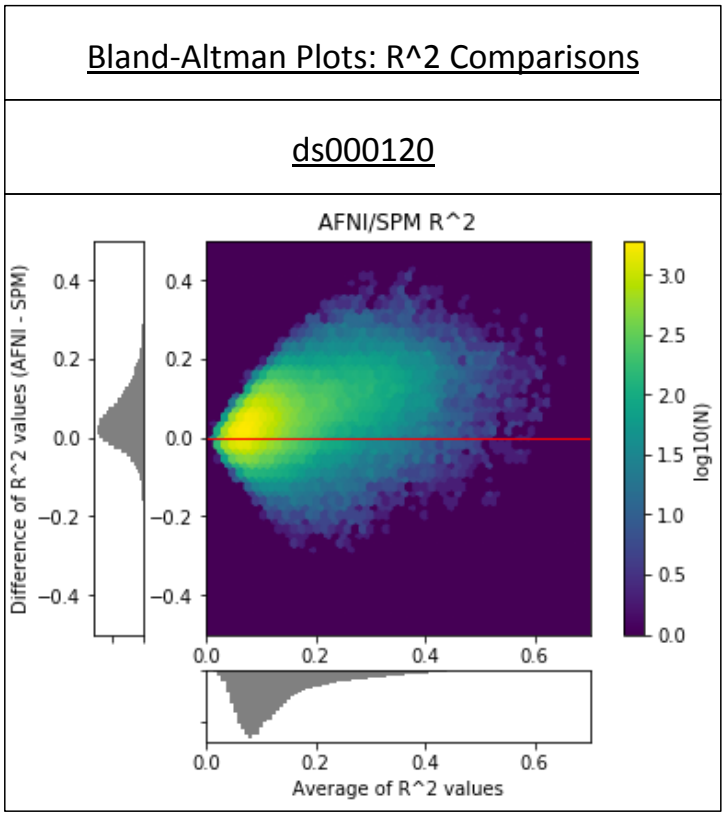


Figure A.14: ds000120 R^2 comparisons

APPENDIX B

%BOLD Confidence Sets Supplementary Material

B.1 Supplementary Human Connectome Project Results

B.2 Supplementary Tables

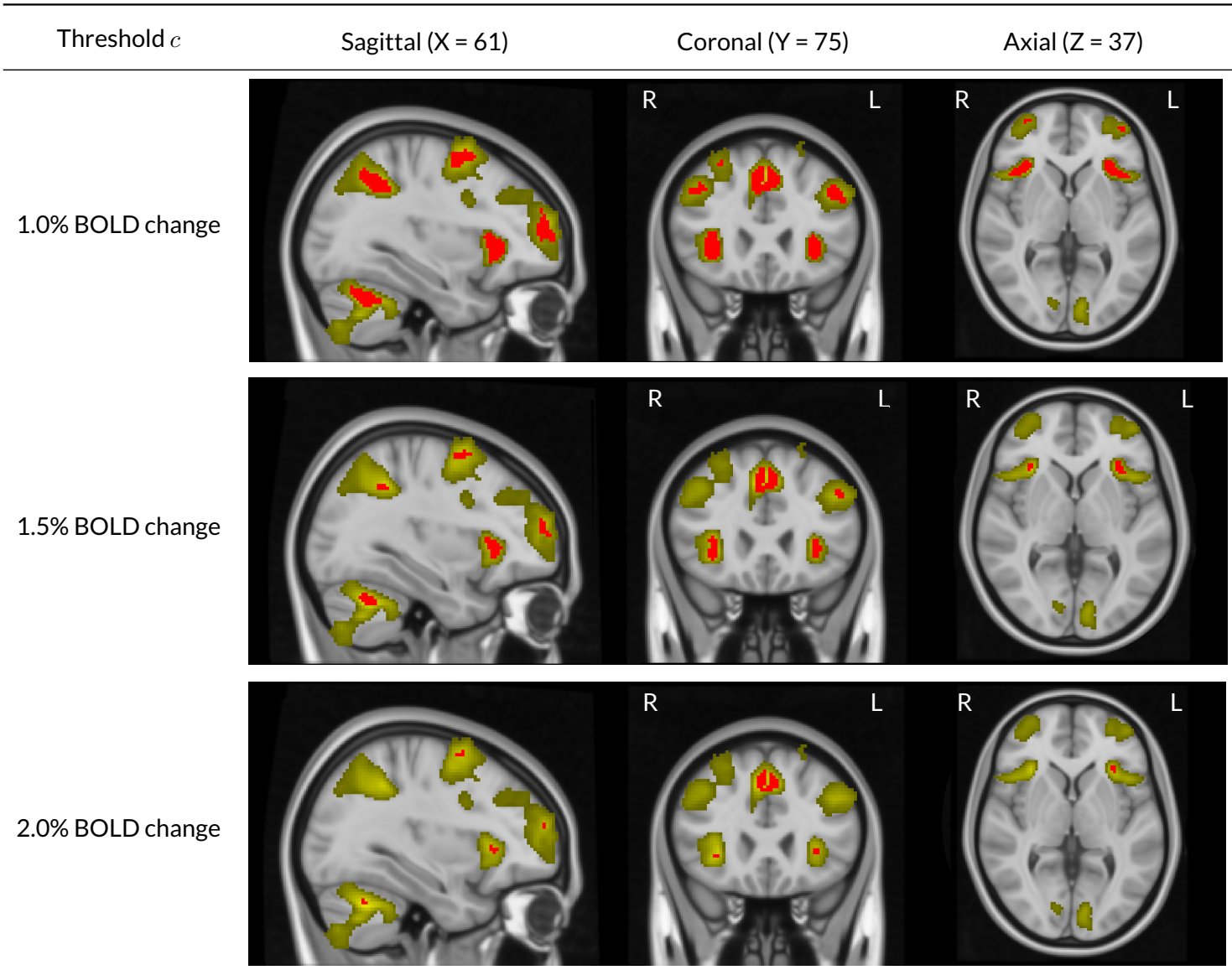


Figure B.1: Comparing the upper Confidences Sets for the HCP working memory task data (same slice views as Fig. 4.15) with the thresholded t-statistic results obtained by applying a traditional group-level one-sample t-test, voxelwise $p < 0.05$ FWE correction (green-yellow voxels). While over 25,000 voxels were determined as statistically significant with the standard inference method, less than 5,000 voxels were asserted to have at least a 1.0% BOLD change by the CSs. In particular, the two statistically significant clusters spanning the left and right side of the frontal lobe contained almost no voxels with a practical effect size of over 1.5% BOLD change.

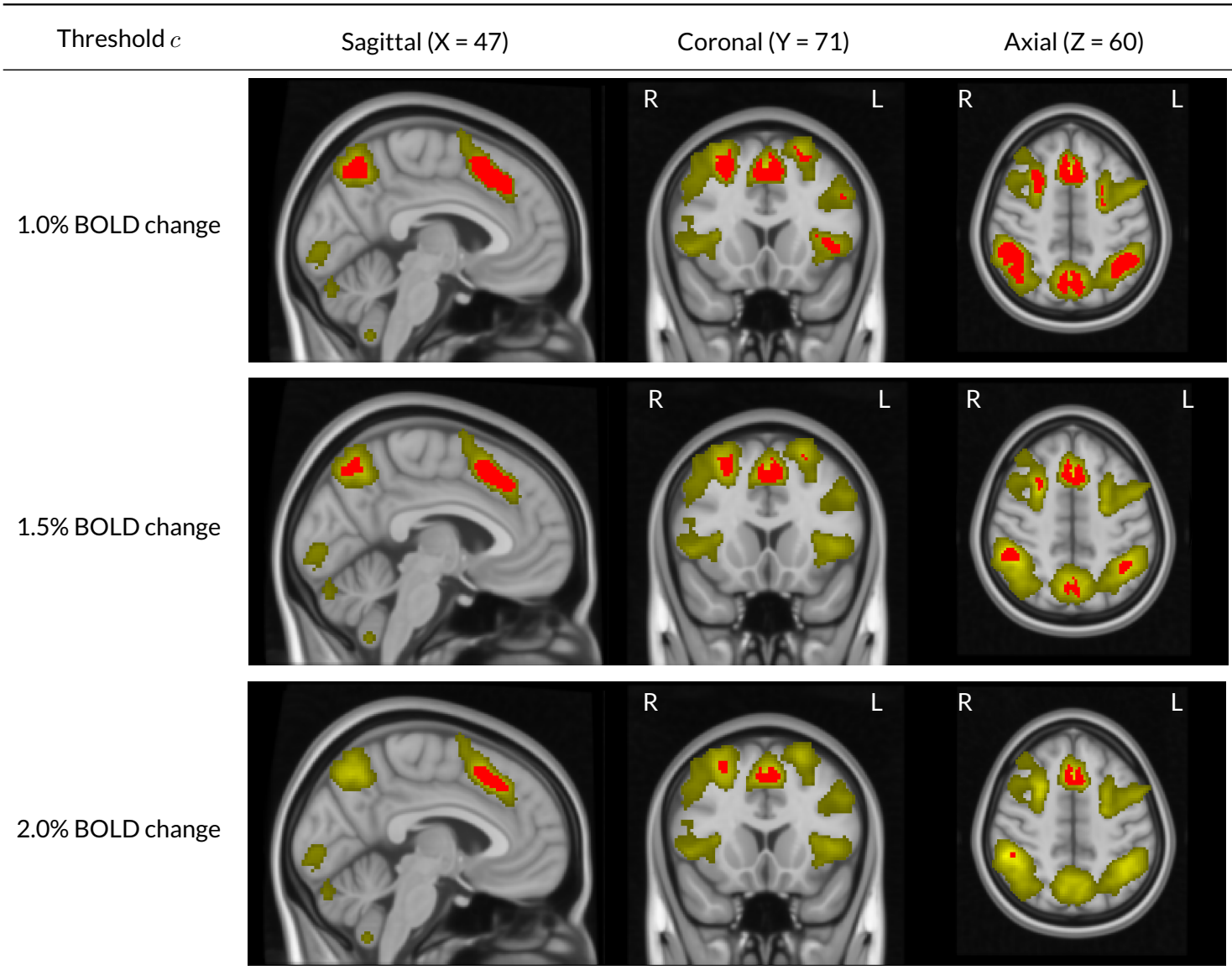


Figure B.2: Comparing the upper Confidences Sets for the HCP working memory task data (same slice views as Fig. 4.16) with the thresholded t-statistic results obtained by applying a traditional group-level one-sample t-test, voxelwise $p < 0.05$ FWE correction (green-yellow voxels). While one large statistically significant cluster covers the supramarginal gyrus, angular gyrus and precuneous, the CSs localize the precise areas with practically significant effect sizes within each of these regions.

Table B.1: Empirical coverage results for the 2D simulations using nominal (nom.) coverage levels $1-\alpha = 80\%, 90\%$ and 95%. Results are shown for applying the Wild t-Bootstrap method to the residual field along the estimated boundary $\partial\hat{\mathcal{A}}_{\mathrm{c}}$ (top) and the true boundary $\partial\mathcal{A}_{\mathrm{c}}$ (bottom).

	2D Signal 1. (Ramp)		2D Signal 2. (Circle)	
	Standard Dev 1.	**Standard Dev 2.**	**Standard Dev 1.**	**Standard Dev 2.**
$\partial\hat{\mathcal{A}}_{\mathrm{c}}$				
80% nom.				
$N = 60$	90.13% ± 0.54%	87.57% ± 0.60%	78.13% ± 0.75%	80.23% ± 0.73%
120	87.53% ± 0.60%	88.40% ± 0.58%	80.53% ± 0.72%	78.70% ± 0.75%
240	87.43% ± 0.61%	87.33% ± 0.61%	79.73% ± 0.73%	79.53% ± 0.74%
480	87.40% ± 0.61%	85.07% ± 0.65%	78.50% ± 0.75%	77.40% ± 0.76%
90% nom.				
$N = 60$	95.53% ± 0.38%	94.83% ± 0.40%	88.90% ± 0.57%	89.90% ± 0.55%
120	94.07% ± 0.43%	93.73% ± 0.44%	90.13% ± 0.54%	89.40% ± 0.56%
240	94.23% ± 0.43%	93.60% ± 0.45%	89.17% ± 0.57%	90.17% ± 0.54%
480	93.50% ± 0.45%	93.33% ± 0.46%	89.30% ± 0.56%	88.40% ± 0.58%
95% nom.				
$N = 60$	97.67% ± 0.28%	97.33% ± 0.29%	94.10% ± 0.43%	94.60% ± 0.41%
120	97.13% ± 0.30%	96.60% ± 0.33%	94.40% ± 0.42%	94.37% ± 0.42%
240	97.30% ± 0.30%	97.07% ± 0.31%	94.43% ± 0.42%	95.53% ± 0.38%
480	96.97% ± 0.31%	97.13% ± 0.30%	94.80% ± 0.41%	93.73% ± 0.44%
$\partial\mathcal{A}_{\mathrm{c}}$				
80% nom.				
$N = 60$	60.27% ± 0.89%	57.30% ± 0.90%	78.17% ± 0.75%	80.23% ± 0.73%
120	66.03% ± 0.86%	68.30% ± 0.85%	80.53% ± 0.72%	78.67% ± 0.75%
240	71.10% ± 0.83%	72.23% ± 0.82%	79.83% ± 0.73%	79.57% ± 0.74%
480	76.27% ± 0.78%	76.17% ± 0.78%	78.57% ± 0.75%	77.40% ± 0.76%
90% nom.				
$N = 60$	78.47% ± 0.75%	76.60% ± 0.77%	88.97% ± 0.57%	90.00% ± 0.55%
120	81.67% ± 0.71%	83.40% ± 0.68%	90.20% ± 0.54%	89.43% ± 0.56%
240	85.20% ± 0.65%	85.83% ± 0.64%	89.17% ± 0.57%	90.17% ± 0.54%
480	88.50% ± 0.58%	87.23% ± 0.61%	89.27% ± 0.57%	88.43% ± 0.58%
95% nom.				
$N = 60$	88.97% ± 0.57%	87.27% ± 0.61%	94.17% ± 0.43%	94.57% ± 0.41%
120	89.87% ± 0.55%	90.67% ± 0.53%	94.47% ± 0.42%	94.30% ± 0.42%
240	92.07% ± 0.49%	92.47% ± 0.48%	94.40% ± 0.42%	95.50% ± 0.39%
480	94.23% ± 0.43%	94.10% ± 0.43 %	94.87% ± 0.40%	93.73% ± 0.44%

Table B.2: Empirical coverage results for the 3D simulations using nominal (nom.) coverage levels $1-\alpha = 80\%, 90\%$ and 95%. Results are shown for applying the Wild t-Bootstrap method to the residual field along the estimated boundary $\partial\hat{\mathcal{A}}_{\mathrm{c}}$ (top) and the true boundary $\partial\mathcal{A}_{\mathrm{c}}$ (bottom).

	3D Signal 1. (Small Sphere)		**3D Signal 2. (Large Sphere)**	
	Standard Dev 1.	**Standard Dev 2.**	**Standard Dev 1.**	**Standard Dev 2.**
$\partial\hat{\mathcal{A}}_{\mathrm{c}}$				
80% nom.				
$N = 60$	$83.40\% \pm 0.68\%$	$83.77\% \pm 0.67\%$	$85.10\% \pm 0.65\%$	$85.73\% \pm 0.64\%$
120	$83.67\% \pm 0.67\%$	$84.03\% \pm 0.67\%$	$85.87\% \pm 0.64\%$	$85.23\% \pm 0.65\%$
240	$84.03\% \pm 0.67\%$	$83.77\% \pm 0.67\%$	$85.23\% \pm 0.65\%$	$85.40\% \pm 0.64\%$
480	$85.03\% \pm 0.65\%$	$82.20\% \pm 0.70\%$	$87.67\% \pm 0.60\%$	$85.30\% \pm 0.65\%$
90% nom.				
$N = 60$	$92.30\% \pm 0.49\%$	$92.87\% \pm 0.47\%$	$92.40\% \pm 0.48\%$	$93.47\% \pm 0.45\%$
120	$92.07\% \pm 0.49\%$	$91.27\% \pm 0.52\%$	$93.00\% \pm 0.47\%$	$93.50\% \pm 0.45\%$
240	$92.33\% \pm 0.49\%$	$92.87\% \pm 0.47\%$	$93.30\% \pm 0.46\%$	$92.90\% \pm 0.47\%$
480	$93.03\% \pm 0.46\%$	$91.53\% \pm 0.51\%$	$93.50\% \pm 0.45\%$	$93.47\% \pm 0.45\%$
95% nom.				
$N = 60$	$96.87\% \pm 0.32\%$	$96.83\% \pm 0.32\%$	$96.40\% \pm 0.34\%$	$96.70\% \pm 0.33\%$
120	$96.07\% \pm 0.35\%$	$95.60\% \pm 0.37\%$	$96.97\% \pm 0.31\%$	$97.10\% \pm 0.31\%$
240	$96.20\% \pm 0.35\%$	$96.83\% \pm 0.32\%$	$96.23\% \pm 0.35\%$	$96.90\% \pm 0.32\%$
480	$96.30\% \pm 0.34\%$	$96.13\% \pm 0.35\%$	$96.83\% \pm 0.32\%$	$93.80\% \pm 0.44\%$
$\partial\mathcal{A}_{\mathrm{c}}$				
80% nom.				
$N = 60$	$83.60\% \pm 0.68\%$	$83.90\% \pm 0.67\%$	$85.20\% \pm 0.65\%$	$85.80\% \pm 0.64\%$
120	$83.80\% \pm 0.67\%$	$83.93\% \pm 0.67\%$	$85.90\% \pm 0.64\%$	$85.23\% \pm 0.65\%$
240	$84.03\% \pm 0.67\%$	$83.90\% \pm 0.67\%$	$85.27\% \pm 0.65\%$	$85.40\% \pm 0.64\%$
480	$85.03\% \pm 0.65\%$	$82.27\% \pm 0.70\%$	$87.73\% \pm 0.60\%$	$85.37\% \pm 0.65\%$
90% nom.				
$N = 60$	$92.43\% \pm 0.48\%$	$92.90\% \pm 0.47\%$	$92.37\% \pm 0.48\%$	$93.40\% \pm 0.45\%$
120	$91.97\% \pm 0.50\%$	$91.43\% \pm 0.51\%$	$92.97\% \pm 0.47\%$	$93.60\% \pm 0.45\%$
240	$92.37\% \pm 0.48\%$	$92.90\% \pm 0.47\%$	$93.33\% \pm 0.46\%$	$92.90\% \pm 0.47\%$
480	$93.03\% \pm 0.46\%$	$91.40\% \pm 0.51\%$	$93.57\% \pm 0.45\%$	$93.47\% \pm 045\%$
95% nom.				
$N = 60$	$96.87\% \pm 0.32\%$	$96.93\% \pm 0.31\%$	$96.37\% \pm 0.34\%$	$96.70\% \pm 0.33\%$
120	$96.07\% \pm 0.35\%$	$95.53\% \pm 0.38\%$	$96.97\% \pm 0.31\%$	$97.13\% \pm 0.30\%$
240	$96.17\% \pm 0.35\%$	$96.93\% \pm 0.31\%$	$96.23\% \pm 0.35\%$	$96.80\% \pm 0.32\%$
480	$96.33\% \pm 0.34\%$	$96.13\% \pm 0.35\%$	$96.77\% \pm 0.32\%$	$96.80\% \pm 0.32\%$

Table 2. (continued)

	3D Signal 3. (Multiple Spheres)		3D Signal 4. (UK Biobank)
	Standard Dev 1.	**Standard Dev 2.**	**UK Biobank SD**
$\partial\hat{\mathcal{A}}_{\mathrm{c}}$			
80% nom.			
$N = 60$	89.47% ± 0.56%	89.20% ± 0.57%	89.17% ± 0.57%
120	87.60% ± 0.60%	88.17% ± 0.59%	87.17% ± 0.61%
240	86.17% ± 0.63%	86.33% ± 0.63%	86.27% ± 0.63%
480	86.13% ± 0.63%	86.10% ± 0.63%	87.67% ± 0.60%
90% nom.			
$N = 60$	95.20% ± 0.39%	94.87% ± 0.40%	95.23% ± 0.39%
120	94.53% ± 0.42%	93.97% ± 0.43%	94.63% ± 0.41%
240	93.67% ± 0.44%	93.17% ± 0.46%	93.73% ± 0.44%
480	93.97% ± 0.43%	93.87% ± 0.44%	93.50% ± 0.45%
95% nom.			
$N = 60$	97.93% ± 0.26%	97.73% ± 0.27%	97.37% ± 0.29%
120	97.37% ± 0.29%	97.47% ± 0.29%	97.73% ± 0.27%
240	97.23% ± 0.30%	96.50% ± 0.34%	96.93% ± 0.31%
480	97.23% ± 0.30%	97.63% ± 0.28%	96.83% ± 0.32%
$\partial\mathcal{A}_{\mathrm{c}}$			
80% nom.			
$N = 60$	84.30% ± 0.66%	85.33% ± 0.65%	83.30% ± 0.68%
120	84.93% ± 0.65%	86.20% ± 0.63%	85.13% ± 0.65%
240	85.73% ± 0.64%	85.60% ± 0.64%	84.97% ± 0.65%
480	86.03% ± 0.63%	85.97% ± 0.63%	87.73% ± 0.60%
90% nom.			
$N = 60$	92.93% ± 0.47%	92.20% ± 0.49%	92.77% ± 0.47%
120	93.20% ± 0.46%	93.27% ± 0.46%	93.50% ± 0.45%
240	93.37% ± 0.45%	93.07% ± 0.46%	92.67% ± 0.48%
480	93.97% ± 0.43%	93.80% ± 0.44%	93.57% ± 0.45%
95% nom.			
$N = 60$	96.80% ± 0.32%	96.50% ± 0.34%	96.70% ± 0.33%
120	96.90% ± 0.32%	96.83% ± 0.32%	97.07% ± 0.31%
240	97.20% ± 0.30%	96.30% ± 0.34%	96.40% ± 0.34%
480	97.27% ± 0.30%	97.63% ± 0.28%	96.77% ± 0.32%

APPENDIX C

Cohen's d Confidence Sets Supplementary Material

C.1 Supplementary Human Connectome Project Results

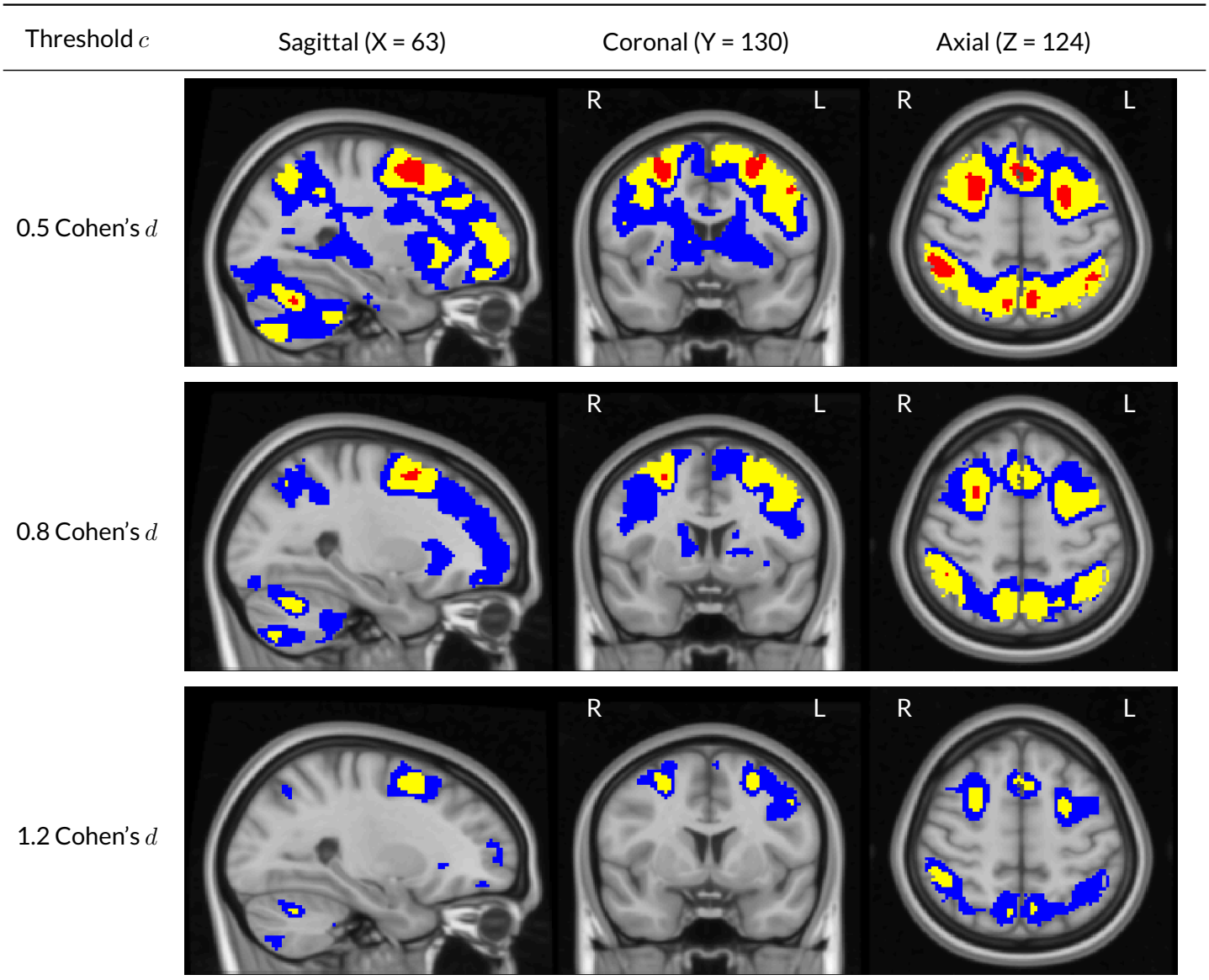


Figure C.1: Slices views of the Cohen's d Confidence Sets obtained from applying Algorithm 1. to the HCP working memory task data, using three Cohen's d effect size thresholds, $c = 0.5, 0.8$ and 1.2. Comparing with Fig. 5.11 and Fig. C.2, the CSs presented here are slightly more conservative than the corresponding CSs obtained with Algorithm 2. and Algorithm 3. (in the sense that the red upper CSs here are smaller, and blue lower CSs are larger). This is consistent with the simulation results obtained in Section 5.2.1 and 5.2.2, where the empirical coverage for Algorithm 1. was consistently larger than the other two methods.

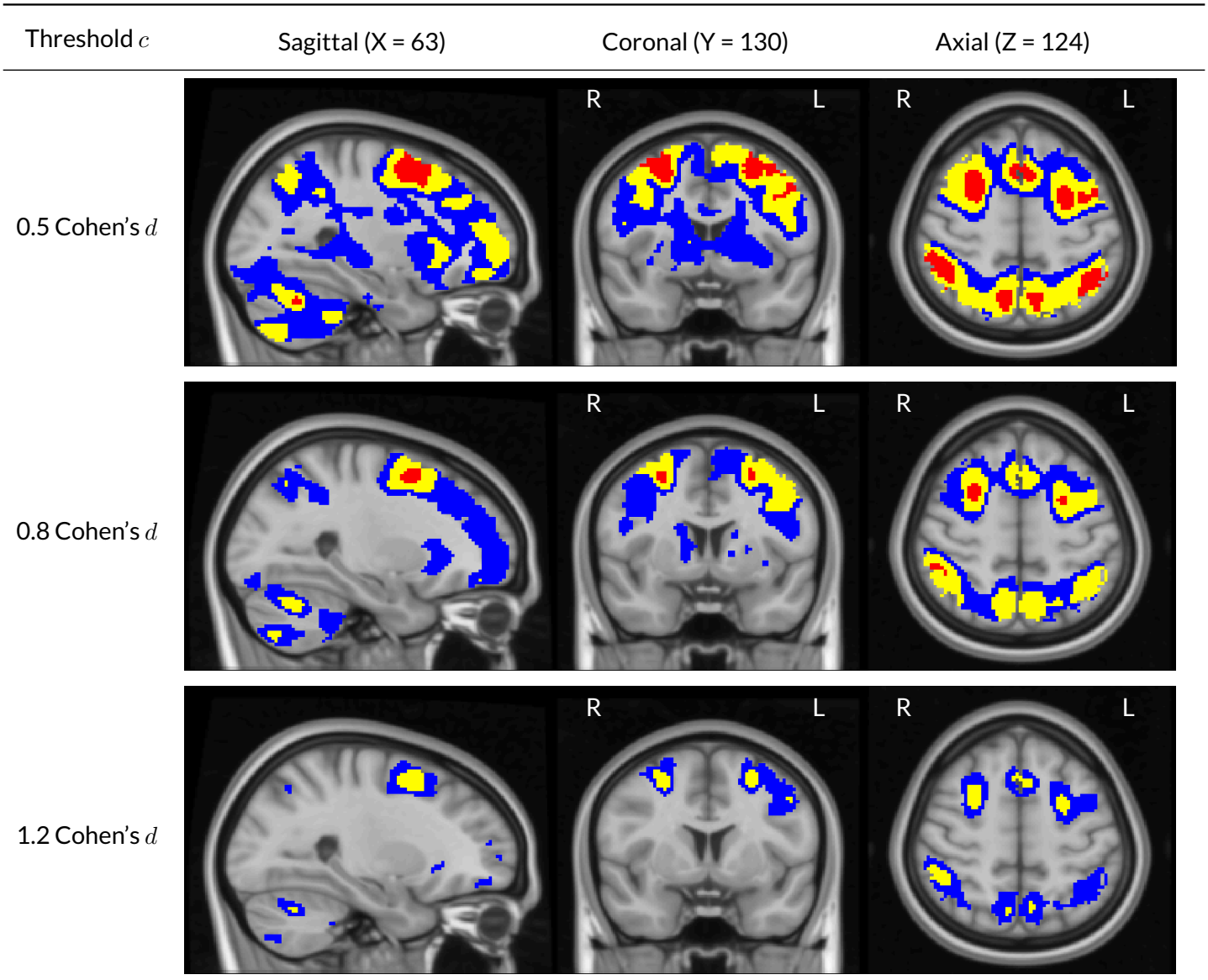


Figure C.2: Slices views of the Cohen's d Confidence Sets obtained from applying Algorithm 2. to the HCP working memory task data, using three Cohen's d effect size thresholds, $c = 0.5, 0.8$ and 1.2. Comparing with Fig. 5.11, the upper and lower CSs presented here are almost identical to the corresponding CSs obtained with Algorithm 3.

Bibliography

Alexandre Abraham, Fabian Pedregosa, Michael Eickenberg, Philippe Gervais, Andreas Mueller, Jean Kossaifi, Alexandre Gramfort, Bertrand Thirion, and Gaël Varoquaux. Machine learning for neuroimaging with scikit-learn. *Front. Neuroinform.*, 8:14, February 2014.

George Adelman and Others. *Encyclopedia of neuroscience*. Birkhäuser, 1987.

Andrew L Alexander, Jee Eun Lee, Mariana Lazar, and Aaron S Field. Diffusion tensor imaging of the brain. *Neurotherapeutics*, 4(3):316–329, July 2007.

Fidel Alfaro-Almagro, Mark Jenkinson, Neal K Bangerter, Jesper L R Andersson, Ludovica Griffanti, Gwenaëlle Douaud, Stamatios N Sotiropoulos, Saad Jbabdi, Moises Hernandez-Fernandez, Emmanuel Vallee, Diego Vidaurre, Matthew Webster, Paul McCarthy, Christopher Rorden, Alessandro Daducci, Daniel C Alexander, Hui Zhang, Iulius Dragonu, Paul M Matthews, Karla L Miller, and Stephen M Smith. Image processing and quality control for the first 10,000 brain imaging datasets from UK biobank. *Neuroimage*, 166:400–424, February 2018.

Grays Anatomy. Gray H. *Barnes &Noble*, 1918.

Jesper L R Andersson, Mark Jenkinson, Stephen Smith, and Others. Non-linear registration, aka spatial normalisation FMRIB technical report TR07JA2. *FMRIB Analysis Group of the University of Oxford*, 2:1–21, 2007.

Frederico A C Azevedo, Ludmila R B Carvalho, Lea T Grinberg, José Marcelo Farfel, Renata E L Ferretti, Renata E P Leite, Wilson Jacob Filho, Roberto Lent, and Suzana

Herculano-Houzel. Equal numbers of neuronal and nonneuronal cells make the human brain an isometrically scaled-up primate brain. *J. Comp. Neurol.*, 513(5):532–541, 2009.

P A Bandettini, E C Wong, R S Hinks, R S Tikofsky, and J S Hyde. Time course EPI of human brain function during task activation. *Magn. Reson. Med.*, 25(2):390–397, June 1992.

Deanna M Barch, Gregory C Burgess, Michael P Harms, Steven E Petersen, Bradley L Schlaggar, Maurizio Corbetta, Matthew F Glasser, Sandra Curtiss, Sachin Dixit, Cindy Feldt, Dan Nolan, Edward Bryant, Tucker Hartley, Owen Footer, James M Bjork, Russ Poldrack, Steve Smith, Heidi Johansen-Berg, Abraham Z Snyder, David C Van Essen, and WU-Minn HCP Consortium. Function in the human connectome: task-fMRI and individual differences in behavior. *Neuroimage*, 80:169–189, October 2013.

Y Benjamini and Y Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc.*, 1995.

Craig M Bennett, Michael B Miller, and George L Wolford. Neural correlates of interspecies perspective taking in the post-mortem atlantic salmon: an argument for multiple comparisons correction. *Neuroimage*, 47(Suppl 1):S125, 2009.

E L Bennett, M C Diamond, D Krech, and M R Rosenzweig. CHEMICAL AND ANATOMICAL PLASTICITY BRAIN. *Science*, 146(3644):610–619, October 1964.

Rotem Botvinik-Nezer, Roni Iwanir, Felix Holzmeister, Jürgen Huber, Magnus Johannesson, Michael Kirchler, Anna Dreber, Colin F Camerer, Russell A Poldrack, and Tom Schonberg. fMRI data of mixed gambles from the neuroimaging analysis replication and prediction study. *Sci Data*, 6(1):106, July 2019.

Alex Bowring, Camille Maumet, and Thomas Nichols. Exploring the impact of analysis software on task fMRI results, 2018a.

Alexander Bowring, Camille Maumet, and Thomas Nichols. NISOx-BDI/Software_Comparison. Zenodo, March 2018b.

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, 2019a. doi: 10.1002/hbm.24603. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/hbm.24603.

Alexander Bowring, Fabian Telschow, Armin Schwartzman, and Thomas E. Nichols. Spatial confidence sets for raw effect size images. *NeuroImage*, 203: 116187, 2019b. ISSN 1053-8119. doi: https://doi.org/10.1016/j.neuroimage.2019.116187. URL http://www.sciencedirect.com/science/article/pii/S1053811919307785.

Matthew Brett, Michael Hanke, Marc-Alexandre Côté, Chris Markiewicz, Satrajit Ghosh, Demian Wassermann, Stephan Gerhard, Eric Larson, Gregory R Lee, Yaroslav Halchenko, Erik Kastman, Cindee M, Félix C Morency, moloney, Ariel Rokem, Michiel Cottaar, Jarrod Millman, jaeilepp, Alexandre Gramfort, Robert D Vincent, Paul McCarthy, Jasper J F van den Bosch, Krish Subramaniam, Nolan Nichols, embaker, markhymers, chaselgrove, Basile, Nikolaas N Oosterhof, and Ian Nimmo-Smith. nipy/nibabel: 2.2.0, 2017.

Matthew J Brookes, Mark Woolrich, Henry Luckhoo, Darren Price, Joanne R Hale, Mary C Stephenson, Gareth R Barnes, Stephen M Smith, and Peter G Morris. Investigating the electrophysiological basis of resting state networks using magnetoencephalography. *Proc. Natl. Acad. Sci. U. S. A.*, 108(40):16783–16788, October 2011.

Katherine S Button, John P A Ioannidis, Claire Mokrysz, Brian A Nosek, Jonathan Flint, Emma S J Robinson, and Marcus R Munafò. Power failure: why small sample

size undermines the reliability of neuroscience. *Nat. Rev. Neurosci.*, 14(5):365–376, May 2013.

Richard B Buxton. Dynamic models of BOLD contrast. *Neuroimage*, 62(2):953–961, August 2012.

Joshua Carp. On the plurality of (methodological) worlds: estimating the analytic flexibility of FMRI experiments. *Front. Neurosci.*, 6:149, October 2012a.

Joshua Carp. The secret lives of experiments: methods reporting in the fMRI literature. *Neuroimage*, 63(1):289–300, October 2012b.

Joshua Carp. Optimizing the order of operations for movement scrubbing: Comment on power et al. *Neuroimage*, 76:436–438, August 2013.

Gang Chen, Paul A Taylor, and Robert W Cox. Is the statistic value all we should care about in neuroimaging? *Neuroimage*, 147:952–959, February 2017.

Gang Chen, Robert W Cox, Daniel R Glen, Justin K Rajendra, Richard C Reynolds, and Paul A Taylor. A tail of two sides: Artificially doubled false positive rates in neuroimaging due to the sidedness choice with t-tests. *Hum. Brain Mapp.*, September 2018.

Victor Chernozhukov, Denis Chetverikov, and Kengo Kato. Gaussian approximations and multiplier bootstrap for maxima of sums of high-dimensional random vectors, 2013.

Justin R Chumbley and Karl J Friston. False discovery rate revisited: FDR and topological inference using gaussian random fields. *Neuroimage*, 44(1):62–70, January 2009.

Jacob Cohen. *Statistical power analysis for the behavioral sciences*. Routledge, 2013.

R W Cox. AFNI: software for analysis and visualization of functional magnetic resonance neuroimages. *Comput. Biomed. Res.*, 29(3):162–173, June 1996.

Robert W Cox, Gang Chen, Daniel R Glen, Richard C Reynolds, and Paul A Taylor. FMRI clustering in AFNI: False-Positive rates redux. *Brain Connect.*, 7(3):152–171, April 2017a.

Robert W Cox, Gang Chen, Daniel R Glen, Richard C Reynolds, and Paul A Taylor. fMRI clustering and false-positive rates. *Proc. Natl. Acad. Sci. U. S. A.*, 114(17): E3370–E3371, April 2017b.

Henk R Cremers, Tor D Wager, and Tal Yarkoni. The relation between statistical power and inference in fMRI. *PLoS One*, 12(11):e0184923, November 2017.

Sean P David, Jennifer J Ware, Isabella M Chu, Pooja D Loftus, Paolo Fusar-Poli, Joaquim Radua, Marcus R Munafò, and John P A Ioannidis. Potential reporting bias in fMRI studies of the brain. *PLoS One*, 8(7):e70104, July 2013.

Matthew F Glasser David C. Van Essen. The human connectome project: Progress and prospects. *Cerebrum*, 2016, 2016.

Russell Davidson and Emmanuel Flachaire. The wild bootstrap, tamed at last. *J. Econom.*, 146(1):162–169, September 2008.

David Degras. Simultaneous confidence bands for the mean of functional data. *Wiley Interdiscip. Rev. Comput. Stat.*, 9(3):e1397, 2017.

M C Diamond, D Krech, and M R Rosenzweig. THE EFFECTS OF AN ENRICHED ENVIRONMENT ON THE HISTOLOGY OF THE RAT CEREBRAL CORTEX. *J. Comp. Neurol.*, 123:111–120, August 1964.

Anders Eklund, Thomas E Nichols, and Hans Knutsson. Cluster failure: Why fMRI inferences for spatial extent have inflated false-positive rates. *Proc. Natl. Acad. Sci. U. S. A.*, 113(28):7900–7905, July 2016.

Cameron Ellis, Christopher Baldassano, Anna C Schapiro, Ming Bo Cai, and

Jonathan D Cohen. Facilitating open-science with realistic fMRI simulation: validation and application. January 2019.

Stephen A Engel and Philip C Burton. Confidence intervals for fMRI activation maps. *PLoS One*, 8(12):e82419, December 2013.

Ariel Deardorff Erin D. Foster. Open science framework (OSF). *J. Med. Libr. Assoc.*, 105(2):203, April 2017.

Guillaume Flandin and Karl J Friston. Analysis of familyâĂŘwise error rates in statistical parametric mapping using random field theory. *Hum. Brain Mapp.*, 40(7): 2052–2054, May 2019.

Tatiana Fomina, Matthias Hohmann, Bernhard Scholkopf, and Moritz Grosse-Wentrup. Identification of the default mode network with electroencephalography. *Conf. Proc. IEEE Eng. Med. Biol. Soc.*, 2015:7566–7569, 2015.

Michael D Fox and Ron L Alterman. Brain stimulation for torsion dystonia. *JAMA Neurol.*, 72(6):713–719, June 2015.

K J Friston, P Fletcher, O Josephs, A Holmes, M D Rugg, and R Turner. Event-related fMRI: characterizing differential responses. *Neuroimage*, 7(1):30–40, January 1998.

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, April 2002.

Matthew F Glasser, Stamatios N Sotiropoulos, J Anthony Wilson, Timothy S Coalson, Bruce Fischl, Jesper L Andersson, Junqian Xu, Saad Jbabdi, Matthew Webster, Jonathan R Polimeni, David C Van Essen, Mark Jenkinson, and WU-Minn HCP Consortium. The minimal preprocessing pipelines for the human connectome project. *Neuroimage*, 80:105–124, October 2013.

T Glatard, L B Lewis, R F da Silva, R Adalat, N Beck, C Lepage, and Others. Reproducibility of neuroimaging analyses across operating systems. front neuroinform. *Frontiers*, 9, 2015.

Gary H Glover. Overview of functional magnetic resonance imaging. *Neurosurg. Clin. N. Am.*, 22(2):133–9, vii, April 2011.

Javier Gonzalez-Castillo, Ziad S Saad, Daniel A Handwerker, Souheil J Inati, Noah Brenowitz, and Peter A Bandettini. Whole-brain, time-locked activation with simple tasks revealed using massive averaging and model-free analysis. *Proc. Natl. Acad. Sci. U. S. A.*, 109(14):5487–5492, April 2012.

Krzysztof J Gorgolewski and Russell A Poldrack. A practical guide for improving transparency and reproducibility in neuroimaging research. *PLoS Biol.*, 14(7): e1002506, July 2016.

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, Tal Yarkoni, and Daniel S Margulies. NeuroVault.org: a web-based repository for collecting and sharing unthresholded statistical maps of the human brain. *Front. Neuroinform.*, 9:8, April 2015.

Krzysztof J Gorgolewski, Tibor Auer, Vince D Calhoun, R Cameron Craddock, Samir Das, Eugene P Duff, Guillaume Flandin, Satrajit S Ghosh, Tristan Glatard, Yaroslav O Halchenko, Daniel A Handwerker, Michael Hanke, David Keator, Xiangrui Li, Zachary Michael, Camille Maumet, B Nolan Nichols, Thomas E Nichols, John Pellman, Jean-Baptiste Poline, Ariel Rokem, Gunnar Schaefer, Vanessa Sochat, William Triplett, Jessica A Turner, Gaël Varoquaux, and Russell A Poldrack. The brain imaging data structure, a format for organizing and describing outputs of neuroimaging experiments. *Sci Data*, 3:160044, June 2016.

C Goutte, F A Nielsen, and K H Hansen. Modeling the hemodynamic response in fMRI

using smooth FIR filters. *IEEE Trans. Med. Imaging*, 19(12):1188–1201, December 2000.

Ed H B M Gronenschild, Petra Habets, Heidi I L Jacobs, Ron Mengelers, Nico Rozendaal, Jim van Os, and Machteld Marcelis. The effects of FreeSurfer version, workstation type, and macintosh operating system version on anatomical volume and cortical thickness measurements. *PLoS One*, 7(6):e38234, June 2012.

Richard J Hargreaves and Michael Klimas. Imaging in drug development. In *Principles of Clinical Pharmacology*, pages 327–341. Elsevier, 2012.

Ahmad R Hariri, Alessandro Tessitore, Venkata S Mattay, Francesco Fera, and Daniel R Weinberger. The amygdala response to emotional stimuli: A comparison of faces and scenes. *Neuroimage*, 17(1):317–323, September 2002.

Yong-Wook Hong, Yejong Yoo, Jihoon Han, Tor D Wager, and Choong-Wan Woo. False-positive neuroimaging: Undisclosed flexibility in testing spatial hypotheses allows presenting anything as a replicated finding. *Neuroimage*, April 2019.

John D Hunter. Matplotlib: A 2D graphics environment. *Comput. Sci. Eng.*, 9(3):90–95, May 2007.

John P A Ioannidis. Why most published research findings are false. *PLoS Med.*, 2(8): e124, August 2005.

John P A Ioannidis, Marcus R Munafò, Paolo Fusar-Poli, Brian A Nosek, and Sean P David. Publication and other reporting biases in cognitive sciences: detection, prevalence, and prevention. *Trends Cogn. Sci.*, 18(5):235–241, May 2014.

Mark Jenkinson, Peter Bannister, Michael Brady, and Stephen Smith. Improved optimization for the robust and accurate linear registration and motion correction of brain images. *Neuroimage*, 17(2):825–841, October 2002.

Mark Jenkinson, Christian F Beckmann, Timothy E J Behrens, Mark W Woolrich, and Stephen M Smith. FSL. *Neuroimage*, 62(2):782–790, August 2012.

Suneil K Kalia, Tejas Sankar, and Andres M Lozano. Deep brain stimulation for parkinson's disease and other movement disorders. *Curr. Opin. Neurol.*, 26(4):374–380, August 2013.

Thomas Kluyver, Benjamin Ragan-Kelley, Fernando Pérez, Brian E Granger, Matthias Bussonnier, Jonathan Frederic, Kyle Kelley, Jessica B Hamrick, Jason Grout, Sylvain Corlay, and Others. Jupyter notebooks-a publishing format for reproducible computational workflows. In *ELPUB*, pages 87–90, 2016.

K K Kwong, J W Belliveau, D A Chesler, I E Goldberg, R M Weisskoff, B P Poncelet, D N Kennedy, B E Hoppel, M S Cohen, and R Turner. Dynamic magnetic resonance imaging of human brain activity during primary sensory stimulation. *Proc. Natl. Acad. Sci. U. S. A.*, 89(12):5675–5679, June 1992.

Nico F Laubscher. Normalizing the noncentral t and F distributions, 1960.

M H Lee, C D Smyser, and J S Shimony. Resting-state fMRI: a review of methods and clinical applications. *AJNR Am. J. Neuroradiol.*, 34(10):1866–1872, October 2013.

Megan H Lee, Carl D Hacker, Abraham Z Snyder, Maurizio Corbetta, Dongyang Zhang, Eric C Leuthardt, and Joshua S Shimony. Clustering of resting state networks. *PLoS One*, 7(7):e40370, July 2012.

Martin A Lindquist, Ji Meng Loh, Lauren Y Atlas, and Tor D Wager. Modeling the hemodynamic response function in fMRI: efficiency, bias and mis-modeling. *Neuroimage*, 45(1 Suppl):S187–98, March 2009.

Francisco López-Muñoz, Jesús Boya, and Cecilio Alamo. Neuron theory, the cornerstone of neuroscience, on the centenary of the nobel prize award to santiago ramón y cajal. *Brain Res. Bull.*, 70(4-6):391–405, October 2006.

Torben E Lund, Minna D Nørgaard, Egill Rostrup, James B Rowe, and Olaf B Paulson. Motion or activity: their role in intra- and inter-subject variation in fMRI. *Neuroimage*, 26(3):960–964, July 2005.

Paul M Matthews, Garry D Honey, and Edward T Bullmore. Neuroimaging: Applications of fMRI in translational medicine and clinical practice. *Nat. Rev. Neurosci.*, 7 (9):732, 2006.

Camille Maumet, Tibor Auer, Alexander Bowring, Gang Chen, Samir Das, Guillaume Flandin, Satrajit Ghosh, Tristan Glatard, Krzysztof J Gorgolewski, Karl G Helmer, Mark Jenkinson, David B Keator, B Nolan Nichols, Jean-Baptiste Poline, Richard Reynolds, Vanessa Sochat, Jessica Turner, and Thomas E Nichols. Sharing brain mapping statistical results with the neuroimaging data model. *Sci Data*, 3:160102, December 2016.

Linda K McEvoy, Christine Fennema-Notestine, J Cooper Roddey, Donald J Hagler, Dominic Holland, David S Karow, Christopher J Pung, James B Brewer, and Anders M Dale. Alzheimer disease: Quantitative structural neuroimaging for detection and prediction of clinical and structural changes in mild cognitive impairment, 2009.

Wes McKinney and Others. Data structures for statistical computing in python. In *Proceedings of the 9th Python in Science Conference*, volume 445, pages 51–56, 2010.

Andrea Mechelli, Cathy J Price, Karl J Friston, and John Ashburner. Voxel-Based morphometry of the human brain: Methods and applications. *Curr. Med. Imaging Rev.*, 1 (2):105–113, 2005.

Paul E Meehl. Theory-Testing in psychology and physics: A methodological paradox. *Philos. Sci.*, 34(2):103–115, June 1967.

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 L R

Andersson, Ludovica Griffanti, Gwenaëlle Douaud, Thomas W Okell, Peter Weale, Iulius Dragonu, Steve Garratt, Sarah Hudson, Rory Collins, Mark Jenkinson, Paul M Matthews, and Stephen M Smith. Multimodal population brain imaging in the UK biobank prospective epidemiological study. *Nat. Neurosci.*, 19(11):1523–1536, November 2016.

W Mohamed. The edwin smith surgical papyrus: Neuroscience in ancient egypt. *IBRO History of Neuroscience*, 2014.

Joseph M Moran, Eshin Jolly, and Jason P Mitchell. Social-cognitive deficits in normal aging. *J. Neurosci.*, 32(16):5553–5561, April 2012.

Malaak N Moussa, Matthew R Steen, Paul J Laurienti, and Satoru Hayasaka. Consistency of network modules in resting-state FMRI connectome data. *PLoS One*, 7(8): e44428, August 2012.

Karsten Mueller, Jöran Lepsien, Harald E Möller, and Gabriele Lohmann. Commentary: Cluster failure: Why fMRI inferences for spatial extent have inflated false-positive rates. *Front. Hum. Neurosci.*, 11:345, June 2017.

Thomas Nichols. SPM plot units, 2012.

Thomas E Nichols and Andrew P Holmes. Nonparametric permutation tests for functional neuroimaging: a primer with examples. *Hum. Brain Mapp.*, 15(1):1–25, January 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, Erika Proal, Bertrand Thirion, David C Van Essen, Tonya White, and B T Thomas Yeo. Best practices in data analysis and sharing in neuroimaging using MRI. *Nat. Neurosci.*, 20(3):299–303, February 2017.

Lars Holm Nielsen and Tim Smith. Zenodo overview. Zenodo, 2014.

Regina Nuzzo. Scientific method: statistical errors. *Nature*, 506(7487):150–152, February 2014.

S Ogawa, T M Lee, A R Kay, and D W Tank. Brain magnetic resonance imaging with contrast dependent on blood oxygenation. *Proc. Natl. Acad. Sci. U. S. A.*, 87(24): 9868–9872, December 1990.

S Ogawa, D W Tank, R Menon, J M Ellermann, S G Kim, H Merkle, and K Ugurbil. Intrinsic signal changes accompanying sensory stimulation: functional brain mapping with magnetic resonance imaging. *Proc. Natl. Acad. Sci. U. S. A.*, 89(13):5951–5955, July 1992.

Wiktor Olszowy, John Aston, Catarina Rua, and Guy B Williams. Accurate autocorrelation modeling substantially improves fMRI reliability, 2019.

Open Science Collaboration. Estimating the reproducibility of psychological science. *Science*, 349(6251):aac4716, August 2015.

Aarthi Padmanabhan, Charles F Geier, Sarah J Ordaz, Theresa Teslovich, and Beatriz Luna. Developmental changes in brain function underlying the influence of reward processing on inhibitory control. *Dev. Cogn. Neurosci.*, 1(4):517–529, October 2011.

O Parker Jones, N L Voets, J E Adcock, R Stacey, and S Jbabdi. Resting connectivity predicts task activation in pre-surgical populations. *NeuroImage: Clinical*, 13:378–385, January 2017.

Ruth Pauli, Alexander Bowring, Richard Reynolds, Gang Chen, Thomas E Nichols, and Camille Maumet. Exploring fMRI results space: 31 variants of an fMRI analysis in AFNI, FSL, and SPM. *Front. Neuroinform.*, 10:24, July 2016.

William D Penny, Karl J Friston, John T Ashburner, Stefan J Kiebel, and Thomas E

Nichols. *Statistical Parametric Mapping: The Analysis of Functional Brain Images*. Elsevier, April 2011.

Russell A Poldrack, Jeanette A Mumford, and Thomas E Nichols. *Handbook of Functional MRI Data Analysis*. Cambridge University Press, August 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. *Front. Neuroinform.*, 7:12, July 2013.

Russell A Poldrack, Chris I Baker, Joke Durnez, Krzysztof J Gorgolewski, Paul M Matthews, Marcus R Munafò, Thomas E Nichols, Jean-Baptiste Poline, Edward Vul, and Tal Yarkoni. Scanning the horizon: towards transparent and reproducible neuroimaging research. *Nat. Rev. Neurosci.*, 18(2):115–126, February 2017.

Marianne C Reddan, Martin A Lindquist, and Tor D Wager. Effect size estimation in neuroimaging, 2017.

William W Rozeboom. The fallacy of the null-hypothesis significance test. *Psychol. Bull.*, 57(5):416–428, 1960.

Tom Schonberg, Craig R Fox, Jeanette A Mumford, Eliza Congdon, Christopher Trepel, and Russell A Poldrack. Decreasing ventromedial prefrontal cortex activity during sequential risk-taking: an FMRI investigation of the balloon analog risk task. *Front. Neurosci.*, *6*:80, June 2012.

Shayle R Searle, George Casella, and Charles E McCulloch. *Variance Components*. John Wiley & Sons, September 2009.

P Skudlarski, R T Constable, and J C Gore. ROC analysis of statistical methods used in functional MRI: individual subjects. *Neuroimage*, 9(3):311–329, March 1999.

Stephen M Smith. Fast robust automated brain extraction. *Hum. Brain Mapp.*, 17(3): 143–155, November 2002.

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, and Christian F Beckmann. Correspondence of the brain's functional architecture during activation and rest. *Proc. Natl. Acad. Sci. U. S. A.*, 106(31):13040–13045, August 2009.

Stephen M Smith, Christian F Beckmann, Jesper Andersson, Edward J Auerbach, Janine Bijsterbosch, Gwenaëlle Douaud, Eugene Duff, David A Feinberg, Ludovica Griffanti, Michael P Harms, Michael Kelly, Timothy Laumann, Karla L Miller, Steen Moeller, Steve Petersen, Jonathan Power, Gholamreza Salimi-Khorshidi, Abraham Z Snyder, An T Vu, Mark W Woolrich, Junqian Xu, Essa Yacoub, Kamil Uğurbil, David C Van Essen, Matthew F Glasser, and WU-Minn HCP Consortium. Resting-state fMRI in the human connectome project. *Neuroimage*, 80:144–168, October 2013.

José M Soares, Paulo Marques, Victor Alves, and Nuno Sousa. A hitchhiker's guide to diffusion tensor imaging. *Front. Neurosci.*, 7:31, March 2013.

Max Sommerfeld, Stephan Sain, and Armin Schwartzman. Confidence regions for spatial excursion sets from repeated random field observations, with an application to climate. *J. Am. Stat. Assoc.*, 113(523):1327–1340, July 2018.

Reisa A Sperling, Dorene M Rentz, Keith A Johnson, Jason Karlawish, Michael Donohue, David P Salmon, and Paul Aisen. The A4 study: stopping AD before symptoms begin? *Sci. Transl. Med.*, 6(228):228fs13, March 2014.

William C Stacey and Brian Litt. Technology insight: neuroengineering and epilepsy-designing devices for seizure control. *Nat. Clin. Pract. Neurol.*, 4(4):190–201, April 2008.

Stephen C Strother, Jon Anderson, Lars Kai Hansen, Ulrik Kjems, Rafal Kustra, John Sidtis, Sally Frutiger, Suraj Muley, Stephen LaConte, and David Rottenberg. The quantitative evaluation of functional neuroimaging experiments: the NPAIRS data analysis framework. *Neuroimage*, 15(4):747–771, April 2002.

I Tavor, O Parker Jones, R B Mars, S M Smith, T E Behrens, and S Jbabdi. Task-free MRI predicts individual differences in brain activity during task performance. *Science*, 352(6282):216–220, April 2016.

Fabian J E Telschow and Armin Schwartzman. Simultaneous confidence bands for functional data using the gaussian kinematic formula. January 2019.

Benjamin O Turner, Erick J Paul, Michael B Miller, and Aron K Barbey. Small sample sizes reduce the replicability of task-based fMRI studies. *Commun Biol*, 1:62, June 2018.

K Uludag, B Müller-Bierl, and K Ugurbil. An integrative model for neuronal activity-induced signal changes for gradient and spin echo functional imaging, 2009.

Tor D Wager, Martin A Lindquist, Thomas E Nichols, Hedy Kober, and Jared X Van Snellenberg. Evaluating the consistency and specificity of neuroimaging data using meta-analysis. *Neuroimage*, 45(1 Suppl):S210–21, March 2009.

Stéfan van der Walt, S Chris Colbert, and Gaël Varoquaux. The NumPy array: A structure for efficient numerical computation. *Comput. Sci. Eng.*, 13(2):22–30, March 2011.

Ronald L Wasserstein, Nicole A Lazar, and Others. The ASA's statement on p-values: context, process, and purpose. *Am. Stat.*, 70(2):129–133, 2016.

Anderson M Winkler, Gerard R Ridgway, Gwenaëlle Douaud, Thomas E Nichols, and Stephen M Smith. Faster permutation inference in brain imaging. *Neuroimage*, 141: 502–516, November 2016.

Choong-Wan Woo, Anjali Krishnan, and Tor D Wager. Cluster-extent based thresholding in fMRI analyses: pitfalls and recommendations. *Neuroimage*, 91:412–419, May 2014.

M W Woolrich, B D Ripley, M Brady, and S M Smith. Temporal autocorrelation in univariate linear modeling of FMRI data. *Neuroimage*, 14(6):1370–1386, December 2001.

Mark W Woolrich, Timothy E J Behrens, Christian F Beckmann, Mark Jenkinson, and Stephen M Smith. Multilevel linear modelling for FMRI group analysis using bayesian inference. *Neuroimage*, 21(4):1732–1747, April 2004.

K J Worsley, C Liao, M Grabove, V Petre, B Ha, and A C Evans. A general statistical analysis for fMRI data, 2000.

Andy W K Yeung. An updated survey on statistical thresholding and sample size of fMRI studies. *Front. Hum. Neurosci.*, 12:16, January 2018.