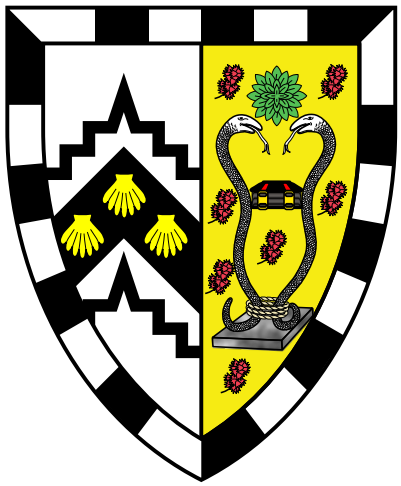
Validation of statistical methods used in task fMRI studies



Wiktor Olszowy

Supervisors: Dr. Guy B. Williams, Prof. John Aston

Wolfson Brain Imaging Centre
Department of Clinical Neurosciences
University of Cambridge

This dissertation is submitted for the degree of
Doctor of Philosophy

Gonville & Caius College

Cambridge, 20 December 2018

Abstract

Validation of statistical methods used in task fMRI studies

Wiktor Olszowy

Functional Magnetic Resonance Imaging (fMRI) is a non-invasive tool used to investigate brain function. The processing of fMRI data consists of multiple steps and the final results often depend greatly on the specific choice of options used: for example, head motion correction, slice timing correction, registration to common space, pre-whitening, hemodynamic response function modelling and multiple comparison correction. As most of these methods were introduced when fMRI was in its infancy, and were initially validated only for small datasets, it is questionable whether the current default methods used in the popular analysis packages are optimal. Despite the huge popularity of fMRI, there have been few studies validating statistical methods. This thesis presents a validation of statistical methods used in task fMRI studies which are related to pre-whitening and to hemodynamic response function modelling. It considers fMRI used with the blood oxygenation level dependent (BOLD) contrast.

Firstly, I compared the most frequently used fMRI analysis packages: AFNI, FSL and SPM, with regard to temporal autocorrelation modelling, often known as pre-whitening. I employed eleven datasets containing 980 scans corresponding to different fMRI protocols and subject populations. Though autocorrelation modelling in AFNI was not perfect, its performance was much higher than the performance of autocorrelation modelling in FSL and SPM. The residual autocorrelated noise in FSL and SPM led to heavily confounded first level results, particularly for low-frequency experimental designs. My results show superior performance of SPM's alternative pre-whitening: `FAST`, over SPM's default algorithm. The reliability of task fMRI studies would increase with more accurate autocorrelation modelling. Furthermore, reliability could increase if the packages provided diagnostic plots. This way the investigator would be aware of pre-whitening problems.

Next, I compared - in terms of specificity-sensitivity trade-offs - a number of hemodynamic response function models which are available in AFNI, FSL and SPM. Again, I used different datasets to represent different fMRI protocols and different experimental designs: altogether scans of 772 subjects from five experiments. In contrast to previous studies, I used real data rather than simulations, investigated methods from more than one software package, and employed scans of many subjects. Among other factors, I found that the use of the temporal and dispersion derivatives led to large sensitivity increases compared to the use of the canonical model, but only when the experimental design was event-related and when the statistical inference was based on an F-test which tested the variance explained by canonical function together with the derivatives rather than a t-test which tested the variance explained by the canonical function only. This was the case both for single subject and for group level analyses.

Finally, I investigated the effect of ageing on the BOLD signal. For this, I used the Cambridge Centre for Ageing and Neuroscience (CamCAN) data of 641 subjects between 18 and 88 years old. I investigated how the shape of the hemodynamic response function changes with age and whether it is on average similar to the canonical function. The CamCAN task fMRI data enabled the estimation of the hemodynamic response function in the auditory, visual and motor regions. I used the biophysical balloon model to investigate whether values of BOLD-derived physiological parameters vary with age and whether these variations can explain the difference of the hemodynamic response function with age. CamCAN Magnetoencephalography (MEG) data enabled a correlation of the results with neural delay estimates. The hemodynamic response function was found to substantially vary with age, with observed response delays in all considered regions. The estimated balloon model parameters were found to vary with age too. A robustness analysis of the SPM's balloon model revealed serious problems with the current SPM's balloon model estimation procedure.

Overall, this thesis presents novel validations of a number of popular statistical methods used in task fMRI studies. I identified several relevant problems related to prewhitening and hemodynamic response function modelling. Importantly, in this thesis I address ways of dealing with such problems so that sensitivity and specificity in task fMRI studies can be improved.

Acknowledgements

First of all, I would like to thank Guy and John, my supervisors, for their patience and continuous support. Richard Henson, Anders Eklund, Karl Friston, Richard Reynolds, Catarina Rua, Thomas Nichols, Oliver Speck and Gang Chen gave me very helpful advice on multiple occasions. Due to my lack of expertise in neuroimaging software at the beginning of my PhD, the AFNI, FSL and SPM mailing lists became for me a major source of knowledge and a place to ask for help. I am deeply grateful to a number of people, who without even knowing me, helped me with technical issues there. This PhD would not have been possible without financial support from Cambridge Trust, the Mateusz Grabowski Fund and the Cambridge Philosophical Society. Guarantors of Brain, Fearnsides Fund, European College of Neuropsychopharmacology, as well as Gonville & Caius College provided funds for conferences and workshops. These meetings together with the ensuing discussions provided me with invaluable feedback on my work.

My special thanks go to Caius friends, who made my life in Cambridge so enjoyable. Together with the Caius Movie Appreciation Society I watched a number of extraordinary movies followed by absorbing discussions, while during the amazing time when I lived together with Beñat Mencia Uranga and Santiago Caño Muñiz I co-hosted possibly the best brunches in town. Together with Beñat and Santiago we were probably making in our Magic House more tortillas de patatas than the rest of Cambridge combined altogether, a truly tremendous achievement. Daniela Peris was the best tennis partner one could wish to have, Julio Song was, really, the funniest person I have been hanging out with, Alexander Shattock and Guo Yu organised terrific Halloween parties, Michał Kosicki always led the least cliché discussions. I have beautiful memories related to Caius trips to Gdańsk and to Ereño. Also, I had extremely interesting discussions about politics with friends from the Polish Society, among others, with Michał Bogdan and Piotr Wieprzowski. During my entire education I have been supported by my family, to whom I am deeply grateful: *Dziękuję Dziadzi Stefanowi, Babci Jucie, Babci Sabinie, Mamie Reni, Tacie Mirkowi, i Siostrze Hani!*

Declaration

This dissertation is the result of my own work and includes nothing which is the outcome of work done in collaboration except as declared in the Preface and specified in the text. It is not substantially the same as any that I have submitted, or, is being concurrently submitted for a degree or diploma or other qualification at the University of Cambridge or any other University or similar institution except as declared in the Preface and specified in the text. I further state that no substantial part of my dissertation has already been submitted, or, is being concurrently submitted for any such degree, diploma or other qualification at the University of Cambridge or any other University or similar institution except as declared in the Preface and specified in the text. This dissertation contains fewer than 60,000 words excluding figures, tables and bibliography.

Cambridge, 20 December 2018 Wiktor Olszowy

Contents

Chapter 1

Introduction

Functional Magnetic Resonance Imaging (fMRI) based on the blood oxygenation level dependent (BOLD) signal is widely used for mapping patterns of activation in the brain. BOLD-fMRI can be used, for example, to detect brain activations due to an experimental condition, to investigate relationships between distinct brain regions, to compare different subject populations in terms of their brain patterns, or to infer causal links between a number of brain regions [Logothetis, 2008]. There are first level studies, also called single-subject studies, which investigate brain patterns in a single subject. However, more popular are second level studies, also called group studies, which investigate average brain patterns across a subject population and where the objective is to make conclusions about the given subject population rather than the individual subjects that were scanned. Furthermore, BOLD-fMRI studies can be divided into task studies and no task (resting state) studies. In the former, an experimental condition is employed, for example a visual or audio stimulus. In the latter, the subject is not performing any task while being in the scanner and her baseline brain activation is the brain state of interest.

Interpretation of the BOLD response in terms of the underlying physiological changes is difficult due to the complexity of the signal [Logothetis, 2008]. Nevertheless, since BOLD-fMRI is non-invasive, safe, widely available and offers relatively high spatiotemporal resolution, it has become a very popular neuroimaging tool used in thousands of studies every year.

1.1 BOLD signal

BOLD-fMRI is based on the principle that the oxygenated and deoxygenated forms of haemoglobin (oxy- and deoxyhaemoglobin) have different magnetic properties [Ogawa

et al., 1990, 1992]. Deoxyhaemoglobin is paramagnetic due to four unpaired electrons at each iron center, while oxyhaemoglobin is diamagnetic [Pauling and Coryell, 1936].

In an MRI scanner there is a constant magnetic field B_0, in which particles precess at the Larmor frequency. In most MRI applications the particles of interest are hydrogen nuclei, as they are abundant in water and, in particular, in the human body. As the hydrogen nuclei have different numbers of protons and neutrons (1 proton, 0 neutrons), they precess. In order to create a detectable signal in the MRI scanner, the spins need to be perturbed so that there is a vector component perpendicular to B_0. This can be done by employing a pulse of an alternating magnetic field (B_1) that oscillates at the Larmor frequency of the spins and which is transverse to B_0. Following application of the B_1 pulse, a component of the net magnetization lies in the plane perpendicular to B_0 and precesses about B_0 at the Larmor frequency. This net magnetization is made up of many individual spins all precessing about B_0, which leads to a decaying oscillating signal in a conducting coil placed about the sample as the net magnetization of the sample changes [Ash, 2011]. The signal decay rate is characterised by the time constants T_2 and T_2*. T_2* describes the effect of an inhomogeneous magnetic field, which causes local changes both in the Larmor frequency and in dephasing. The paramagnetic properties of deoxyhaemoglobin affect the T_2* around blood vessels, which constitutes the BOLD signal. Spins located near deoxyhaemoglobin experience locally variable Larmor frequencies and will dephase more rapidly, decreasing the local T_2*. Deoxyhaemoglobin-induced T2*-shortening effect is strongest near larger veins and is strengthened by the use of gradient echo (GRE) sequences with echo times (TEs) close to T_2* [Elster and Burdette, 2001].

Following neural activation, cerebral blood flow (CBF) increases, but the cerebral metabolic rate of oxygen consumption ($CMRO_2$) does not increase proportionately. This "uncoupling" relationship between blood flow and oxygen demand was first described in Fox and Raichle [1986], who refuted a previous belief that the cerebral hemodynamics were directly linked to the brain's short-term metabolic needs. The authors showed that more oxygenated blood is supplied to the active brain region than is actually required for this region's immediate metabolism. The oversupply of oxygenated blood causes the relative concentration of deoxyhaemoglobin in the active brain region to decrease. Thus, the BOLD signal in the active brain region will generally increase.

The BOLD response following a brief experimental stimulus is called the hemodynamic response function (HRF). This function often exhibits a small initial dip, during which the BOLD signal is below its baseline value. This is followed by the main response - a tall peak. The main response is followed by the so-called post-stimulus undershoot, during

which the BOLD signal is again below its baseline value. While the mechanism generating the main BOLD response is clear, there is much controversy related to the initial dip and to the post-stimulus undershoot [Buxton, 2012]. The former might reflect physiological changes occurring immediately after neural activation: for example an increase in $CMRO_2$ or an increase in the cerebral blood volume (CBV). Perhaps both these factors contribute to the initial dip. The post-stimulus undershoot likely results from a slow CBV recovery and a CBF decrease. The processes by which neural activity changes CBF, CBV and $CMRO_2$ are termed neurovascular coupling.

While the BOLD response is linked to the number of "firing" neurons, Logothetis et al. [2001] showed that it is more related to the extracellular local field potentials (LFPs), which are electrophysiological signals generated by the summed electric current of large neuron populations, but within a small volume. LFPs correspond to the total activity of regional neural networks including neural discharges, as well as the sum of positive and negative post-synaptic potentials at multiple dendritic connections [Elster and Burdette, 2001]. This activity reflects slowly changing voltages.

BOLD-fMRI data is acquired in k-space, which is the Fourier transform of the image measured. After reconstruction, the image of interest is four-dimensional: three dimensions correspond to space, and one dimension corresponds to time. Often, the spatial resolution is around 3 mm in each direction, while the temporal resolution is normally around 2 s.

1.2 Statistical pipelines for fMRI data

Task fMRI data is usually analysed using a statistical parametric mapping framework [Friston et al., 1994a]. In this approach, the data are first preprocessed to account, among others, for head motion and scanner-induced drifts in the data. Then, a general linear model (GLM) is fitted to fMRI time series from each voxel [Friston et al., 1994b, 1995b, Worsley and Friston, 1995]. Finally, the univariate analysis results are combined using a multiple comparison correction, for example the cluster inference, which is based on Gaussian random field theory [Friston et al., 1994c, Forman et al., 1995]. This analysis framework is used in the most popular fMRI data analysis packages [Yeung, 2018]: AFNI [Cox, 1996], FSL [Jenkinson et al., 2012] and SPM [Penny et al., 2011]. For resting state fMRI, there is more variability in the statistical pipelines.

1.3 fMRI reliability

Logothetis [2008] discussed a number of physical, biophysical and engineering issues that can confound BOLD-fMRI studies. Interestingly, strong confounding effect of some statistical methods had not been much recognised until recently. Power et al. [2012] showed that residual head motion often leads to spurious correlations in functional connectivity MRI networks. While that study referred to resting state, some other recent studies pointed to a number of important statistical problems in task studies. Bennett and Miller [2010] discussed a number of possible factors confounding BOLD-fMRI task studies, including statistical methods. Eklund et al. [2012] showed that SPM's pre-whitening leads to many false positives, whereas Eklund et al. [2015] compared AFNI, FSL and SPM, and found that all packages lead to high familywise error rates in single subject task analyses. Even more importantly, Eklund et al. [2016] pointed to high familywise error rates in AFNI, FSL and SPM for group task analyses. As the investigator uses implementations of the statistical procedures that are available in different software packages, the choice of the package might become a confounder in the study [Poline et al., 2006, Carp, 2012, Bowring et al., 2018].

In spite of huge scientific interest in fMRI, Carmichael et al. [2017] reports there are few industry-sponsored clinical trials with sufficiently rigorous fMRI data for regulatory agencies like FDA and EMA to consider when reviewing an application for a new therapeutic. So far, no requests have been made to qualify fMRI as a drug development tool.

However, currently there is an increasing interest in making the interpretation of fMRI experiments more demonstrably robust. There are more and more method validation studies, guidelines regarding the reporting of fMRI results were recently introduced [Nichols et al., 2017], code sharing is becoming more and more popular [Gorgolewski and Poldrack, 2016] and data organisation standards are being adopted [Gorgolewski et al., 2016]. This will increase confidence in the interpretation of fMRI findings, probably greatly enhancing its utility as a major tool for investigating brain function.

1.4 Motivation for current work

Given the above mentioned concerns related to some of the statistical methods used in fMRI studies, there is need to further investigate the fMRI image processing pipelines used in popular software packages. This thesis addresses this need for task BOLD fMRI studies. In particular, this work validated some of the statistical methods available in

AFNI, FSL and SPM, the most popular packages used in fMRI research.

Eklund et al. [2012] showed problems related to modelling temporal autocorrelation in task fMRI studies: pre-whitening in SPM was shown to remove only part of the autocorrelated noise. This led to inflated false positive rates. However, that study only referred to specificity and SPM. Two other studies about pre-whitening investigated data corresponding to only one fMRI protocol [Woolrich et al., 2001, Lenoski et al., 2008]. Lack of studies investigating pre-whitening which would employ data representing a wide range of fMRI protocols, as well as lack of studies investigating pre-whitening across AFNI, FSL and SPM, gave rise to the first project described in this thesis.

Accurate modelling of the shape characteristics of the BOLD response, the co-called hemodynamic response function (HRF), is known to be crucial when analysing fMRI data due to possible large sensitivity benefits [Handwerker et al., 2004]. However, there have been very few studies comparing HRF models. Perhaps, the most relevant comparisons are Lindquist et al. [2007, 2009], but the conclusions of these studies were based on simulated data rather than acquired images. The second project of this thesis employed acquired data to investigate specificity and sensitivity resulting from the use of a number of widely-used HRF models from AFNI, FSL and SPM.

In the above mentioned project, among others, the canonical HRF model was compared against some alternative methods. The canonical model, which assumes fixed shape of the BOLD response both across brain regions and across subjects, is the most popular approach to model the hemodynamic response function and it is often used to compare subject populations which might differ in age. While most fMRI studies aim to investigate neural activity, the BOLD signal is a result of neural activity combined with neurovascular coupling, which is known to vary with age [D'Esposito et al., 2003, Wright and Wise, 2018]. It is crucial to know how age affects the HRF, since in some studies inferred neural differences could actually reflect vascular differences. However, previous studies investigating the impact of age on the task-evoked hemodynamic response function used small samples and had conflicting results. For example, D'Esposito et al. [1999] and Grinband et al. [2017] did not find significant differences in the HRF shape between younger and older subjects, whereas Buckner et al. [2000] and West et al. [2018] did. The third project of this thesis used data from 641 healthy individuals sampled approximately uniformly from 18-88 years of age and investigated the impact of age on the task-evoked BOLD response, as well as on the BOLD-derived physiological parameters of the balloon model.

1.5 Thesis structure

Chapter 2 presents a validation study of the temporal autocorrelation modelling approaches used in AFNI, FSL and SPM. Implications both for first and for second level studies are investigated. Chapter 3 compares - in terms of specificity-sensitivity trade-offs - hemodynamic response function models available in AFNI, FSL and SPM. Again, both first and second level analyses are performed. Chapter 4 investigates the relationship between the hemodynamic response function and subject's age. The biophysical balloon model is used to investigate whether values of BOLD-derived physiological parameters vary with age and whether these physiological parameters are linked to Magnetoencephalography (MEG)-derived measures, as well as to cardiovascular health markers. Chapter 5 discusses the results and presents ways fMRI studies could lead to more reliable findings.

Chapter 2

Comparison of pre-whitening methods*

2.1 Introduction

Functional Magnetic Resonance Imaging (fMRI) data is known to be positively autocorrelated in time [Bullmore et al., 1996]. These correlations arise from neural sources, but also from scanner-induced low-frequency drifts, respiration and cardiac pulsation, as well as from movement artefacts not accounted for by motion correction [Lund et al., 2006]. If this autocorrelation is not accounted for, spuriously high fMRI signal at one time point can be affecting the subsequent time points, which increases the likelihood of obtaining false positives in task studies [Purdon and Weisskoff, 1998]. As a result, parts of the brain might erroneously appear active during an experiment. The degree of temporal autocorrelation is different across different regions of the brain [Worsley et al., 2002]. In particular, autocorrelation in grey matter is stronger than in white matter and cerebrospinal fluid, but it also varies within grey matter.

AFNI [Cox, 1996], FSL [Jenkinson et al., 2012] and SPM [Penny et al., 2011], the most popular packages used in fMRI research, first remove the signal at very low frequencies (for example, using a high-pass filter), after which they estimate the residual temporal autocorrelation and remove it in a process called pre-whitening. In AFNI temporal autocorrelation is modelled voxel-wise. For each voxel, an autoregressive-moving-average ARMA(1,1) model is estimated. The two ARMA(1,1) parameters are estimated only on a discrete grid and are not spatially smoothed. For FSL, a Tukey taper is used to smooth

*This chapter is an extension of Olszowy et al. [2019]. The study was fully conducted by me, but the study design and the results were thoroughly discussed with Guy Williams, Catarina Rua and John Aston.

Table 2.1: Overview of the employed datasets.

Study	Experiment	Place	Design	No. subjects	Field [T]	TR [s]	Voxel size [mm]	Time points
FCP	resting state	Beijing	N/A	198	3	2	3.1x3.1x3.6	225
	resting state	Cambridge, US	N/A	198	3	3	3x3x3	119
NKI	resting state	Orangeburg, US	N/A	30	3	1.4	2x2x2	404
	resting state	Orangeburg, US	N/A	30	3	0.645	3x3x3	900
CRIC	resting state	Cambridge, UK	N/A	73	3	2	3x3x3.8	300
neuRosim	resting state	(simulated)	N/A	100	NA	2	3.1x3.1x3.6	225
NKI	checkerboard	Orangeburg, US	20s off+20s on	30	3	1.4	2x2x2	98
	checkerboard	Orangeburg, US	20s off+20s on	30	3	0.645	3x3x3	240
BMMR	checkerboard	Magdeburg	12s off+12s on	21	7	3	1x1x1	80
CRIC	checkerboard	Cambridge, UK	16s off+16s on	70	3	2	3x3x3.8	160
CamCAN	sensorimotor	Cambridge, UK	event-related	200	3	1.97	3x3x4.44	261

FCP = Functional Connectomes Project. NKI = Nathan Kline Institute. BMMR = Biomedical Magnetic Resonance. CRIC = Cambridge Research into Impaired Consciousness. CamCAN = Cambridge Centre for Ageing and Neuroscience. For the Enhanced NKI data, only scans from release 3 were used. Out of the 46 subjects in release 3, scans of 30 subjects were taken. For the rest, at least one scan was missing. For the BMMR data, there were 7 subjects at 3 sessions, resulting in 21 scans. For the CamCAN data, 200 subjects were considered only.

the spectral density estimates voxel-wise. These smoothed estimates are then additionally smoothed within tissue type. Woolrich et al. [2001] showed the applicability of the FSL's method in two fMRI protocols: with repetition time of 1.5 s and of 3 s, and with voxel size 4 x 4 x 7 mm^3. Repetition time (TR), which is the acquisition time difference between two consecutive volumes, corresponds to the temporal resolution at which the scan was acquired and largely affects the temporal characteristics of the fMRI signal [Eklund et al., 2012]. By default, SPM estimates temporal autocorrelation globally as an autoregressive AR(1) plus white noise process [Friston et al., 2002]. SPM has an alternative approach: `FAST`, but a literature review reveals only three studies which have used it [Todd et al., 2016, Bollmann et al., 2018, Corbin et al., 2018]. `FAST` uses a dictionary of covariance components based on exponential covariance functions [Corbin et al., 2018]. More specifically, the dictionary is of length $3p$ and is composed of p different exponential time constants along their first and second derivatives. By default, `FAST` employs 18 components ($p = 6$). Like SPM's default pre-whitening method, `FAST` is based on a global noise model.

Lenoski et al. [2008] compared several fMRI autocorrelation modelling approaches for one fMRI protocol (TR = 3 s, voxel size 3.75 x 3.75 x 4 mm^3). The authors found that the use of the global AR(1), of the spatially smoothed AR(1) and of the spatially smoothed FSL-like noise models resulted in worse whitening performance than the use of the non-spatially smoothed noise models. Eklund et al. [2012] showed that in SPM the shorter the TR, the more likely it is to get false positive results in first level (also known as single subject) analyses. It was argued that SPM often does not remove a substantial part of the autocorrelated noise. The relationship between shorter TR and increased false

positive rates was also shown for the case when autocorrelation is not accounted [Purdon and Weisskoff, 1998].

In this study I investigated the whitening performance of AFNI, FSL and SPM for a wide variety of fMRI protocols. I analysed both the default SPM's method and the alternative one: `FAST`. Furthermore, I analysed the resulting specificity-sensitivity trade-offs in first level fMRI results, and I investigated the impact of pre-whitening on second level analyses. I observed better whitening performance for AFNI and SPM tested with option `FAST` than for FSL and SPM. Imperfect pre-whitening heavily confounded first level analyses.

2.2 Data

In order to explore a range of parameters that may affect autocorrelation, I investigated 11 fMRI datasets (Table 2.1). These included resting state and task studies, healthy subjects and a patient population, different TRs, magnetic field strengths and voxel sizes. I also used anatomical MRI scans, as they were needed for the registration of brains to the MNI (Montreal Neurological Institute) atlas space. FCP [Biswal et al., 2010], NKI [Nooner et al., 2012] and CamCAN data [Shafto et al., 2014] are publicly shared anonymised data. Data collection at the respective sites was subject to their local institutional review boards (IRBs), who approved the experiments and the dissemination of the anonymised data. For the 1,000 Functional Connectomes Project (FCP), collection of the Beijing data was approved by the IRB of State Key Laboratory for Cognitive Neuroscience and Learning, Beijing Normal University; collection of the Cambridge data was approved by the Massachusetts General Hospital partners' IRB. For the Enhanced NKI Rockland Sample, collection and dissemination of the data was approved by the NYU School of Medicine IRB. For the analysis of an event-related design dataset, I used the CamCAN dataset (Cambridge Centre for Ageing and Neuroscience, `http://www.cam-can.org`). Ethical approval for the study was obtained from the Cambridgeshire 2 (now East of England - Cambridge Central) Research Ethics Committee. The study from Magdeburg, "BMMR checkerboard" [Hamid et al., 2015], was approved by the IRB of the Otto von Guericke University. The study of Cambridge Research into Impaired Consciousness (CRIC) was approved by the Cambridge Local Research Ethics Committee (99/391). In all studies all subjects or their consultees gave informed written consent after the experimental procedures were explained. One rest dataset consisted of simulated data generated with the `neuRosim` package in `R` [Welvaert et al., 2011].

2.2.1 Simulation details

One rest dataset consisted of simulated data generated with the `neuRosim` package in `R`. I used it to simulate 100 resting state scans. The `neuRosim` simulations account for white noise, temporal noise, low-frequency scanner-induced noise, physiological noise, task-related noise and spatial noise. Spatial noise captures spatial relationships in the data: that time series from voxels next to each other tend to be similar. The user specifies the weights of different noises. I arbitrarily chose a weight of 25% corresponding to white noise, a weight of 50% corresponding to temporal noise and a weight of 25% corresponding to spatial noise. For several other tested weights, I could not detect significant activation in any of the 100 simulated scans. `neuRosim` provides AR(m) models to account for temporal autocorrelation. The same model, in other words with the same parameters, is used for each voxel. I decided to generate the temporally autocorrelated noise with the help of an AR(1) model. For the simulation procedure, a 3-dimensional baseline image must be provided by the user. The voxel-wise means in the simulated scans are equal to this baseline image. I chose a subject from the "FCP Beijing" dataset, subject ID "sub98617", as the baseline subject. The baseline image used for the simulation was the average of the real scan over time. Scanning parameters are shown in Table 2.1. The number of time points was also chosen as in "FCP Beijing". For the real "FCP Beijing" scan, I arbitrarily chose a cuboidal region of interest, where I calculated the average parameter of voxel-wise AR(1) models. In the simulation procedure it was not possible to directly use the AR(1) parameter from the real "FCP Beijing" scan, as white noise and spatial noise influence the effective value of the parameter of the AR(1) model. That is why, a parameter for the `neuRosim`'s AR(1) model was found so that the resulting average AR(1) parameter in the simulated scans in the same cuboidal region of interest was very similar.

2.2.2 Data availability

FCP, NKI and CamCAN data are publicly shared anonymised data. CRIC and BMMR scans could be obtained from me upon request. The simulated data can be generated again using script https://github.com/wiktorolszowy/fMRI_temporal_autocorrelation/blob/master/simulate_4D.R.

2.3 Methods

For AFNI, FSL and SPM analyses, the preprocessing, brain masks, brain registrations to the 2 mm isotropic MNI atlas space, and multiple comparison corrections were kept consistent (Figure 2.1). This way I limited the influence of possible confounders on the results. In order to investigate whether my results are an artefact of the comparison approach used for assessment, AFNI, FSL and SPM were compared by investigating (1) the power spectra of the GLM residuals, (2) the Q-Q plots of the GLM residuals, (3) the spatial distribution of significant clusters, (4) the average percentage of significant voxels within the brain mask, and (5) the positive rate: proportion of subjects with at least one significant cluster. The power spectrum represents the variance of a signal that is attributable to an oscillation of a given frequency. When calculating the power spectra of the GLM residuals, I considered voxels in native space using the same brain mask for AFNI, FSL and SPM. For each voxel, I normalised the time series to have unit variance and calculated the power spectra as the square of the discrete Fourier transform. Without variance normalisation, different signal scaling across voxels and subjects would make it difficult to interpret power spectra averaged across voxels and subjects.

Apart from assuming dummy designs for resting state data as in recent studies [Eklund et al., 2012, 2015, 2016], I also assumed wrong (dummy) designs for task data, and I used resting state scans simulated using the `neuRosim` package in `R` [Welvaert et al., 2011]. I treated such data as null data. For null data, the positive rate is the familywise error rate, which was investigated in a number of recent studies [Eklund et al., 2012, 2015, 2016]. I use the term "significant voxel" to denote a voxel that is covered by one of the clusters returned by the multiple comparison correction.

The analyses employed AFNI 16.2.02, FSL 5.0.10 and SPM 12 (v7219). All the processing scripts needed to fully replicate the current study are at `https://github.com/wiktorolszowy/fMRI_temporal_autocorrelation`.

2.3.1 Preprocessing

Slice timing correction was not performed as part of my main analysis pipeline, since for some datasets slice timing information was not available. In each of the three packages I performed motion correction, which resulted in six parameters that I considered as confounders in the consecutive statistical analysis. As the 7T scans from the "BMMR checkerboard" dataset were prospectively motion corrected [Thesen et al., 2000], I did not perform motion correction on them. The "BMMR checkerboard" scans were also

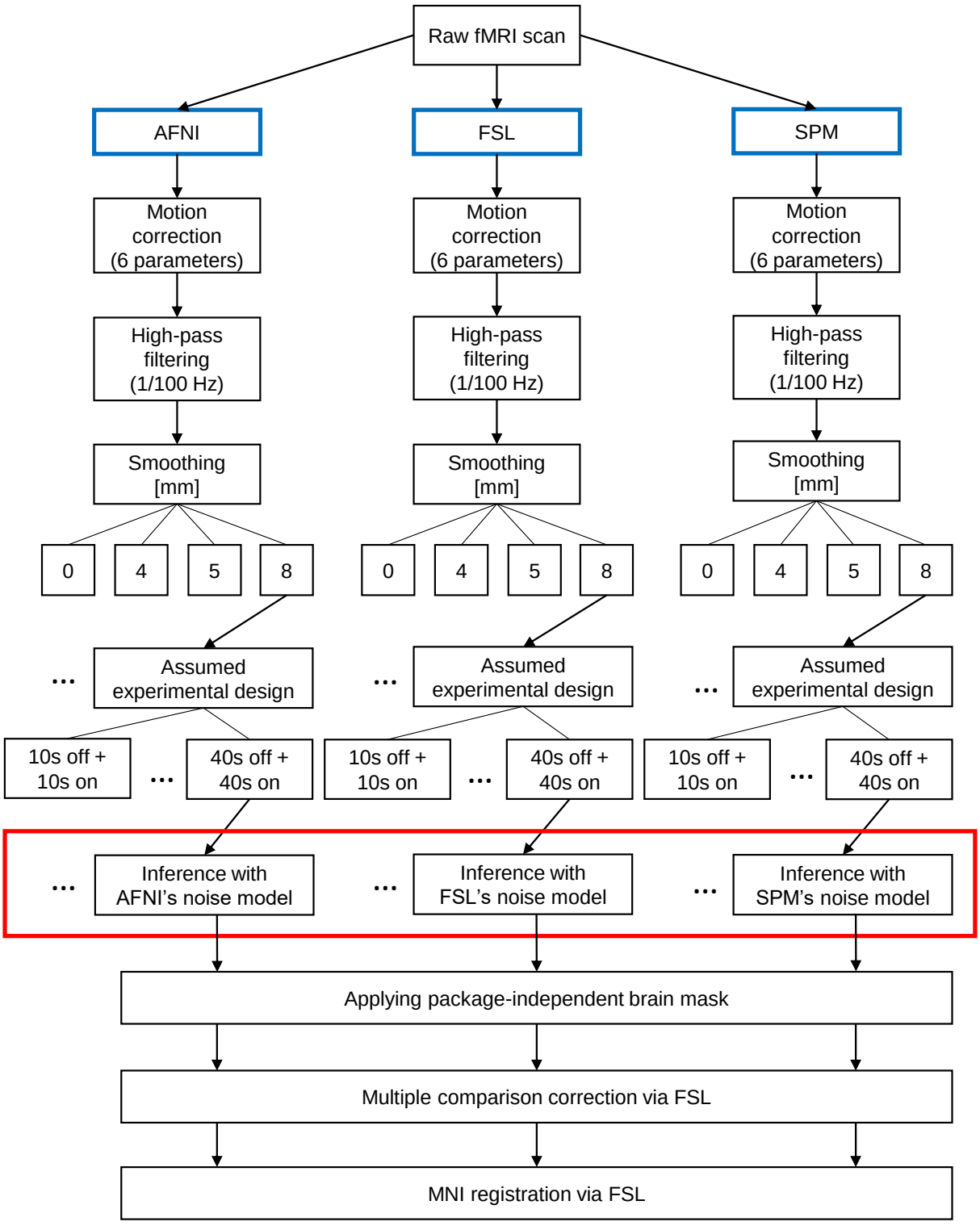


Figure 2.1: The employed analyses pipelines. For SPM, I investigated both the default noise model and the alternative noise model: `FAST`. The noise models used by AFNI, FSL and SPM were the only relevant difference (marked in a red box).

prospectively distortion corrected [In and Speck, 2012]. For all the datasets, in each of the three packages I conducted high-pass filtering with frequency cut-off of 1/100 Hz. I performed registration to MNI space only within FSL. For AFNI and SPM, the results of the multiple comparison correction were registered to MNI space using transformations generated by FSL. First, anatomical scans were brain extracted with FSL's brain extraction tool (BET) [Smith, 2002]. Then, FSL's boundary based registration (BBR) was used for registration of the fMRI volumes to the anatomical scans. The anatomical scans were aligned to 2 mm isotropic MNI space using affine registration with 12 degrees of freedom. The two transformations were then combined for each subject and saved for later use in all analyses, including in those started in AFNI and SPM. Gaussian spatial smoothing was performed in each of the packages separately.

2.3.2 Statistical analysis

For analyses in each package, I used the canonical hemodynamic response function (HRF) model, also known as the double gamma model. It is implemented the same way in AFNI, FSL and SPM: the response peak is set at 5 seconds after stimulus onset, while the post-stimulus undershoot is set at around 15 seconds after onset. This function was combined with each of the assumed designs using the convolution function. To account for possible response delays and different slice acquisition times, I used in the three packages the first derivative of the double gamma model, also known as the temporal derivative. I did not incorporate physiological recordings to the analysis pipeline, as these were not available for most of the datasets used.

I estimated the statistical maps in each package separately. AFNI, FSL and SPM use Restricted Maximum Likelihood (ReML), where autocorrelation is estimated given the residuals from an initial Ordinary Least Squares (OLS) model estimation. The ReML procedure then pre-whitens both the data and the design matrix, and estimates the model. I continued the analysis with the statistic maps corresponding to the t-test with null hypothesis being that the full regression model without the canonical HRF explains as much variance as the full regression model with the canonical HRF. All three packages produced brain masks. The statistic maps in FSL and SPM were produced within the brain mask only, while in AFNI the statistic maps were produced for the entire volume. I masked the statistic maps from AFNI, FSL and SPM using the intersected brain masks from FSL and SPM. I did not confine the analyses to a grey matter mask, because autocorrelation is at strongest in grey matter [Worsley et al., 2002]. In other words, false positives caused by imperfect pre-whitening can be expected to occur mainly in grey

matter. By default, AFNI and SPM produced t-statistic maps, while FSL produced both t- and z-statistic maps. In order to transform the t-statistic maps to z-statistic maps, I extracted the degrees of freedom from each analysis output.

Next, I performed multiple comparison correction in FSL for all the analyses, including for those started in AFNI and SPM. First, I estimated the smoothness of the brain-masked four-dimensional residual maps using the `smoothest` function in FSL. Knowing the `DLH` parameter, which describes image roughness, and the number of voxels within the brain mask (`VOLUME`), I then ran the `cluster` function in FSL on the z-statistic maps using a cluster defining threshold of 3.09 and significance level of 5%. This is the default multiple comparison correction in FSL and it refers to one-sided testing. Finally, I applied previously saved MNI transformations to the binary maps which were showing the location of the significant clusters.

2.4 Results

2.4.1 Whitening performance of AFNI, FSL and SPM

To investigate the whitening performance resulting from the use of noise models in AFNI, FSL and SPM, I plotted the power spectra of the GLM residuals. Figure 2.2 shows the power spectra averaged across all brain voxels and subjects for smoothing of 8 mm and assumed boxcar design of 10 s of rest followed by 10 s of stimulus presentation. The dips at 0.05 Hz are due to the assumed design period being 20 s (10 s + 10 s). For some datasets, the dip is not seen as the assumed design frequency was not covered by one of the sampled frequencies. The frequencies on the x-axis go up to the Nyquist frequency, which is 0.5/TR. The statistical inference in AFNI, FSL and SPM relies on the assumption that the residuals after pre-whitening are white. For white residuals, the power spectra should be flat. However, for all the datasets and all the packages, there was some visible structure. The strongest artefacts were visible for FSL and SPM at low frequencies. At high frequencies, power spectra from `FAST` were closer to 1 than power spectra from the other pre-whitening methods. Figure 2.2 does not show respiration-induced spikes which one could expect to see. This is because the figure refers to averages across subjects. I observed respiration-induced spikes when analysing power spectra for the first subject in each dataset (Figure 2.3). Figure 2.4 shows power spectra of the GLM residuals for the four task datasets when the true designs were assumed. Again, the most distinctive deviations of the power spectra from a flat line occurred for FSL and SPM at low frequencies. As pre-whitening could influence the distribution of the residuals, I also investigated Q-Q plots

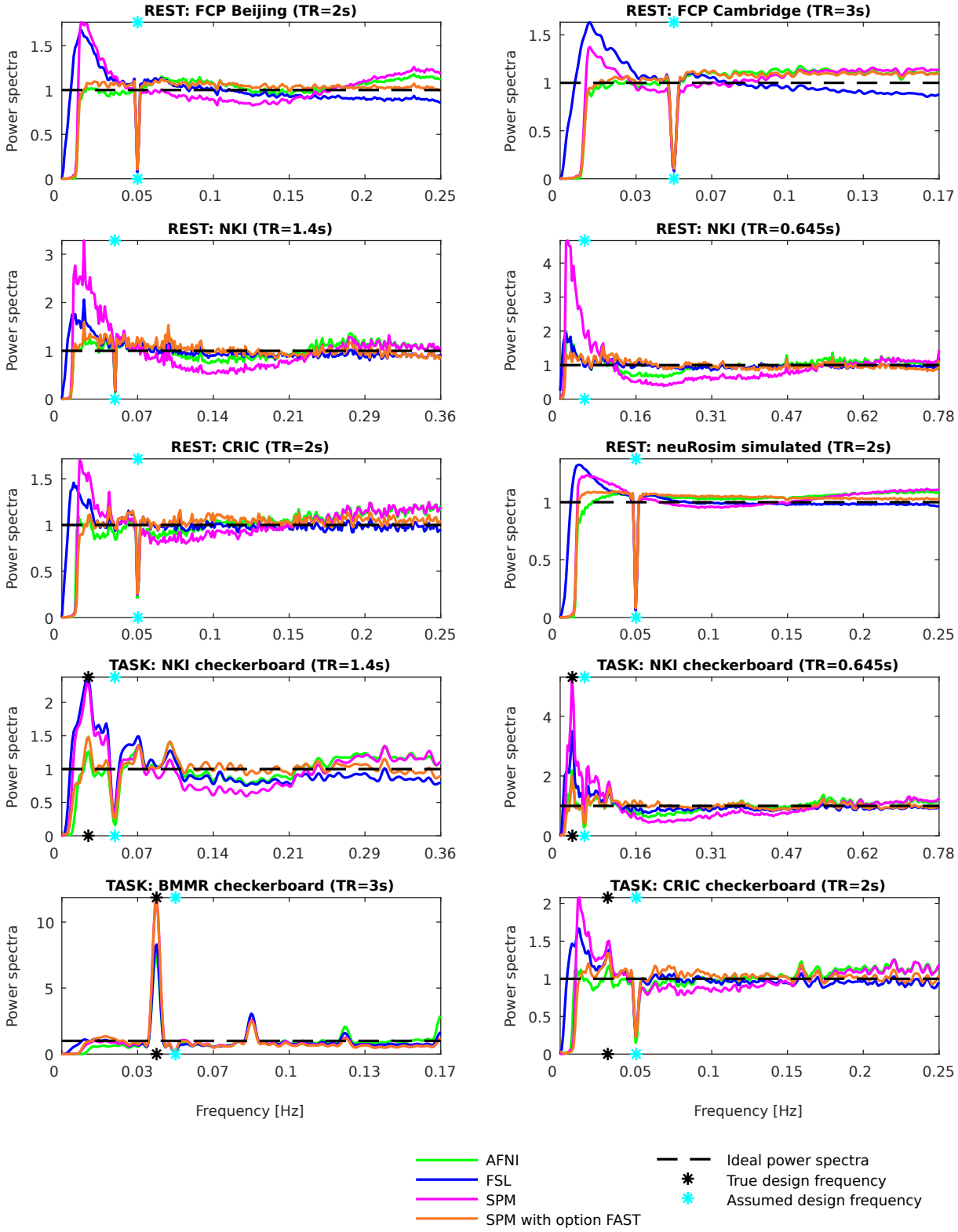


Figure 2.2: Power spectra of the GLM residuals in native space averaged across brain voxels and across subjects for the assumed boxcar design of 10 s of rest followed by 10 s of stimulus presentation ("boxcar10"). Following smoothing with FWHM of **8 mm**.

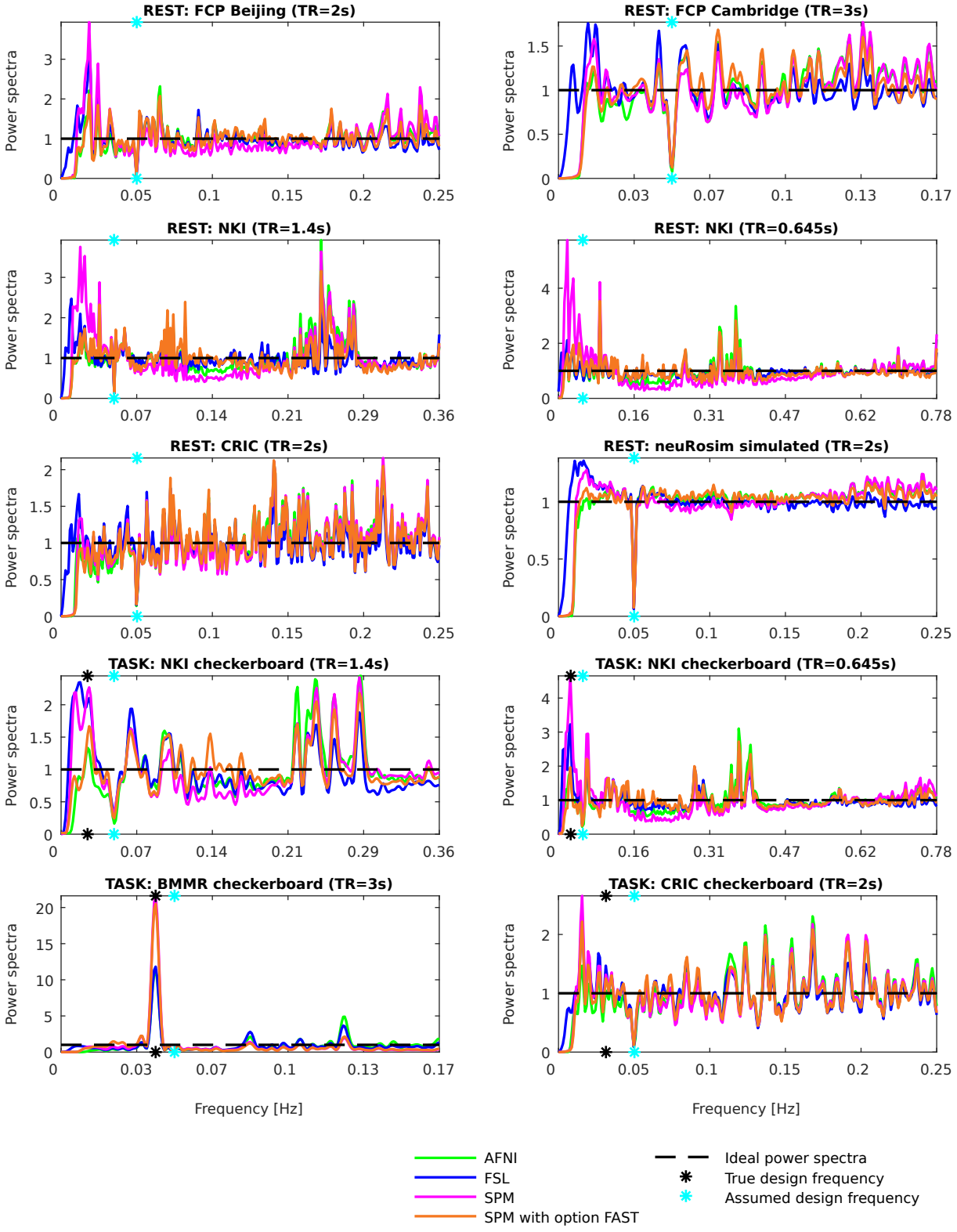


Figure 2.3: Power spectra of the GLM residuals in native space averaged across brain voxels for the first subject in each dataset and for the assumed boxcar design of 10 s of rest followed by 10 s of stimulus presentation ("boxcar10"). Following smoothing with FWHM of **8 mm**.

of the GLM residuals for the first subject in each of the 10 datasets (Figure 2.5). This way I compared the distribution of the GLM residuals to a normal distribution. For the resting state datasets, I assumed boxcar design of 10 s of rest followed by 10 s of stimulus presentation, while for the task datasets, I assumed the true designs. For all datasets, I observed substantial deviations from a normal distribution. On the other hand, different pre-whitening algorithms affected the distribution of the GLM residuals in a limited way only.

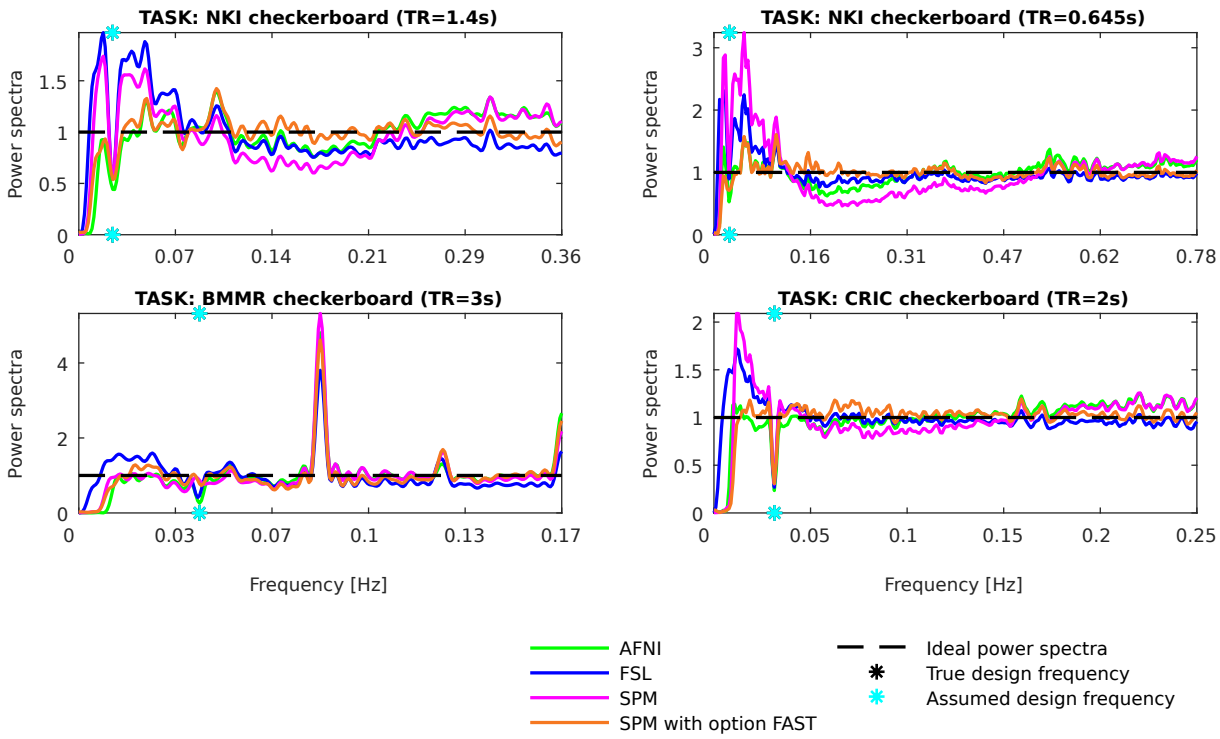


Figure 2.4: Power spectra of the GLM residuals in native space averaged across brain voxels and across subjects for the task datasets tested with the true designs. Following smoothing with FWHM of **8 mm**.

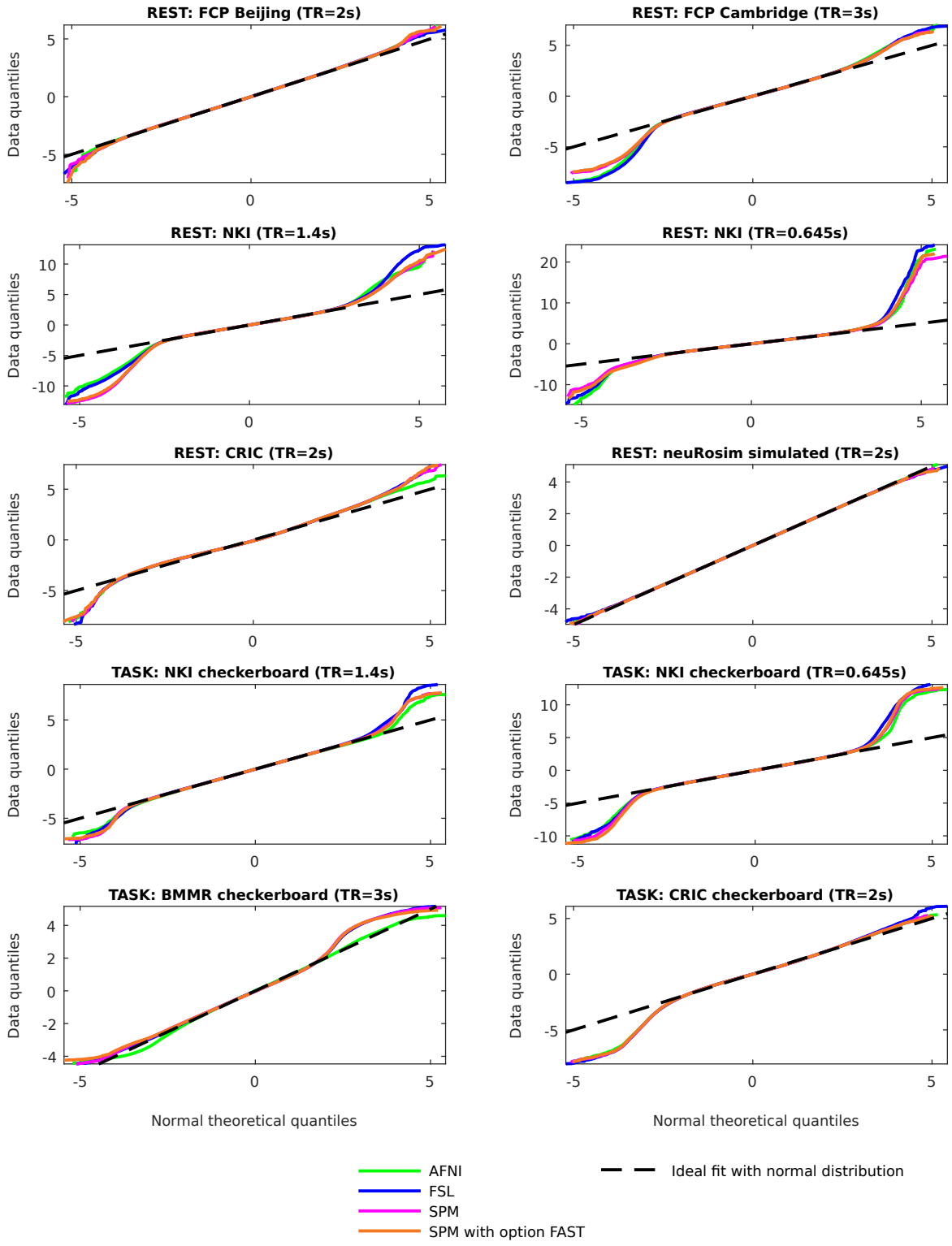


Figure 2.5: Q-Q plots of the GLM residuals in native space averaged across brain voxels for the first subject in each dataset. For the rest datasets, the "boxcar10" design was assumed, while for the task datasets, the true designs were assumed. Following smoothing with FWHM of **8 mm**.

2.4.2 Resulting specificity-sensitivity trade-offs

In order to investigate the impact of the whitening performance on first level results, I analysed the spatial distribution of significant clusters in AFNI, FSL and SPM. Figures 2.6-2.7 show an exemplary axial slice in the MNI space for 8 mm smoothing. It was made through the imposition of subjects' binarised significance masks on each other. Scale refers to the percentage of subjects within a dataset where significant activation was detected at the given voxel. The x-axis corresponds to four assumed designs. Resting state data was used as null data. Thus, low numbers of significant voxels were a desirable outcome, as this was suggesting high specificity. Task data with assumed wrong designs was used as null data too. Thus, clear differences between the true design (indicated with red boxes) and the wrong designs were a desirable outcome. The clearest cuts between the true and the wrong/dummy designs were obtained with AFNI's noise model and with `FAST`. For FSL and SPM, often the relationship between lower assumed design frequency ("boxcar40" vs. "boxcar12") and an increased number of significant voxels was visible, in particular for the resting state datasets: "FCP Beijing", "FCP Cambridge" and "CRIC". For null data, significant clusters in AFNI were scattered primarily within grey matter. For FSL and SPM, many significant clusters were found in the posterior cingulate cortex, while most of the remaining significant clusters were scattered within grey matter across the brain. False positives in grey matter occur due to the stronger positive autocorrelation in this tissue type compared to white matter [Worsley et al., 2002]. For the task datasets: "NKI checkerboard TR=1.4s", "NKI checkerboard TR=0.645s", "BMMR checkerboard" and "CRIC checkerboard" tested with the true designs, the majority of significant clusters were located in the visual cortex. This resulted from the use of visual experimental designs for the fMRI task. For the impaired consciousness patients ("CRIC"), the registrations to MNI space were imperfect, as the brains were often deformed.

The above analysis referred to the spatial distribution of significant clusters on an exemplary axial slice. As the results can be confounded by the comparison approach, I additionally investigated two other comparison approaches: the percentage of significant voxels and the positive rate. Since smoothing implicitly affects the voxel size, I considered different smoothing kernel sizes. I chose 4, 5 and 8 mm, as these are the defaults in AFNI, FSL and SPM. No smoothing was also considered, as for 7T data this preprocessing step is sometimes avoided [Walter et al., 2008, Polimeni et al., 2017]. Figures 2.8-2.9 show the average percentage of significant voxels across subjects in 10 datasets for smoothing of 4 mm and 8 mm, and for 16 assumed boxcar experimental designs. Resting state data was used as null data. Thus, a low percentage of significant voxels was a desirable outcome, as

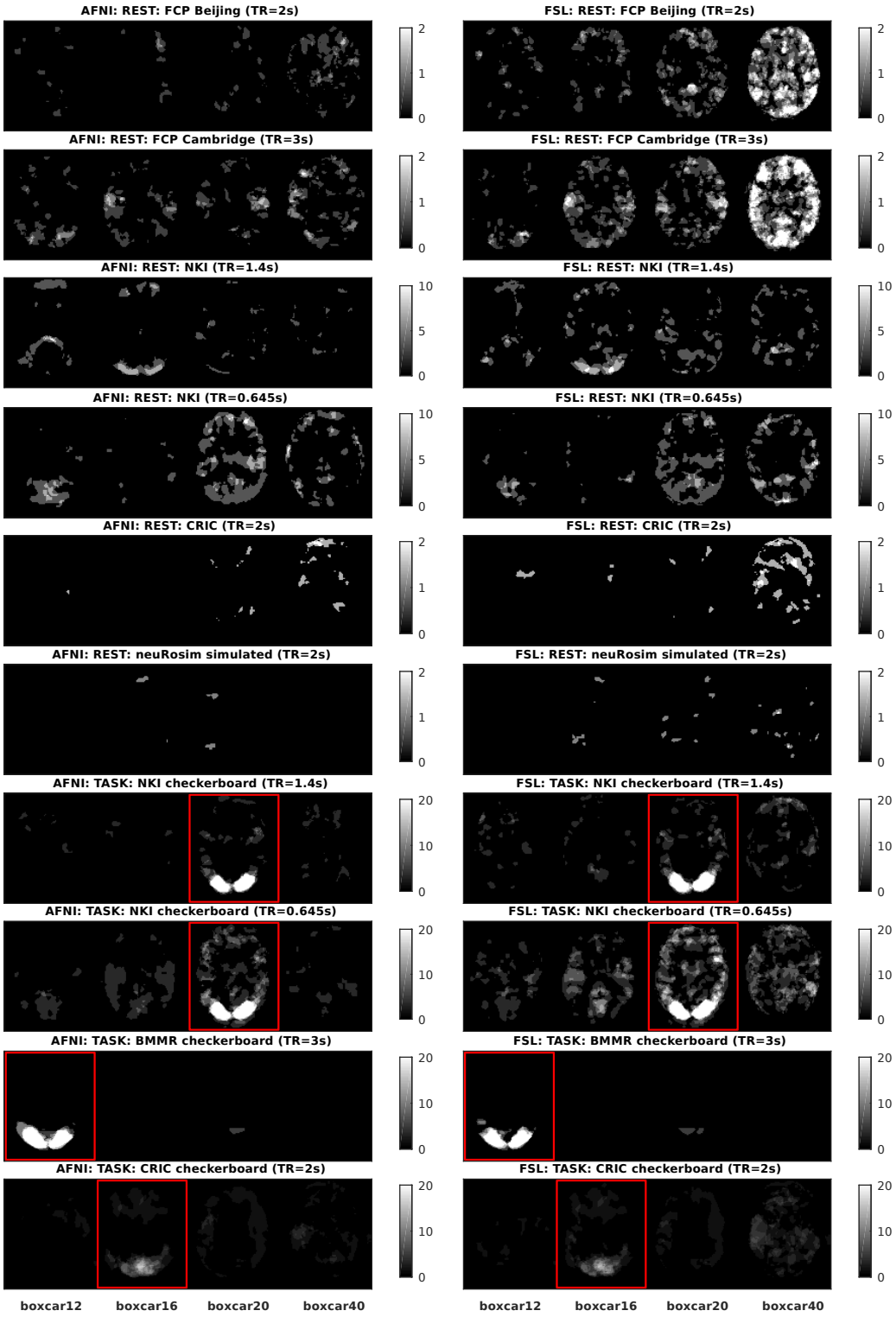


Figure 2.6: Spatial distribution of significant clusters for **AFNI (left)** and **FSL (right)**. On the x-axis the assumed experimental designs are listed. Scale refers to the percentage of subjects where significant activation was detected at the given voxel. The red boxes indicate the true designs (for task data). Following smoothing with FWHM of **8 mm**.

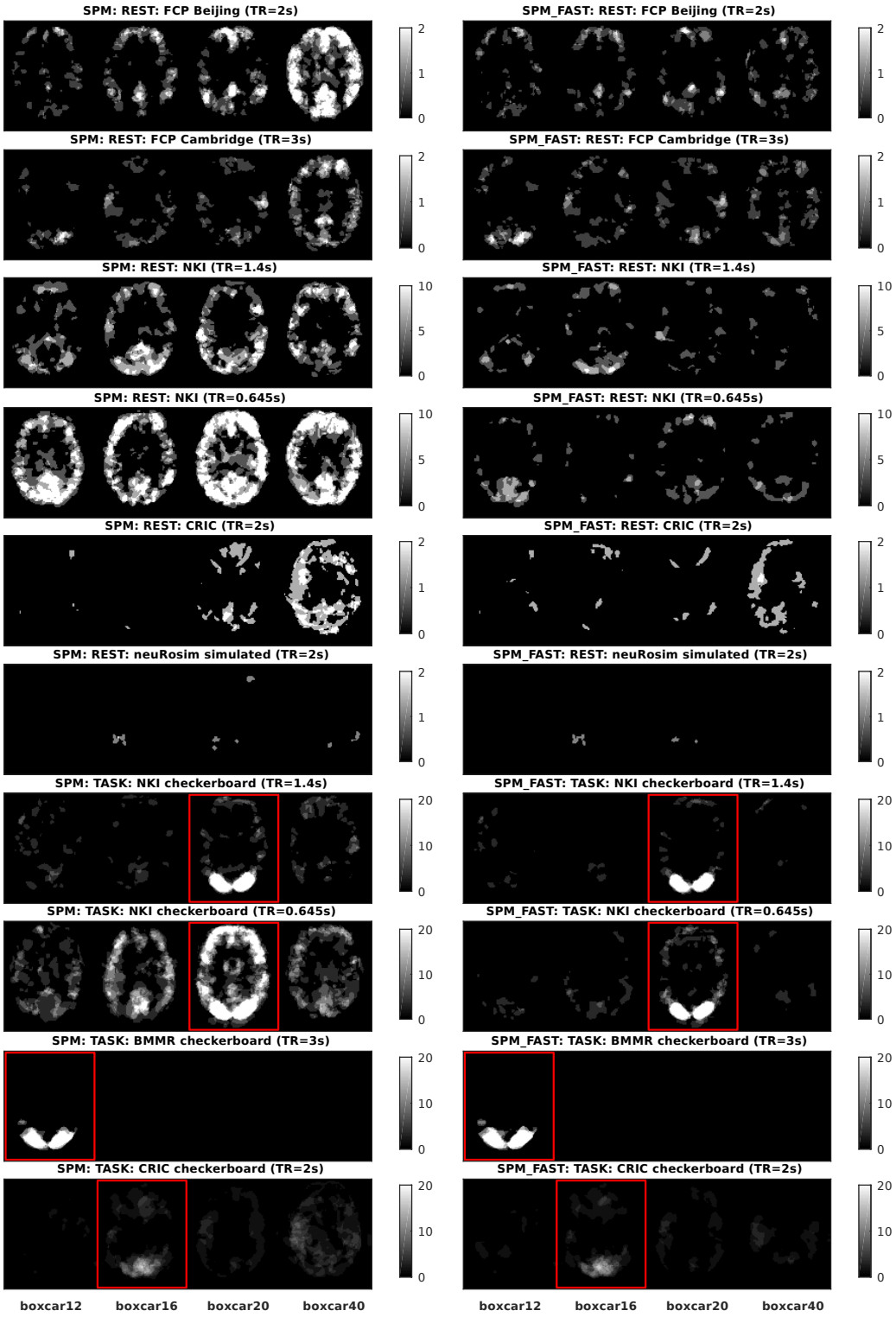


Figure 2.7: Spatial distribution of significant clusters for **SPM (left)** and **FAST (right)**. On the x-axis the assumed experimental designs are listed. Scale refers to the percentage of subjects where significant activation was detected at the given voxel. The red boxes indicate the true designs (for task data). Following smoothing with FWHM of **8 mm**.

it was suggesting high specificity. Task data with assumed wrong designs was used as null data too. Thus, large positive differences between the true design and the wrong designs were a desirable outcome. As more designs were considered, the relationship between lower assumed design frequency and an increased percentage of significant voxels in FSL and SPM (discussed before for Figures 2.6-2.7) was even more apparent. This relationship was particularly interesting for the "CRIC checkerboard" dataset. When tested with the true design, the percentage of significant voxels for AFNI, FSL, SPM and `FAST` was similar: for example, at an 8 mm smoothing level, 1.2%, 1.2%, 1.5% and 1.3%, respectively. However, AFNI and `FAST` returned much lower percentages of significant voxels for the assumed wrong designs. For the assumed wrong design "40" and an 8 mm smoothing level, FSL and SPM returned on average a higher percentage of significant voxels than for the true design: 1.4% and 2.2%, respectively. Results for AFNI and `FAST` for the same design showed only 0.3% and 0.4% of significantly active voxels.

Overall, AFNI and `FAST` outperformed FSL and SPM showing a lower average percentage of significant voxels in tests with the wrong designs: on average across 10 datasets, at an 8 mm smoothing level and across the wrong designs, the average percentage of significant voxels was 0.4% for AFNI, 0.9% for FSL, 1.9% for SPM and 0.4% for `FAST`. The percentage of significant voxels following 8 mm of smoothing was much higher than following 4 mm of smoothing.

Figures 2.10-2.11 show the positive rate for smoothing of 4 and 8 mm. The general patterns resemble those already discussed for the percentage of significant voxels, with AFNI and `FAST` consistently returning lowest positive rates (familywise error rates) for resting state scans and task scans tested with wrong designs. For task scans tested with the true designs, the positive rates for the different pre-whitening methods were similar. With a wider smoothing kernel, the positive rate decreased. The black horizontal lines show the 5% false positive rate, which is the expected proportion of scans with at least one significant cluster if in reality there was no experimentally-induced signal in any of the subjects' brains. The dashed horizontal lines are the confidence intervals for the proportion of false positives. These were calculated knowing that variance of a Bernoulli(p) distributed random variable is $p(1-p)$. Thus, the confidence intervals were $0.05 \pm \sqrt{0.05 \cdot 0.95/n}$, with n denoting the number of subjects in the dataset.

As multiple comparison correction depends on the smoothness level of the residual maps, I also checked the corresponding differences between AFNI, FSL and SPM. The residual maps seemed to be similarly smooth. At an 8 mm smoothing level, the average geometric mean of the estimated FWHMs of the residual maps in x-, y-, and z-dimensions

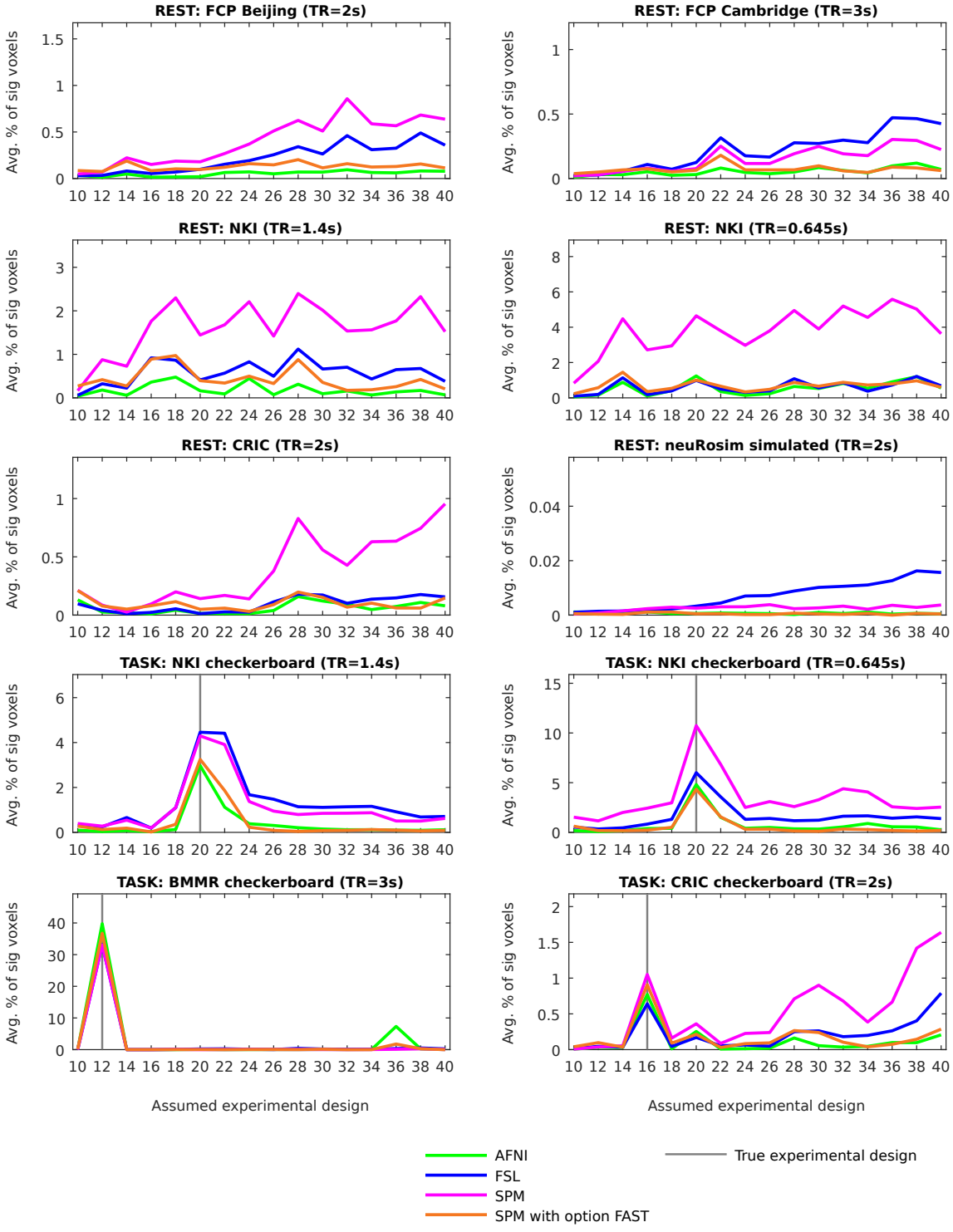


Figure 2.8: Average percentage of significant voxels across subjects for different packages. x-axis shows the assumed designs, e.g. "10" refers to the boxcar design of 10 s of rest followed by 10 s of stimulus presentation. Following smoothing with FWHM of **4 mm**.

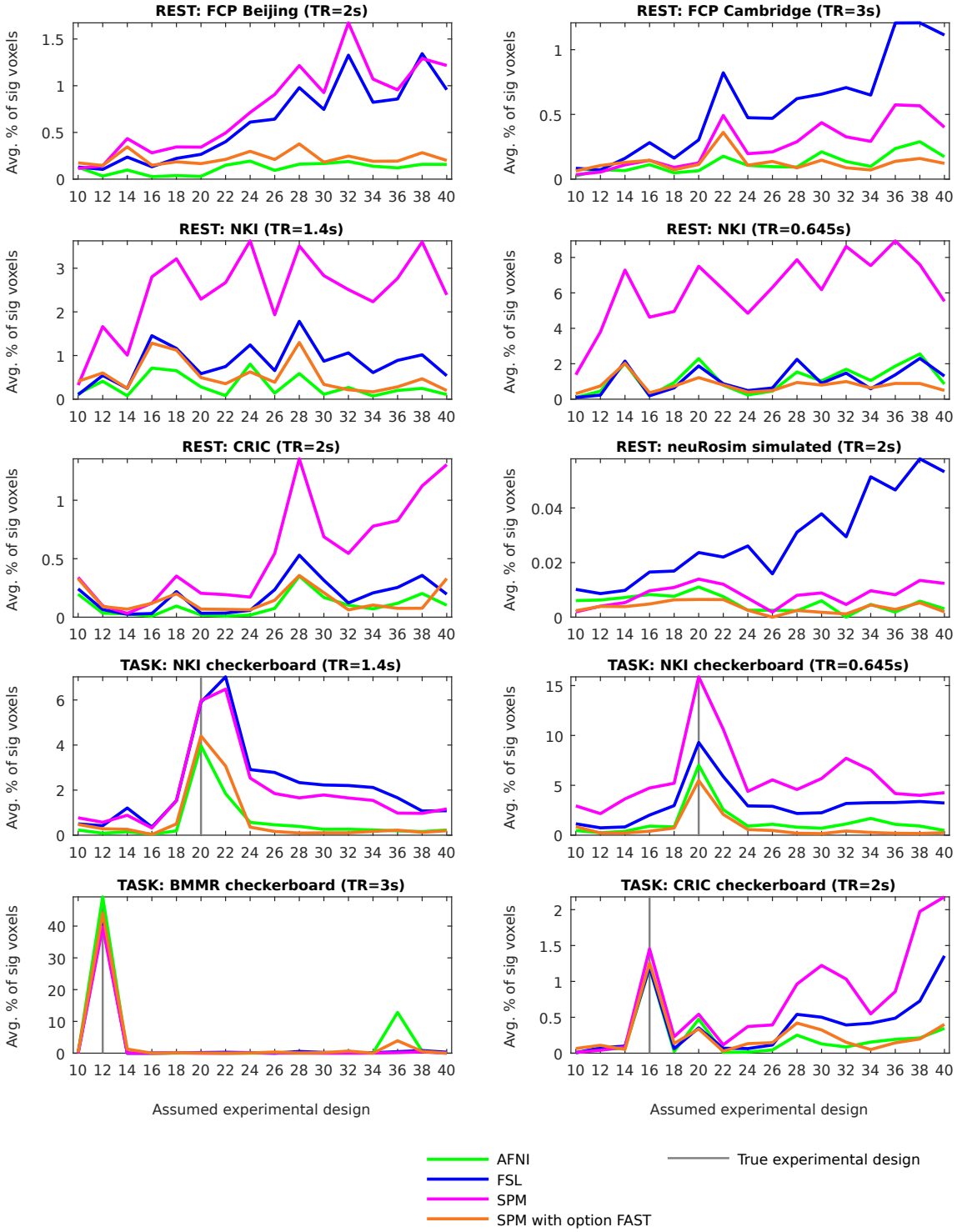


Figure 2.9: Average percentage of significant voxels across subjects for different packages. x-axis shows the assumed designs, e.g. "10" refers to the boxcar design of 10 s of rest followed by 10 s of stimulus presentation. Following smoothing with FWHM of **8 mm**.

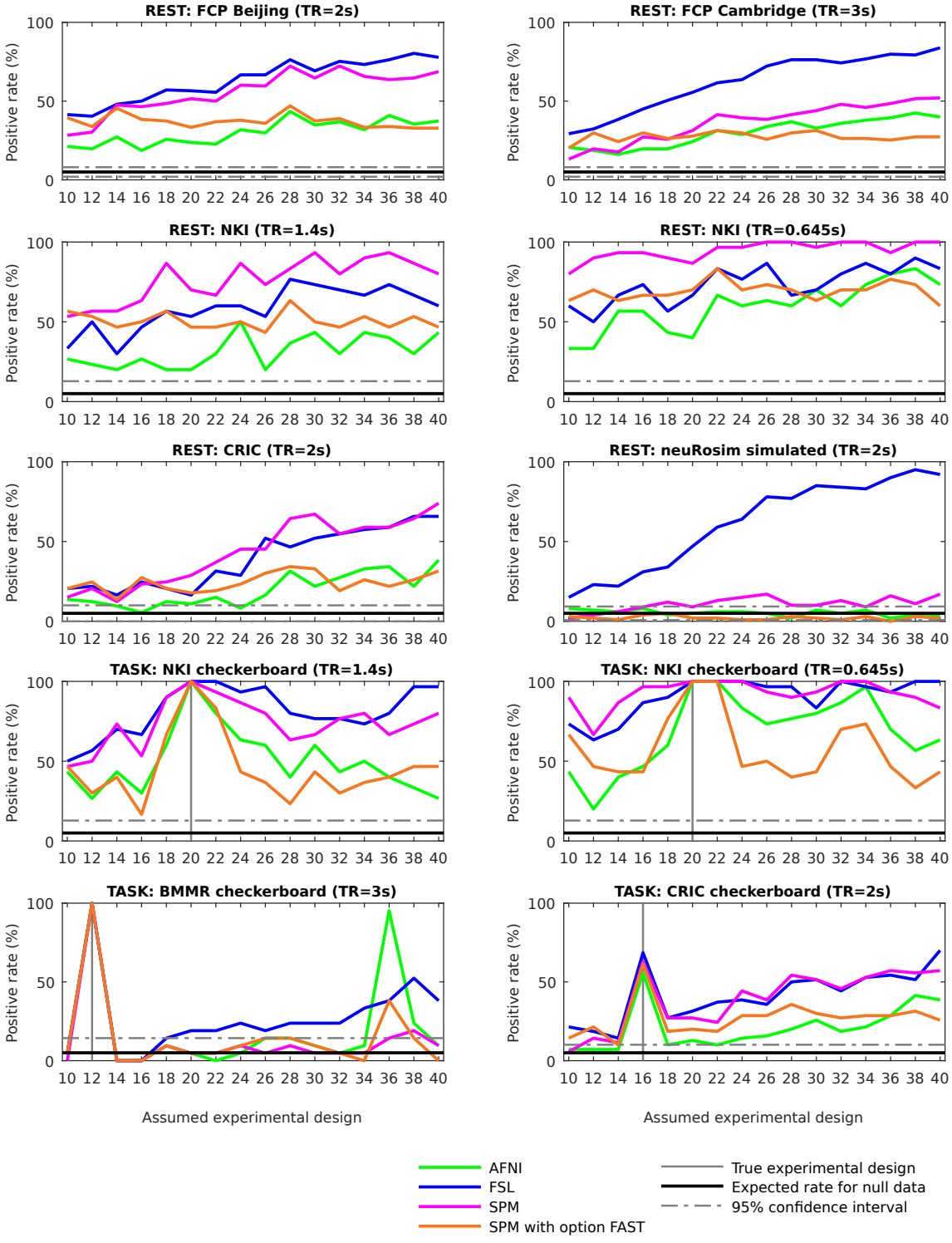


Figure 2.10: Positive rate for different packages. x-axis shows the assumed designs, e.g. "10" refers to the boxcar design of 10 s of rest followed by 10 s of stimulus presentation. Following smoothing with FWHM of **4 mm**. For null data, the positive rate is the familywise error rate.

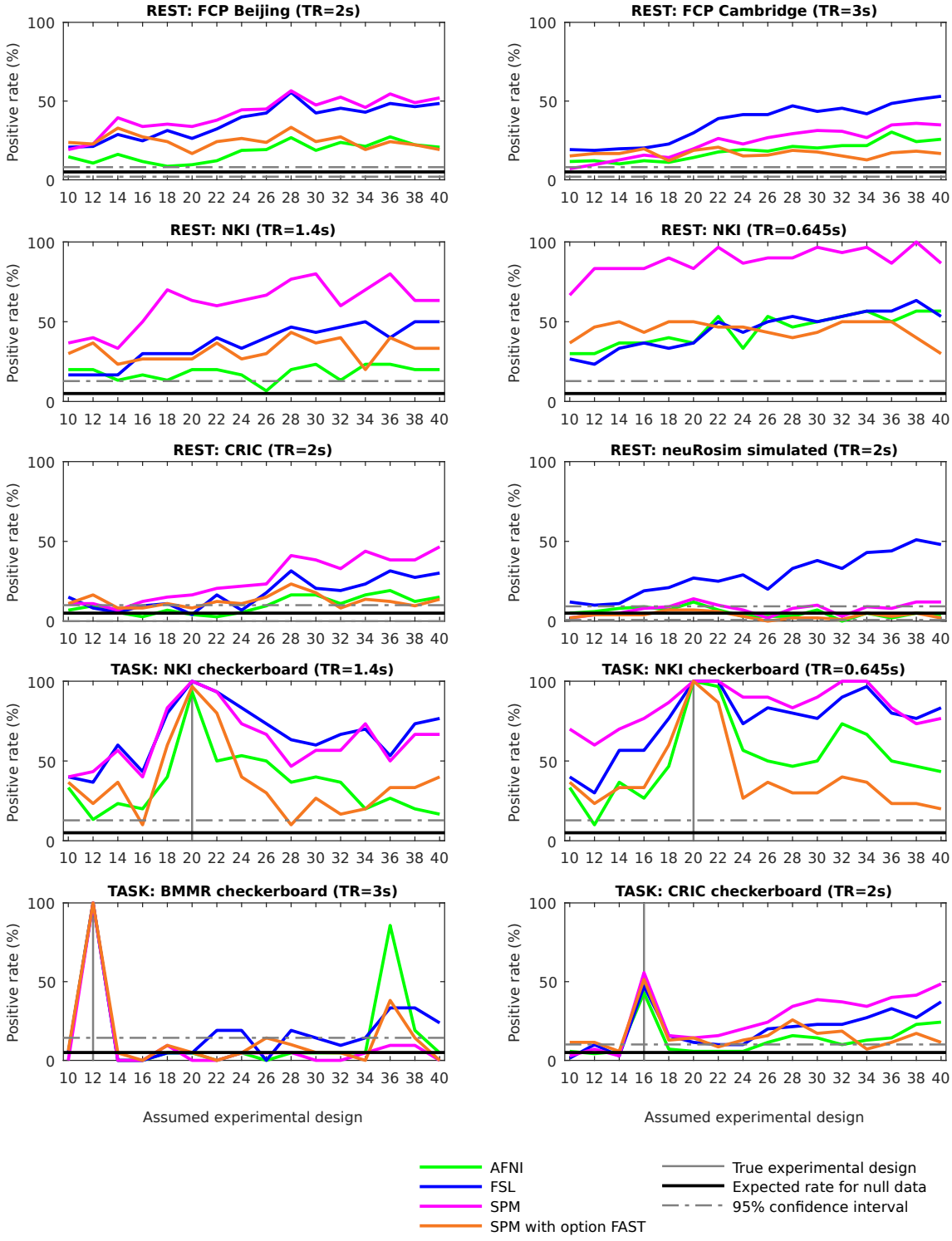


Figure 2.11: Positive rate for different packages. x-axis shows the assumed designs, e.g. "10" refers to the boxcar design of 10 s of rest followed by 10 s of stimulus presentation. Following smoothing with FWHM of **8 mm**. For null data, the positive rate is the familywise error rate.

across the 10 datasets and across the 16 assumed designs was 10.9 mm for AFNI, 10.3 mm for FSL, 12.0 mm for SPM and 11.8 mm for `FAST`. Moreover, I investigated the percentage of voxels with z-statistic above 3.09. This value is the 99.9% quantile of the standard normal distribution and is often used as the cluster defining threshold. For null data, this percentage should be 0.1%. At an 8 mm smoothing level, the average percentage across the 10 datasets and across the wrong designs was 0.6% for AFNI, 1.2% for FSL, 2.1% for SPM and 0.4% for `FAST`.

Further results are located at https://github.com/wiktorolszowy/fMRI_temporal_autocorrelation/tree/master/figures.

2.4.3 Event-related design studies

In order to check if differences in autocorrelation modelling in AFNI, FSL and SPM lead to different first level results for event-related design studies, I analysed the CamCAN dataset. The task was a sensorimotor one with visual and audio stimuli, to which the participants responded by pressing a button. The design was based on an m-sequence [Buračas and Boynton, 2002]. Figure 2.12 shows (1) the power spectra of the GLM residuals in native space averaged across brain voxels and across subjects for the assumed true design ("E1"), (2) the average percentage of significant voxels for three wrong designs and the true design, (3) the positive rate for the same four designs, and (4) the spatial distribution of significant clusters for the assumed true design ("E1"). Only smoothing of 8 mm was considered. The dummy event-related design ("E2") consisted of relative stimulus onset times generated from a uniform distribution with limits 3 s and 6 s. The stimulus duration times were 0.1 s.

For the assumed low-frequency design ("B2"), AFNI's autocorrelation modelling led to the lowest familywise error rate as residuals from FSL and SPM again showed a lot of signal at low frequencies. However, residuals from SPM tested with option `FAST` were similar at low frequencies to AFNI's residuals. As a result, the familywise error rate was similar to AFNI. For high frequencies, power spectra from SPM tested with option `FAST` were more closely around 1 than power spectra corresponding to the standard three approaches (AFNI/FSL/SPM). For an event-related design with very short stimulus duration times (around zero), residual positive autocorrelation at high frequencies makes it difficult to distinguish the activation blocks from the rest blocks, as part of the experimentally-induced signal is in the assumed rest blocks. This is what happened with AFNI and SPM. As their power spectra at high frequencies were above 1, I observed for the true design a lower percentage of significant voxels compared to SPM tested with

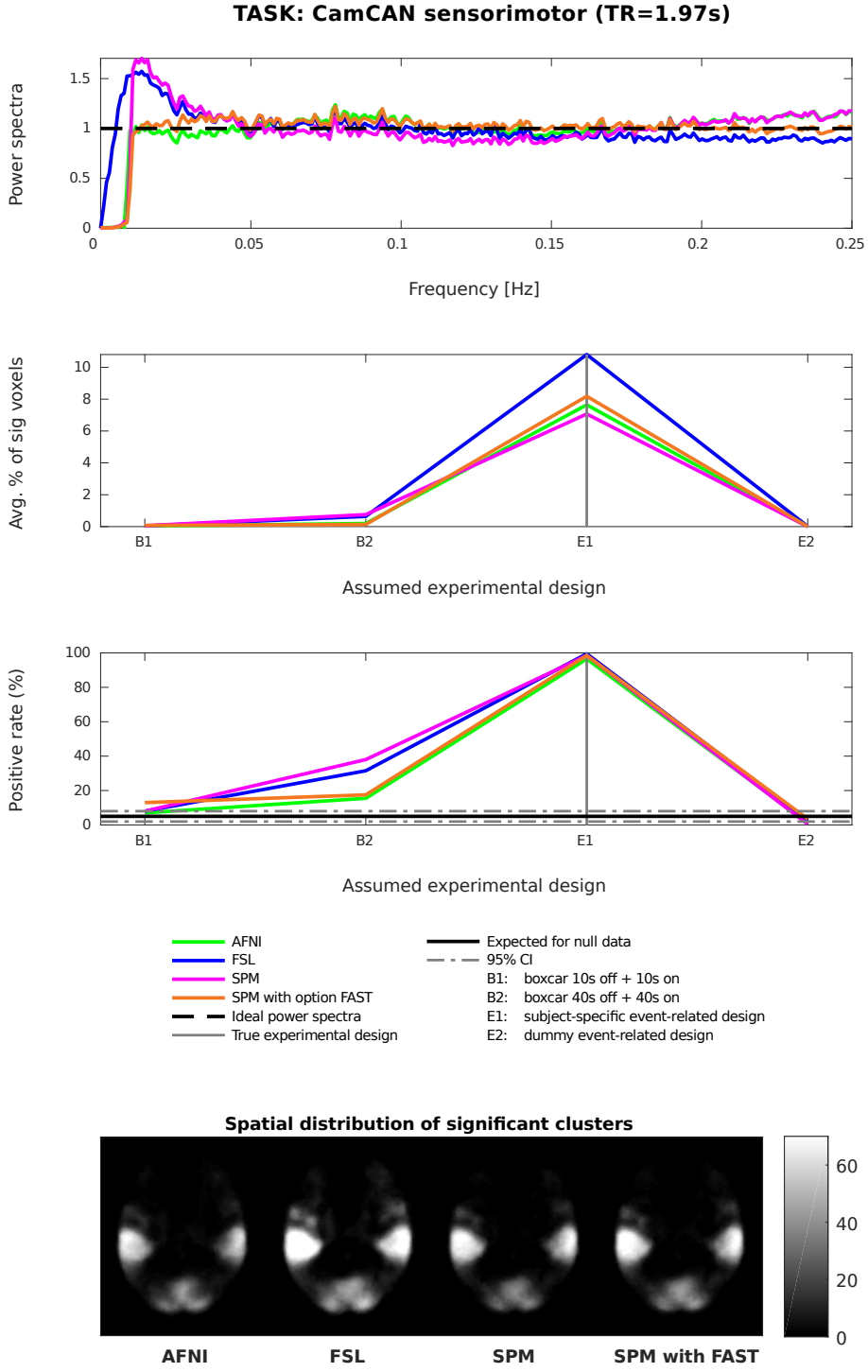


Figure 2.12: Differences between AFNI, FSL, SPM and `FAST` for the event-related design task dataset "CamCAN sensorimotor". From top to bottom: (1) power spectra of the GLM residuals in native space averaged across brain voxels and across subjects for the assumed true design ("E1"), (2) average percentage of significant voxels for three wrong designs and the true design, (3) positive rate for the same four designs, and (4) spatial distribution of significant clusters for the assumed true design ("E1") on an exemplary MNI axial slice. Following smoothing with FWHM of 8 mm.

option `FAST`. On the other hand, FSL's power spectra at high frequencies were below 1. As a result, FSL decorrelated activation blocks from rest blocks possibly introducing negative autocorrelations at high frequencies, leading to a higher percentage of significant voxels than SPM tested with option `FAST`. Though I do not know the ground truth, I might expect that AFNI and SPM led for this event-related design dataset to more false negatives than SPM with option `FAST`, while FSL led to more false positives. Alternatively, FSL might have increased the statistic values above their nominal values for the truly but little active voxels.

2.4.4 Slice timing correction

As slice timing correction is an established preprocessing step, which often increases sensitivity [Sladky et al., 2011], I analysed its impact on pre-whitening for two datasets for which I knew the acquisition order: "CRIC checkerboard" and "CamCAN sensorimotor". "CRIC checkerboard" scans were acquired with an interleave acquisition starting with the second axial slice from the bottom (followed with fourth slice, etc.), while "CamCAN sensorimotor" scans were acquired with a descending acquisition with the most upper axial slice being scanned first. I considered only the true designs. For the two datasets and for the four pre-whitening methods, slice timing correction changed the power spectra of the GLM residuals in a very limited way (Figure 2.13). Regardless of whether slice timing correction was performed or not, pre-whitening approaches from FSL and SPM left substantial positive autocorrelated noise at low frequencies, while `FAST` led to even more flat power spectra than AFNI. I also investigated the average percentage of significant voxels (Table 2.2). Slice timing correction changed the amount of significant activation only negligibly, with the exception of AFNI's pre-whitening in the "CamCAN sensorimotor" scans. In the latter case, the apparent sensitivity increase (from 7.64% to 13.45% of the brain covered by significant clusters) was accompanied by power spectra of the GLM residuals falling below 1 for the highest frequencies. This suggests negative autocorrelations were introduced at these frequencies, which could have led to statistic values being on average above their nominal values.

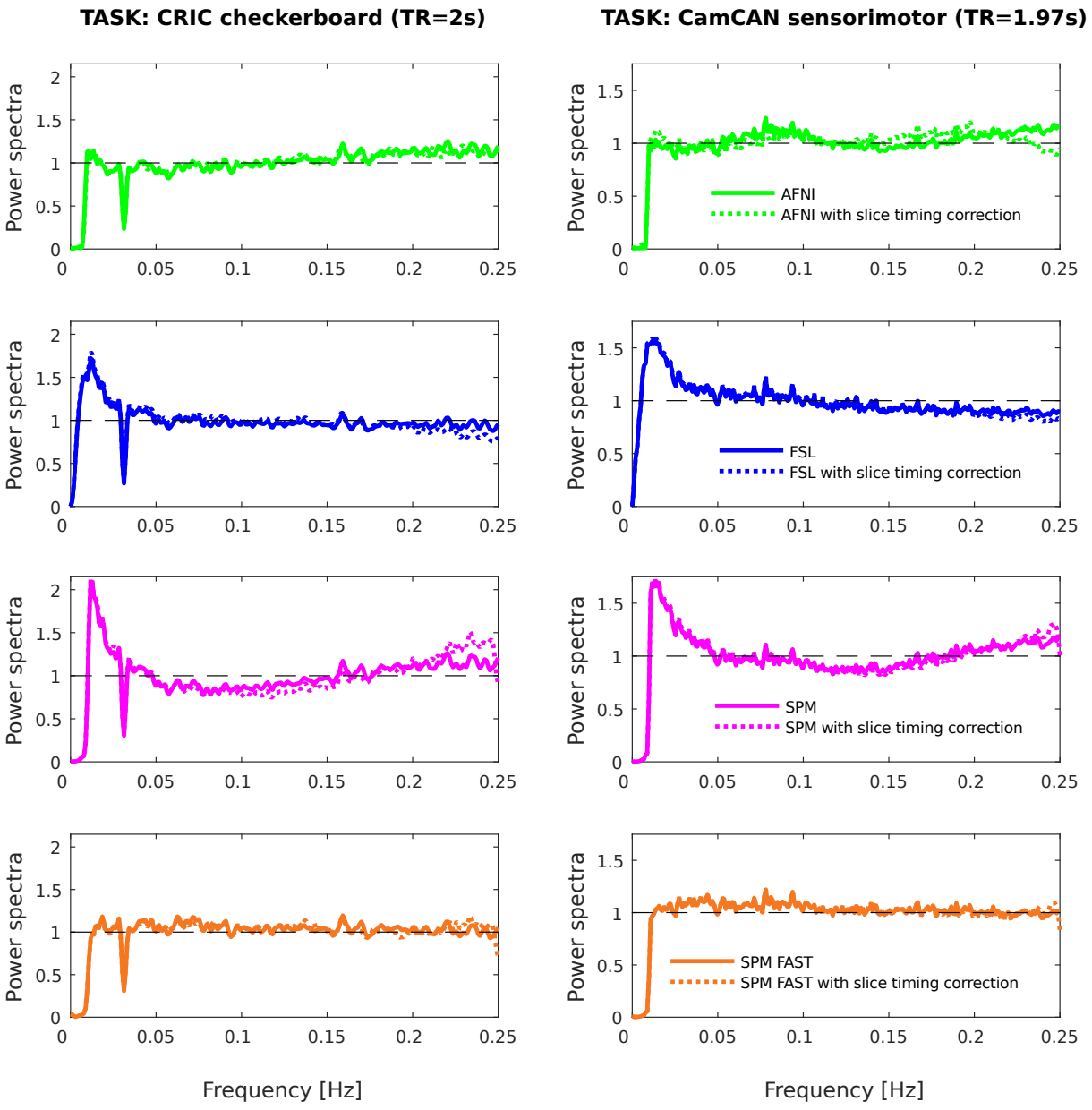


Figure 2.13: Power spectra of the GLM residuals in native space averaged across brain voxels and across subjects for the assumed true designs. Slice timing correction changed the power spectra in a very limited way. Following smoothing with FWHM of 8 mm.

CRIC checkerboard (TR=2s)		
Pre-whitening	No slice timing correction	Slice timing correction
AFNI	1.19%	1.08%
FSL	1.20%	1.24%
SPM	1.45%	1.35%
SPM with `FAST`	1.26%	1.12%

CamCAN sensorimotor (TR=1.97s)		
Pre-whitening	No slice timing correction	Slice timing correction
AFNI	7.64%	13.45%
FSL	10.80%	10.68%
SPM	7.07%	6.69%
SPM with `FAST`	8.18%	7.78%

Table 2.2: Average percentage of significant voxels across subjects for different packages. Results without slice timing correction are compared to results with slice timing correction. For each dataset, the true design was assumed. Following smoothing with FWHM of 8 mm.

2.4.5 Group studies

To investigate the impact of pre-whitening on the group level, I performed via SPM random effects analyses and via AFNI's `3dMEMA` [Chen et al., 2012] I performed mixed effects analyses. To be consistent with a previous study on group analyses [Eklund et al., 2016], I considered one-sample t-test with sample size 20. For each dataset, I considered the first 20 subjects. I exported coefficient maps and t-statistic maps (from which standard errors can be derived) following 8 mm spatial smoothing and pre-whitening from AFNI, FSL, SPM and `FAST`. Both for the random effects analyses and for the mixed effects analyses, I employed cluster inference with cluster defining threshold of 0.001 and significance level of 5%. Altogether, I performed 1312 group analyses: 2 (for random/mixed) $\times$ 4 (for pre-whitening) $\times$ $(10 \times 16 + 4)$ (for the first 10 datasets tested with 16 boxcar designs each and for the 11th dataset tested with four designs).

For each combination of group analysis model and pre-whitening (2×4), I ran 164 analyses. As five datasets were task datasets, 159 analyses ran on null data. Table 2.3 shows familywise error rate (FWER) for the random effects and mixed effects null data analyses, and for the four pre-whitening approaches. On average, FWER for the mixed effects analyses was almost twice higher than FWER for the random effects analyses. The use of AFNI's pre-whitening led to highest FWER, while `FAST` led to lower FWER than the SPM's default approach.

Pre-whitening	FWER for random effects	FWER for mixed effects
AFNI	15.72%	29.56%
FSL	9.43%	17.61%
SPM	11.95%	18.87%
SPM with `FAST`	8.81%	16.35%

Table 2.3: Familywise error rate (FWER) for the SPM's random effects model (summary statistic approach) and for the AFNI's mixed effects model (`3dMEMA`) following the use of noise models from AFNI, FSL, SPM and `FAST`. FWER was estimated as the number of null data group analyses with any significant result, divided by the number of null data group analyses (159 for each of the 8 combinations of the group analysis type and of the pre-whitening). Following smoothing with FWHM of 8 mm.

Figure 2.14 shows the percentage of significant voxels for four task datasets with assumed true designs. Results for the "CRIC checkerboard" dataset are not shown, as no significant clusters were found at the group level. This occurred due to several of the subjects having deformed brains, which led to the group brain mask not covering the primary visual cortex. For the "BMMR checkerboard" dataset, the brain mask was limited mainly to the occipital lobe and the percentage relates to the field of view that was used. Both for the random effects analyses and for the mixed effects analyses, I observed little effect of pre-whitening. For task data tested with the true designs, I found only negligible differences between the random effects analyses and the mixed effects analyses.

Noteworthily, for the event-related task dataset "CamCAN sensorimotor" tested with the true design, the use of `FAST` led to slightly higher amount of significant activation compared to the default SPM's method, while FSL led to much higher amount of significant activation. This means that for this event-related design dataset, the sensitivity differences from the first level analyses propagated to the second level. This happened both for the random effects model and for the mixed effects model.

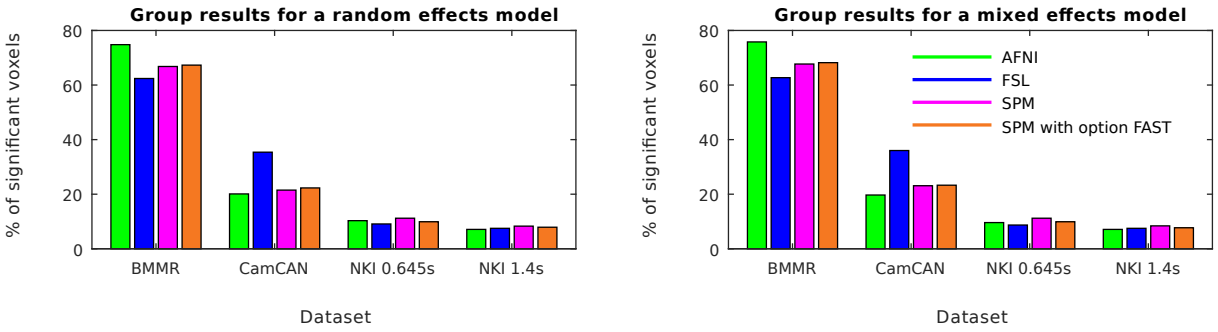


Figure 2.14: Group results for four task datasets with assumed true designs. Random effects analyses and mixed effects analyses led to only negligibly different average percentages of significant voxels. Following smoothing with FWHM of 8 mm.

As the above results suggest that the use of standard error maps changes the group results in a very limited way only, I investigated AFNI's `3dMEMA` by artificially re-scaling the t-statistic maps for one false positive analysis: "NKI rest (TR=1.4s)" dataset with assumed design 36s off + 36s on. For each subject, I multiplied the value of each voxel with 0.01, 0.1, 0.5, 2, 5 and 10. I observed a surprising negative relationship between the magnitude of the t-statistic maps and the amount of significant activation (Table 2.4). Even when the t-statistics were extremely small (standard errors 100 times bigger compared to the original values), `3dMEMA` found significant activation.

Pre-whitening	**Percentage of significant voxels as returned by 3dMEMA**						
	T × 0.01	T × 0.1	T × 0.5	T × 1	T × 2	T × 5	T × 10
AFNI	2.06%	2.06%	2.06%	1.69%	1.13%	0.81%	0.77%
FSL	0.92%	0.92%	0.92%	0.58%	0.37%	0.23%	0.23%
SPM	1.94%	1.94%	1.94%	0.97%	0.76%	0.82%	0.84%
SPM with `FAST`	0.9%	0.9%	0.9%	0.67%	0.61%	0.51%	0.53%

Table 2.4: Negative relationship between the magnitude of the t-statistic map and the amount of significant activation as returned by `3dMEMA`. The analyses were run on the NKI TR=1.4 s resting state scans with the assumed boxcar experimental design 36s off + 36s on. The re-scaling was done for each subject and each voxel, so that T × 0.01 means that the value of each voxel in the t-statistic map was multiplied for each subject with 0.01 before the mixed effects analysis was run. Following smoothing with FWHM of 8 mm.

2.5 Discussion

An analysis of the power spectra of the GLM residuals revealed whitening problems in FSL and in SPM (when using SPM's default method). While AFNI and SPM with option `FAST` performed particularly better for scans with short TRs, all considered fMRI protocols were affected.

In the case of FSL and SPM for the datasets "FCP Beijing", "FCP Cambridge", "CRIC RS" and "CRIC checkerboard", there was a clear relationship between lower assumed design frequency and an increased percentage of falsely significant voxels. This relationship exists when positive autocorrelation is not removed from the data [Purdon and Weisskoff, 1998]. It is caused by the spurious signal spillage. If during the assumed activation period the noise process spuriously takes high values and the assumed design frequency is high, due to the residual positive autocorrelation one can expect higher signal values during the beginning of the assumed rest period. Thus, it will be difficult to distinguish the assumed activation period from the assumed rest period, and the spuriously high signal during the

former period will likely not result in detected significance. On the other hand, if such a spuriously high signal occurs in the middle of a long assumed activation period, there will be enough time for the signal to return to its baseline level, so that there will be a larger difference between the mean signal during the assumed activation period and the mean signal during the assumed rest period. As a result, detection of significant activation will be more likely. Alternatively, the above phenomenon can be explained with regard to variances. Autocorrelated processes show increasing variances at lower frequencies. Thus, when the frequency of the assumed design decreases, the mismatch between the true autocorrelated residual variance and the incorrectly estimated white noise variance grows. In this mismatch, the variance is underestimated, which results in a larger number of false positives.

An interesting case was the checkerboard experiment conducted with impaired consciousness patients, where FSL and SPM found a higher percentage of significant voxels for the design with the assumed lowest design frequency than for the true design. As this subject population was unusual, one might suspect weaker or inconsistent response to the stimulus. However, positive rates for this experiment for the true design were all around 50%, substantially above other assumed designs.

Compared to FSL and SPM, the use of AFNI's and `FAST` noise models for task datasets resulted in larger differences between the true design and the wrong designs in the first level results. This occurred because of more accurate autocorrelation modelling in AFNI and in `FAST`. In my analyses, FSL and SPM left a substantial part of the autocorrelated noise in the data and the statistics were biased. For none of the pre-whitening approaches, were the positive rates around 5%, which was the significance level used in the cluster inference. This is likely due to imperfect cluster inference in FSL. High familywise error rates in first level FSL analyses were already reported [Eklund et al., 2015]. In my study the familywise error rate following the use of AFNI's and `FAST` noise models was consistently lower than the familywise error rate following the use of FSL's and SPM's noise models. Opposed to the average percentage of significant voxels, high familywise error rate directly points to problems in the modelling of many subjects.

In my main analysis pipeline I did not perform slice timing correction. For two datasets, I additionally considered slice timing correction and observed very similar first level results compared to the case without slice timing correction. The observed little effect of slice timing correction is likely a result of the temporal derivative being modelled within the GLM framework. This way a large part of the slice timing variation might have been captured without specifying the exact slice timing. For the only case where

slice timing correction led to noticeably higher amount of significant activation, I observed negative autocorrelations at high frequencies in the GLM residuals. If one did not see the power spectra of the GLM residuals, slice timing correction in this case could be thought to directly increase sensitivity, while in fact pre-whitening confounded the comparison.

2.5.1 Temporal and spatial resolution

The highly significant responses for the NKI datasets are in line with previous findings [Eklund et al., 2012], where it was shown that for fMRI scans with short TR it is more likely to detect significant activation. The NKI scans that I considered had TR of 0.645 s and 1.4 s, in both cases much shorter than the usual repetition times. Such short repetition times are now possible due to multiband sequences [Larkman et al., 2001]. The shorter the TR, the higher the correlations between adjacent time points [Purdon and Weisskoff, 1998]. If positive autocorrelation in the data is higher than the estimated level, then false positive rates will increase. The former study [Eklund et al., 2012] only referred to SPM. In addition to the previous study, I observed that the familywise error rate for short TRs was substantially lower in FSL than in SPM, though still much higher than for resting state scans at TR = 2 s ("FCP Beijing" and "CRIC RS"). FSL models autocorrelation more flexibly than SPM, which seems to be confirmed by my study. For short TRs, AFNI's performance deteriorated too, as most of the autocorrelation results from signal beyond one TR and an ARMA(1,1) noise model can only partially capture it.

Apart from the different TRs, I analysed the impact of spatial smoothing. If more smoothing is applied, the signal from grey matter will be often mixed with the signal from white matter. As autocorrelation in white matter is lower than in grey matter [Worsley et al., 2002], autocorrelation in a primarily grey matter voxel will likely decrease following stronger smoothing. The observed relationships of the percentage of significant voxels and of the positive rate from the smoothing level can be surprising, as random field theory is believed to account for different levels of data smoothness. The relationship for the positive rate (familywise error rate) was already known [Eklund et al., 2012, 2015]. The impact of smoothing and spatial resolution was investigated in a number of previous studies [Geissler et al., 2005, Weibull et al., 2008, Mueller et al., 2017]. I considered smoothing only as a confounder. Importantly, for all four levels of smoothing, AFNI and `FAST` outperformed FSL and SPM.

2.5.2 Links to previous studies

My results confirm Lenoski et al. [2008], insofar as my study also showed problems with SPM's default pre-whitening. Interestingly, Eklund et al. [2015] already compared AFNI, FSL and SPM in the context of first level fMRI analyses. AFNI resulted in substantially lower false positive rates than FSL and slightly lower false positive rates than SPM. I observed lowest false positive rates for AFNI too. Opposed to that study [Eklund et al., 2015], which compared the packages in their entirety, I compared the packages only with regard to pre-whitening. It is possible that pre-whitening is the most crucial single difference between AFNI, FSL and SPM, and that the relationships described in Eklund et al. [2015] would look completely different if AFNI, FSL and SPM employed the same pre-whitening. For one dataset, Eklund et al. [2015] also observed that SPM led to worst whitening performance.

The differences in first level results between AFNI, FSL and SPM which I found could have been smaller if physiological recordings had been modelled, for example, with the help of RETROICOR [Glover et al., 2000]. The modelling of physiological noise is known to improve whitening performance, particularly for short TRs [Lund et al., 2006, Bollmann et al., 2018, Corbin et al., 2018]. Unfortunately, cardiac and respiratory signals are not always acquired in fMRI studies. Even less often are the physiological recordings incorporated to the analysis pipeline. Interestingly, a recent report suggested that the FSL's tool ICA FIX applied to task data can successfully remove most of the physiological noise [Eklund et al., 2018]. This was shown to lower the familywise error rate. Such an approach corresponds to more accurate pre-whitening, although this was not mentioned in Eklund et al. [2018]. The use of independent components to remove artefacts in task fMRI data was also discussed in Kelly Jr et al. [2010].

2.5.3 How to explain pre-whitening problems in FSL and SPM?

FSL is the only package with a benchmarking paper of its pre-whitening approach [Woolrich et al., 2001]. The study employed data corresponding to two fMRI protocols. For one protocol, TR was 1.5 s, while for the other protocol, TR was 3 s. For both protocols, the voxel size was 4 x 4 x 7 mm^3. These were large voxels. I suspect that the FSL's pre-whitening approach could have been overfitted to this data. Regarding SPM, pre-whitening with simple global noise models was found to result in profound bias in at least two previous studies [Friston et al., 2000a, Lenoski et al., 2008]. SPM's default is a simple global noise model. However, SPM's problems could be partially related to the estimation procedure. Firstly, the estimation is approximative as it uses a Taylor expansion [Friston

et al., 2002]. Secondly, the estimation is based on a subset of the voxels. Only voxels with $p < 0.001$ following inference with no pre-whitening are selected. This means that the estimation strongly depends both on the TR and on the experimental design [Purdon and Weisskoff, 1998].

2.5.4 Impact on group studies

If the second level analysis is performed with a random effects model, the standard error maps are not used. Thus, random effects models like the summary statistic approach in SPM should not be affected by imperfect pre-whitening [Friston et al., 2005]. On the other hand, residual positive autocorrelated noise decreases the signal differences between the activation blocks and the rest blocks. This is relevant for event-related designs. Bias from confounded coefficient maps can be expected to propagate to the group level. I showed that pre-whitening indeed confounds group analyses performed with a random effects model. However, more relevant is the case of mixed effects analyses, for example, when using `3dMEMA` in AFNI [Chen et al., 2012] or `FLAME` in FSL [Woolrich et al., 2004a]. These approaches additionally employ standard error maps, which are directly confounded by imperfect pre-whitening. Bias in mixed effects fMRI analyses resulting from non-white noise at the first level was already reported in Bianciardi et al. [2004]. Surprisingly, I did not observe pre-whitening-induced specificity problems for analyses using `3dMEMA`, including for very short TRs. While this means that imperfect pre-whitening does not meaningfully affect group results when using `3dMEMA`, it is surprising that the AFNI's mixed effects model makes so little use of the standard error maps. For task datasets tested with the true designs, the results from random effects analyses differed very little compared to `3dMEMA` results. Furthermore, I observed for `3dMEMA` a worrying negative relationship between the magnitude of the t-statistic maps and the amount of significant activation. This is particularly surprising given that subject heterogeneity in that analysis was kept constant. I think `3dMEMA` does not always work as well as it was shown in the simulations in Chen et al. [2012]. In fact, Chen et al. [2012] compared `3dMEMA` with `FLAME` and found lower FWER for `3dMEMA` than for `FLAME`, although this conflicts with Eklund et al. [2016] (cf. Figure 1 in Eklund et al. [2016]).

`FLAME` was also shown to have similar sensitivity compared to random effects analyses [Mumford and Nichols, 2009]. However, mixed effects models should be more optimal than random effects models as they employ more information. Although group analysis modelling in task fMRI studies needs to be investigated further, it is beyond the scope of this work. As mixed effects models use standard errors, bias in them should be avoided.

2.5.5 Diagnostic plots

Unfortunately, although the vast majority of task fMRI analyses is conducted with linear regression, the popular analysis packages do not provide diagnostic plots. For old versions of SPM, the external toolbox `SPMd` generated them [Luo and Nichols, 2003]. It provided a lot of information, which paradoxically could have limited its popularity. I believe that task fMRI analyses would strongly benefit if AFNI, FSL and SPM provided some basic diagnostic plots. This way the investigator would be aware, for example, of residual autocorrelated noise in the GLM residuals. I provide a simple MATLAB tool for the fMRI researchers to check if their analyses might be affected by imperfect pre-whitening. It is available at https://github.com/wiktorolszowy/fMRI_temporal_autocorrelation/blob/master/plot_power_spectra_of_GLM_residuals.m.

2.5.6 Problems with motion correction

In the initial analyses, which this chapter does not present results from, I experienced problems with motion correction in SPM for the "BMMR checkerboard" dataset. Although scans in this dataset were prospectively motion corrected using the approach from Thesen et al. [2000], at first I additionally performed on them standard motion correction within AFNI, FSL and SPM. My rationale behind it was that no motion correction is perfect, so additionally applying retrospective motion correction could slightly improve motion correction performance and it would keep the processing pipeline the same across all the datasets. However, I found much less significant activation for the SPM analyses than for the AFNI and FSL analyses. Surprisingly, in SPM I found a lot of experimentally-induced activation in the motion regressors. These were used as confounders in the GLM analyses. As the statistical inference was based on t-test on the canonical function (rather than F-test on all regressors), SPM led to very little significant activation in my original analyses. Motion correction algorithms from AFNI and FSL did not lead to experimentally-induced activation in the motion regressors. It is surprising given study Oakes et al. [2005], where it was shown that different motion correction algorithms lead to only negligibly different analysis results, though a recent study found similar problems with SPM as I did [Yakupov et al., 2017]. In the latter study it was found that SPM's motion correction works much less accurately than AFNI and FSL for the special case of ultra high field data and limited acquisition field of view, a situation which was not covered in Oakes et al. [2005]. The "BMMR checkerboard" scans are ultra high field data and were acquired with a limited acquisition field of view. In the final analyses, which this chapter is based on, I did not employ retrospective motion correction for the "BMMR checkerboard" scans.

2.6 Conclusions

To conclude, I showed that AFNI and SPM tested with option `FAST` had the best whitening performance, followed by FSL and SPM. Pre-whitening in FSL and SPM left substantial residual autocorrelated noise in the data, primarily at low frequencies. Though the problems were most severe for short repetition times, different fMRI protocols were affected. I showed that the residual autocorrelated noise led to heavily confounded first level results. Low-frequency boxcar designs were affected the most. Due to better whitening performance, it was much easier to distinguish the assumed true experimental design from the assumed wrong experimental designs with AFNI and `FAST` than with FSL and SPM. This suggests superior specificity-sensitivity trade-off resulting from the use of AFNI's and `FAST` noise models. False negatives can occur when the design is event related and there is residual positive autocorrelated noise at high frequencies. In my analyses, such false negatives propagated to the group level both when using a random effects model and a mixed effects model, although only to a small extent. Surprisingly, pre-whitening-induced false positives did not propagate to the group level when using AFNI's mixed effects model `3dMEMA`. My results suggest that `3dMEMA` makes very little use of the standard error maps and does not differ much from the SPM's random effects model.

Results derived from FSL could be made more robust if a different autocorrelation model was applied. However, currently there is no alternative pre-whitening approach in FSL. For SPM, my findings support more widespread use of the `FAST` method.

Chapter 3

Comparison of HRF models*

3.1 Introduction

When neural activity is increased in one part of the brain, an increased amount of cerebral blood flow to that area can be expected, enabling the delivery of nutrients, including oxygen and glucose, to active tissues. This is the basis of the hemodynamic response (Figure 3.1). Different regions of the brain can respond differently to a stimulus, for example, following different hemodynamic response function (HRF) peak times [Lu et al., 2006, 2007, Badillo et al., 2013]. Saad et al. [2001] estimated that half of the BOLD response delay variance is the result of differences between brain regions, while the other half of this variance results from fMRI noise. Handwerker et al. [2004] showed that more variation occurs across subjects than across different brain regions.

There are populations where the hemodynamic response could differ due to neurological disorders [LeVan et al., 2010]. Importantly, the hemodynamic response could be also affected by the experimental design. Becker et al. [2011] and Scheeringa et al. [2011] used simultaneous Electroencephalography (EEG)-fMRI recordings to show that stimuli arriving at the peak of the alpha cycle result in lower BOLD response than stimuli which arrive at the trough of the cycle. This relationship was proved to exist in extrastriate, thalamic and cerebellar areas. Furthermore, Levin et al. [2001] showed that the blood hematocrit level (volume percentage of red blood cells in blood) influences the BOLD response too. Larger BOLD response was observed in subjects with higher baseline levels of hematocrit. Rombouts et al. [2005] showed HRF differences between healthy elderly subjects, mild cognitive impairment subjects and Alzheimer's disease patients. Turner

*Preliminary results of this study were published in Olszowy et al. [2018]. The study was fully conducted by me, but the study design and the results were thoroughly discussed with Guy Williams, Catarina Rua, John Aston and Richard Henson.

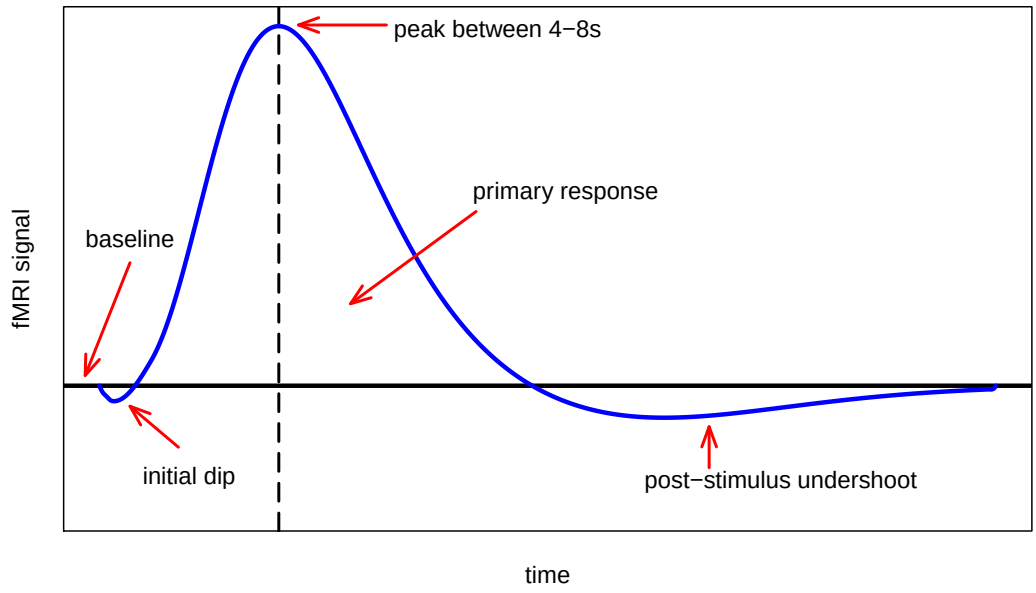


Figure 3.1: Blood flows to active tissue carrying, among others, oxygen and glucose. The resulting BOLD response resembles a wave. Hemodynamic response function models are used to capture the shape characteristics of the BOLD response.

et al. [2018] suggested that the canonicality of the HRF corresponds to the health of the neural-glial-vascular system, which is crucial for optimal cognitive performance. The authors showed that compared to multiple sclerosis subjects, healthy subjects displayed HRFs that were more similar to the canonical HRF.

Regardless of the large number of studies showing HRF variation both across different regions of the brain and across subjects, almost all fMRI studies are based on a fixed HRF model: the canonical model [Grinband et al., 2008, Monti, 2011]. If the experimentally-induced response was, for example, delayed for some subjects, a more flexible HRF model could lead to higher statistical sensitivity with which the experimentally-induced neural activity is detected [Handwerker et al., 2004, Loh et al., 2008]. Small HRF misestimates were found not to be serious for single-subject studies. However, for random effects analyses, even small misestimates like 1 s influence model parameter estimation [Handwerker et al., 2004]. Lindquist and Wager [2007] and Lindquist et al. [2009] compared a number of HRF models and showed that a superposition of three inverse logit functions performs best among these considered models, followed by a finite impulse response (FIR) model.

The current study investigated the performance of a number of HRF models which are available in AFNI [Cox, 1996], FSL [Jenkinson et al., 2012] and SPM [Penny et al., 2011]. In particular, I analysed the specificity-sensitivity trade-offs which result from the use of these HRF models.

3.2 Data

In order to cover a wide range of fMRI data, I investigated five fMRI task datasets (Table 3.1). These included healthy subjects and a patient population, different experimental designs, magnetic field strengths, TRs and voxel sizes. I also used anatomical MRI scans, as they were needed for the registration of brains to the MNI (Montreal Neurological Institute) atlas space. CamCAN (Cambridge Centre for Ageing and Neuroscience, http://www.cam-can.org, Shafto et al. [2014]) and Enhanced NKI (Nathan Kline Institute) data [Nooner et al., 2012] are publicly shared anonymised data. Data collection at the respective sites was subject to their local institutional review boards (IRBs), who approved the experiments and the dissemination of the anonymised data. For the CamCAN dataset, ethical approval for the study was obtained from the Cambridgeshire 2 (now East of England - Cambridge Central) Research Ethics Committee. The study of Cambridge Research into Impaired Consciousness (CRIC) was approved by the Cambridge Local Research Ethics Committee (99/391). For the Enhanced NKI Rockland Sample, collection and dissemination of the data was approved by the NYU School of Medicine IRB. The study from Magdeburg, "BMMR checkerboard" [Hamid et al., 2015], was approved by the IRB of the Otto von Guericke University. In all studies all subjects or their consultees gave informed written consent after the experimental procedures were explained.

Table 3.1: Overview of the employed datasets.

Study	Experiment	Place	Design	No. subjects	Field [T]	TR [s]	Voxel size [mm]	Time points
CamCAN	sensorimotor	Cambridge, UK	event-related	621	3	1.97	3x3x4.44	261
CRIC	checkerboard	Cambridge, UK	16s off+16s on	70	3	2	3x3x3.8	160
NKI	checkerboard	Orangeburg, US	20s off+20s on	30	3	1.4	2x2x2	98
	checkerboard	Orangeburg, US	20s off+20s on	30	3	0.645	3x3x3	240
BMMR	checkerboard	Magdeburg	12s off+12s on	21	7	3	1x1x1	80

CamCAN = Cambridge Centre for Ageing and Neuroscience. CRIC = Cambridge Research into Impaired Consciousness. NKI = Nathan Kline Institute. BMMR = Biomedical Magnetic Resonance. For the Enhanced NKI data, only scans from release 3 were used. Out of the 46 subjects in release 3, scans of 30 subjects were taken. For the rest, at least one scan was missing. For the BMMR data, there were 7 subjects at 3 sessions, resulting in 21 scans.

3.2.1 Data availability

CamCAN and NKI data are publicly shared anonymised data. CRIC and BMMR scans could be obtained from me upon request.

3.3 Methods

This chapter presents a specificity-sensitivity comparison of the HRF models listed in Table 3.2. The most popular HRF model used in fMRI studies is the canonical one, sometimes also called the double gamma model. Its use was first suggested in Glover [1999], where a previously discussed model based on a curve of a gamma distribution density function [Boynton et al., 1996] was extended by another gamma curve. The response peak of the canonical HRF is set at 5 seconds after stimulus onset, while its post-stimulus undershoot is set at around 15 seconds after onset. This HRF model is available in all the considered packages: AFNI, FSL and SPM. Friston et al. [1998b] introduced the temporal and dispersion derivatives, which are the partial derivatives of the canonical function with regard to time and with regard to duration, respectively. They are used along the canonical function to increase sensitivity. AFNI, FSL and SPM all enable the addition of the temporal derivative to the analysis pipeline, while AFNI and SPM additionally enable the addition of the dispersion derivative.

In FSL the default HRF model is the single gamma model, which reflects only the first gamma-like curve of the double gamma model. This fixed function can be accompanied by its temporal derivative, which is the default option in FSL. Furthermore, all three packages provide more flexible HRF models, in particular the Finite Impulse Response (FIR) model, where a number of bins are used to model the signal within a pre-specified post-stimulus period. For each voxel, the signal for each time bin is averaged across the trials. For FSL and SPM analyses, a FIR model with six time bins covering 18 s of the post-stimulus period was used. In AFNI there is no FIR model, but there are two closely-

Table 3.2: Overview of the employed HRF models.

Package	HRF model	Abbreviated as	No. of parameters
AFNI	double gamma	gam2	1
	double gamma with temporal derivative	gam2+T	2
	double gamma with temporal and dispersion derivatives	gam2+TD	3
	tent: variation of Finite Impulse Response	tent	6
	csplin: cubic spline function expansion of tent	csplin	6
FSL	double gamma	gam2	1
	double gamma with temporal derivative	gam2+T	2
	single gamma	gam1	1
	single gamma with temporal derivative	gam1+T	2
	Finite Impulse Response	FIR	6
SPM	double gamma	gam2	1
	double gamma with temporal derivative	gam2+T	2
	double gamma with temporal and dispersion derivatives	gam2+TD	3
	Fourier: windowed sine and cosine functions	Fourier	11
	Finite Impulse Response	FIR	6

related models: `tent` and `csplin`. The former is an extension of the FIR model to enable the estimated HRF to be a continuous rather than a step-wise function, while the latter is an extension of `tent` as it uses cubic splines to smooth the estimated HRF. Both `tent` and `csplin` were used in the current study with the same parameters as the FIR models in FSL and SPM: six parameters modelled 18 s of the post-stimulus period. Moreover, in SPM the Fourier set of order five was used to model HRF within approximately 24 s of the post-stimulus period. The use of this model resulted in 11 HRF-related covariates.

A literature review revealed that AFNI is the only package which was provided with a report on the available HRF models [Ward, 1998/2006], although this report is not complete. The choice of HRF models in the current study is hoped to reflect the particularly popular models used in fMRI studies.

The current study was primarily about differences between HRF models within AFNI, FSL and SPM. However, to enable comparisons between the packages, the processing pipelines for AFNI, FSL and SPM were aligned to each other in such a way that differences in results across the packages can be expected to be driven almost only by the use of the HRF model and the package-specific pre-whitening approach (Figure 3.2).

I compared the HRF models both at the first and at the second level by investigating (1) the spatial distribution of significant clusters and (2) the percentage of significant voxels within the brain mask. Additionally, for first level analyses, I investigated the positive rate: proportion of subjects with at least one significant cluster. I use the term "significant voxel" to denote a voxel that is covered by one of the clusters returned by the cluster inference. In order to investigate specificity, I followed Chapter 2 and Olszowy et al. [2019], and assumed wrong designs when analysing the task data. If for such null data, two HRF models perform comparably, while for task-based data tested with the true design, the use of one HRF model leads to detection of more significant activation, this could be treated as evidence of this HRF model being more sensitive.

For each of the five datasets, I employed three boxcar experimental designs: one where it was assumed that after 12 s of rest a stimulus was presented for 12 s, the second design assumed that after 16 s of rest a stimulus was presented for 16 s, while the third design was a boxcar 20 s off + 20 s on. Also, for each dataset, I employed two event-related experimental designs: one from the CamCAN sensorimotor task and one which consisted of relative stimulus onset times generated from a uniform distribution with limits 3 s and 6 s. The assumed stimulus duration time for the event-related designs was 0.1 s. For each dataset, one design was the true one, while the remaining ones were wrong.

Given previous work showing problems related to pre-whitening (Chapter 2 and

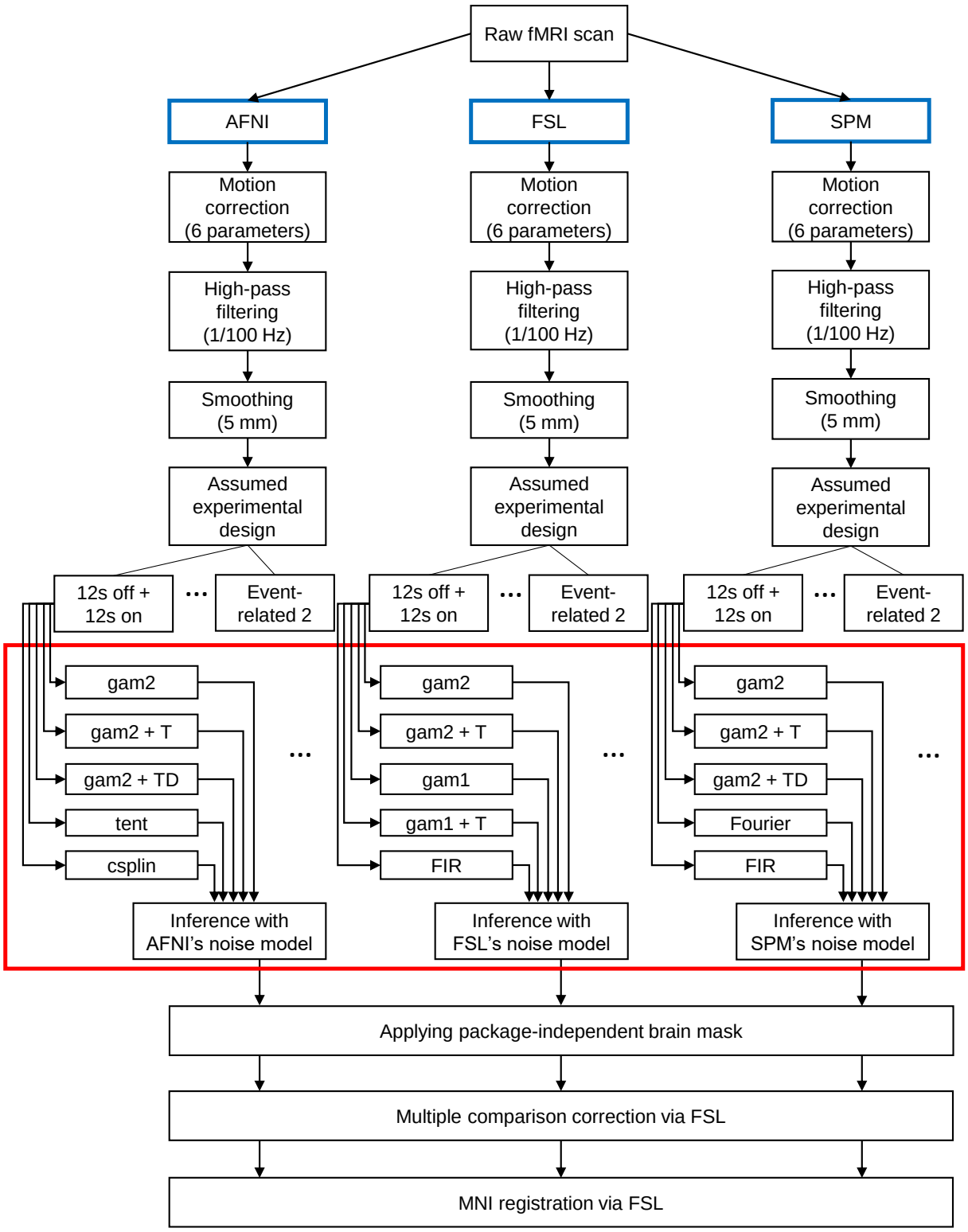


Figure 3.2: The employed analyses pipelines. HRF models and the noise models used by AFNI, FSL and SPM were the only relevant difference (marked in a red box).

Olszowy et al. [2019]), the whitening performance resulting from the use of the aforementioned HRF models was considered. Both the power spectra and the Q-Q plots of the GLM residuals were investigated. The power spectrum represents the variance of a signal that is attributable to an oscillation of a given frequency. When calculating the power spectra of the GLM residuals, I considered voxels in native space using the same brain mask for AFNI, FSL and SPM. For each voxel, I normalised the time series to have unit variance and calculated the power spectra as the square of the discrete Fourier transform. Without variance normalisation, different signal scaling across voxels and subjects would make it difficult to interpret power spectra averaged across voxels and subjects.

The analyses employed AFNI 18.0.11, FSL 5.0.10 and SPM 12 (v7219). All the processing scripts needed to fully replicate the current study can be found at https://github.com/wiktorolszowy/fMRI_HRFs_comparison.

3.3.1 Preprocessing

I performed slice timing correction for the "CamCAN sensorimotor", "CRIC checkerboard" and "BMMR checkerboard" datasets. For the NKI datasets, no slice timing correction was performed, as the slice timing information was not directly available. Besides, TRs of the NKI datasets were short: 0.645 s and 1.4 s, which decreases the possible sensitivity benefits of this correction. In each of the three packages I performed motion correction, which resulted in six parameters that I considered as confounders in the consecutive statistical analysis. As the 7T scans from the "BMMR checkerboard" dataset were prospectively motion corrected [Thesen et al., 2000], I did not perform motion correction on them. The "BMMR checkerboard" scans were also prospectively distortion corrected [In and Speck, 2012]. For all the datasets, in each of the three packages I conducted high-pass filtering with frequency cut-off of 1/100 Hz. I performed registration to MNI space only within FSL. For AFNI and SPM, the results of the multiple comparison correction were registered to MNI space using transformations generated by FSL. First, anatomical scans were brain extracted with FSL's brain extraction tool (BET) [Smith, 2002]. Then, FSL's boundary based registration (BBR) was used for registration of the fMRI volumes to the anatomical scans. The anatomical scans were aligned to 2 mm isotropic MNI space using affine registration with 12 degrees of freedom. The two transformations were then combined for each subject and saved for later use in all analyses, including in those started in AFNI and SPM. Gaussian spatial smoothing with full width at half maximum (FWHM) of 5 mm was performed in each of the packages separately.

3.3.2 Statistical analysis

For analyses in each package, I used five HRF models (Table 3.2). I did not incorporate physiological recordings to the analysis pipeline, as these were not available for most of the datasets used. I estimated the statistical maps in each package separately. For AFNI and FSL analyses, I employed default pre-whitening, while for SPM analyses, I used the `FAST` pre-whitening. This alternative method is more accurate than SPM's default (Chapter 2 and Olszowy et al. [2019]).

Statistical inference was conducted with the use of t- and F-tests. The former were used for the double gamma related models only, while F-tests were performed for all HRF models. All three packages produced brain masks. The statistic maps in FSL and SPM were produced within the brain mask only, while in AFNI the statistic maps were produced for the entire volume. I masked the statistic maps from AFNI, FSL and SPM using the intersected brain masks from FSL and SPM. I did not confine the analyses to a grey matter mask. In order to transform the t- and F-statistic maps to z-statistic maps, I extracted the degrees of freedom from each analysis output. First level analyses employed FSL's multiple comparison correction, while second level analyses employed SPM's random effects model. Both for first and for second level analyses, cluster inference was used with cluster defining threshold of 3.09 and significance level of 5%.

The t-tests referred to a null hypothesis that the full regression model without the canonical function explained as much variance as the full regression model with the canonical function. The F-tests referred to a null hypothesis that the full regression model without the HRF-related covariates explained as much variance as the full regression model with the HRF-related covariates. By default, t-tests in FSL and SPM are one-sided. In order to reliably compare the performance of t- and F-tests at the first level, the input of the FSL's function `cluster` for the t-test was the absolute value of the z-statistic map, and the significance level used by the `cluster` function for the t-test was adjusted to 2.5%. Importantly, this procedure reflects a two-sided rather than a bi-sided test [Chen et al., 2018a], as a single cluster could be comprised of both positive and negative effects. At the second level, t-tests were one-sided, so that comparisons between t- and F-tests could have been confounded. However, task fMRI studies are primarily expected to cause positive responses. To be consistent with a previous study on group analyses [Eklund et al., 2016], the group level analyses in the current study referred to the first 20 subjects in each dataset only. For first level analyses, I applied previously saved MNI transformations to the binary maps showing the location of the significant clusters. For group analyses, registration to MNI space was performed before the random effects analyses.

3.4 Results

3.4.1 Single subject analyses

Figure 3.3 shows the spatial distribution of significant clusters for the "CamCAN sensorimotor" dataset and the true design. In the upper part of the figure results for t-test on the canonical function are shown, while in the lower part there are results for F-test on all HRF-related covariates. For t-test, there was much less detected activation for the canonical model used along its two derivatives than for the canonical model tested alone. However, this relationship reversed when an F-test on all HRF-related covariates was used. For the "CRIC checkerboard" images, differences in the spatial distribution of significant clusters between HRF models were much smaller (Figure 3.4). For `tent` and `csplin`, the flexible HRF models in AFNI, less activation was detected than for the canonical basis sets. These two models returned less activation also for the NKI datasets (Figures 3.5-3.6). For the BMMR images, most perceived activation was observed for the extended canonical models following the use of the F-test (Figure 3.7).

Differences between HRF models were also seen when a wrong design was assumed: Figure 3.8 shows the spatial distribution of significant clusters for the "CamCAN sensorimotor" dataset and the randomised event-related design. Most of the false positives occurred for SPM. This likely results from accurate pre-whitening. As Chapter 2 showed, the use of `FAST` leads to power spectra being very close to one, which for an event-related design might lead to more false positives when compared to a pre-whitening method which left in the data positive residual autocorrelated noise at high frequencies. For the latter, it is more difficult to distinguish assumed activation blocks from assumed rest blocks.

Opposed to the above analyses, which referred to four exemplary MNI slices only and, with one exception, to the true designs, Figures 3.9-3.10 show the average percentage of significant voxels for all the considered designs, tested with t-test and F-test, respectively. For the true designs, the differences between HRF models are in line with observations made above for Figures 3.3-3.7. For the event-related design ("CamCAN sensorimotor"), the inclusion of derivatives increased the amount of perceived activation, but only when the statistical inference was based on an F-test testing all HRF-related covariates together. For the remaining datasets, which were boxcar, the addition of the derivatives to the canonical function showed smaller effect. The flexible HRF models: `tent`, `csplin`, FIR and the Fourier set, displayed on average slightly worse specificity-sensitivity trade-offs than the canonical model. An analysis of the positive rate for t- and F-tests did not reveal relevant differences between the considered HRF models (Figures 3.11-3.12).

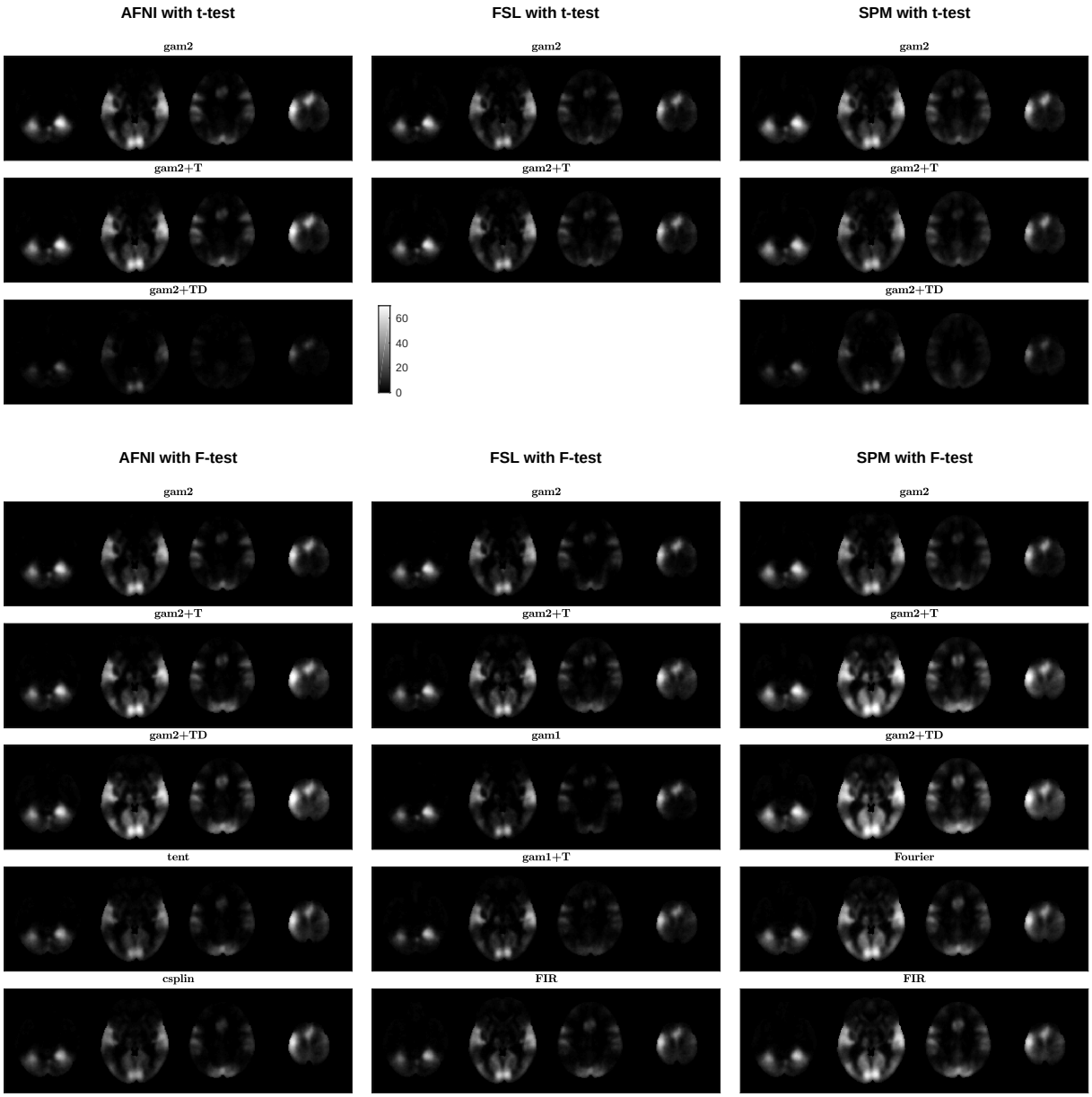


Figure 3.3: Single subject analyses: spatial distribution of significant clusters for the **"CamCAN sensorimotor"** dataset, HRF models from AFNI, FSL and SPM and the true experimental design. At the top there are results for t-test on the canonical function and at the bottom there are results for F-test on all HRF-related covariates. Four exemplary MNI axial slices from the bottom to the top of the head were selected (left to right). Scale refers to the percentage of subjects where significant activation was detected at the given voxel.

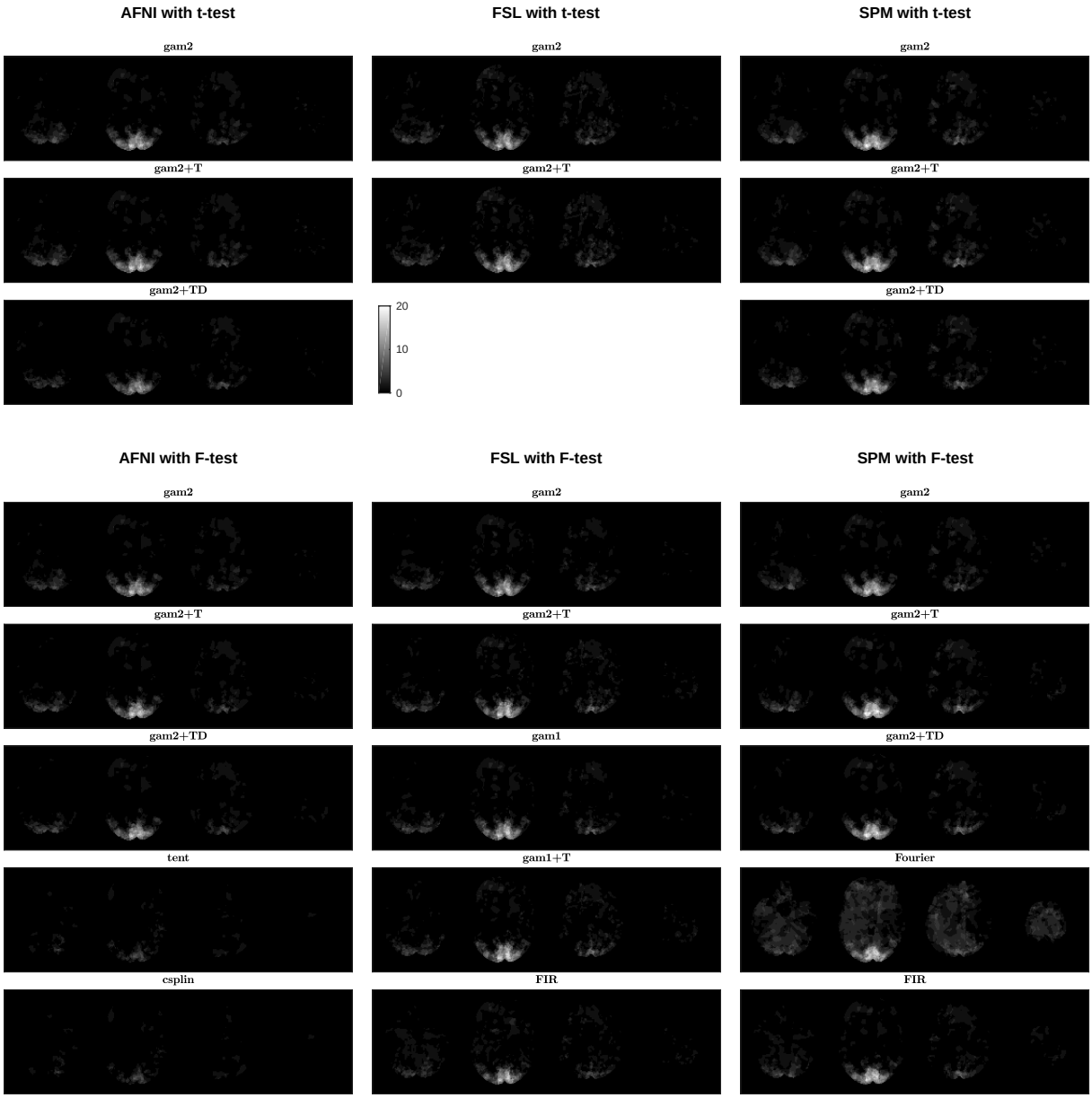


Figure 3.4: Single subject analyses: spatial distribution of significant clusters for the **"CRIC checkerboard"** dataset, HRF models from AFNI, FSL and SPM and the true experimental design. At the top there are results for t-test on the canonical function and at the bottom there are results for F-test on all HRF-related covariates. Four exemplary MNI axial slices from the bottom to the top of the head were selected (left to right). Scale refers to the percentage of subjects where significant activation was detected at the given voxel.

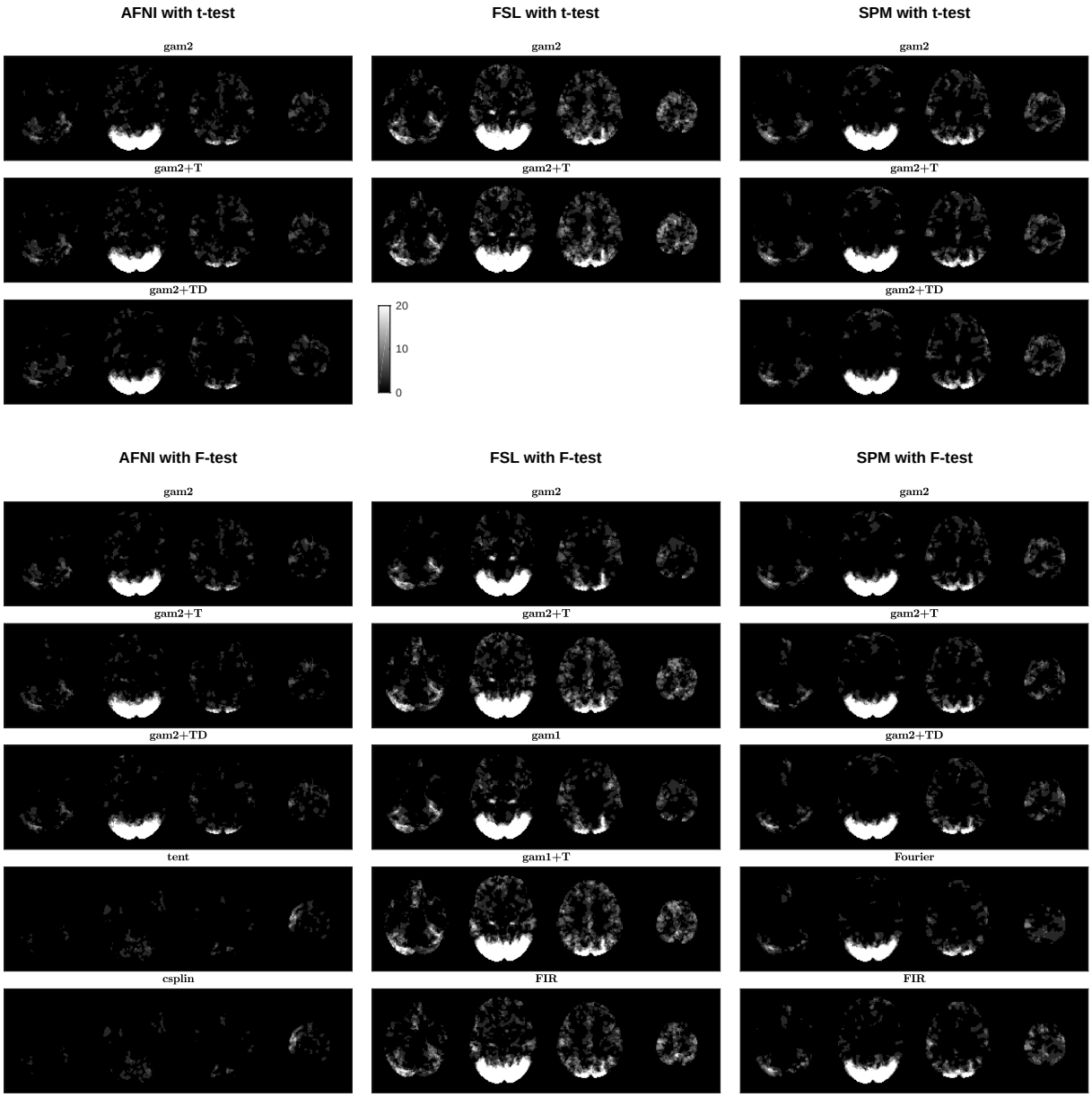


Figure 3.5: Single subject analyses: spatial distribution of significant clusters for the **"NKI checkerboard (TR=1.4s)"** dataset, HRF models from AFNI, FSL and SPM and the true experimental design. At the top there are results for t-test on the canonical function and at the bottom there are results for F-test on all HRF-related covariates. Four exemplary MNI axial slices from the bottom to the top of the head were selected (left to right). Scale refers to the percentage of subjects where significant activation was detected at the given voxel.

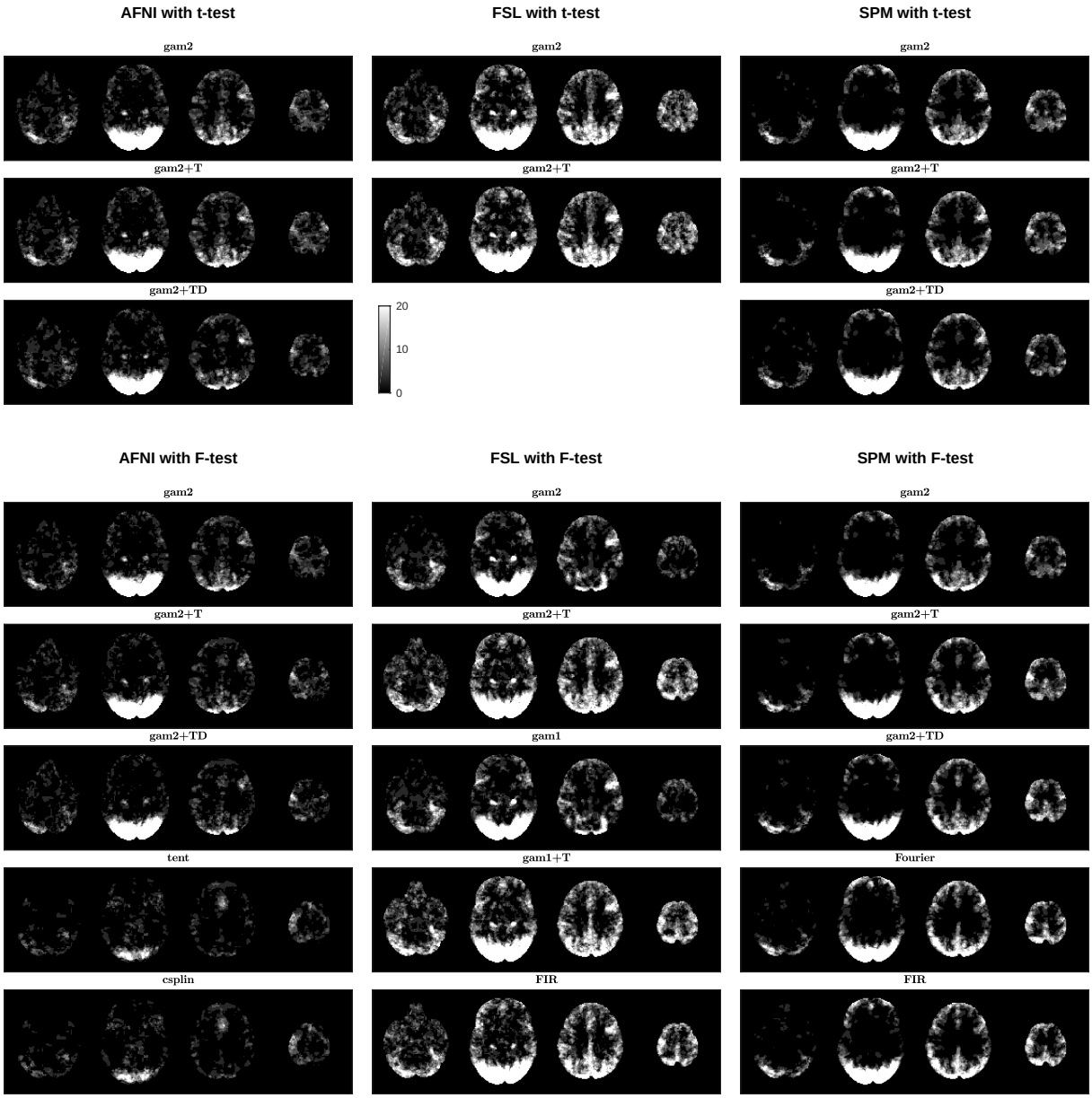


Figure 3.6: Single subject analyses: spatial distribution of significant clusters for the **"NKI checkerboard (TR=0.645s)"** dataset, HRF models from AFNI, FSL and SPM and the true experimental design. At the top there are results for t-test on the canonical function and at the bottom there are results for F-test on all HRF-related covariates. Four exemplary MNI axial slices from the bottom to the top of the head were selected (left to right). Scale refers to the percentage of subjects where significant activation was detected at the given voxel.

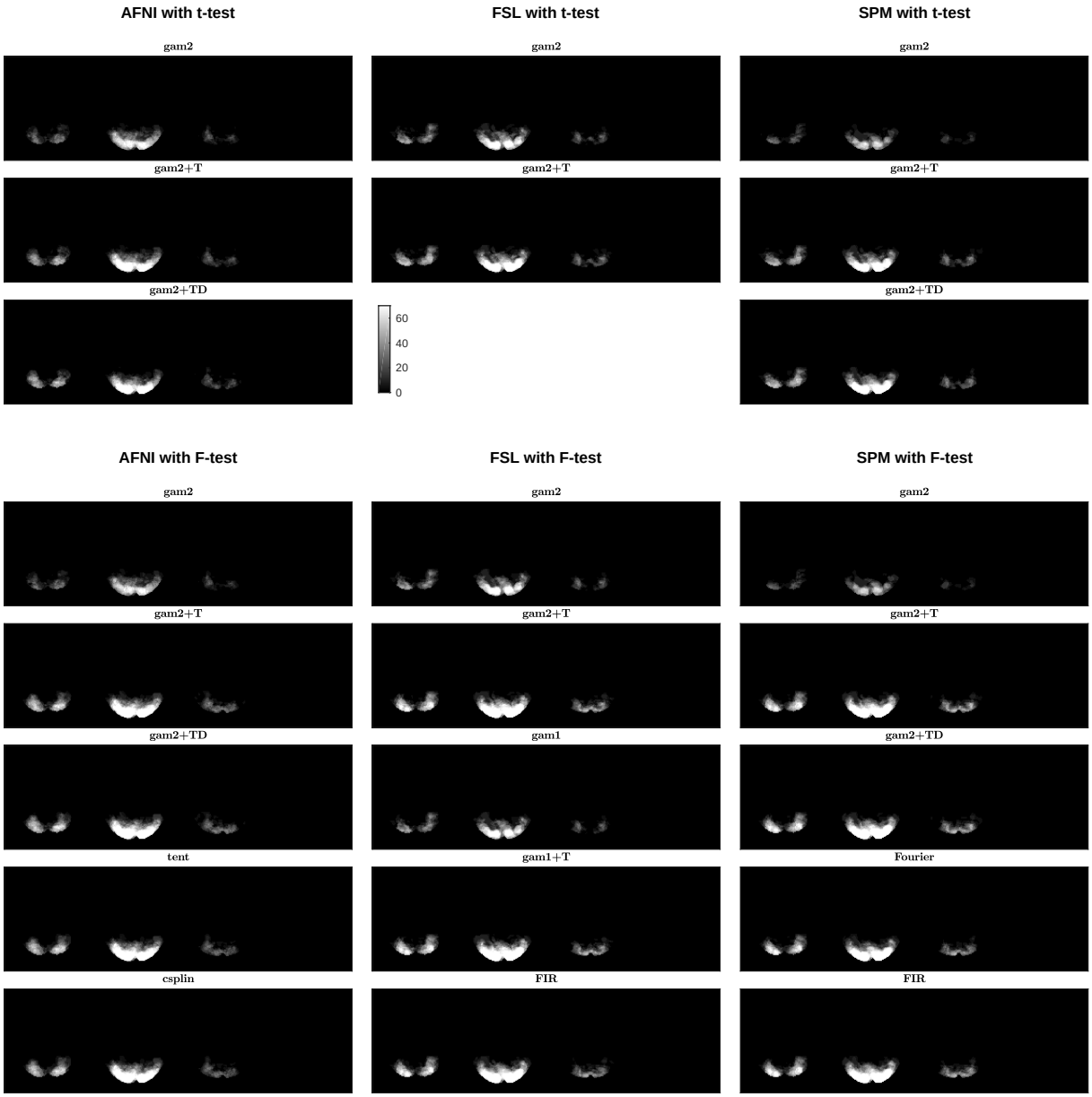


Figure 3.7: Single subject analyses: spatial distribution of significant clusters for the **"BMMR checkerboard"** dataset, HRF models from AFNI, FSL and SPM and the true experimental design. At the top there are results for t-test on the canonical function and at the bottom there are results for F-test on all HRF-related covariates. Four exemplary MNI axial slices from the bottom to the top of the head were selected (left to right). Scale refers to the percentage of subjects where significant activation was detected at the given voxel.

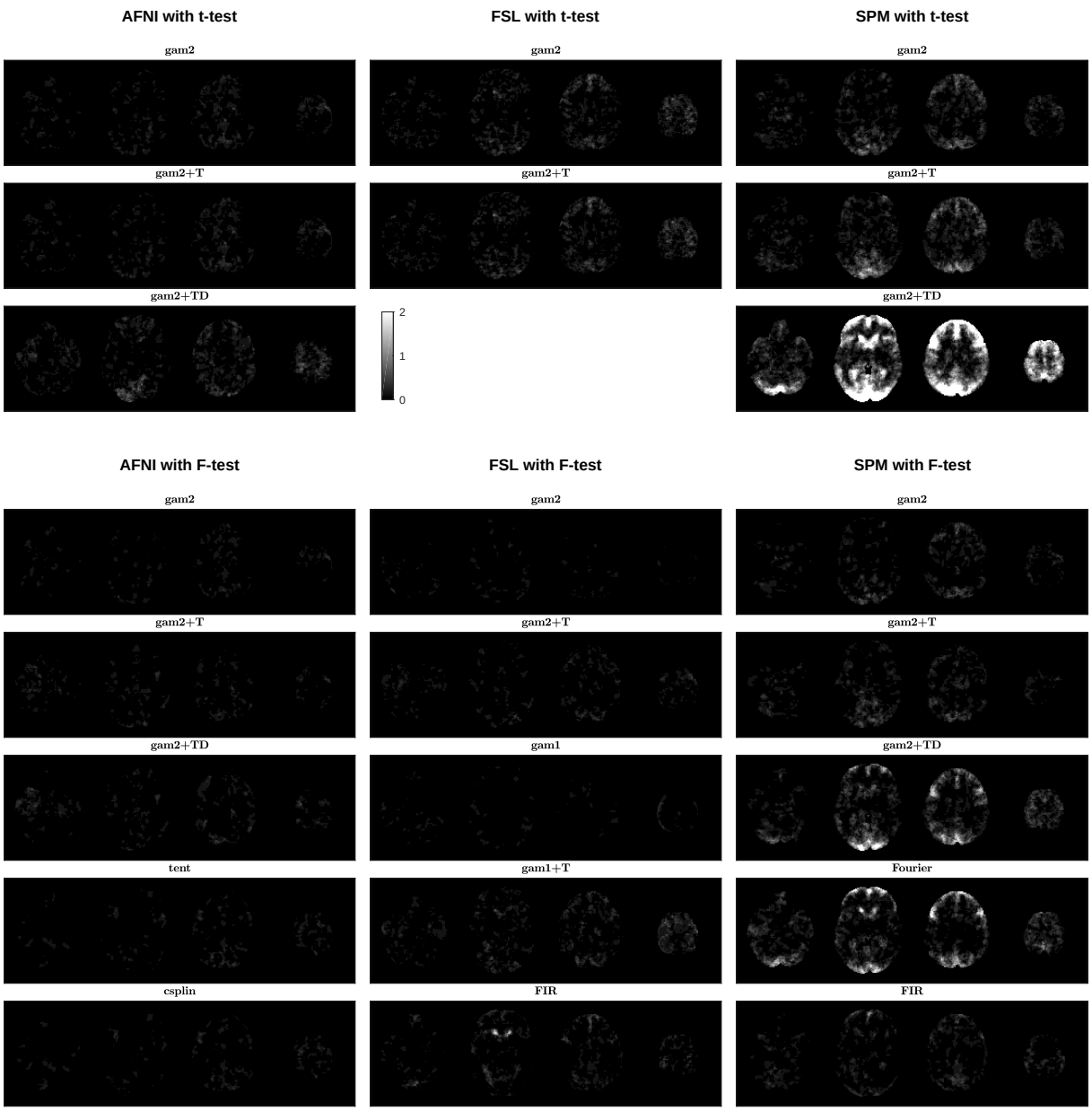


Figure 3.8: Single subject analyses: spatial distribution of significant clusters for the **"CamCAN sensorimotor"** dataset, HRF models from AFNI, FSL and SPM and a wrong experimental design (the randomised event-related design). At the top there are results for t-test on the canonical function and at the bottom there are results for F-test on all HRF-related covariates. Four exemplary MNI axial slices from the bottom to the top of the head were selected (left to right). Scale refers to the percentage of subjects where significant activation was detected at the given voxel.

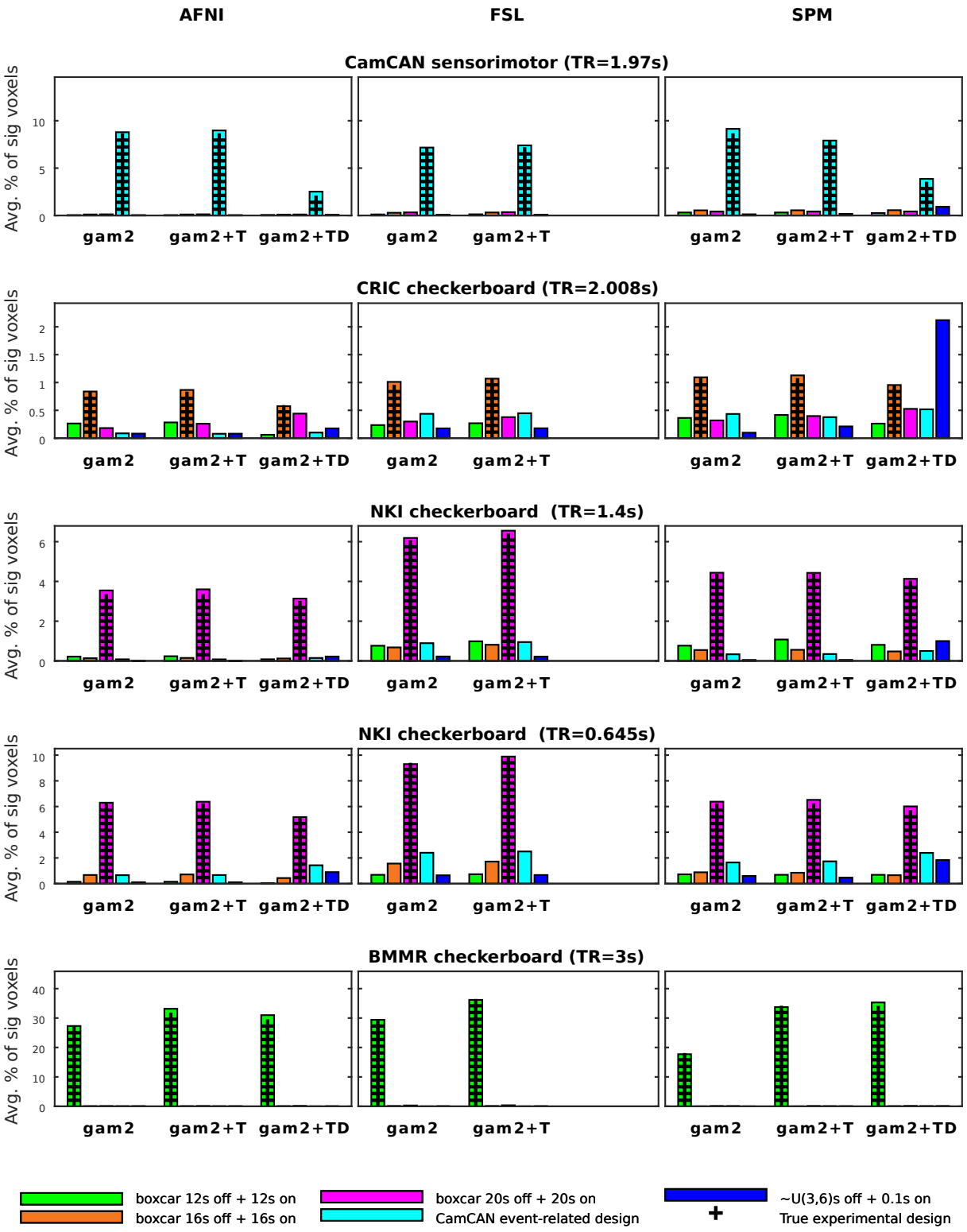


Figure 3.9: Single subject analyses: average percentage of significant voxels resulting from t-test on the canonical function only. For each dataset, five designs were assumed, one of which was the true design.

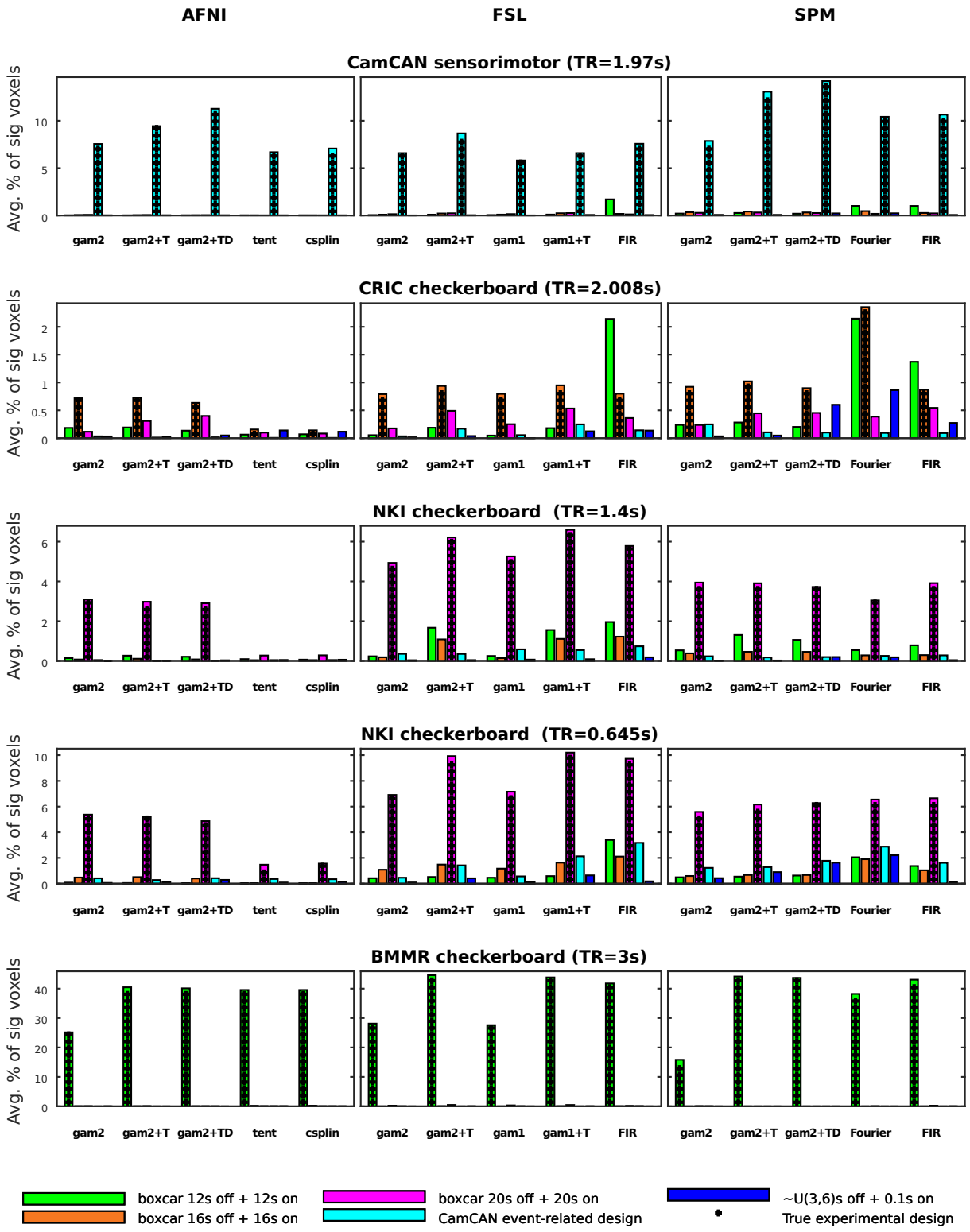


Figure 3.10: Single subject analyses: average percentage of significant voxels resulting from F-test on all HRF-related covariates. For each dataset, five designs were assumed, one of which was the true design.

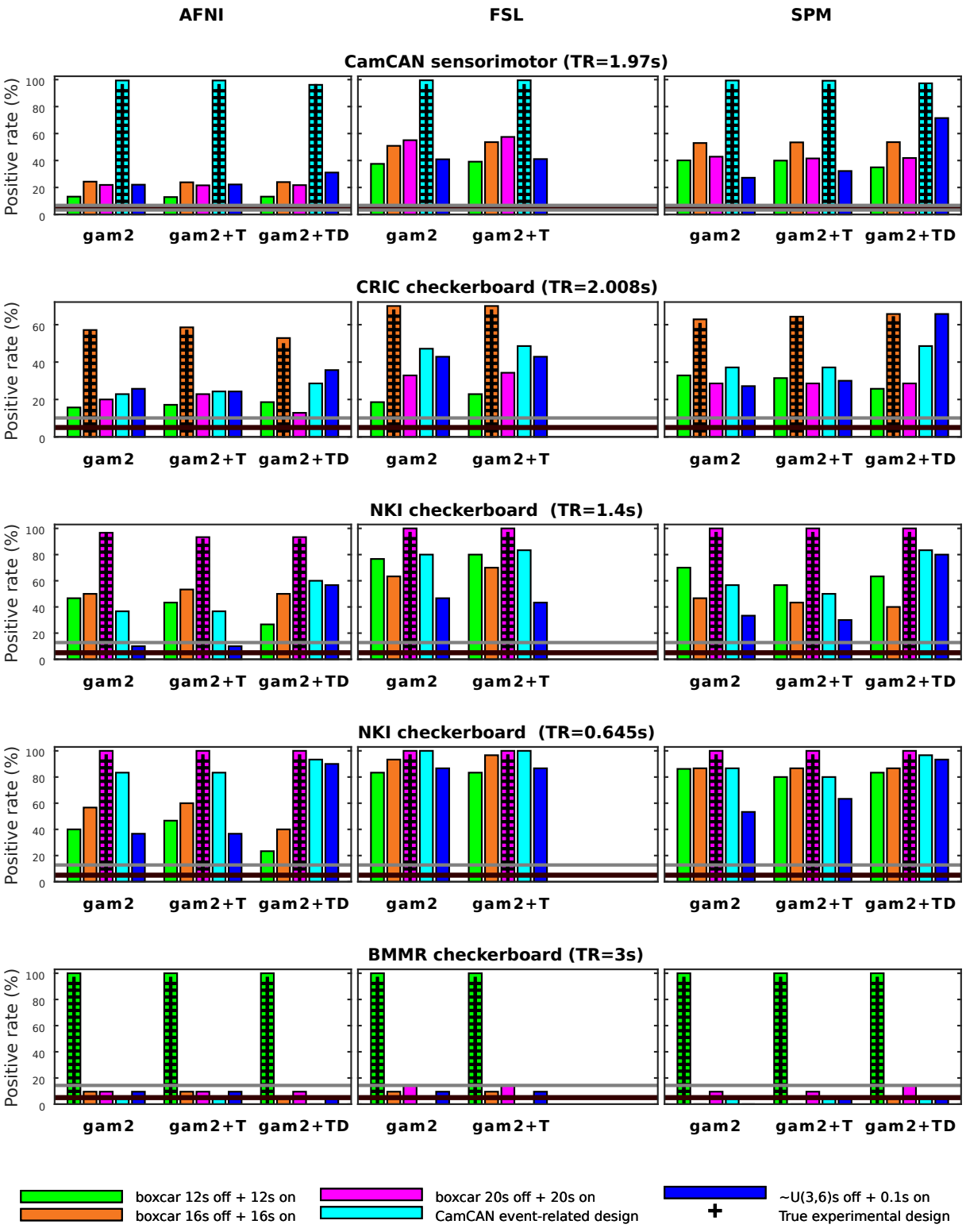


Figure 3.11: Single subject analyses: positive rate resulting from t-test on the canonical function only. For each dataset, five designs were assumed, one of which was the true design. The brown and grey lines indicate the expected positive rate together with the confidence interval (for null data).

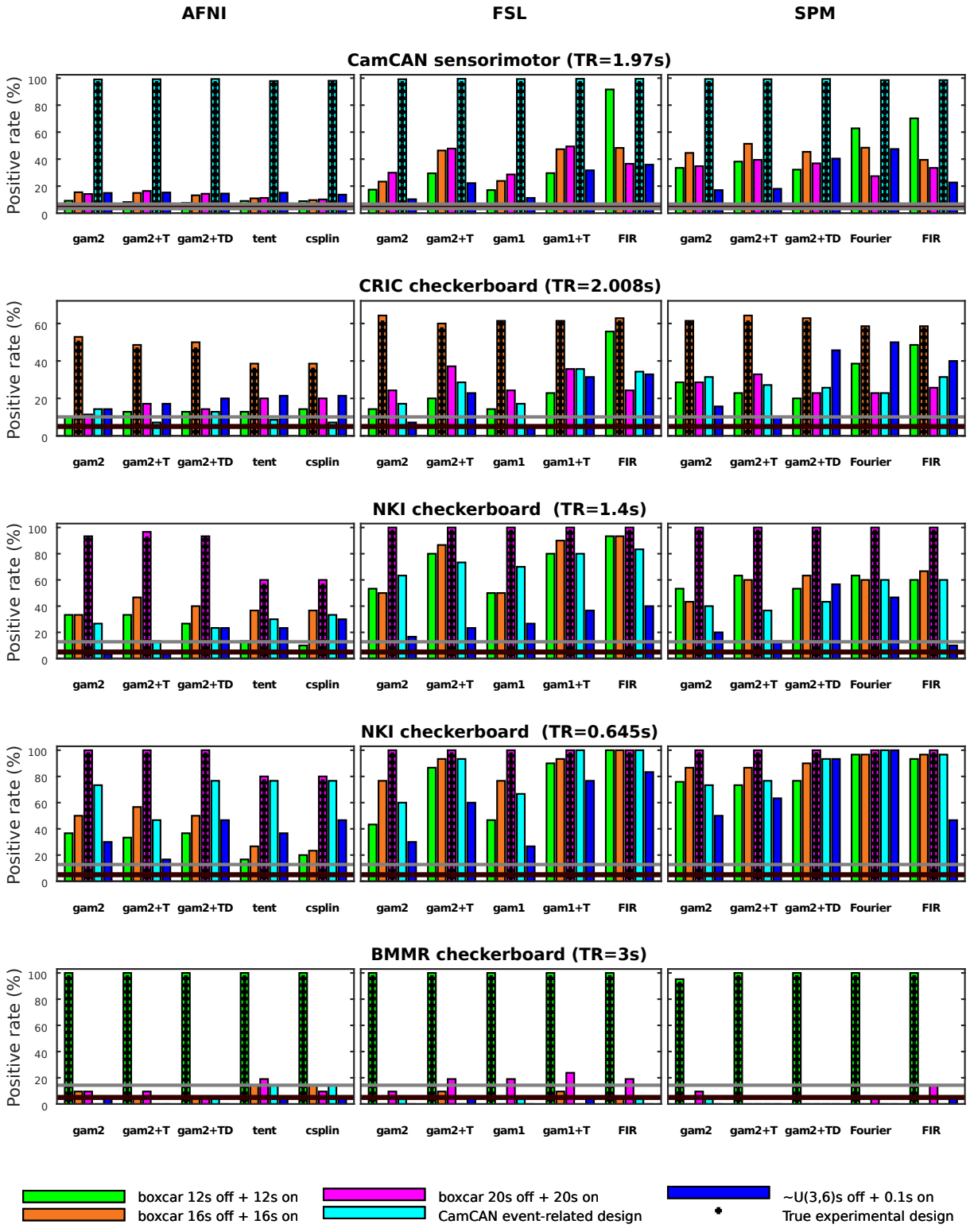


Figure 3.12: Single subject analyses: positive rate resulting from F-test on all HRF-related covariates. For each dataset, five designs were assumed, one of which was the true design. The brown and grey lines indicate the expected positive rate together with the confidence interval (for null data).

3.4.2 Group level analyses

An analysis of the spatial distribution of significant clusters resulting from group level analyses with assumed true designs (Figures 3.13-3.16) reveals similar patterns as from first level analyses. Importantly, for the event-related design dataset, the addition of the derivatives led to much higher amount of significant activation, but only for the F-test. For the t-test, sensitivity deteriorated when the partial derivatives were added to the model. Across the different datasets, the use of the flexible HRF models led to many significant clusters scattered across the brain. The "CRIC checkerboard" dataset was not investigated, as the corresponding group mask did not cover visual cortex. For the subjects from this dataset, some brains were deformed and registrations to MNI space were imperfect.

An investigation of the percentage of significant voxels (Figures 3.17-3.18) confirmed that the addition of the derivatives for an event-related design dataset can only increase sensitivity if statistical inference is conducted with an F-test. Figure 3.18 showed that the flexible HRF models: `tent`, `csplin`, FIR and the Fourier set, often displayed much significant activation for the true design, but also displayed much significant activation for the wrong designs. This suggests problems with specificity, though these problems might be related to the employed random effects model.

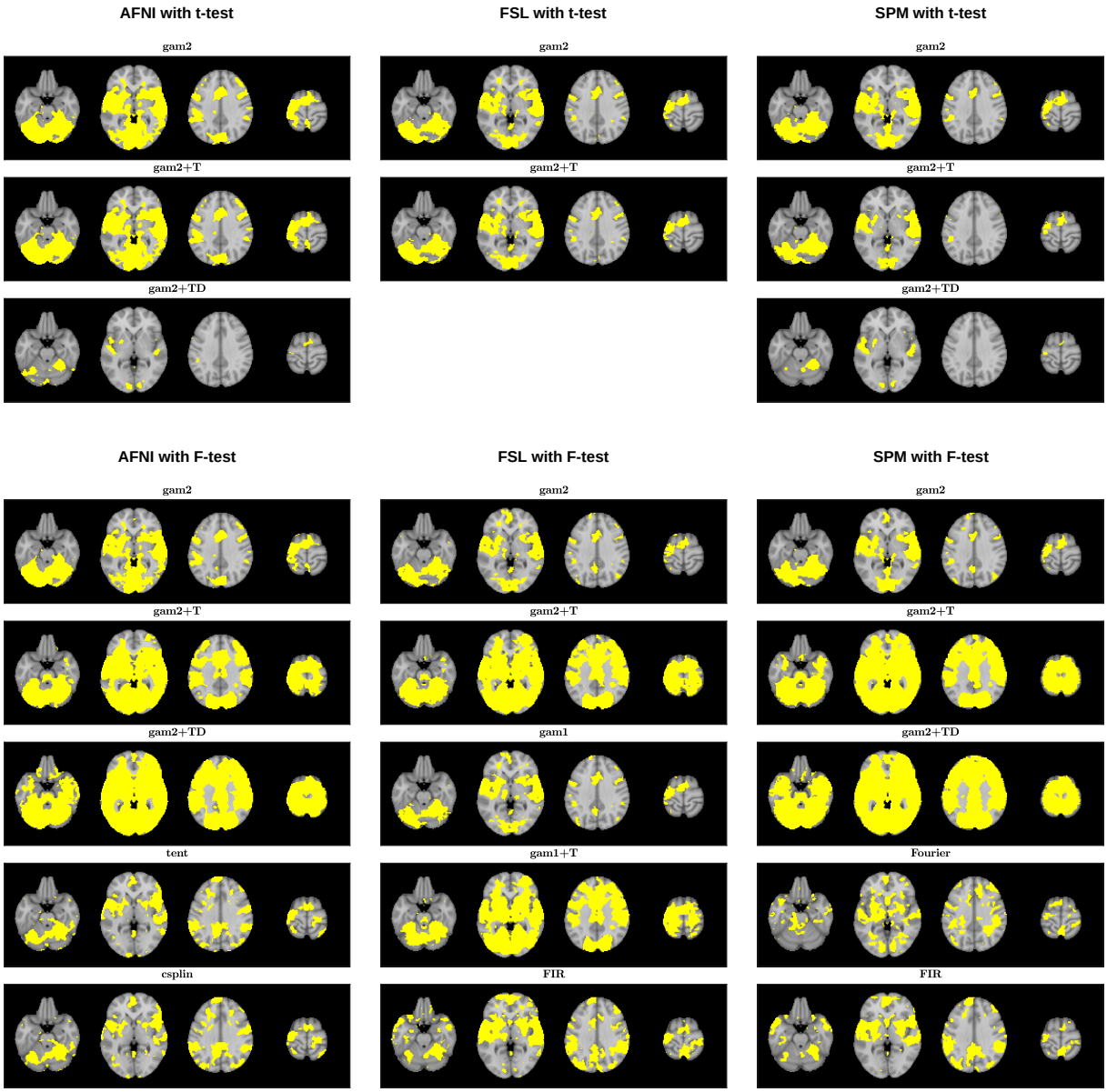


Figure 3.13: Group level analyses: spatial distribution of significant clusters for the **"CamCAN sensorimotor"** dataset, HRF models from AFNI, FSL and SPM and the true experimental design. At the top there are results for t-test on the canonical function and at the bottom there are results for F-test on all HRF-related covariates. Four exemplary MNI axial slices from the bottom to the top of the head were selected (left to right).

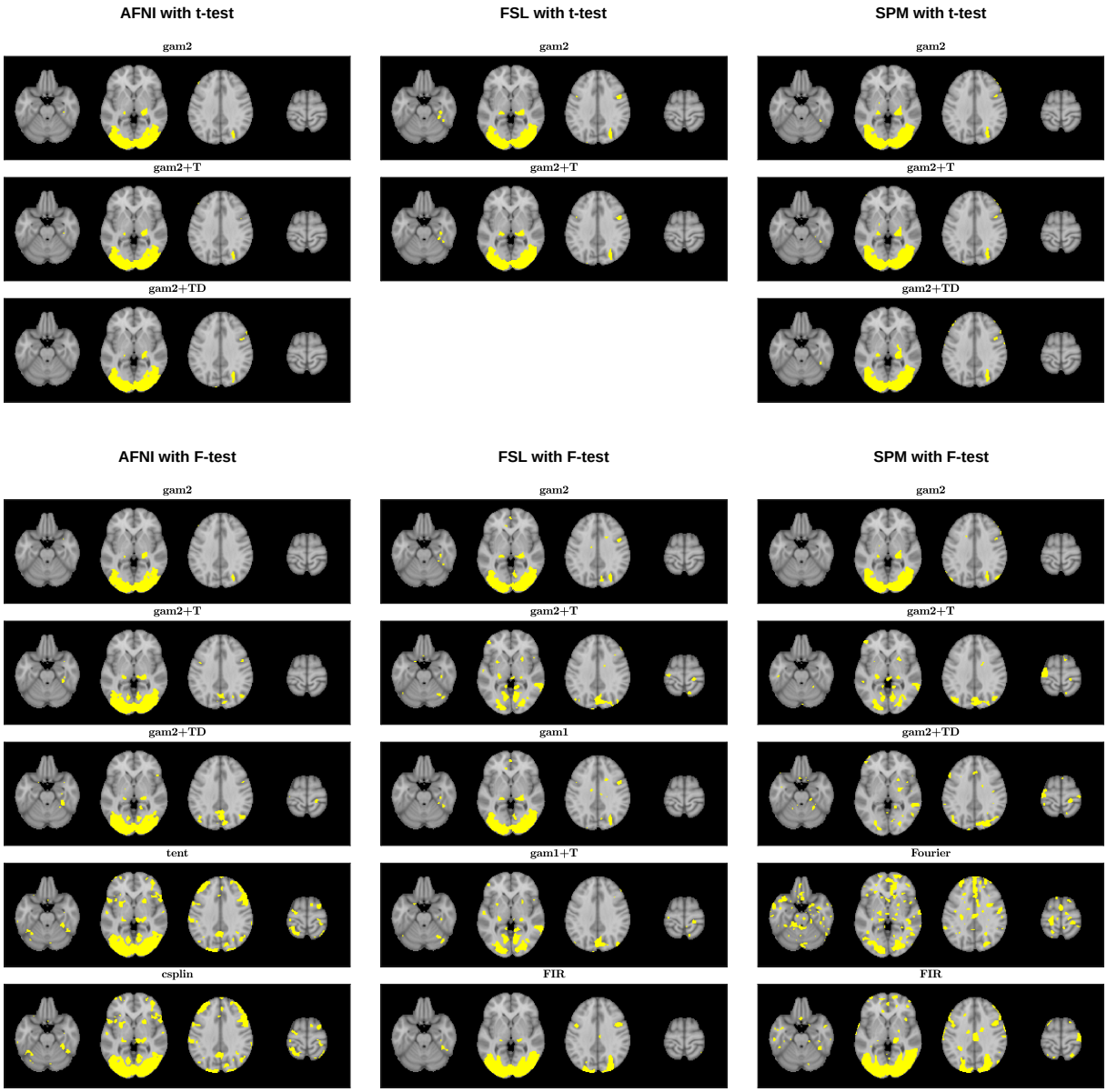


Figure 3.14: Group level analyses: spatial distribution of significant clusters for the **"NKI checkerboard (TR=1.4s)"** dataset, HRF models from AFNI, FSL and SPM and the true experimental design. At the top there are results for t-test on the canonical function and at the bottom there are results for F-test on all HRF-related covariates. Four exemplary MNI axial slices from the bottom to the top of the head were selected (left to right).

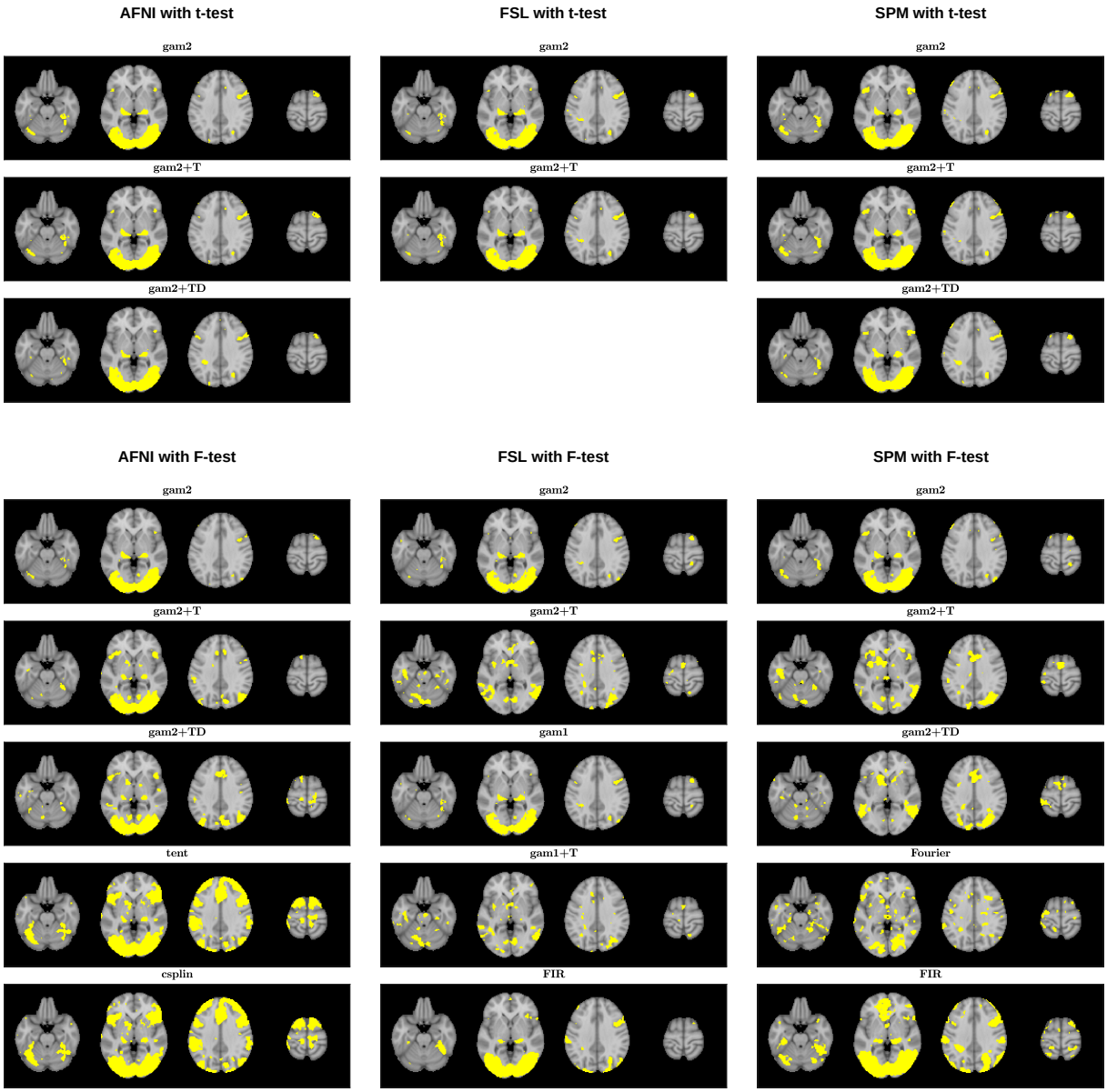


Figure 3.15: Group level analyses: spatial distribution of significant clusters for the **"NKI checkerboard (TR=0.645s)"** dataset, HRF models from AFNI, FSL and SPM and the true experimental design. At the top there are results for t-test on the canonical function and at the bottom there are results for F-test on all HRF-related covariates. Four exemplary MNI axial slices from the bottom to the top of the head were selected (left to right).

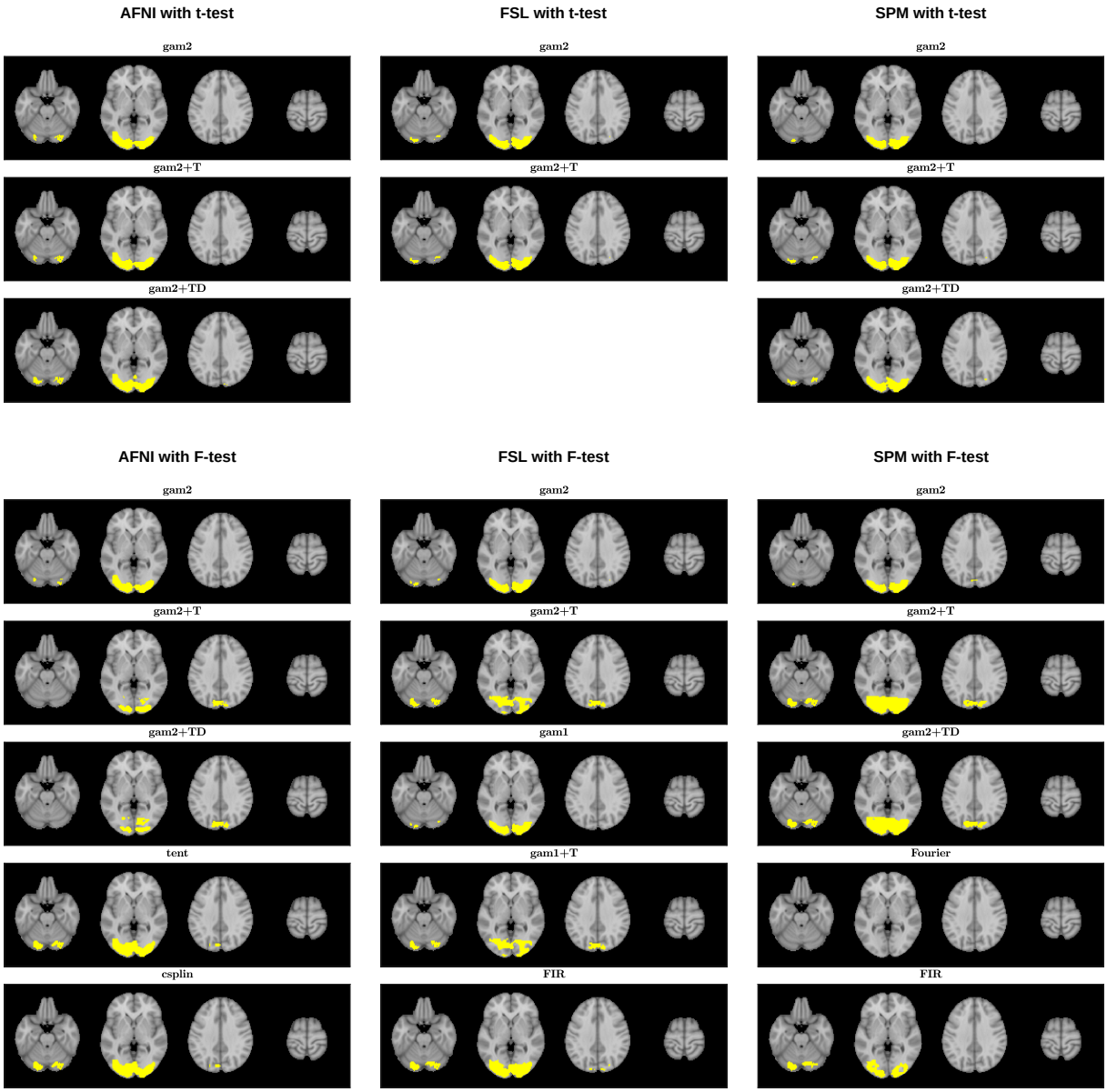


Figure 3.16: Group level analyses: spatial distribution of significant clusters for the **"BMMR checkerboard"** dataset, HRF models from AFNI, FSL and SPM and the true experimental design. At the top there are results for t-test on the canonical function and at the bottom there are results for F-test on all HRF-related covariates. Four exemplary MNI axial slices from the bottom to the top of the head were selected (left to right).

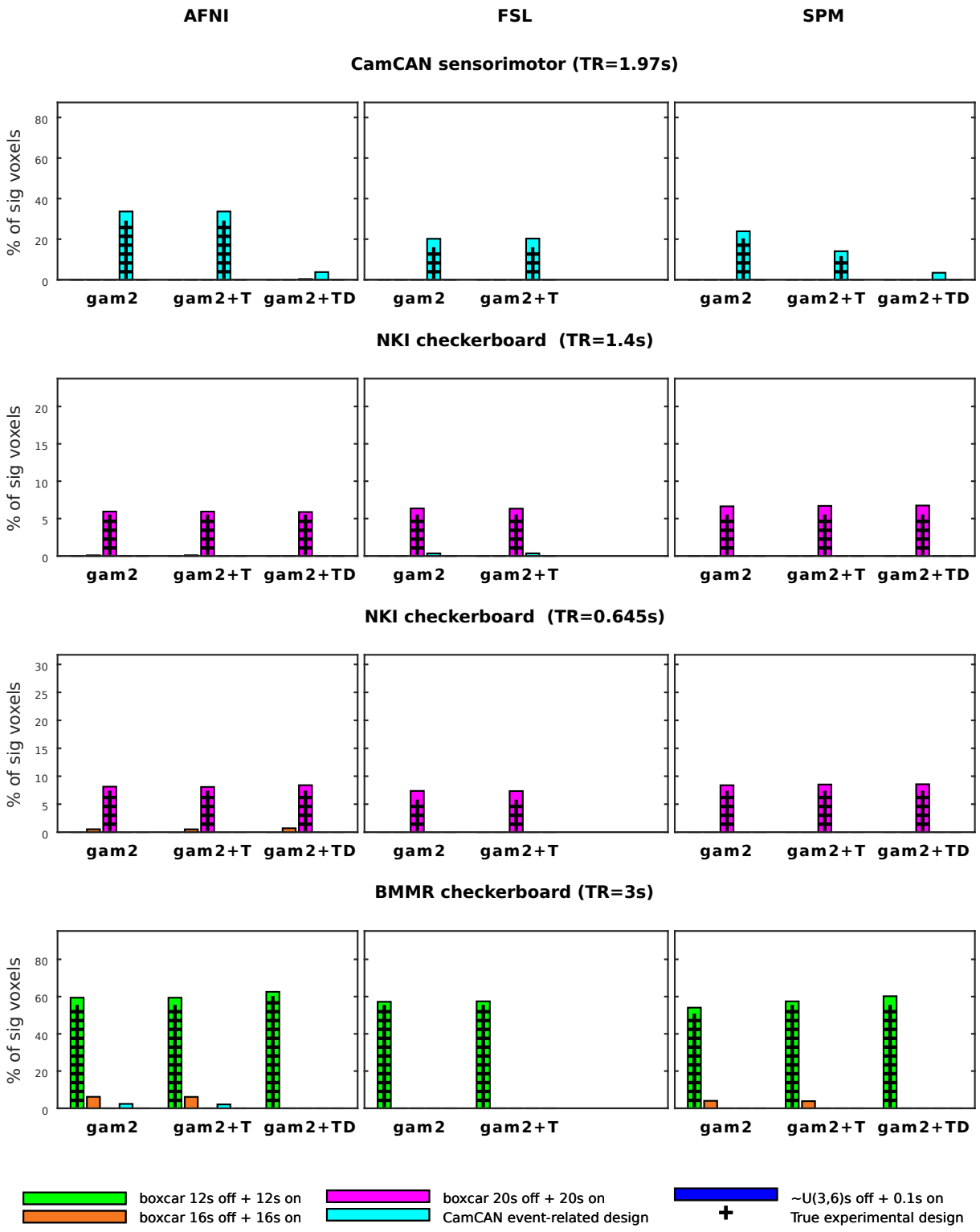


Figure 3.17: Group level analyses: percentage of significant voxels resulting from t-test on the canonical function only. For each dataset, five designs were assumed, one of which was the true design. For the "CRIC checkerboard" dataset, several of the subjects had deformed brains, which led to the group brain mask not covering the primary visual cortex. Thus, I excluded this dataset from the group analyses.

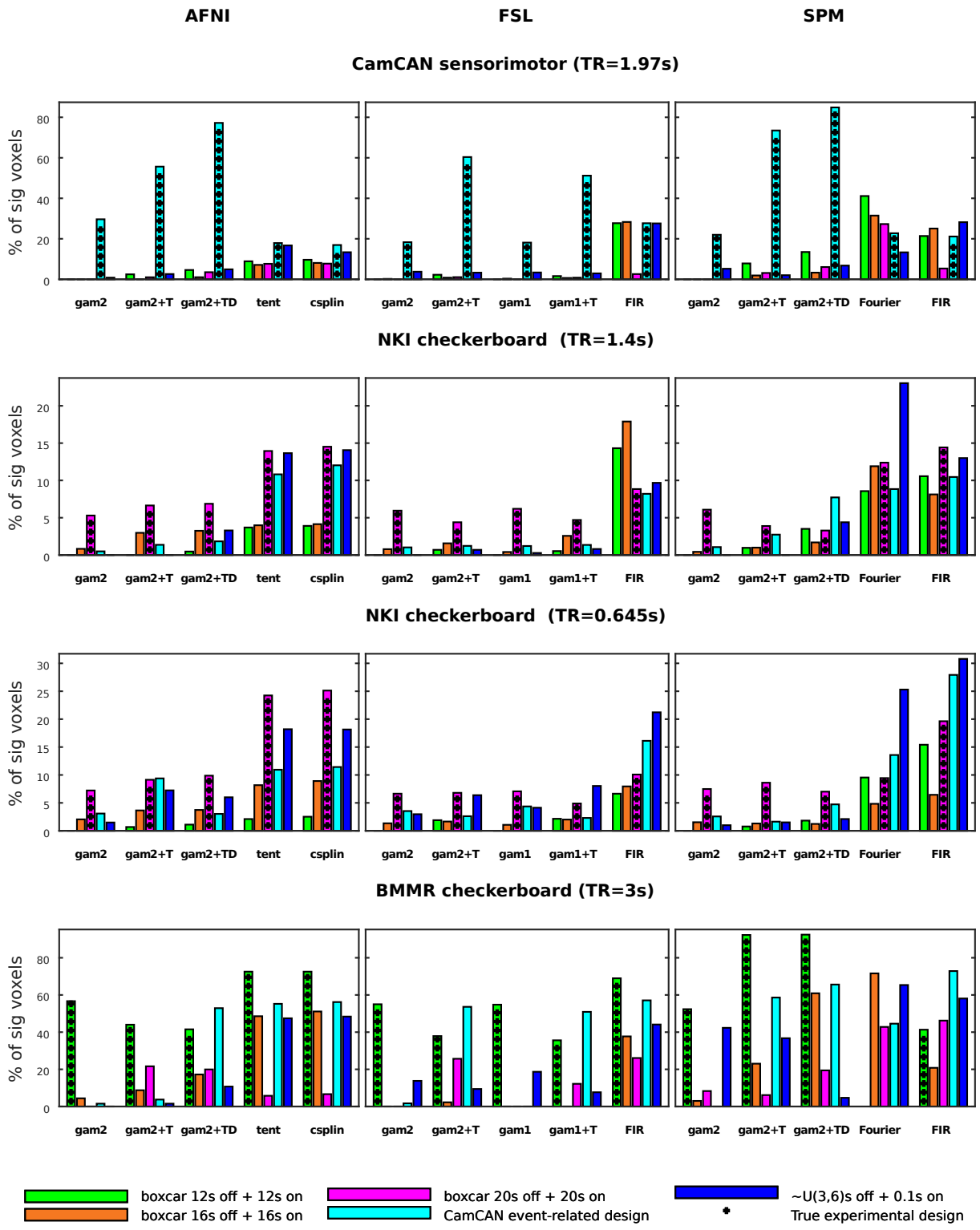


Figure 3.18: Group level analyses: percentage of significant voxels resulting from F-test on all HRF-related covariates. For each dataset, five designs were assumed, one of which was the true design. For the "CRIC checkerboard" dataset, several of the subjects had deformed brains, which led to the group brain mask not covering the primary visual cortex. Thus, I excluded this dataset from the group analyses.

3.4.3 Whitening performance for different HRF models

Figure 3.19 shows the power spectra of the GLM residuals for each of the datasets tested with each of the HRF models. For these analyses, only the true designs were assumed and the power spectra reflect averages across all subjects. On average, GLM residuals from FSL displayed more residual signal than GLM residuals from AFNI or SPM. This is in line with results shown in Chapter 2, as in the current analyses, SPM was used only with the `FAST` pre-whitening option. While for the "CamCAN sensorimotor" dataset, the power spectra resulting from the use of the different HRF models looked similar, they were different for the other datasets, which were based on boxcar designs. For the flexible HRF models: `tent`, `csplin`, FIR and the Fourier set, the GLM residuals displayed patterns of dips for some of the datasets. This relationship was not consistent across AFNI, FSL and SPM. For example, for AFNI the only dataset where this pattern emerged, was "CRIC checkerboard". For FSL and SPM, this pattern was visible for the "CRIC checkerboard", "NKI checkerboard (TR=1.4s)" and "NKI checkerboard (TR=0.645s)" datasets, although it differed slightly between FSL and SPM.

Particularly strong was the effect of including the temporal derivative in the analysis of the "BMMR checkerboard" images. Across all the three packages, there was a peak rather than an expected dip at the design frequency (approximately 0.0417 Hz) when using the canonical function without the derivatives. However, after the inclusion of the temporal derivative, the power spectrum at the design frequency became much smaller, resembling a dip. It is an unexpected behaviour given that the design for this study was a boxcar with stimulus duration time of as much as 12 s. While this points to problems in the processing pipeline of the "BMMR checkerboard" images, it is not sure what the reason is. Possibly, there are some acquisition artefacts in this dataset, for example related to the prospective motion correction with which these images were acquired. Problems with this dataset were already discussed in Subsection 2.5.6.

Following the procedure from Chapter 2, the Q-Q-plots of the GLM residuals were investigated (Figure 3.20). These refer to the first subject in each dataset only and to the assumed true designs. GLM residuals from none of the HRF models and none of the packages resembled a normal distribution, with "CamCAN sensorimotor" and "NKI checkerboard (TR=0.645s)" residuals being least normal. Distribution differences between HRF models were only visible for the "BMMR checkerboard" dataset.

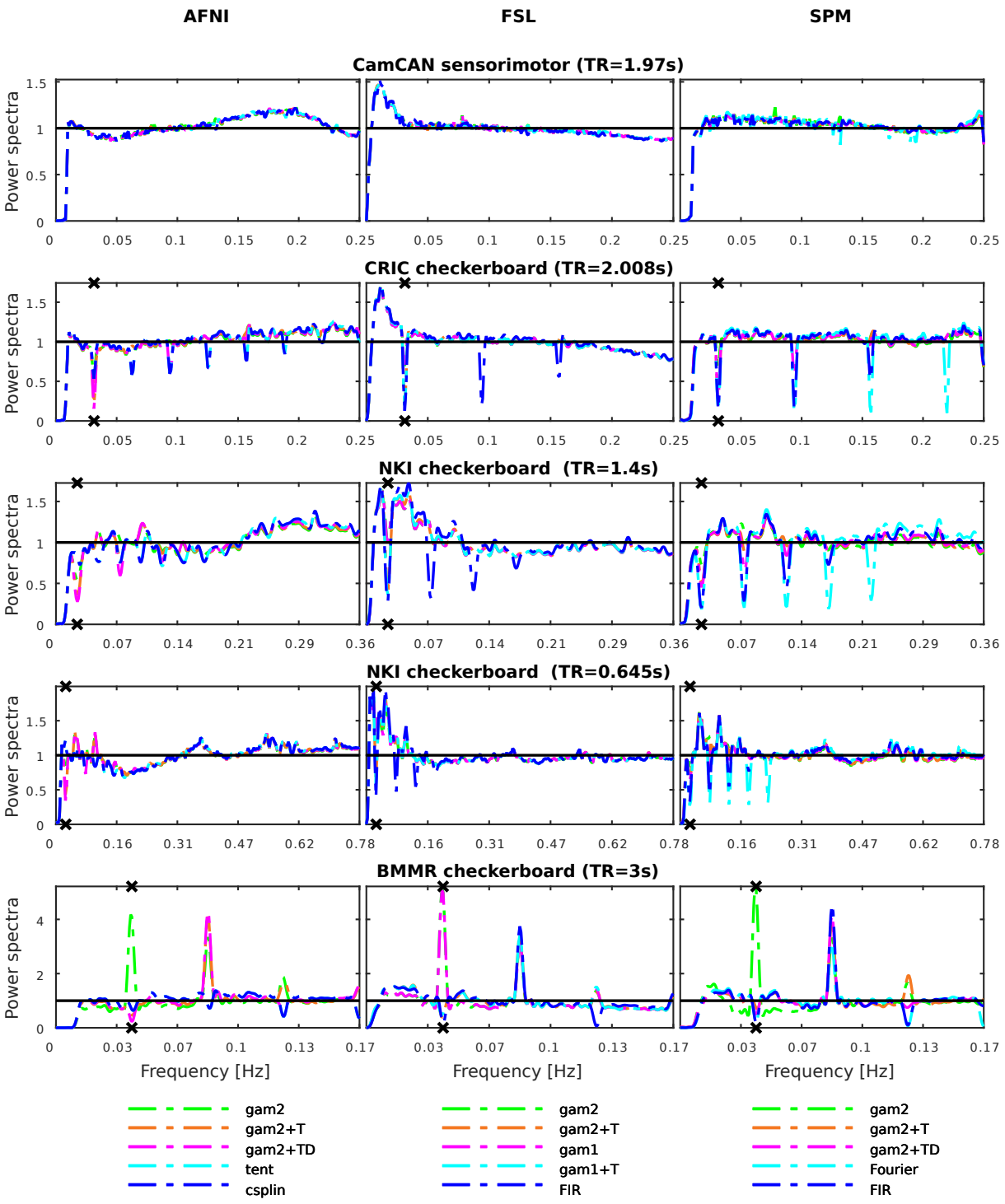


Figure 3.19: Power spectra of the GLM residuals in native space averaged across brain voxels and across subjects. For each dataset, the true design was assumed and HRF models from AFNI, FSL and SPM were employed. The black line indicates perfect power spectra, while the crosses indicate the frequency of the experimental design.

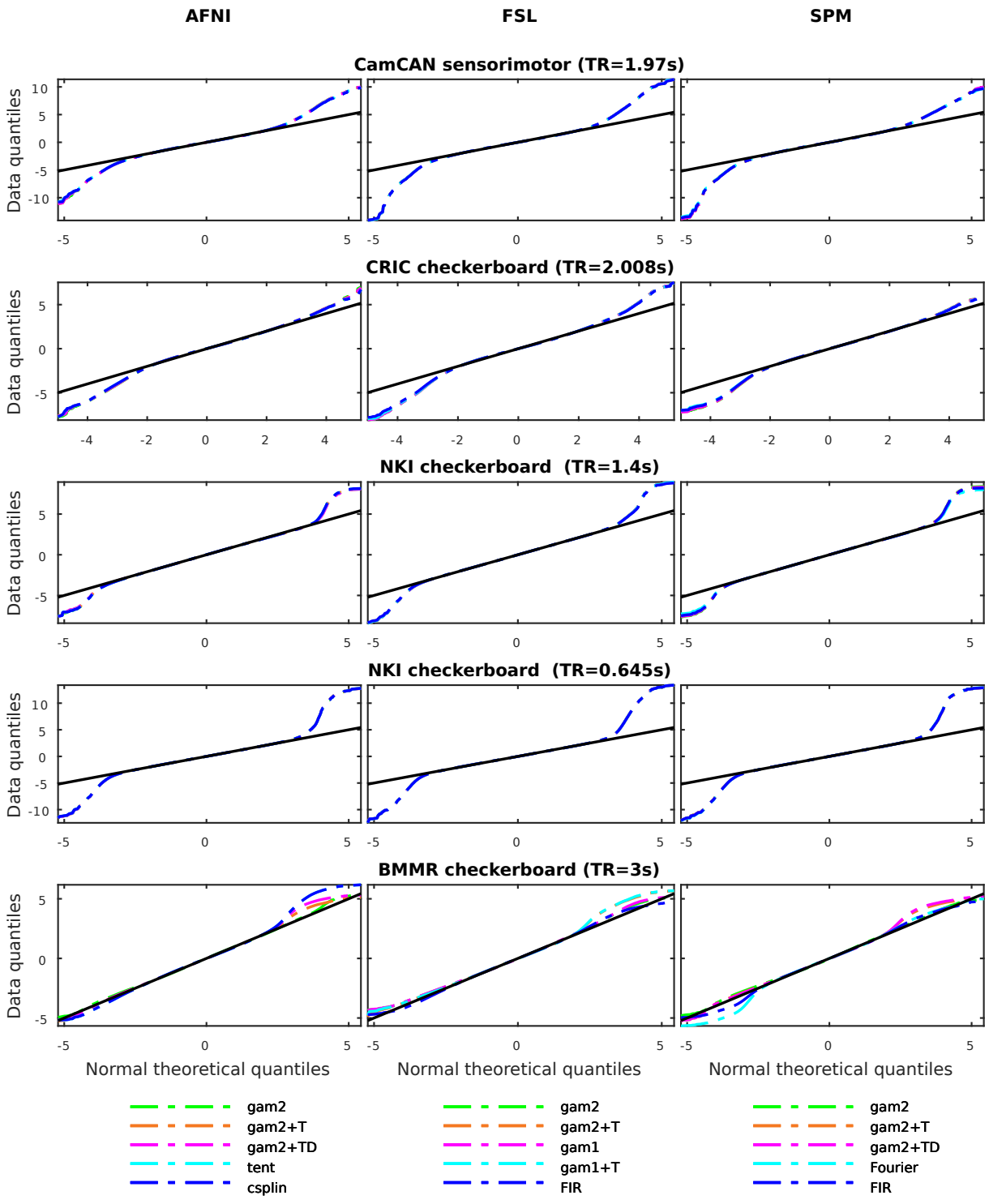


Figure 3.20: Q-Q plots of the GLM residuals in native space averaged across brain voxels for the first subject in each dataset. For each dataset, the true design was assumed and HRF models from AFNI, FSL and SPM were employed. The black line indicates a perfect match with the normal distribution.

3.5 Discussion

On average, the use of temporal and dispersion derivatives did not lead to higher sensitivity when using a t-test on the canonical function. For the event-related design dataset "CamCAN sensorimotor", the use of derivatives even led to a lower percentage of significant voxels, both at the first and at the second level. Importantly, F-test which was run on all HRF-related covariates led to higher sensitivity for this dataset, both at the first and at the second level. For the other datasets, which corresponded to boxcar designs, the impact of including derivatives was more limited. However, on average, for boxcar design datasets tested with t-test, the inclusion of derivatives did not increase the amount of perceived activation.

Sladky et al. [2011] discussed that the canonical HRF and its derivative may lose their orthogonality when convolved with a stimulus function. It was noted that following the use of the convolution function, part of the variance initially explained by the canonical HRF regressor can be explained by its temporal derivative. Importantly, this can reduce sensitivity at the group level as usually only the estimates of the canonical regressor enter the random-effects group analysis. Sladky et al. [2011] noted that this problem can be addressed by specifying F-contrasts at the group level. Similar reasoning can be applied to the first level analyses, where usually statistical inference is performed with the help of a t-test on the canonical function only. When for a boxcar design, the temporal derivative after convolution is correlated with the canonical function after convolution, the temporal derivative might capture variance that is not linked to HRF delay, but represents the main BOLD response. If inference is performed with a t-test rather than an F-test, this part of the explained variance might actually decrease sensitivity both at the first and at the second analysis level.

The default HRF model in FSL does not account for the post-stimulus undershoot of around 15 seconds. For the boxcar design datasets, I observed only negligible differences in results corresponding to the "gam1" and "gam2" HRF models. However, for the event-related design dataset, the use of the canonical HRF ("gam2") led to a higher percentage of significant voxels than FSL's default model. This sensitivity improvement occurred both at the single subject level and at the group level. As the modelling of the post-stimulus undershoot by the canonical HRF model does not require additional parameters in the regression model, it can be considered a better way of modelling the HRF than the FSL's default method.

The analysis of the power spectra of the GLM residuals suggested problems for complex HRF models used for data with boxcar designs. For AFNI, the use of the `tent` and `csplin`

models led to dips visible in the power spectra for multiples of the design frequency. For FSL, the use of the FIR model led to the same problem. For SPM, such dips were visible for the Fourier set and the FIR models. For boxcar designs, accurate modelling of the hemodynamic response function becomes less relevant, as a long stimulus duration time means that a possible deviation from the canonical HRF is only a small fraction of the model after convolution with the stimulus duration time vector. Thus, there is less need to model the hemodynamic response function with such flexible models. These complex HRF models led to worst specificity-sensitivity trade-offs, sometimes leading to higher percentage of significant voxels for a wrong design than for the true design.

3.5.1 Confusion about the shape of the canonical HRF

Although this chapter compared HRF models available in AFNI, FSL and SPM, it is worth noting that there is much confusion in the fMRI literature with regard to the shape of the canonical HRF. If the canonical function is assumed to reflect the BOLD response in healthy subjects, such ambiguities can, among others, confound conclusions of studies where claims are made about BOLD responses in a particular subject population compared to healthy subjects.

The use of the canonical HRF was first suggested in Glover [1999], where a previously discussed model based on a curve of a gamma distribution density function [Boynton et al., 1996] was extended by another gamma curve. The main response peak of the canonical HRF is at 5 s, while the undershoot peak is around 15 s. These values are reminiscent of the estimated HRFs from Glover [1999], can be seen when plotting the canonical function with SPM script `spm_hrf.m` and can be derived from an SPM wiki plot: https://en.wikibooks.org/wiki/SPM/Haemodynamic_Response_Function, though the same SPM wiki article confusingly also says: *"The SPM HRF is shown above, and exhibits a rise peaking around 6 sec"*. For example, Lindquist and Wager [2007] states that the main response peak of the canonical HRF occurs at 6 s, while the undershoot peak occurs at 16 s. Henson and Friston [2007] also notes the times 6 and 16 s, for which Friston et al. [1998b] is cited. However, Friston et al. [1998b] does not seem to define the canonical HRF this way. Instead, it considers different HRF models with most of them resulting in the main response peak being below 6 s.

Furthermore, Handwerker et al. [2004] says that the undershoot peak of the canonical HRF is around 16 rather than around 15 s. Moeller et al. [2008], Hamandi et al. [2006] mention the canonical HRF and describe the peak to be at 6 s and the undershoot to be at 16 s. Ford et al. [2005] says: *"Although our latency delays were small (∼500 ms), we*

may have underestimated the height of the activation in the controls, whose hemodynamic responses tended to peak earlier than the expected 6-s peak of the canonical HRF.". Here, the confusion affected the conclusions. Laufs et al. [2006], Szaflarski et al. [2010] also mention that the peak of the canonical HRF is at 6 s. Ritzl et al. [2003] states that the peak is about 6 s and cites Friston et al. [1998a]. Friston et al. [1998a] does not refer to a peak at 6 s. On the contrary, Figure 1 in Friston et al. [1998a] shows SPM's canonical HRF with a clear peak at 5 s.

Moreover, this misconception also appears in discussions on neuroimaging mailing lists, where it is not clarified, for example:
https://www.jiscmail.ac.uk/cgi-bin/webadmin?A2=ind1711&L=SPM&P=R36362&1=SPM&9=A&J=on&d=No+Match%3BMatch%3BMatches&z=4
https://www.jiscmail.ac.uk/cgi-bin/webadmin?A2=ind0706&L=SPM&D=0&1=SPM&9=A&J=on&d=No+Match%3BMatch%3BMatches&z=4&P=246837
https://www.jiscmail.ac.uk/cgi-bin/webadmin?A2=ind02&L=SPM&P=R192361&1=SPM&9=A&J=on&d=No+Match%3BMatch%3BMatches&z=4
https://www.jiscmail.ac.uk/cgi-bin/webadmin?A2=ind03&L=FSL&D=0&P=341967

This worrisome confusion in many fMRI studies might have resulted from a misleading comment in SPM script `spm_hrf.m`, where it is stated that the first parameter of the double gamma HRF model (= 6) is the main response peak and that seconds are the units. The second parameter (= 16) is described as the undershoot peak and it is also given in seconds. However, these parameters are used to combine curves from two gamma distribution density functions, where the first two HRF parameters used by `spm_hrf.m` are the modes of the two gamma distribution density functions plus 1 (as dispersion parameters equal 1) [Hoel et al., 1954]. Alternatively, the confusion might have arisen from one of the early papers from the SPM team. For example, Friston et al. [2000a] says: *"The HRF in this instance comprises the sum of two gamma functions modeling a peak at 6 s and a subsequent undershoot."*

The modelling of the hemodynamic response function is a crucial part of the processing pipeline of task fMRI data. As most studies use the canonical function, it is worrying that there are so many misconceptions about the way this canonical function looks like. Surprisingly, a literature review did not reveal any study which discussed this confusion.

3.5.2 Alternative HRF models

There was a large number of studies proposing novel HRF modelling techniques. The HRF models in AFNI, FSL and SPM which I used were not all the HRF models available

in these packages, but just a subset of them. For example, in Lange and Zeger [1997] it was shown how to analyse fMRI data in the frequency-domain using a double gamma HRF model which was allowed to vary across voxels. However, the method suffered from the identifiability problem. Marchini and Ripley [2000] suggested a frequency-domain method to detect significant activation in voxel-wise time series. This approach does not involve any assumptions regarding the HRF shape, though, like most HRF models, it relies on the assumption that the voxel's HRF looks the same way for each stimulus repetition. In Olszowy et al. [2016] I showed that this method performs similarly to the single gamma HRF model, which is the default HRF model in FSL. Processing scripts needed to fully replicate that study are at `https://github.com/wiktorolszowy/fMRI_Marchini_method`. The Marchini method was tested only for block designs. For event-related designs, its performance would likely deteriorate. If the design of the study (timing of the stimuli/responses) is not clear or the duration of psychological events can largely vary, an alternative approach to using a fixed HRF model can be to apply the change-point theory [Lindquist et al., 2007, Nam et al., 2012]. However, change-point detection approaches are computationally intensive and were primarily developed for volume of interest analyses rather than for voxel-wise analyses. As change-point detection approaches are complex, it is difficult to perform proper validation and sensitivity analyses of them.

3.6 Conclusions

Using data of 772 subjects corresponding to five different fMRI protocols and different experimental designs, a number of HRF models from AFNI, FSL and SPM were compared in terms of specificity and sensitivity. For the event-related design dataset ("CamCAN sensorimotor"), inclusion of the temporal and dispersion derivatives in the analysis pipeline improved sensitivity both at the first and at the second level, but only when the statistical inference was conducted with an F-test on all the HRF-related covariates. Importantly, when the partial derivatives of the canonical function were used only as confounders, sensitivity was lower than for the canonical function used alone. As the dispersion derivative improved sensitivity for the event-related design dataset when using F-test, it might be considered unfortunate that FSL does not provide this basis function. Moreover, the canonical model, which in FSL is called "double gamma", displayed higher sensitivity than FSL's default HRF model: "single gamma". As both models simply employ a fixed curve, the "double gamma" model perhaps should replace the "single gamma" model as FSL's default option. For the boxcar design datasets, the inclusion of derivatives affected the

results only to a small extent. While the use of more flexible HRF models: `tent`, `csplin`, FIR and the Fourier set led in some cases for the assumed true designs to detection of more activation than the canonical HRF model used along its partial derivatives, the specificity resulting from the use of these methods was often poor. Interestingly, an analysis of the power spectra of the GLM residuals revealed that the choice of the HRF model affected the whitening performance. This reaffirms conclusions from Chapter 2 that pre-whitening is a crucial processing step, results of which should be investigated more often with the help of diagnostic plots.

Chapter 4

Effect of ageing on the BOLD signal*

4.1 Introduction

The blood oxygenation level dependent (BOLD) signal depends on neurovascular coupling, which denotes the processes by which neural activity influences the hemodynamic properties of the surrounding vasculature. Neurovascular coupling changes with healthy ageing, though the underlying mechanisms are not fully understood [D'Esposito et al., 2003, Wright and Wise, 2018]. Ageing affects arterial stiffness and cerebrovascular reactivity [Peng et al., 2018], which leads to lower vasodilation (widening of blood vessels) in older subjects. Compared to young subjects, older subjects display lower cerebral blood flow (CBF) and lower cerebral blood volume (CBV) increases. This leads to decreases in deoxyhaemoglobin content in brains of older subjects, and in consequence to their lower BOLD response amplitude [Wright and Wise, 2018].

To accurately infer neural activity from the BOLD signal, the dynamics between blood flow changes, blood volume changes and deoxyhaemoglobin content changes have to be accounted for. While there is still no consensus how the different components of the BOLD signal exactly influence each other [Buxton, 2012], the so-called balloon model [Buxton et al., 1998] has become a standard approach to link neural activity with the resulting BOLD response. It is the basis of the popular BOLD-based dynamic causal modelling framework [Friston et al., 2003], which is used to investigate possible causal neural connections across the brain. Figure 4.1 shows a recent variant of the balloon model, where additionally the dynamic transients between steady states are accounted for. Buxton et al. [1998] proved the original balloon model to be flexible enough to account for differ-

*The study was conducted by me and it was supervised by Richard Henson. Preliminary analyses were conducted by Richard Henson. The study design and the results were discussed with Guy Williams.

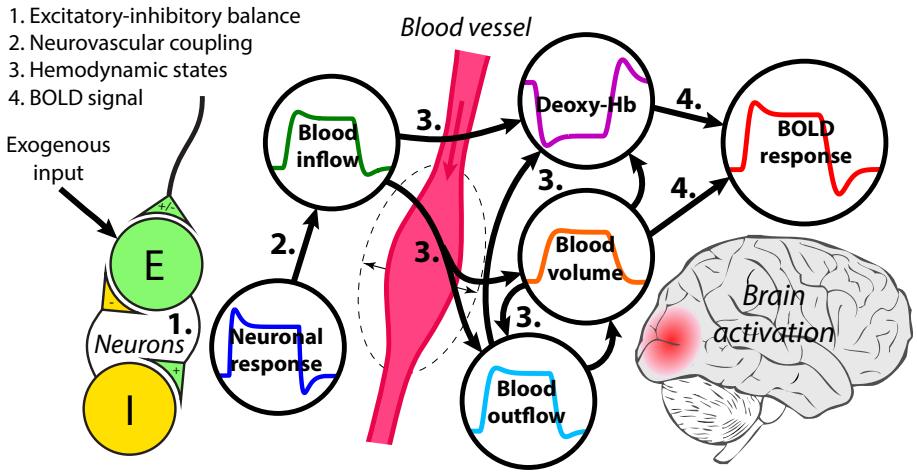


Figure 4.1: To accurately infer neural activity from the BOLD signal, the dynamics between blood flow changes, blood volume changes and deoxyhaemoglobin content changes have to be accounted for, for example following a variant of the original balloon model as presented in Havlicek et al. [2015] (distributed under Creative Commons Attribution-NonCommercial-NoDerivatives Licence (CC BY NC ND 4.0)).

ent characteristics of the BOLD signal, in particular the initial dip [Malonek and Grinvald, 1996] and the post-stimulus undershoot [Davis et al., 1994]. The model allowed simulation of BOLD responses that were similar to acquired BOLD data. The original balloon model was extended in Friston et al. [2000b] and subsequently it was implemented in SPM: the most popular neuroimaging software used for research applications. Friston [2002] described the SPM's implementation of the balloon model, which uses a Bayesian estimation framework and an expectation maximisation (EM) algorithm.

There are studies about differences in the BOLD response between young and older subjects [D'Esposito et al., 1999, Buckner et al., 2000, Ances et al., 2009, Gauthier et al., 2013, Grinband et al., 2017, West et al., 2018], as well as studies investigating young and older subjects with regard to differences in BOLD-related physiological parameters [Ances et al., 2009, De Vis et al., 2015]. Here, using CamCAN (Cambridge Centre for Ageing and Neuroscience, Shafto et al. [2014]) data, I utilised task fMRI images of 641 subjects (18-88 years old) and discovered a relationship between age and the task-evoked hemodynamic response function, as well as the BOLD-derived physiological parameters. For the latter, SPM's balloon model was employed, parameters of which were later related to Magnetoencephalography (MEG)-derived measures and to cardiovascular variables. The MEG-derived measures came from Price et al. [2017], who, using CamCAN data corresponding to the same task as the fMRI study, investigated the impact of age on constant and cumulative delays in auditory and visual evoked fields. Constant delay affects all time points equally, corresponding to a temporal shift of the whole evoked response, including

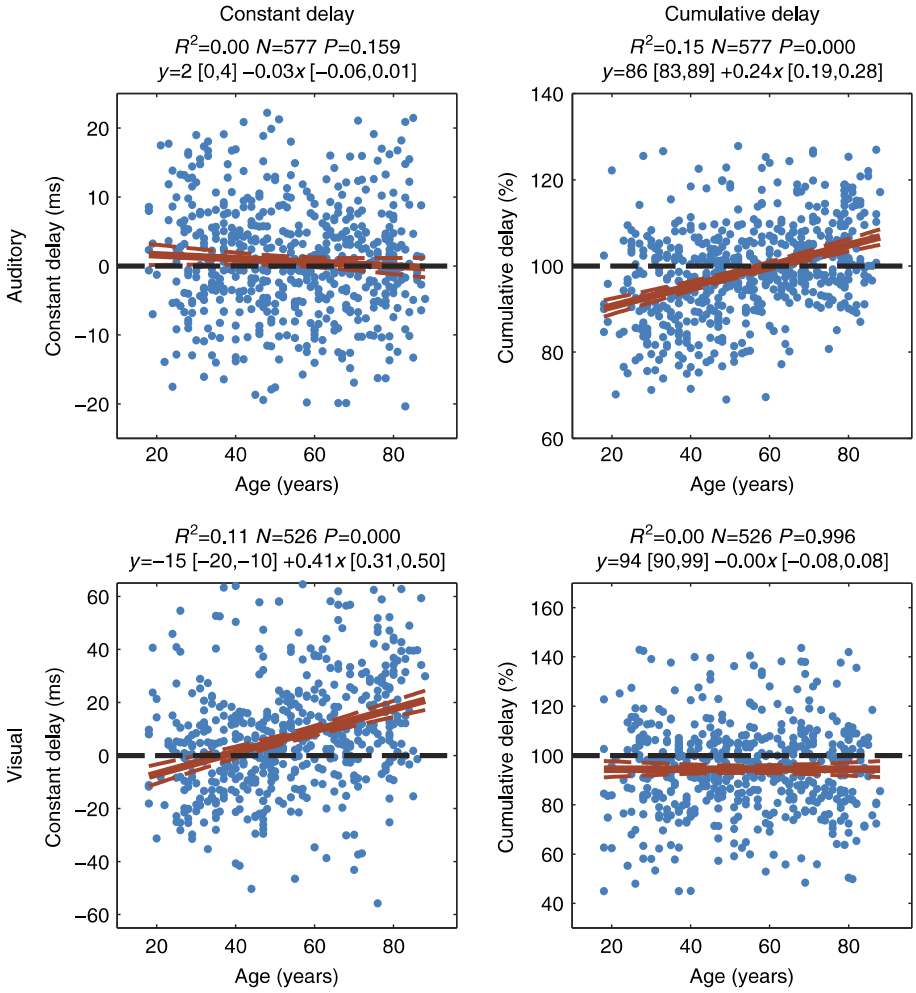


Figure 4.2: Price et al. [2017] found a significant effect of age on cumulative but not constant delay in the auditory evoked field, and a significant effect of age on constant but not cumulative delay in the visual evoked field. From Price et al. [2017] (distributed under Attribution 4.0 International Licence (CC BY 4.0)).

both early and late components. On the other hand, cumulative delay increases with post-stimulus time. Thus, it is easier to detect it for late than for early components. The study found a significant effect of age on cumulative but not constant delay in auditory evoked field, as well as a significant effect of age on constant but not cumulative delay in visual evoked field (Figure 4.2).

4.2 Data

For this study, sensorimotor task fMRI images of 641 (18-88 years old) from Cam-CAN [Shafto et al., 2014] (`http://www.cam-can.org`) were used. The age of the subjects

was approximately uniformly distributed between 18 and 88 years. Ethical approval for the study was obtained from the Cambridgeshire 2 (now East of England - Cambridge Central) Research Ethics Committee. Shafto et al. [2014] described this task. Subjects responded to 129 trials consisting of an initial practice trial, 120 bimodal audio-visual trials, and eight unimodal trials used to discourage strategic responding to one modality (four visual only and four auditory only). The timing of trials was optimised for the estimation of the hemodynamic response function by generating a sequence of stimulation and null trials using a 255-length m-sequence [Buračas and Boynton, 2002] with $m = 2$ and minimal stimulus onset asynchrony (SOA) of 2 s (resulting in SOAs ranging from 2-26 s). Additionally, 0.1-0.3 s jitter was added to the stimulus onsets. Effective sampling rate was around 120 ms. For each bimodal trial, subjects saw two checkerboards presented to the left and right of a central fixation for 34 ms and simultaneously heard a 300 ms binaural tone at one of three frequencies: 300, 600 or 1200 Hz. Each tone frequency was used the same number of times and the order was selected pseudorandomly. For unimodal trials, subjects either only heard a tone or saw the checkerboards. For each trial, subjects responded by pressing a button with their right index finger when they heard or saw any stimuli.

The data were acquired using a T2*-weighted echo-planar imaging sequence with the following parameters: TR = 1970 ms; TE = 30 ms; flip angle = 78 degrees; total of 261 volumes; field of view = 192 x 192 mm^2; 32 axial slices; slice thickness = 3.7 mm; interslice gap = 20%; voxel size = 3 x 3 x 4.44 mm^3; acquisition time = 8 minutes and 40 seconds. The BOLD images that were used in the current study came from 641 subjects and were preprocessed at the MRC Cognition and Brain Sciences Unit following the pipeline described in Taylor et al. [2017]. All preprocessing steps employed SPM 12 (v6906). The 1 mm isotropic T1+T2 images were coregistered to each other, and approximately (rigid-body) aligned to MNI template, which was at 3 mm isotropic resolution. The T1 image was then bias-corrected, and T1+T2 images were segmented into grey matter, white matter, cerebrospinal fluid and three other classes using multimodal segmentation. `DARTEL` was then used to create a sample-specific template of grey matter, and the 12-parameter affine transformation from this template to MNI space was calculated. The fMRI EPI images were undistorted using fieldmaps, and then realigned and further corrected for movement-by-distortion interactions. The data in each slice were interpolated to match the acquisition time of the middle slice. The mean fMRI image was coregistered with the T1 image, and the `DARTEL` warps and MNI-affine transformations were applied to move the fMRI images in the MNI space. Then, the data were smoothed with a Gaussian

filter with 10 mm FWHM and outlying wavelet coefficients were removed using wavelet despiking [Patel et al., 2014].

In the current study, CamCAN MEG-derived measures were employed. These came from study Price et al. [2017], where they are described in detail. This MEG experiment was conducted with the same experimental paradigm as the CamCAN task fMRI study. Both for auditory and visual evoked fields, estimates of MEG constant delay, MEG cumulative delay and MEG amplitude were used. Also, CamCAN cardiovascular health measures were employed: mean systolic pressure, mean diastolic pressure, mean pulse and pulse pressure.

To improve the signal-to-noise ratio for the analysis of the MEG data, Price et al. [2017] performed principal component analysis (PCA) on the trial-averaged event-related fields (ERFs) for both the auditory and the visual tasks. The weights of each principal component reflected the degree to which each channel contributed to this component. The authors estimated the constant and cumulative delays for each subject, both in the auditory and in the visual evoked fields. Constant delay affects all time points equally, corresponding to a temporal shift of the whole evoked response, including both early and late components. On the other hand, cumulative delay increases with post-stimulus time. Thus, it is easier to detect it for late than for early components. For each principal component, a template ERF was calculated as the trial-averaged ERF using data of all the subjects. Then, an iterative procedure based on a gradient ascent algorithm was employed to estimate the constant and cumulative delays for each subject and for both fields. This estimation procedure relied on starting parameters which corresponded to the null hypothesis of no age effect, for which no delays would be observed compared to the group average. In such a case the constant delay would be 0, while the cumulative delay would be 1. In each iteration four new fits were examined: decreasing/increasing the subject's constant delay by 20 ms, and decreasing/increasing the subject's cumulative delay by 10%. In each iteration the fit leading to the highest percentage of explained variance was selected until the improvement over the current best fit was below 1e-6. The estimation procedure was based on a regression model, amplitude scaling factor of which corresponded to the MEG amplitude parameter. Overall, Price et al. [2017] derived the three MEG metrics from stretching the group averages towards subject's individual ERFs.

4.2.1 Data availability

CamCAN data are publicly shared anonymised data.

4.3 Methods

As the CamCAN sensorimotor task was expected to induce activation in auditory, visual and motor regions, the analyses referred to VOIs representing superior temporal gyrus (STG), calcarine cortex, precentral gyrus and supplementary motor cortex (SMC). Each of these regions was analysed separately for the left and right hemispheres. The anatomical masks for these eight VOIs were derived from the SPM's Neuromorphometrics atlas. Superior temporal gyrus corresponds to the auditory region, calcarine cortex corresponds to the visual region, while precentral gyrus and supplementary motor cortex correspond to the motor region, respectively.

The CamCAN data were preprocessed prior to this study as explained in Section 4.2. For each subject, a whole-brain analysis was conducted: the experimental design was convolved with the HRF model, both the data and the model were high-pass filtered using SPM's default cut-off frequency of 1/128 Hz, six motion correction covariates were added to the GLM, and statistical inference was based on an F-test on all HRF-related covariates, which tested the null hypothesis of no experimentally-induced activation. Different HRF models were applied, as discussed later. The stimuli were modelled together with an assumed stimulus duration time of 0.1 s. The default SPM's pre-whitening was used.

The analyses employed SPM 12 (v7219) [Penny et al., 2011]. All the processing scripts needed to fully replicate this study can be found at `https://github.com/wiktorolszowy/fMRI_HRF_vs_age`.

4.3.1 HRF estimation

The HRF shape can be estimated with the canonical model used along its partial derivative with respect to time ("temporal derivative") and its partial derivative with respect to duration ("dispersion derivative") [Henson et al., 2002]. Following first level analyses on all subjects, I estimated HRF for each subject and each voxel multiplying the canonical function with its coefficient and adding it to the temporal and dispersion derivatives multiplied with the coefficients of the temporal and dispersion derivatives, respectively. Furthermore, I estimated HRFs with the help of the Finite Impulse Response (FIR) model in SPM. This approach is based on a pre-specified number of time bins covering a pre-specified time window, where the signal for each voxel is separately averaged in each time bin. Opposed to the canonical model, the FIR model does not produce continuous HRF estimates. In order to estimate HRF at high temporal resolution, 32 time bins of width 0.5 s were used to cover 16 s of the post-stimulus period. For both HRF models, the VOI-

wise HRF estimates were obtained averaging the voxel-wise estimated HRFs across voxels with uncorrected $p < 0.001$ following first-level F-test on all HRF-related covariates. For both HRF models, VOI analyses for which the subject displayed an HRF peak before 2 s or after 8 s were removed as outliers. Throughout this chapter, the canonical HRF along its two derivatives is referred to as "canonical + TD", while the above mentioned FIR model is referred to as "FIR (32 x 0.5s bins)".

4.3.2 Balloon model

For the analysis of age impact on brain physiology, SPM's balloon model was used. It employs seven parameters, which are described in detail in Friston et al. [2000b]. The signal decay parameter (τ_s) corresponds to signal elimination. The effect of an increase is not clear from Friston et al. [2000b], where it is stated: *"Increases in this parameter dampen the rCBF response to any input and will also suppress the undershoot"*. However, Figure 8 in Friston et al. [2000b] shows that a decrease of the signal decay parameter suppressed the post-stimulus undershoot. The confusion might have arisen from a re-formulation of the parameter at some point, as the parameter value of 1.54 from Friston et al. [2000b] corresponds via inversion to the parameter value of 0.65 from Friston [2002] ($1/1.54 \approx 0.65$). The autoregulation parameter (τ_f) refers to the balloon model's coupled differential equations, which correspond to a damped oscillator with a resonance frequency of $\omega = 1/(2\pi\sqrt{\tau_f})$. The physiological nature of this parameter remains unspecified. Increasing the value of the autoregulation parameter reduces the post-stimulus undershoot. The transit time parameter (τ_0) corresponds to the resting venous volume divided by the resting flow. It is the time a blood cell needs to traverse the venous compartment. Higher values correspond to slower dynamics of the BOLD signal. Grubb's exponent (α) models the outflow as a function of volume $f_{\text{out}}(v) = v^{1/\alpha}$. Increasing this parameter increases the degree of nonlinearity in the volume-flow behaviour of the balloon model, but this affects the evoked BOLD responses only negligibly. Increasing the oxygen extraction parameter (E_0) increases the initial dip. Oxygen extraction fraction is high in regions with very low blood flow and in tissues with endogenously high extraction. The "intra:extra ratio" parameter refers to the ratio of intra- to extravascular components of the gradient echo signal. The neural efficacy parameter (ϵ) corresponds to the increase in perfusion signal evoked by neural activity, expressed as the number of evoked transients per second. An increase of this parameter elevates the amplitude of the BOLD response.

The balloon model as expanded in Friston et al. [2000b] is a nonlinear model linking single input with single output. The input is the experimentally-induced neural activ-

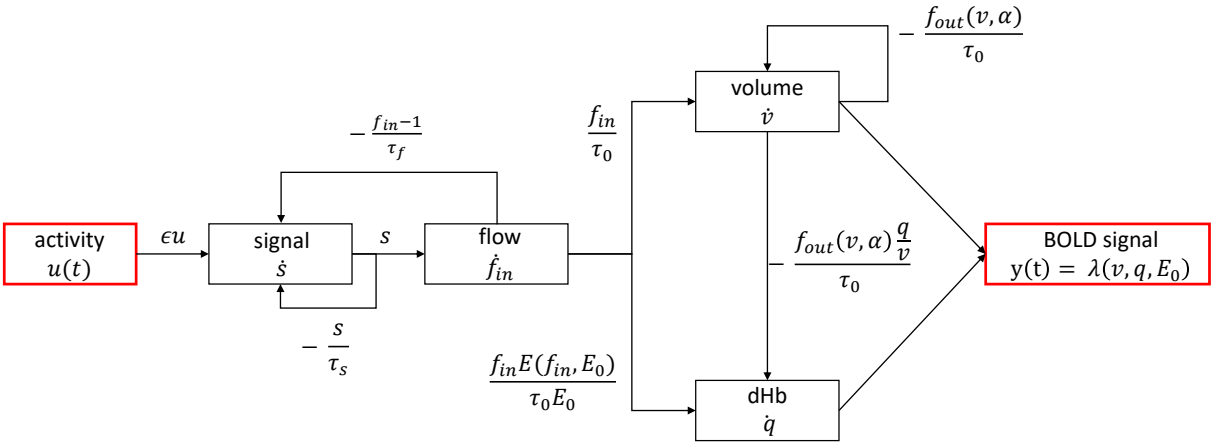


Figure 4.3: Diagram showing the way SPM's balloon model [Friston et al., 2000b] links experimentally-induced neural activity to the BOLD response.

ity, while the output is the BOLD response. Within this framework there are four state variables: neuronally-induced perfusion signal s, inflow f_{in}, venous volume v and deoxyhaemoglobin content q. The state variables are modelled with the help of differential equations as summarised in Fig. 4.3. $\dot{s}$, $\dot{f}_{\text{in}}$, $\dot{v}$ and $\dot{q}$ are the derivatives of the state variables with regard to time.

As BOLD signal is noisy, the estimation of the nonlinear seven-parameter balloon model can be complicated. SPM's estimation is based on an EM algorithm which optimises the parameter values with regard to the model's free energy value [Friston et al., 2003]. One problem is related to the choice of the prior expected values. By default, SPM uses 0.65 as the value of the prior expected value for the signal decay parameter, 0.41 for the autoregulation parameter, 0.98 for the transit time parameter, 0.32 for the Grubb's exponent, 0.34 for the resting oxygen extraction parameter, -1 for the ratio of intra- to extra-vascular components of the gradient echo signal, and 0 for the neural efficacy parameter. These values can be found in SPM's script `spm_hdm_priors.m`, they are listed in Friston et al. [2003] (Table 1) and they are approximately the same as in Friston [2002]. They were derived from a number of studies on rodents and humans, though there is not a paper discussing the appropriateness of this particular set of values. Two recent studies suggested that the prior expected value for the transit time parameter should be 2, while the prior expected value for the resting oxygen extraction parameter should be 0.40 (cf. Table 1A in Havlicek et al. [2015] and Table 2 in Friston et al. [2017]).

In the current study, SPM was used to estimate the balloon model for each of the eight considered VOIs. For each subject and each VOI, voxels with uncorrected $p < 0.001$ following first-level F-test on all HRF-related covariates were selected. For these seemingly

active voxels, SPM's VOI generation utility tool was used, which calculated the first eigenvariate from singular value decomposition [Friston et al., 2006]. The balloon model was estimated only with regard to this summary of the response within functional volumes of interest.

The balloon model parameters were compared to MEG-derived measures and to cardiovascular health markers. In this part of the study, I removed outliers following the procedure from Price et al. [2017]. A subject was considered an outlier if either his estimated MEG constant or MEG cumulative delay was away from the first and third quartiles across all subjects by more than 1.5 times the interquartile range (IQR). In this part of the study, I also removed outliers from the balloon model estimates. A subject was removed as an outlier if for any of the balloon model parameters, the estimated parameter was more than 2.5×IQR away from the first and third quartiles. 2.5 was chosen as a value above the standard 1.5, because the balloon model estimation employs seven parameters, so it is more likely to obtain an extreme value in at least one of them. For the comparison of the balloon model parameters with the cardiovascular health markers, outliers with regard to cardiovascular health were not removed, but outliers with regard to the balloon model estimation were.

4.4 Results

4.4.1 Estimated HRFs

Figure 4.4 presents estimated HRFs for the "canonical + TD" and "FIR (32 x 0.5s bins)" models, for all considered VOIs, averaged across subjects. Estimated HRFs following the use of the canonical model along temporal and dispersion derivatives were much smoother than the estimates following the use of the FIR model. Right superior temporal gyrus was the region with the highest average activation, while right precentral gyrus was the region with the lowest average activation. Figures 4.5-4.6 show estimated HRFs resulting from the use of the same HRF models as above, plotted for all eight VOIs and for subjects sorted by age. For better visualisation, the HRF estimates were smoothed in these figures in the y-direction with a Gaussian kernel of bandwidth 5 subjects. For the "canonical + TD" model, for all VOIs there was a peak delay of 0.5 s to 1 s between the oldest and the youngest subjects. Also, the response width increased with age. These HRF shape characteristics were changing continuously with age. For the FIR model, the estimated hemodynamic responses were much less smooth and though the HRF peak time seemed to have been nearly constant across the lifespan, the response width increased with age.

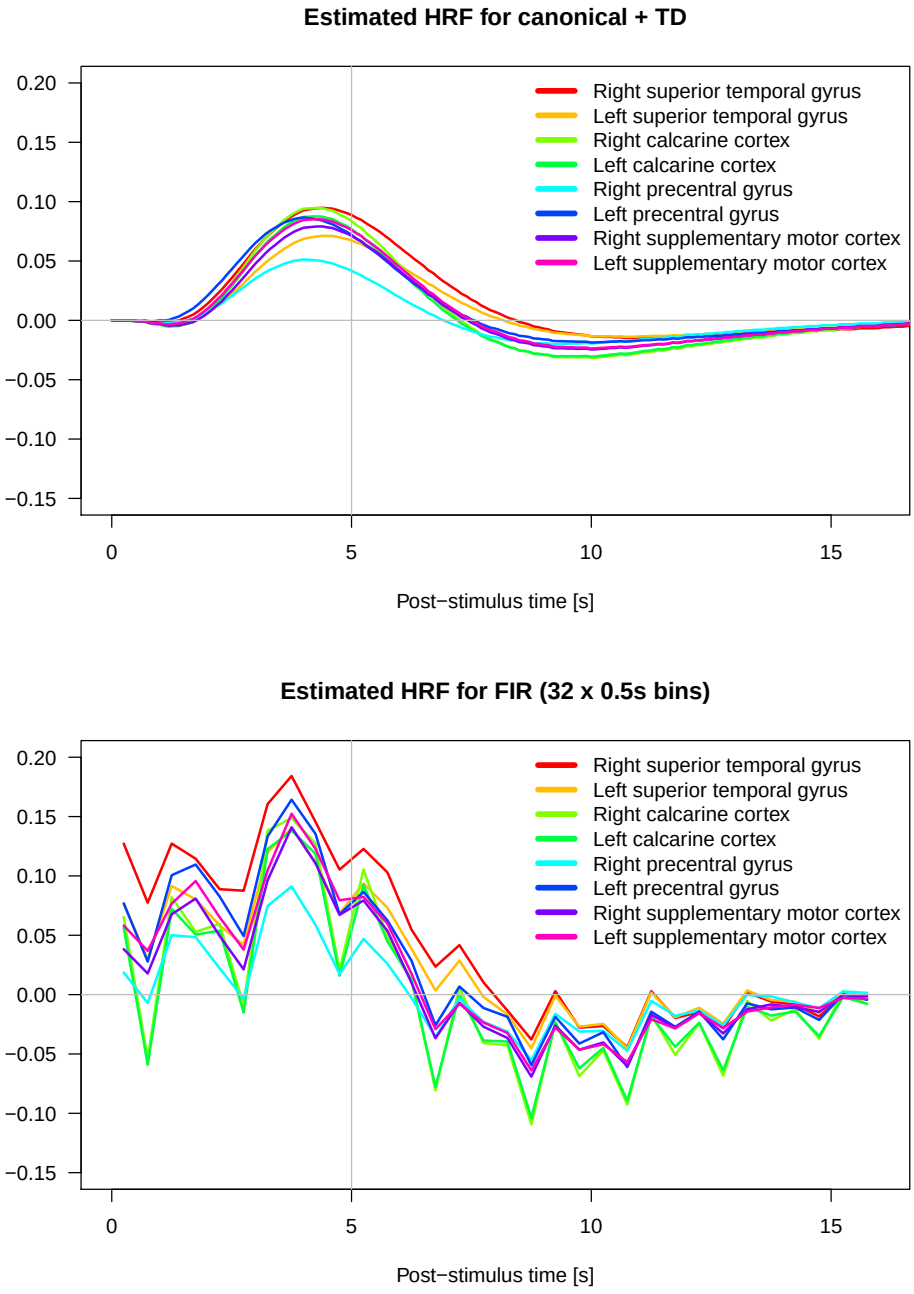


Figure 4.4: Estimated HRF for two HRF models and all eight VOIs.

Across both HRF models, for right precentral gyrus and for right SMC, some of the older subjects displayed more activation than the younger subjects. Across all VOIs, for the vast majority of cases, the HRF peak occurred within 5 s of stimulus onset. Also, there was a negative relationship between age and the magnitude of the post-stimulus undershoot.

Figure 4.7 shows the percentage of seemingly active voxels ($p < 0.001$) to all VOI mask voxels for each of the eight VOIs and for the two HRF models. Furthermore, it

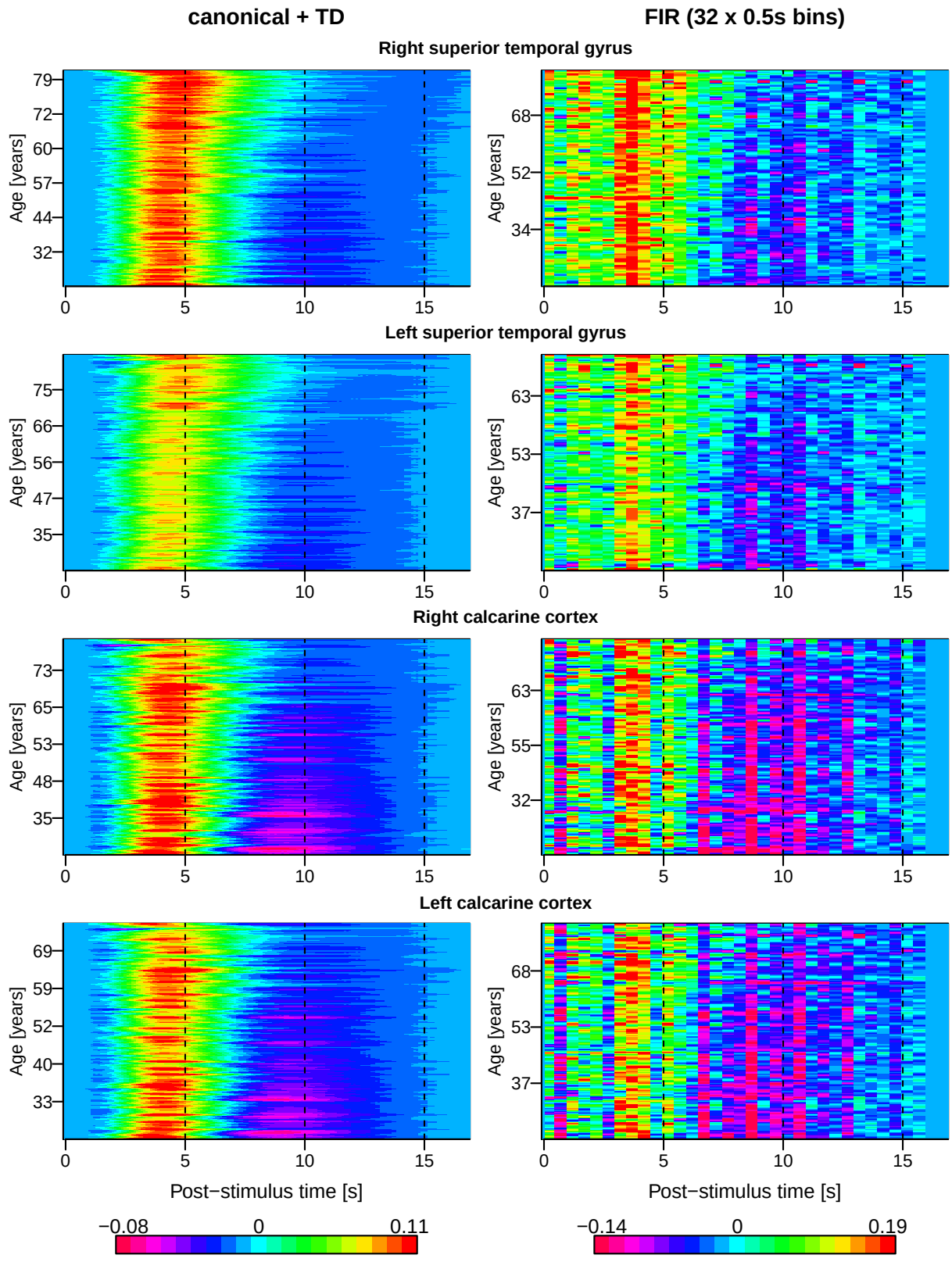


Figure 4.5: Estimated HRF for two HRF models and four VOIs, plotted for subjects sorted by age. Purple refers to lowest activation, light blue to baseline activation, while red to highest activation.

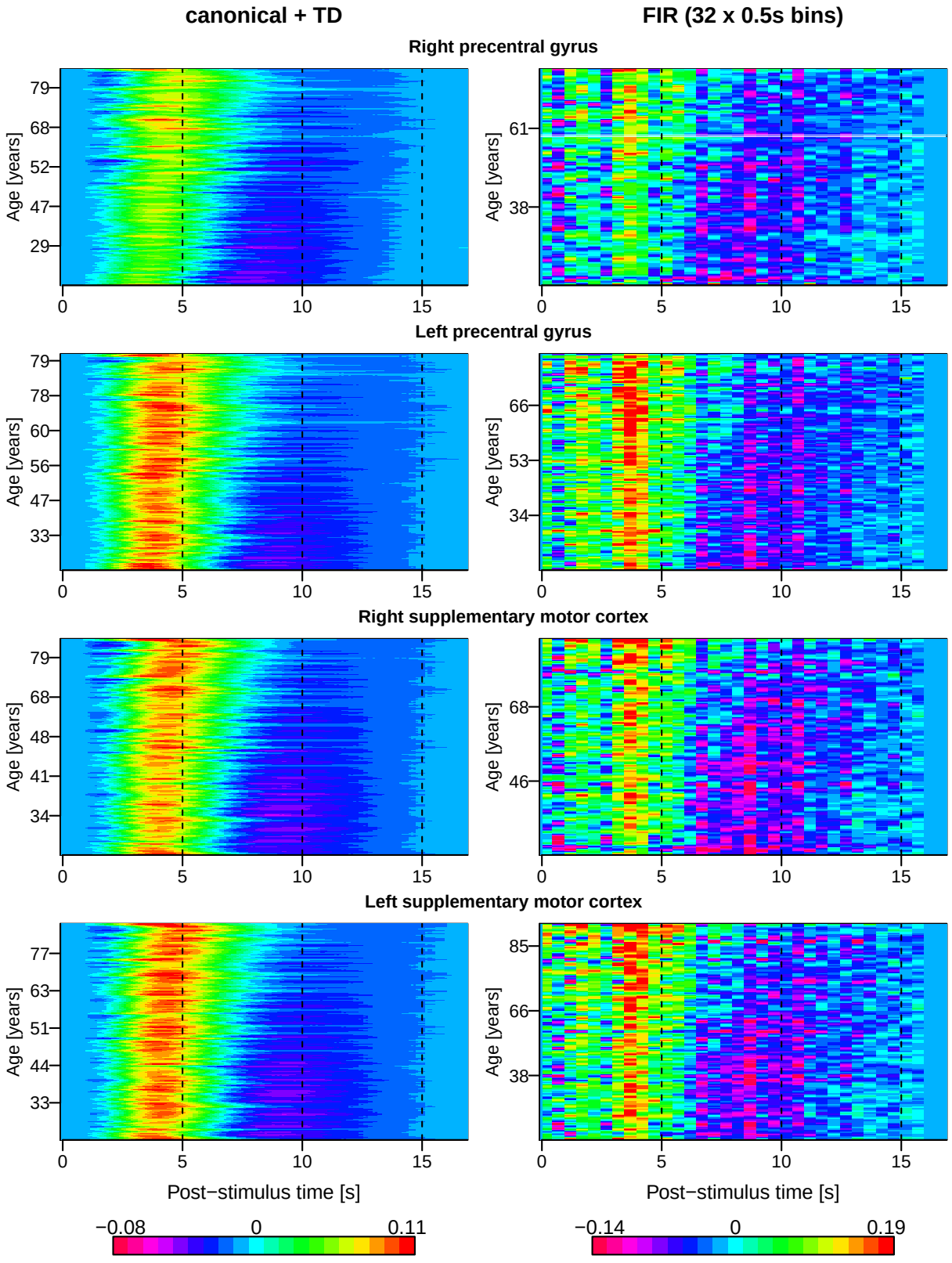


Figure 4.6: Estimated HRF for two HRF models and four VOIs, plotted for subjects sorted by age. Purple refers to lowest activation, light blue to baseline activation, while red to highest activation.

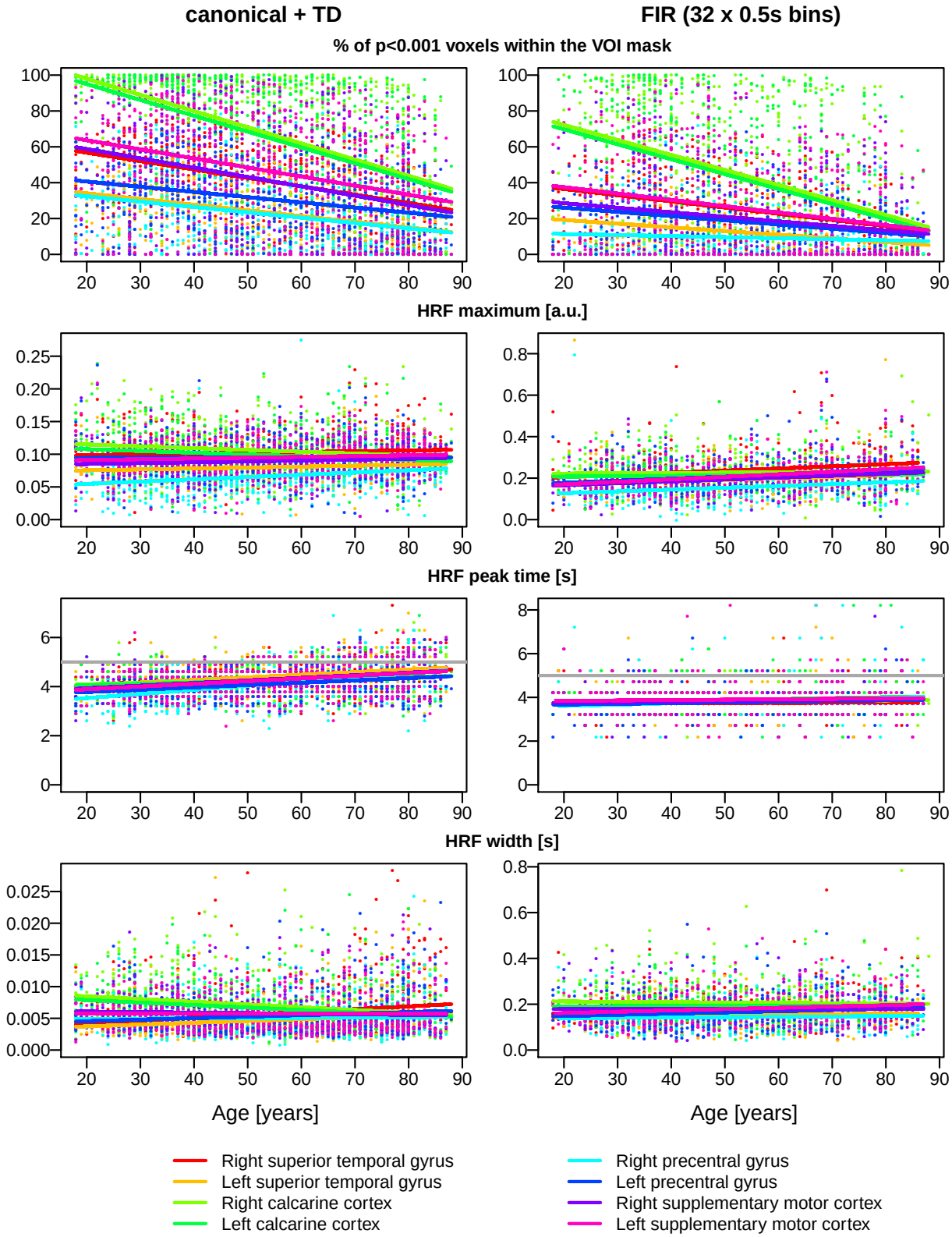


Figure 4.7: Percentage of active voxels to all mask voxels and HRF shape characteristics for two HRF models and different VOIs. The grey horizontal lines refer to the canonical model's peak time.

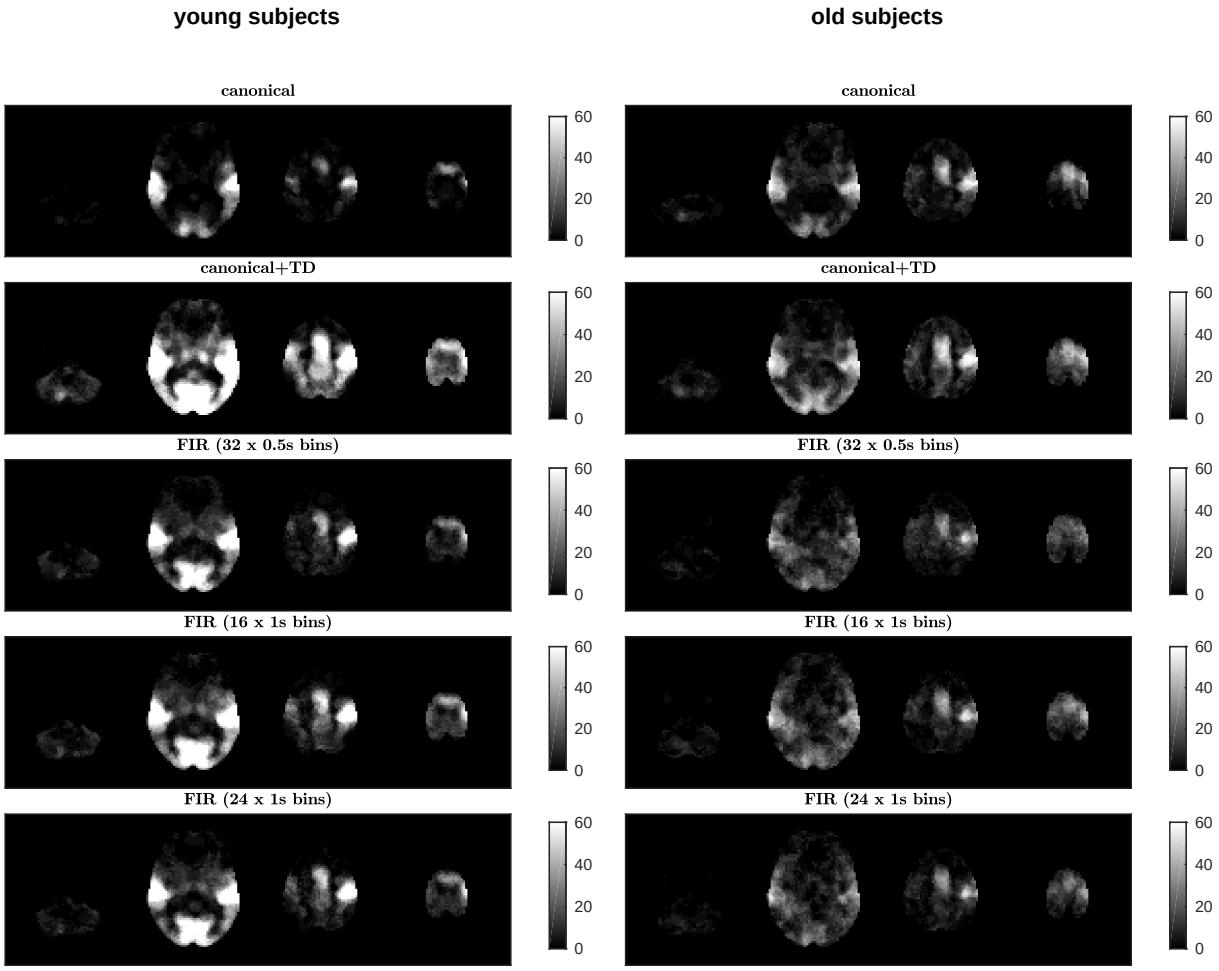


Figure 4.8: Single subject analyses: spatial distribution of significant clusters for six HRF models and an F-test on all HRF-related covariates. 30 youngest subjects (18-25 years old) were compared to 30 oldest subjects (82-88 years old). Four exemplary MNI axial slices from the bottom to the top of the head were selected (left to right). Scale refers to the percentage of subjects where significant activation was detected at the given voxel.

shows changes in HRF shape characteristics plotted against age. "HRF maximum" is the highest value in the estimated HRF, "HRF peak time" is the corresponding location of the HRF maximum, while "HRF width" is approximated with the full width at half maximum (FWHM) of the estimated HRF. Each point refers to one subject and one VOI, and the straight lines are linear regression fits linking age with the variable of interest. For all VOIs, there was a strong link between age and the percentage of seemingly active voxels, with strongest age-related decreases visible for both halves of the calcarine cortex, left superior temporal gyrus and right precentral gyrus. HRF maximum increased with age for most of the VOIs. HRF peak time increased with age for all VOIs, though for the FIR analyses only to a small extent. HRF width increased with age for all regions but the

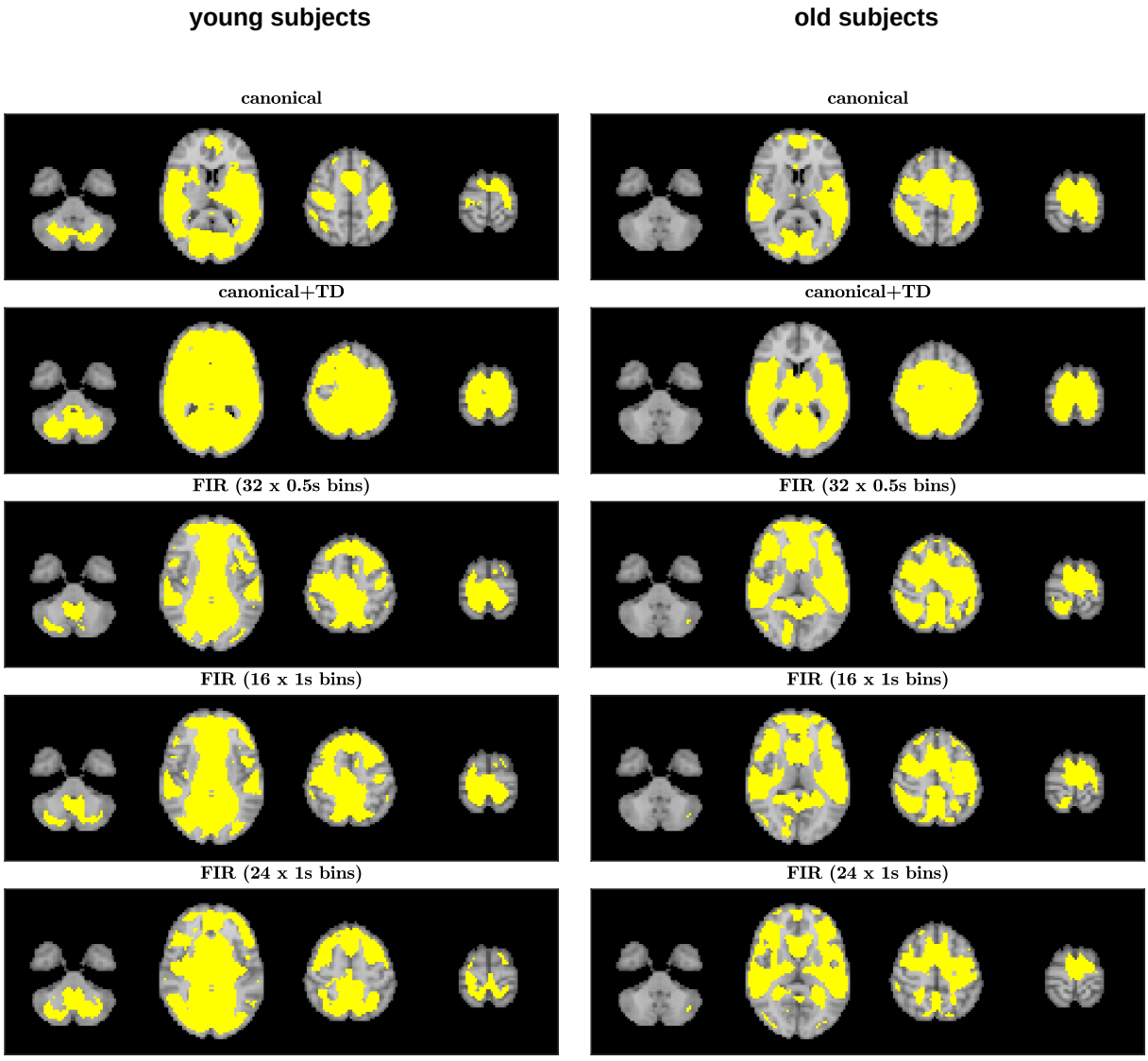


Figure 4.9: Group level analyses: spatial distribution of significant clusters for six HRF models and an F-test on all HRF-related covariates. 30 youngest subjects (18-25 years old) were compared to 30 oldest subjects (82-88 years old). Four exemplary MNI axial slices from the bottom to the top of the head were selected (left to right).

calcarine cortex. The grid-like pattern of points, which is visible for the "HRF peak time" metric, is a result of the temporal resolution of the HRF estimate. For the "canonical + TD" model, HRF was estimated every 0.1 s, while for the "FIR (32 x 0.5s bins)" model, the HRF was estimated every 0.5 s.

In order to investigate whether flexible HRF modelling decreases differences in perceived activation between young and older subjects, a number of HRF models were compared. Figure 4.8 shows the spatial distribution of significant clusters from first level

analyses combined across the 30 youngest (18-25 years old) and across the 30 oldest subjects (82-88 years old) for five HRF models. Apart from the two HRF models for which the HRFs were estimated and discussed above, this analysis also covered the canonical model without derivatives, the FIR model covering 16 s of the post-stimulus period with 1 s wide bins and the FIR model covering 24 s of the post-stimulus period with 1 s wide bins. These results correspond to F-tests on all HRF-related covariates testing the null hypothesis of no experimentally-induced activation. Multiple testing correction was performed with the cluster inference employing a cluster defining threshold of 3.1 and a significance level of 5%. Both for the young and for the older subjects, there was much more significant activation for the canonical model along the temporal and dispersion derivatives than for the canonical model alone. On the contrary, both for the young and for the older subjects, there were very little differences in results across the three FIR models. For all HRF models, the young subjects displayed much more significant activation than the older subjects. Figure 4.9 presents analogous results as Figure 4.8, but for the second level. The yellow blobs refer to significant clusters. The results were similar as for the first level. The addition of the temporal and dispersion derivatives substantially increased the amount of detected significant activation, while the results across FIR models were similar. For all HRF models, the young subjects displayed much more significant activation than the older subjects.

4.4.2 Linking balloon parameters with age and other covariates

Figures 4.10-4.11 show the estimated balloon model parameters for all the eight considered VOIs plotted against age. Each point refers to one subject. In all cases the balloon model was estimated based on the seemingly active voxels ($p < 0.001$) following GLM with the canonical HRF along its two derivatives. The red lines show linear regression fits, while the dashed black lines show SPM's prior expected values of the balloon model parameters. Values above each plot show the Pearson's correlation coefficient along the corresponding p-value. The signal decay parameter was the only parameter which showed positive relationships with age across all VOIs. For all eight VOIs, I found negative relationships between age and the autoregulation parameter, and the neural efficacy parameter. The Grubb's exponent was negatively correlated with age for the auditory and visual regions, while for the motor regions the correlations were not consistent: three VOIs correlated positively and one VOI correlated negatively. For the oxygen extraction parameter, all VOIs except for the right precentral gyrus and the right supplementary motor cortex showed negative relationships with age. The transit time parameter was positively

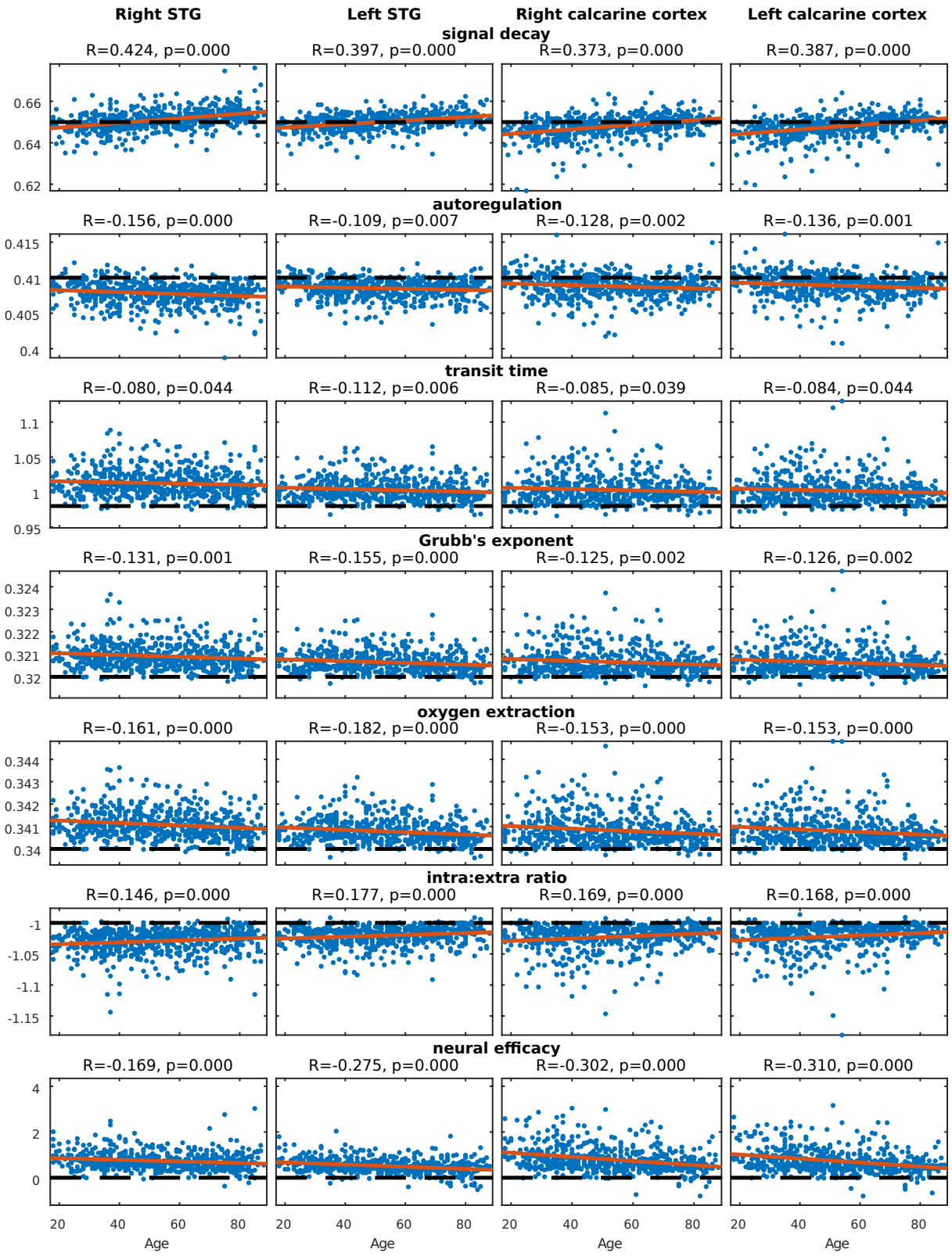


Figure 4.10: Estimated balloon model parameters for four VOIs, plotted for subjects sorted by age. The red lines show linear regression fits, while the dashed black lines show SPM's prior expected values.

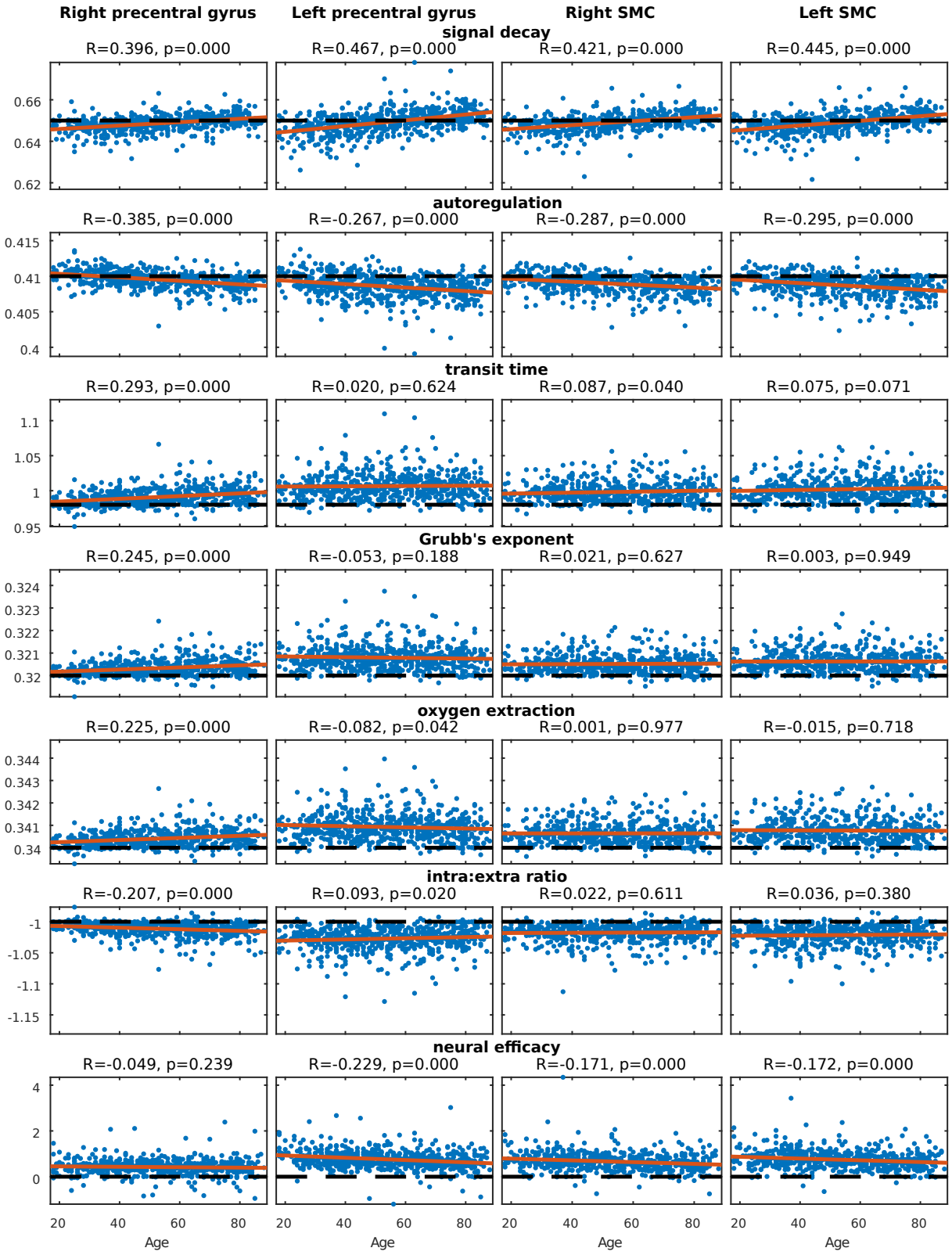


Figure 4.11: Estimated balloon model parameters for four VOIs, plotted for subjects sorted by age. The red lines show linear regression fits, while the dashed black lines show SPM's prior expected values.

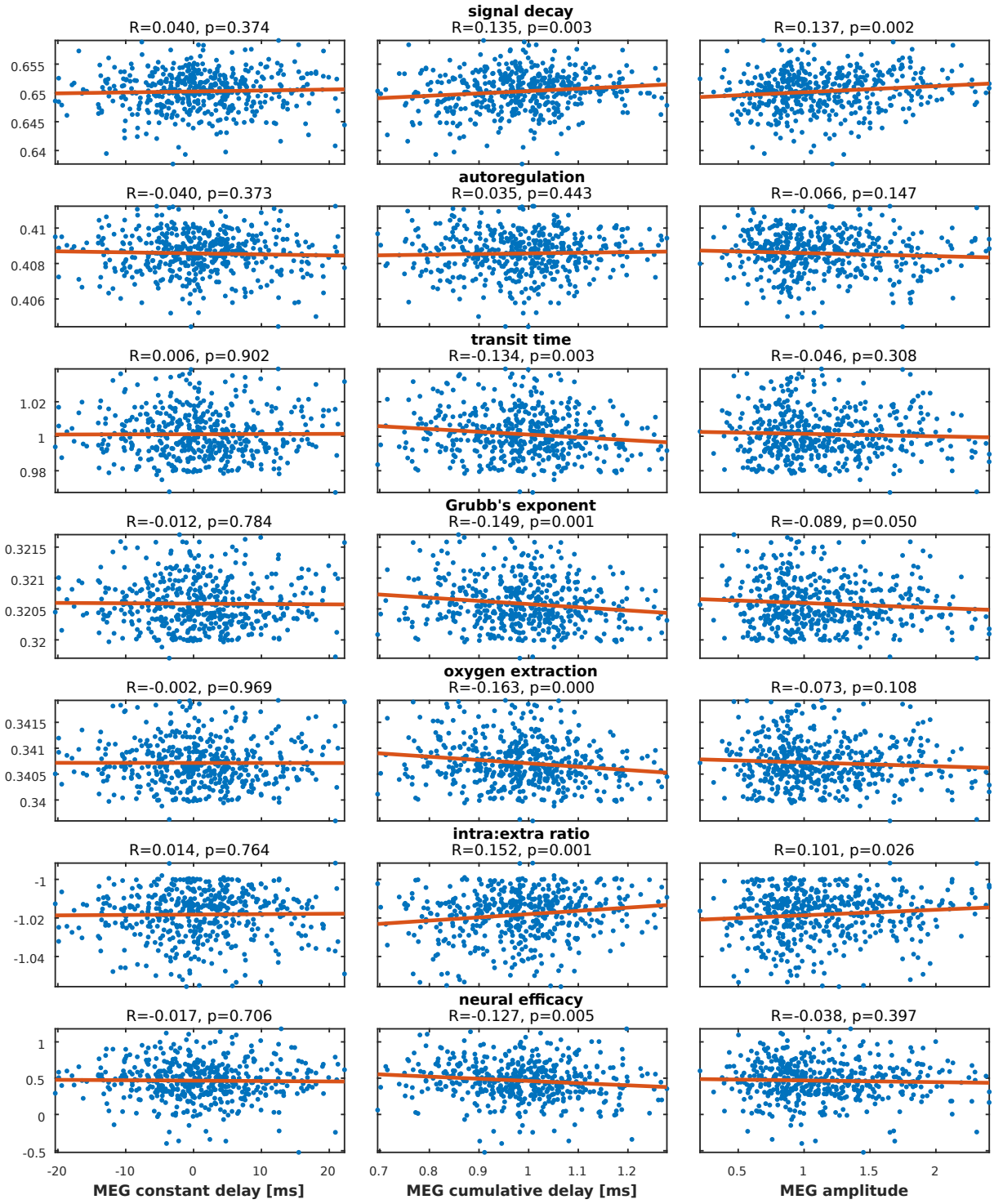


Figure 4.12: Estimated balloon model parameters for left superior temporal gyrus plotted against MEG-derived measures from the auditory evoked field.

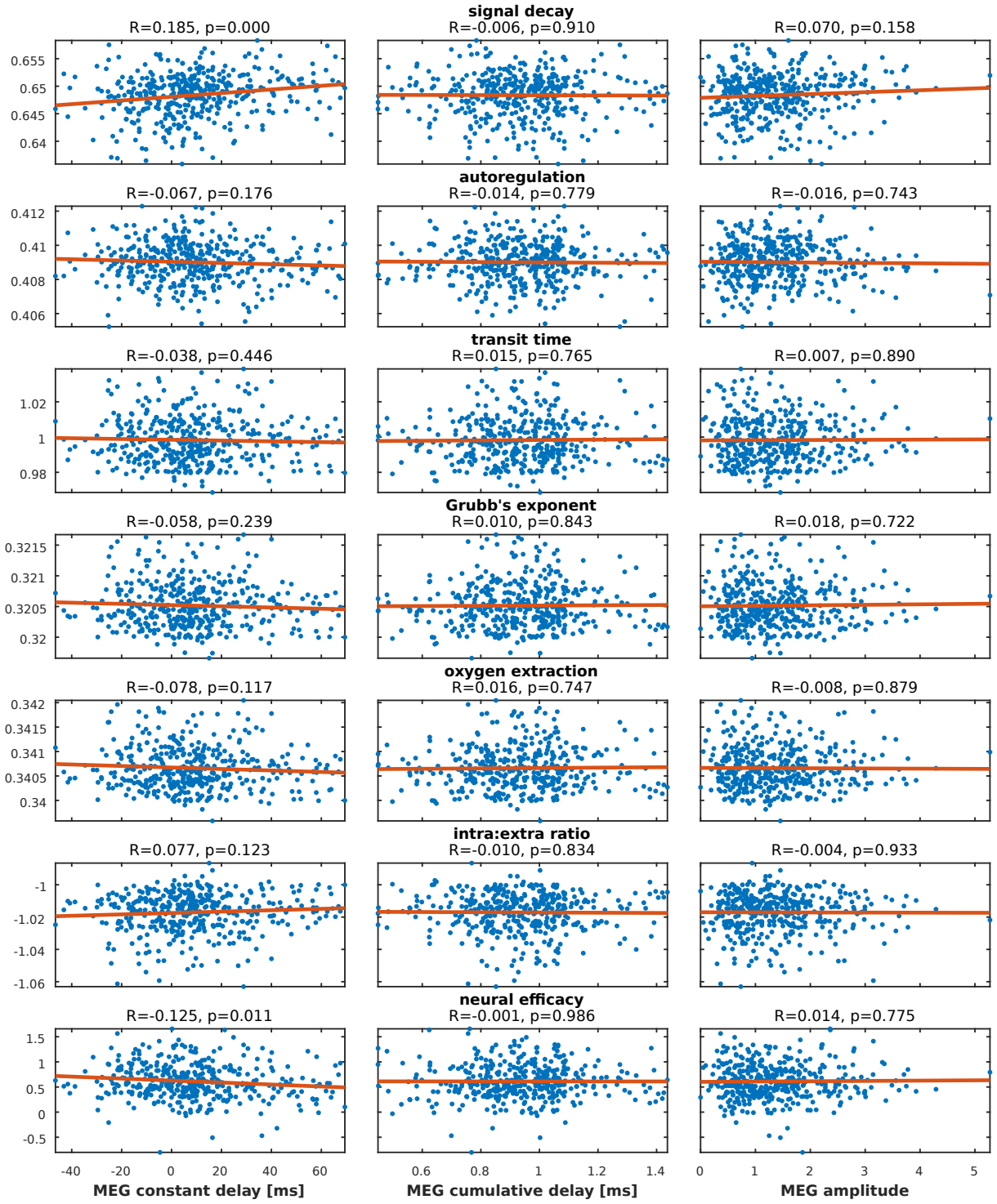


Figure 4.13: Estimated balloon model parameters for left calcarine cortex plotted against MEG-derived measures from the visual evoked field.

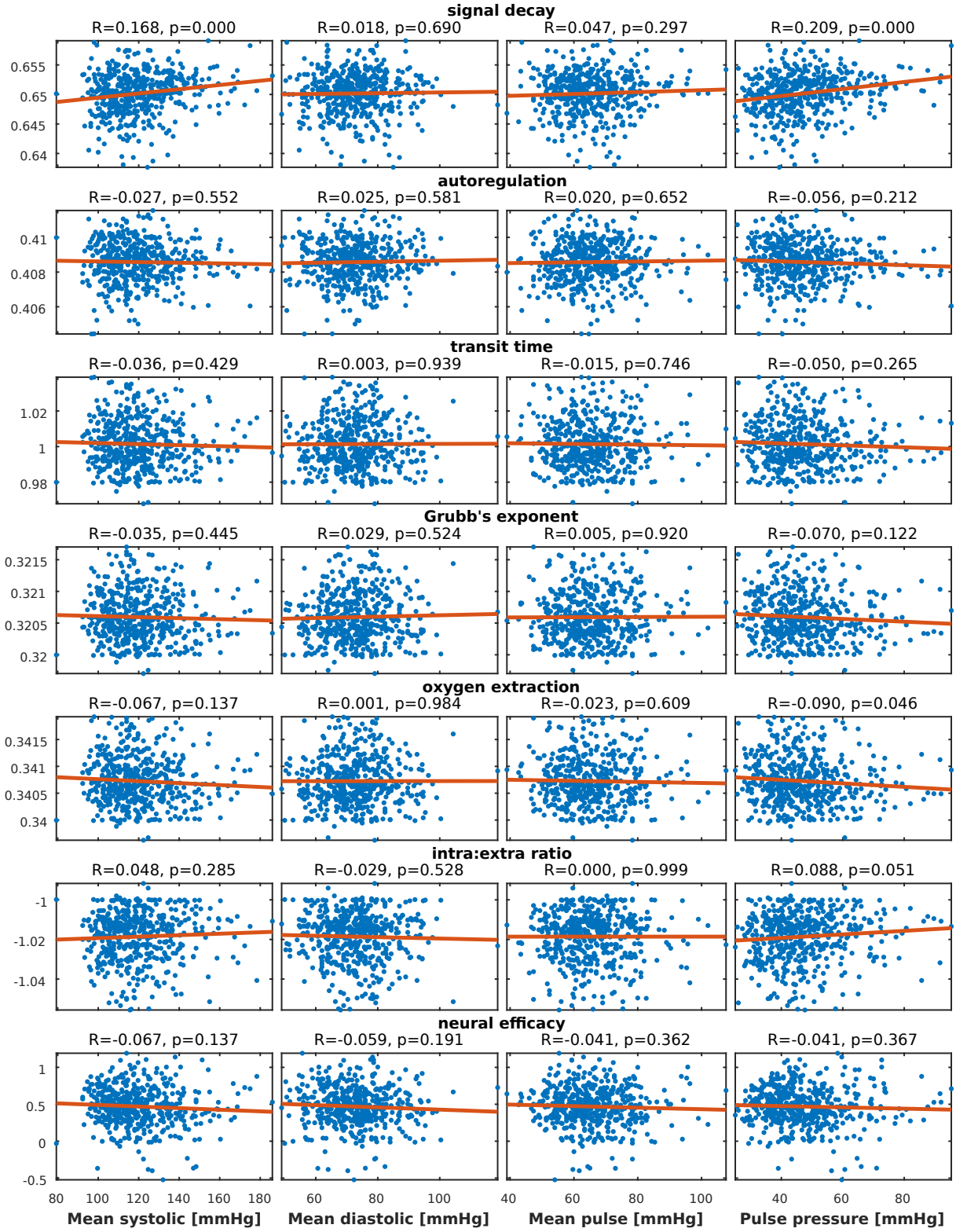


Figure 4.14: Estimated balloon model parameters for left superior temporal gyrus plotted against cardiovascular measures.

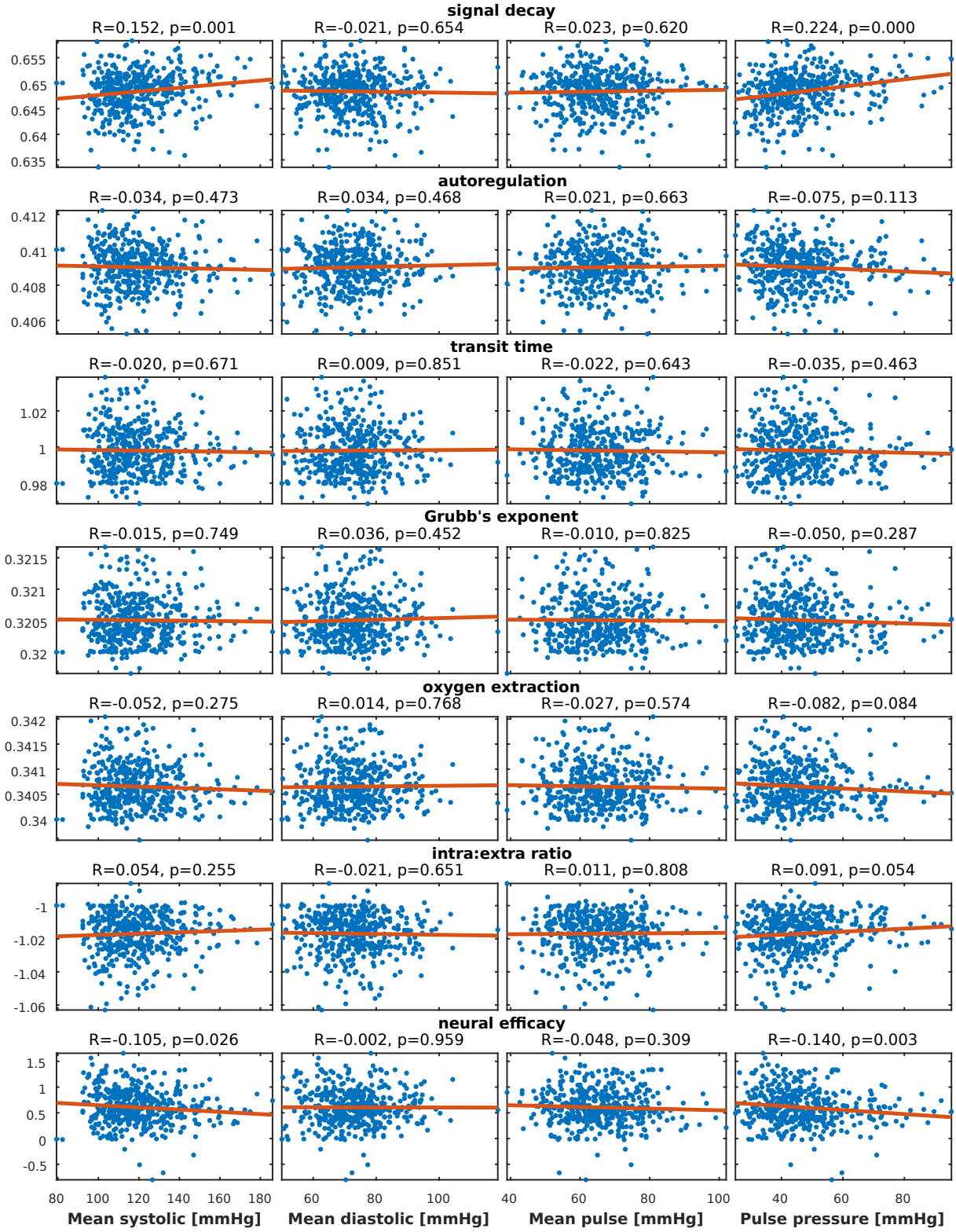


Figure 4.15: Estimated balloon model parameters for left calcarine cortex plotted against cardiovascular measures.

correlated with age for the motor regions and negatively correlated with age for the auditory and visual regions. Finally, for the intra:extra ratio parameter, all VOIs except for the right precentral gyrus showed positive relationships with age. Though the majority of the above relationships were significant ($p < 0.05$), most of the correlations were weak. The strongest correlations were those for the signal decay parameter - for example for left precentral gyrus $R = 0.467$.

For left superior temporal gyrus and left calcarine cortex, I investigated links between the estimated balloon model parameters and the MEG-derived measures, as well as the cardiovascular measures. The estimated balloon model parameters correlated with MEG cumulative delays for left superior temporal gyrus (Figure 4.12). For left calcarine cortex (Figure 4.13), the strongest correlations with the balloon model parameters were for the MEG constant delays, although most of these correlations were not significant ($p > 0.05$).

The signal decay parameter correlated positively with mean systolic pressure and pulse pressure for left superior temporal gyrus (Figure 4.14). The same relationships were observed for left calcarine cortex, for which additionally negative relationships between the same two cardiovascular measures and the neural efficacy parameter were observed (Figure 4.15).

4.4.3 Robustness analysis of SPM's balloon model

Given the above mentioned weak correlations, as well as lack of a sensitivity analysis of the SPM's balloon model estimation procedure in literature, I analysed how robust SPM's balloon model estimation is with regard to the choice of priors. In this subsection, I investigated left superior temporal gyrus only. One confusion is related to the prior expected value of the transit time parameter. While Friston et al. [2000b, 2003] suggested to use 0.98 s, Havlicek et al. [2015], Friston et al. [2017] suggested to use 2 s. Increasing the prior expected value in SPM from 0.98 to 2 caused the posterior values of the transit time parameter to increase from around 0.98 to around 2 (cf. first column in Figure 4.16). The correlation coefficient linking age with the transit time parameter changed sign and increased in magnitude from -0.112 ($p = 0.006$) to 0.385 ($p < 0.001$). Similar confusion is related to the prior expected value of the oxygen extraction parameter. While Friston et al. [2000b, 2003] suggested to use 0.34, Havlicek et al. [2015], Friston et al. [2017] suggested to use 0.40. Increasing the prior expected value in SPM from 0.34 to 0.40 caused the posterior values to increase from around 0.34 to around 0.40 (cf. second column in Figure 4.16). The correlation coefficient linking age with the oxygen extraction parameter changed only negligibly: from -0.182 to -0.181. The above results suggest that

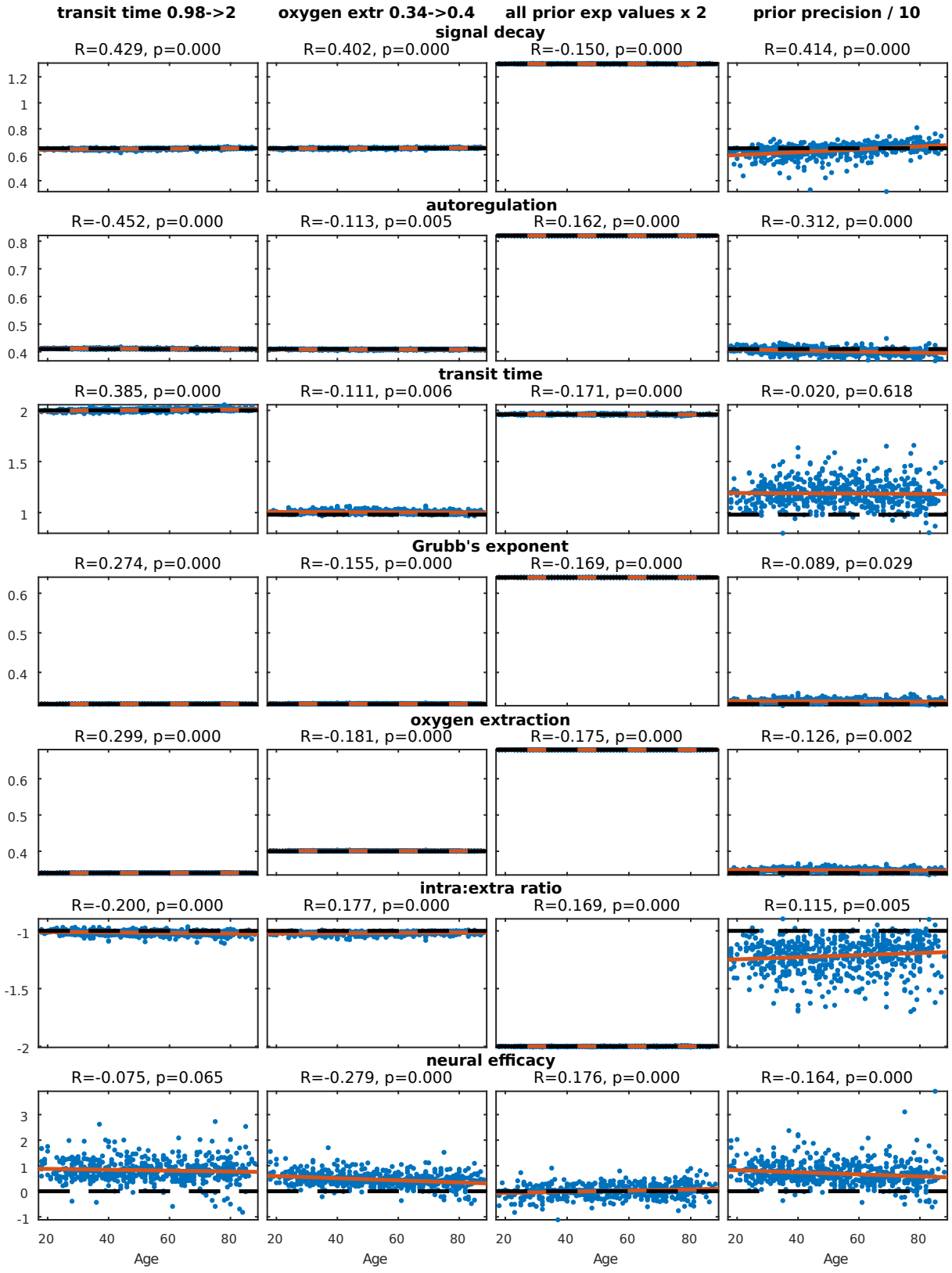


Figure 4.16: Estimated balloon model parameters for left superior temporal gyrus following changes of the priors used in SPM's balloon model estimation procedure.

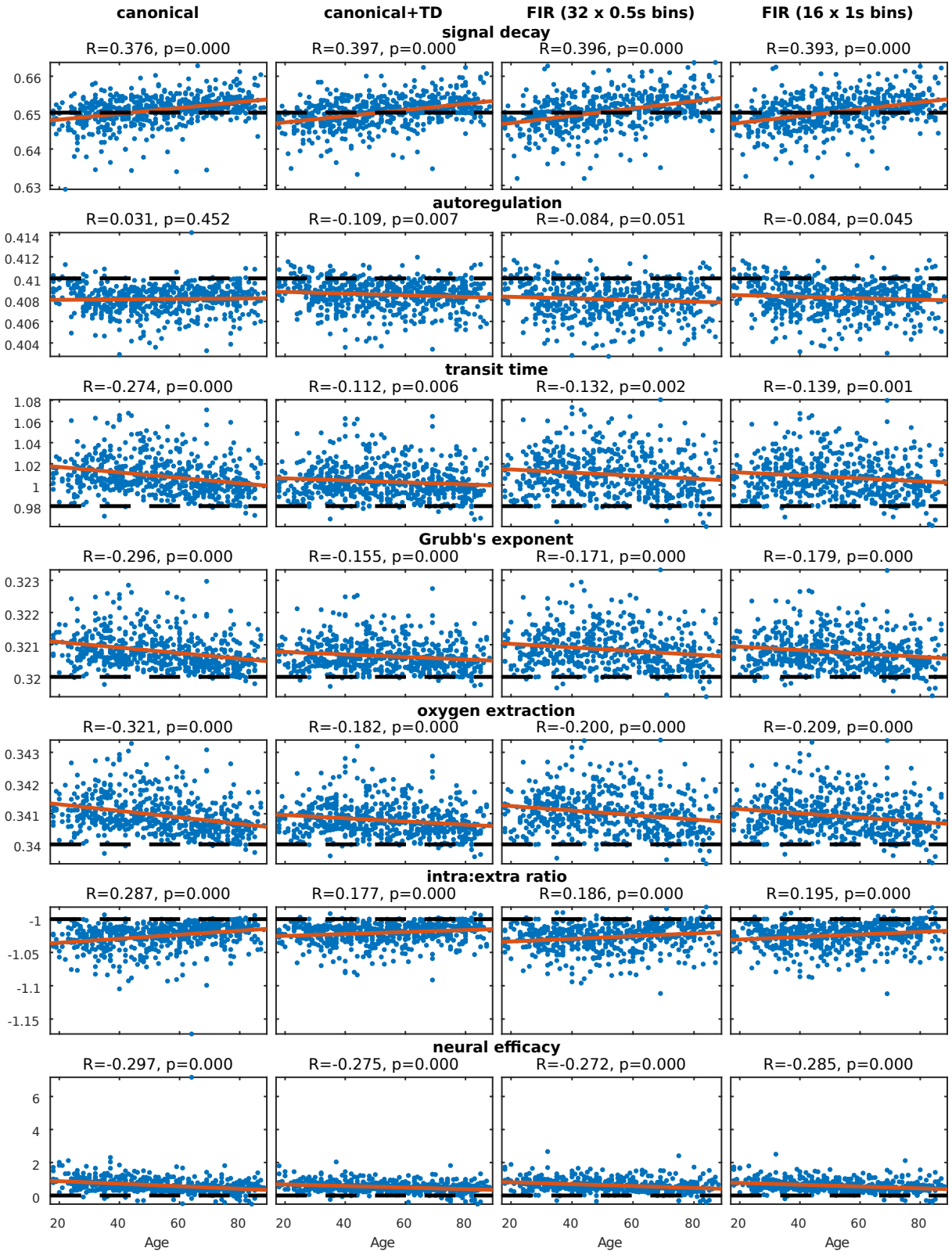


Figure 4.17: Estimated balloon model parameters for left superior temporal gyrus following choice of active voxels obtained with different HRF models. For other analyses, "canonical + TD" was used.

either SPM's balloon model estimation is too dependent on the choice of the priors or that the above two parameters are currently modelled in SPM with very inaccurate prior expected values.

To understand the reasons of the above described SPM's behaviour, I multiplied all default balloon model's prior expected values with 2. The result was that all new posterior values of the balloon model parameters were around the new prior expected values (cf. third column in Figure 4.16). While the above analyses referred to the prior expected values, the SPM's robustness problems might be also related to the employed prior precision. If the prior precision is too high, the estimation procedure might not be flexible enough to divert much from the prior expected values even if the data strongly speak against the prior expected values. That is why, I also estimated the balloon model with SPM's prior precisions divided by 10 (cf. fourth column in Figure 4.16). Interestingly, the majority of the resulting posterior values of the transit time parameter were above the prior expected value of 0.98 and the majority of the resulting posterior values of the oxygen extraction parameter were above the prior expected value of 0.34.

As for all the above balloon model analyses, the balloon model was estimated based on the seemingly active voxels following GLM with the canonical HRF along the two derivatives, the procedure could have been confounded by the choice of the HRF model. Figure 4.17 shows estimated balloon model parameters for left superior temporal gyrus following choice of seemingly active voxels obtained with different HRF models. The differences across HRF models were negligible. The only parameter for which the sign of the correlation coefficient was not the same across all the four HRF models was the autoregulation parameter. Here, for the canonical HRF model, the correlation with age was positive, though not significant, while for the other HRF models, this correlation was negative.

4.5 Discussion

4.5.1 HRF shape variation across the lifespan

Age was found to continuously affect the HRF shape through an increase in the response delay. This relationship was consistent across auditory, visual and motor regions. The FIR estimates were noisy, which probably resulted from averaging FIR estimates across seemingly active voxels rather than estimating the FIR model on an averaged signal from these voxels. Smallest levels of activation were observed for right precentral gyrus. As the motor part of the task involved pressing a button with the right index finger and the vast

majority of subjects were right-handed, lack of strong motor response for the ipsilateral side was expected. The observation that older subjects displayed more activation in right motor regions than younger subjects likely reflects the ageing-linked inability to inhibit the ipsilateral side by the contralateral side. Positive effect of age on the HRF maximum reflects much lower numbers of seemingly active voxels for the older subjects. For these subjects, voxels with $p < 0.001$ displayed on average HRFs with higher amplitudes. While for the "canonical + TD" HRF model, there was a relationship both between age and the peak time, and between age and the response width, for the "FIR (32 x 0.5s bins)" HRF model, the peak time shift occurred only to a very small extent. This might be partially due to the FIR time bins having a width of 0.5 s, which could have precluded observation of a peak shift of less than 0.5 s. Alternatively, it might have been the response width that was primarily changing with age, and the apparent peak shift observed in the "canonical + TD" analyses was due to the inability of that model to differentiate the response width change from the peak time change, a problem already discussed in Lindquist et al. [2007].

D'Esposito et al. [1999] did not find significant difference between young and elderly subjects in the shape of the hemodynamic response function, though the difference was almost significant. However, that study analysed the data of many fewer subjects than available from CamCAN. The former study included 32 young subjects (18-32 years old) and 20 elderly subjects (61-82 years old). For the current study, fMRI data from more than 600 subjects were collected. A higher number of subjects increases statistical power. Moreover, D'Esposito et al. [1999] analysed the HRF shape in the primary sensorimotor cortex only. The finding from D'Esposito et al. [1999] that it was easier to find significant activation in the younger subjects is in line with my observation of a negative relationship between subject's age and the percentage of seemingly active voxels.

A very recent study investigated the link between age and the hemodynamic response function using the CamCAN data [West et al., 2018]. This coincides with the first part of my study. However, opposed to my analyses, West et al. [2018] did not investigate all CamCAN subjects. The study compared a group of 74 younger CamCAN subjects (18-30 years old) to a group of 173 older CamCAN subjects (54-74 years old) and rather than analysing the continuous impact of age on HRF shape, the authors only investigated differences in the HRF shape between these two groups. Furthermore, in West et al. [2018] the analyses were run in native space rather than in MNI space, in FSL rather than in SPM, with much weaker spatial smoothing (FWHM of 5 mm rather than of 10 mm) and employing FMRIB's Linear Optimal Basis Sets (`FLOBS`) HRF model [Woolrich et al., 2004b] rather than the canonical HRF model along the derivatives, or a FIR model.

Nevertheless, the main conclusions regarding the HRF shape are similar as presented in this chapter. Older subjects displayed delayed HRF: the HRF peak was delayed by around 0.5 s compared to younger subjects, and older subjects displayed a wider BOLD response, albeit this difference was not significant. The differences between younger and older subjects with regard to HRF shape characteristics were very similar across the auditory, visual and motor regions. The estimated HRF peaks were below 5 s for the vast majority of subjects. I observed these relationships too. It is reassuring that the results from my analysis of the continuous impact of age on the HRF shape characteristics and the results from the two-group analysis presented in West et al. [2018] converge despite different processing pipelines. Opposed to the current study, West et al. [2018] investigated the possible impact of the anatomical VOI volumes on the detection of age-related HRF changes. Although the study found a significant link between age and the anatomical VOI volumes, age was found to affect HRF shape without anatomical VOI volume mediation. While the recent study clearly shows differences in HRF shape between younger and older subjects, this chapter additionally showed that the impact of age on the HRF shape is continuous.

Interestingly, Grinband et al. [2017] did not find significant differences in the HRF shape between a group of 55 younger subjects (18-30 years old) and a group of 34 older subjects (54-74 years old). However, Figure 1 in Grinband et al. [2017] shows that some of the estimated HRFs had unusual shapes. It is possible that for some subjects, the HRF estimation was confounded by the noisy BOLD signal, especially as the HRF estimation procedure employed the very flexible `FLOBS` model [Woolrich et al., 2004b]. Importantly, while differences in HRF shape were not found to be significant, Figure 1 in Grinband et al. [2017] points to slightly delayed peak time and a larger response width for the older subjects compared to the younger subjects. These differences were consistent across the considered brain regions: the visual and auditory cortices, and agree with the results of the current study.

Buckner et al. [2000] compared the HRF shape for 14 young subjects (18-24 years old, mean age: 21.1) to 14 older subjects (66-89 years old, mean age: 74.9). For the motor cortex, the study found negligible differences between the two groups, while for the visual cortex, lower response amplitudes were observed for the older subjects. Differences in response delays were not observed, but the HRF was estimated only at the temporal resolution of the repetition time, which in that study was 2.68 s. Such a low temporal resolution could have hidden differences between the two subject populations. Figure 2 in Buckner et al. [2000] shows that in both considered regions the HRF peak occurred

around 7 s. The peak could have been higher than in my analyses as the stimulus in the former study was applied for a longer time (1.5 s) and the BOLD signal was not deconvolved. Also, the peak time differences between Buckner et al. [2000] and the current study could have resulted from different temporal resolution at which the HRF was estimated. Moreover, given high levels of noise in the BOLD signal, the samples employed in Buckner et al. [2000] were small.

Large HRF dependence on age supports the use of flexible HRF models in fMRI studies. However, even for the very flexible Finite Impulse Response model covering 16 s of the post-stimulus period with 0.5 s wide bins, "FIR (32 x 0.5s bins)", there were large differences in the amount of significant activation between young and older subjects. These differences existed both at the first and at the second analysis level and indicate that apart from neurovascular changes, grey matter volume and neural changes substantially affect comparisons of subject populations of different age too.

4.5.2 Balloon parameters linked to age, MEG and cardiovascular measures

Compared to West et al. [2018], the current study also endeavoured to explain the differences in HRF shape with the help of physiological parameters. That is why, the balloon model was estimated for every subject and its parameters were compared to age, MEG-derived measures and cardiovascular health markers. Although I found links between balloon model parameters and age across different VOIs, the correlations were not very strong. Given that data of more than 600 subjects were used, statistical significance is not necessarily indicative of the links being relevant. The observed negative relationship between age and the balloon model's oxygen extraction parameter confirms previous studies. For example, De Vis et al. [2015] compared 20 younger subjects (24-33 years old) with 45 older subjects (60-78 years old) and found that while for the young subjects the whole-brain oxygen extraction fraction was on average 0.43, for the older subjects this parameter was on average 0.39 ($p = 0.066$). However, Peng et al. [2014] found a positive relationship between global $CMRO_2$ and subject's age. Puzzlingly, although the balloon model's autoregulation parameter was negatively correlated with age, I observed a negative relationship between age and the magnitude of the post-stimulus undershoot. Friston et al. [2000b] discussed that a decrease of the autoregulation parameter increases the magnitude of the post-stimulus undershoot.

The observed links between balloon model parameters and MEG-derived measures could be explained as an age effect. Price et al. [2017] showed that for the auditory region

age affects the MEG cumulative delays, while for the visual region age affects the MEG constant delays. I found that age affected balloon model parameters across different VOIs, so my findings linking balloon model parameters with MEG-derived parameters only confirm the previous results. Further analyses on this data should be based on a joint regression model which would combine balloon model parameters, age, MEG-derived measures and cardiovascular measures, so that the impact of age could be separated from the impact of MEG-derived measures and other covariates. In such an extended analysis the effect of gender should be analysed too. Weak relationships between balloon model parameters and the MEG-derived measures should not be very surprising as six of the balloon model parameters are hemodynamic. However, it is surprising that I found such weak relationships between the balloon model parameters and the cardiovascular measures, as cardiovascular health could be expected to influence neurovascular coupling to a high extent [Tsvetanov et al., 2015].

4.5.3 Robustness problems of SPM's balloon model

The lack of strong relationships in the above discussed analyses could be related to robustness problems of the SPM's balloon model estimation procedure. Posterior estimates of the balloon model parameters were found to be very close to the prior expected values. When priors were changed, some of the relationships between age and the balloon model parameters changed their direction. It is likely that both the SPM's prior expected values are inaccurate and the SPM's prior precisions are too high. SPM's balloon model estimation problems could explain why the autoregulation parameter was initially found to be negatively correlated with age, although the estimated HRFs indicated a negative relationship between age and the magnitude of the post-stimulus undershoot. The robustness analysis revealed that when all prior expected values were multiplied with 2, the correlation between age and the autoregulation parameter became positive ($R = 0.162$, $p < 0.001$). A positive relationship between age and the autoregulation parameter would agree with the lower magnitude of the post-stimulus undershoot for older subjects.

Possibly, the SPM's estimation algorithm converges too quickly and the posterior values correspond to some local optima. Balloon model estimation problems should not be surprising given high levels of noise in the BOLD signal, high complexity of the balloon model and the lack of previous studies investigating SPM's balloon model estimation robustness. Interestingly, the study which described the SPM's estimation of the balloon model, Friston [2002], stated: *"Normally priors play a critical role in inference; indeed the traditional criticism leveled at Bayesian inference reduces to reservations about the*

validity of the priors employed". The robustness analysis of the SPM's balloon model estimation procedure which was presented in this chapter emphasises that comment.

Hu and Shi [2010] noted the popularity of sensitivity analyses related to complex models in physics, chemistry, economics and social sciences, and to the lack of such studies on biomedical models. The authors mentioned the balloon model, for which some preliminary analyses were presented. For example, it was found that the Grubb's exponent parameter only negligibly affects the estimation procedure. This corresponds to Figure 8 in Friston et al. [2000b], where the impact of the Grubb's exponent parameter on the evoked BOLD response was also shown to be limited. While I am not aware of a study investigating the robustness of SPM's balloon model estimation, Zayane and Laleg-Kirati [2015] investigated robustness of a balloon model similar to the one described in Friston et al. [2000b]. Among others, the authors analysed correlations of the estimated parameters. Also, it was shown that large variations in the signal decay parameter change the resulting BOLD response in a negligible way. This points to possible large balloon model estimation uncertainties and it is surprising as this finding contradicts Friston et al. [2000b], where Figure 8 shows that a change of the signal decay parameter affects the BOLD signal in a pronounced way. Importantly, the data used in Zayane and Laleg-Kirati [2015] came from a block design task, which, as the authors noted, hindered accurate estimation of the balloon model parameters. What is more, the authors used balloon model parameters that are physiologically implausible (cf. Table 1 in Zayane and Laleg-Kirati [2015]). These values appeared in Friston et al. [2000b], where Figure 3 shows a BOLD response that results from such balloon model parameters. For instance, the value of the oxygen extraction parameter used in this example was 0.80, as Friston and colleagues noted: *"We have used a very high value for oxygen extraction to accentuate the early dip"*. The estimated oxygen extraction parameter values, which appear later in Friston et al. [2000b] were much below 0.80 (cf. Figure 7 in Friston et al. [2000b]). Given that a block design task and physiologically implausible values were used in Zayane and Laleg-Kirati [2015], the applicability of their findings to the current work is limited.

4.5.4 Physiological plausibility of the balloon model

Apart from the transit time and the oxygen extraction parameters, confusion also exists with regard to the prior expected value of the Grubb's exponent parameter. Grubb Jr et al. [1974] suggested that at steady state for total CBV $\alpha = 0.38$. On the other hand, when flow and volume are changing dynamically, this value is smaller and Friston et al. [2000b] suggested $\alpha = 0.18$. Recently, Chen and Pike [2009] found that during neural activation

the BOLD-specific Grubb's exponent $\alpha = 0.23$. In that study differences between the visual and sensorimotor areas were not significant. It was argued that the use of $\alpha = 0.38$ in BOLD modelling results in an underestimation of differences in $CMRO_2$. SPM uses $0.32 > 0.23$ as the prior expected value of the Grubb's exponent parameter.

While this chapter investigated issues related to the balloon model estimation, physiological assumptions of the balloon model might have confounded the analyses too. For example, Havlicek et al. [2015] showed how the modelling of dynamic transients between steady states improves the physiological appropriateness of the balloon model. Furthermore, in Buxton [2012] the author of the balloon model discussed physiological limitations of his own model. The balloon model links blood flow, oxygen metabolism and venous blood volume under assumptions of limited oxygen delivery at baseline and a slow recovery of venous blood volume following the stimulus. While this worked well enough in Buxton et al. [1998] to simulate BOLD responses that resembled experimentally acquired data, the balloon model does not account for blood flow and oxygen metabolism being driven in parallel, which possibly reflects different aspects of neural activity [Buxton, 2012]. Moreover, there is still no consensus whether the post-stimulus undershoot is a hemodynamic or a metabolic phenomenon [Buxton, 2012]. Problematically, the coupling of cerebral blood flow and oxygen metabolism differs across the brain, for example Ances et al. [2008] showed that the ratio of fractional changes in CBF to $CMRO_2$ in cortical regions is much higher than in subcortical regions. Also, Peng et al. [2018] showed that cerebrovascular reactivity declines in ageing heterogeneously across the brain. While some previous physiological studies measured CBF and CBV along the BOLD signal using arterial spin labelling (ASL) and vascular space occupancy (VASO), respectively, most of these studies measured total CBV: a weighted sum of arterial, capillary and venous CBV [Havlicek et al., 2015]. Unfortunately, total CBV is not directly embedded in the balloon model. There are advanced techniques to measure venous CBV [Lu and van Zijl, 2012], though these are still in their infancy [Hua et al., 2018].

Overall, given that most fMRI studies try to infer neural activity from the BOLD signal, more work on the physiological underpinnings of the BOLD signal is needed.

4.5.5 Implications for dynamic causal modelling

The findings showing poor robustness of SPM's balloon model estimation are particularly worrying as the balloon model is the basis of the popular BOLD-based dynamic causal model available in SPM [Friston et al., 2003]. Prior expected values of the five hemodynamic parameters employed in SPM's DCM estimation are specified in SPM's script

`spm_fx_fmri.m`. Surprisingly, they differ from those used in the estimation of the balloon model and which appeared in the original DCM paper [Friston et al., 2003]. The only hemodynamic parameter, the prior expected value of which was not changed is the Grubb's exponent. For the signal decay parameter, the prior expected value was changed from 0.65 to 0.64, for the autoregulation parameter the value was changed from 0.41 to 0.32, for the transit time parameter the value was changed from 0.98 to 2.00 and for the oxygen extraction parameter the prior expected value was changed from 0.34 to 0.40, respectively. It might be expected that the prior expected values that appeared in Havlicek et al. [2015] and in Friston et al. [2017], and which differed from SPM's balloon model estimation, referred to the balloon model as implemented in the current SPM's DCM estimation routine.

Nevertheless, I am not aware of a comprehensive robustness analysis of SPM's BOLD-based dynamic causal modelling framework. However, Handwerker et al. [2012] showed an example of how a change of the hemodynamic response can dramatically alter BOLD-based DCM results. In that study a two-node analysis was presented where the hemodynamic response in node 2 had a 1 s delay compared to node 1. The resulting DCM analysis suggested that node 1 predicted node 2. When the hemodynamic response in node 2 was changed in such a way that it displayed a larger post-stimulus undershoot, the DCM analysis suggested that node 2 predicted node 1. This opposed the initial findings. Handwerker et al. [2012] mentioned the relevance of the priors' choice examination, which the findings from this chapter reaffirm.

4.6 Conclusions

I found a continuous relationship between age and the shape of the task-evoked HRF. For older subjects, the estimated HRF was more delayed. This relationship held for all considered VOIs across auditory, visual and motor regions. I explained the relationship between age and the shape of the hemodynamic response function with the help of the balloon model, where BOLD-derived physiological parameters were shown to vary with age too. Linking the estimated balloon model parameters with MEG-derived estimates revealed several relationships: between balloon model estimates and MEG cumulative delays for left superior temporal gyrus, and between balloon model estimates and MEG constant delays for left calcarine cortex. Links between balloon model estimates and cardiovascular health markers were very weak. Given the surprising lack of strong relationships in analyses involving the balloon model, I conducted a sensitivity analysis of

the SPM's balloon model estimation procedure. I found that the balloon model estimates were very dependent on the assumed prior expected values and prior precision. In particular, the CamCAN sensorimotor data support use of higher prior expected values both for the transit time parameter and for the oxygen extraction parameter. Possibly, the estimation procedure could be further improved if the prior precisions were lower. All in all, poor robustness of the SPM's balloon model estimation could have weakened some relationships in my analyses. As the balloon model is the basis of SPM's BOLD-based dynamic causal modelling framework (DCM), my findings support more caution with the interpretation of both the balloon model and of the BOLD-based DCM results, and speak to the need of investigating robustness of both the balloon model and the BOLD-based DCM estimation routines in SPM.

Chapter 5

Discussion

Chapter 2 investigated pre-whitening methods available in AFNI, FSL and SPM. It was shown that the default method in SPM performed worse than FSL, while FSL performed worse than AFNI and SPM's alternative method: `FAST`. FSL and SPM's default method performed worse for all considered fMRI protocols. Differences in pre-whitening performance were shown to affect both first and second level analyses, though second level analyses were affected only to a limited extent. Primarily, poor pre-whitening introduced false positives. False negatives can result from positive residual autocorrelation at high frequencies when an event-related design is used.

Chapter 3 compared a number of popular HRF models which are available in AFNI, FSL and SPM. It was shown that including the temporal and dispersion derivatives along the canonical HRF model increases sensitivity only if the consecutive statistical inference tests all HRF-related covariates. For boxcar designs, the sensitivity benefits resulting from the use of more flexible HRF models decreased. For some boxcar analyses, the amount of perceived activation following the more flexible HRF models was even lower than when the canonical model was used alone.

Chapter 4 showed a continuous link between age and the shape of the hemodynamic response function. In particular, the BOLD response width was found to increase with age, and the magnitude of the post-stimulus undershoot was found to decrease with age. Furthermore, balloon model was employed to link age with BOLD-derived physiological parameters. The balloon model parameters were also linked with MEG-derived neural delay estimates and with cardiovascular health markers. A number of relationships was found, though most of them were very weak. Thus, a basic sensitivity analysis of the SPM's balloon model was performed. It revealed problems both with the prior expected values and with the prior precisions. Robustness problems of the SPM's balloon model

estimation procedure could have confounded the above analyses. They are particularly worrying as the balloon model is the basis of the popular BOLD-based dynamic causal model (DCM) available in SPM [Friston et al., 2003]. It might be expected that SPM's BOLD-based dynamic causal model can suffer from similar problems as SPM's balloon model estimated alone, even though the physiological parameters of the balloon model are only nuisance parameters within the BOLD-based DCM framework, and the hemodynamic priors used in the DCM estimation in SPM were found to slightly differ from priors used in SPM's balloon model estimation.

5.1 Relevance of the findings

The above findings are either novel or indicate that problems previously discussed in the literature are more severe than initially suspected. All analyses were based on large samples: 980 subjects for the pre-whitening study, 772 subjects for the comparison of HRF models and 641 subjects for the ageing study, respectively. Importantly, the comparisons of the pre-whitening methods and of the HRF models employed data corresponding to different fMRI protocols. Recommendations made in this thesis do not involve investing in new hardware, buying new software licences or implementing new processing methods. Instead, it is suggested to use some of the already available fMRI statistical methods instead of some of their alternatives. As a result, both specificity and sensitivity of task fMRI studies can be increased at no additional cost.

While work presented in this thesis suggests that reliability of task fMRI studies is seriously affected by a number of statistical methods, it is difficult to estimate how many task fMRI studies came to wrong conclusions due to imperfect pre-whitening, little sensitive HRF modelling or poor robustness of SPM's balloon model estimation procedure. Recently, Eklund et al. [2018] investigated how many task fMRI studies employed cluster inference with the cluster defining threshold of 2.3, which was shown to lead to high familywise error rates in Eklund et al. [2016]. The authors suggested that at least 10% of task fMRI studies used cluster inference with the problematic cluster defining threshold. However, such an estimate does not refer to the proportion of studies which led to wrong conclusions due to the choice of this threshold: the estimated 10% corresponds merely to an upper boundary of the proportion of interest. Since in fMRI research one rarely knows the ground truth, it is impossible to accurately quantify the severity of the specificity and sensitivity problems which result from the choice of the statistical methods, including methods that were discussed in this thesis. As the use of more accurate pre-whitening

methods and of more sensitive HRF models does not incur any additional costs, their use should become more prevalent.

5.2 Data and code sharing

Unfortunately, the vast majority of previous studies can not be repeated to check if a change in the processing pipeline, for example following a different HRF model, or a different set of prior expected values used in the estimation of the balloon model, changes the study's conclusions. This is due to poor archiving and data-sharing practices, a problem widely recognised now [Poldrack and Gorgolewski, 2014, Eklund et al., 2016, Gorgolewski and Poldrack, 2016, Eklund et al., 2017, 2018]. Furthermore, data processing codes are rarely shared [Baker, 2016, Gorgolewski and Poldrack, 2016]. This is worrying given that fMRI data processing pipelines are complex and there are no quality control procedures in neuroimaging labs related to code production. The example of this thesis might serve as an anecdote. While all the projects were computational and led to the creation of several GitHub repositories, no-one tested any of my codes. Luckily, at least two fMRI researchers from other institutes successfully used one of my tools to plot the power spectra of the GLM residuals in their own analyses.

Usually, not all of the data processing steps are explained in a manuscript. For example, Chen et al. [2018a] notes that most task fMRI papers do not report on the sidedness of the statistical tests involved. Whether a statistical test was one-sided or two-sided is crucial for an investigator who wants to repeat someone's study with one processing step changed: for example using a different HRF model. Although Chen et al. [2018a] states that in most such cases default options of the software are used, transparency would increase if all the data processing codes were available on a platform like GitHub. In such a case, all deviations from the software's default options would be clear. All in all, findings of a task fMRI study where data or codes were not made public should be treated with particular caution, particularly as it is estimated that most scientific studies lead to wrong conclusions [Ioannidis, 2005]. Importantly, all analyses presented in this thesis can be fully repeated using codes available from GitHub (`https://github.com/wiktorolszowy/`). Most of the data used for this thesis are publicly shared data. The other datasets (BMMR and CRIC) can be obtained from me upon request, although a permission from the appropriate principal investigators will be needed. These datasets can not be made public due to restrictions implicitly imposed by the IRBs.

5.3 Limitations and future work

5.3.1 What is the best null data for fMRI methods validation studies?

Problematically, for resting state data treated as task data, it is possible to observe activation both in the posterior cingulate cortex and in the frontal cortex, since these regions belong to the default mode network [Raichle et al., 2001]. In fact, in Supplementary Figure 18 in Eklund et al. [2016] the spatial distribution plots of significant clusters indicate that the significant clusters appeared mainly in the posterior cingulate cortex, even though the assumed design for that analysis was a randomised event-related design. The rest activity in these regions can occur at different frequencies and can underlie different patterns [Stark and Squire, 2001]. Resting state provides an opportunity for day-dreaming, self-reflection and other active mental states, so that the analysis procedures can find activity related to these mental states and reject the null hypothesis of no task-induced activation in favour of the alternative hypothesis of task-induced activation due to a non-existent stimulus. Thus, resting state data are not perfect null data for task fMRI analyses, especially if one uses an approach where a subject with one small cluster in the posterior cingulate cortex enters an analysis with the same weight as a subject with a number of large clusters spread throughout the entire brain. Nevertheless, such an approach was used in Eklund et al. [2012, 2015, 2016]. Task fMRI data tested with a wrong design are not perfect null data either, as an assumed wrong design might be confounded by the underlying true design.

For simulated data, a consensus is needed how to model autocorrelation, spatial dependencies, physiological noise, scanner-induced low-frequency drifts and head motion. Some of the current simulation toolboxes [Welvaert and Rosseel, 2014] enable the modelling of all these aspects of fMRI data, but as the later analyses might heavily depend on the specific choice of parameters, more work is needed to understand how the different sources of noise influence each other. My results for simulated resting state data (Chapter 2) were substantially different compared to acquired real resting state scans. In particular, the percentage of significant voxels for the simulated data was much lower, indicating that the simulated data did not appropriately correspond to the underlying brain physiology. Considering resting state data where the posterior cingulate cortex and the frontal cortex are masked out could be an alternative null. Stark and Squire [2001] suggests treating a mindless task, for example odd-even judgments, as null data. For such task, the location and the magnitude of activity can be exactly predicted. Because there is no perfect fMRI

null data, in Chapter 2 both resting state data with assumed dummy designs and task data with assumed wrong designs were used. As results for both approaches coincided, specificity analyses in Chapter 3 were conducted with task data only.

5.3.2 Choice of software packages

Work presented in this thesis referred to AFNI, FSL and SPM, though there are some research groups that use `BrainVoyager` [Goebel, 2012], `fmristat` [Worsley et al., 2002] or `FreeSurfer` [Fischl, 2012]. However, `BrainVoyager` is a commercial package, `fmristat` has not been updated since 2006, while `FreeSurfer` is primarily used for cortical surface analyses rather than volume analyses. A recent report on statistical thresholding in fMRI studies [Yeung, 2018] investigated 388 task fMRI studies and found that 52.1% of these task studies were conducted with SPM, 20.4% with FSL and 18.3% with AFNI. Only 9.2% of the considered studies were performed with other packages. All in all, the restriction of my work towards AFNI, FSL and SPM was a natural choice, although some of the findings and recommendations presented in this thesis might not be helpful for users of other packages.

5.3.3 Preprocessing

In all the three studies I used fMRI data which were sequentially preprocessed. In each study I only considered one preprocessing order. A recent study [Lindquist et al., 2018] showed that for sequential preprocessing, later steps can reintroduce artefacts removed in prior preprocessing steps. It was argued that combining all preprocessing steps into a single filter would improve the preprocessing performance. Alternatively, covariates/filters could be orthogonalized to each other. Sequential preprocessing without orthogonalization is currently a standard and I did not thoroughly analyse the impact of preprocessing on my results.

In Chapters 2 and 3 I compared AFNI, FSL and SPM with regard to pre-whitening and with regard to hemodynamic response function models, respectively. In these studies I performed preprocessing in each package separately. Preprocessing was not exactly the same, as high-pass filtering/detrending is applied both to the data and to the model, and the use of the very same high-pass filter would require hard-coding in two packages. Then, the possible resulting numerical problems could have confounded the results. However, motion correction was performed in each of the packages using the same number of parameters (six), high-pass filtering employed the same frequency cut-off (1/100 Hz), while the

spatial smoothing applied the same sizes of the kernel. All in all, I expect the confounding effect of the slightly different preprocessing in my studies to be negligible. Importantly, the brain mask, the multiple comparison correction, and the MNI registration were kept exactly the same across the AFNI/FSL/SPM pipelines.

Subsection 2.5.6 discussed problems in SPM related to retrospective motion correction of ultra high field data acquired with a limited acquisition field of view. It would be advantageous if problems related to the SPM's motion correction algorithm were investigated further. Motion correction is particularly relevant for fMRI protocols with very small voxels, as such data are more susceptible to motion-related problems than low resolution data [Yakupov et al., 2017]. Unfortunately, motion correction algorithms in AFNI, FSL and SPM were developed using old fMRI protocols [Cox and Jesmanowicz, 1999, Jenkinson et al., 2002, Friston et al., 1995a] and have not been thoroughly validated, for example, for high spatiotemporal resolution data.

In each of the three studies there were datasets on which slice timing correction was applied. Ideally, slice timing correction should be applied always [Sladky et al., 2011], although if the TR is very short, the sensitivity benefits diminish. Problematically, slice timings are not always known or sometimes might be wrongly specified. For example, for the Functional Connectomes Project, slice timings are not shared, and the paper describing the "BMMR checkerboard" study [Hamid et al., 2015] does not mention slice timings. In such cases, the study authors have to be contacted and sometimes it might happen that they are not sure which slice timings had been applied as proper experimental documentation is missing. The most common fMRI data format, NIfTI, includes space for slice timing information (cell `slice_code`), but this part of the header is usually kept empty. Even if the fMRI scan is in the DICOM format, information on slice timing is often missing. Wrong slice timings are even more problematic than the lack of them. For FSL analyses, a timing file can be used with values between -0.5 and 0.5, but there is confusion whether the highest value refers to the slice acquired first or to the slice acquired last (cf. FSL mailing list `https://www.jiscmail.ac.uk/cgi-bin/webadmin?A2=ind1703&L=FSL&D=0&1=FSL&9=A&J=on&d=No+Match%3BMatch%3BMatches&z=4&P=365734`). Moreover, although several common slice acquisitions are specified in FSL's GUI, there is only one interleave acquisition included, so that for Siemens standard acquisitions with even numbers of slices, slice timings have to be manually specified by the investigator. This increases risk of a mistake. There is a tool which detects slice timings from the fMRI scan [Parker et al., 2014], but it does not handle the case of multiband sequences, and in my experience it sometimes detected wrong slice timings. However, the tool's underlying

principle that slices acquired close in time tend to be more correlated than slices acquired much apart might be used when developing new diagnostic tools.

5.3.4 Multiple comparison correction

A major challenge facing fMRI statistics as a field is establishment of a very reliable multiple comparison correction method. Recently, there have been many studies on multiple comparison correction, for example Smith and Nichols [2009], Chen et al. [2012], Eklund et al. [2016, 2018], Chen et al. [2018b], Lohmann et al. [2018]. However, there is not currently an approach widely considered optimal for all fMRI protocols and most studies employ cluster inference where an arbitrarily selected cluster defining threshold is used. What is more, cluster inference is based on the assumption of smooth statistic maps, for which spatial smoothing is applied during preprocessing. The kernel size of the Gaussian smoothing is chosen by the investigator in an arbitrary way, usually following the package default: for AFNI the default is 4 mm, for FSL it is 5 mm, while for SPM it is 8 mm, respectively. Smoothing strongly affects the results, as was shown for example in Eklund et al. [2015, 2016], as well as in this thesis. Cluster defining threshold also strongly affects the results, which was most notably shown in Eklund et al. [2016]. It is a pitfall of the fMRI analysis pipelines that arbitrarily chosen parameters can heavily distort the results.

5.3.5 Balloon model

The presented robustness analysis of the balloon model (Chapter 4) pointed to several problems in its SPM estimation, but that analysis was only preliminary. In particular, possible convergence problems were not discussed. Only posterior expected values were analysed, though posterior covariances, as well as the values of the objective function (free energy) are of interest too. Importantly, no robustness analysis of SPM's dynamic causal model was performed. The estimation of the balloon model and of the DCM model should be tested for a wide range of physiologically-plausible priors. Furthermore, the convergence stopping rule should be investigated.

5.3.6 Diagnostic tools

I believe that the single most important challenge facing scientists interested in fMRI statistics is the development of tools that diagnose problems related to the processing of fMRI data. For example, my study on pre-whitening led to surprising results only because AFNI, FSL and SPM do not plot the power spectra of the GLM residuals. If

these packages plotted them, pre-whitening-related problems would have been known for a long time. Visual inspection of GLM residuals does not help, as residual autocorrelation primarily occurs at low frequencies. For old versions of SPM, the external toolbox `SPMd` generated a number of diagnostic plots [Luo and Nichols, 2003]. Possibly, it provided too much information, which could have limited its popularity. Also, the `SPMd` package did not investigate whether the specified slice timings were correct. There is need for a comprehensive set of diagnostic tools which can point the fMRI investigator to serious problems in the statistical modelling of data in a straightforward way.

Given how much variability there is with regard to fMRI protocols and how quickly the hardware is improving, diagnostic tools could guarantee reliability of fMRI statistical methods for novel fMRI protocols. For example, the FSL's pre-whitening method was shown to perform well in Woolrich et al. [2001], but this study referred to only two fMRI protocols, where the voxel sizes were large. I showed that the FSL's pre-whitening method is imperfect for many contemporary fMRI protocols. Perhaps, future fMRI protocols will allow image acquisitions with much smaller voxels and much shorter TRs than the current fMRI protocols, and my results supporting appropriateness of pre-whitening methods in AFNI and in SPM when using option `FAST` might not hold. Given the ever-increasing interest in multiband acquisitions, it might be presumed that spatiotemporal resolution of fMRI data will be constantly improving. The shortest TR that I considered was 0.645 s, so already more than the TR in some recent studies, for example in Corbin et al. [2018]. If there were diagnostic tools that many fMRI investigators use, novel problems related to the statistical modelling of fMRI data could get attention sooner and they might be addressed faster.

References

Beau M Ances, Oleg Leontiev, Joanna E Perthen, Christine Liang, Amy E Lansing, and Richard B Buxton. Regional differences in the coupling of cerebral blood flow and oxygen metabolism changes in response to activation: implications for BOLD-fMRI. *NeuroImage*, 39(4):1510–1521, 2008.

Beau M Ances, Christine L Liang, Oleg Leontiev, Joanna E Perthen, Adam S Fleisher, Amy E Lansing, and Richard B Buxton. Effects of aging on cerebral blood flow, oxygen metabolism, and blood oxygenation level dependent responses to visual stimulation. *Human Brain Mapping*, 30(4):1120–1132, 2009.

Thomas William John Ash. *Use of statistical classifiers in the analysis of fMRI data*. PhD thesis, University of Cambridge, 2011.

Solveig Badillo, Thomas Vincent, and Philippe Ciuciu. Group-level impacts of within-and between-subject hemodynamic variability in fMRI. *NeuroImage*, 82:433–448, 2013.

Monya Baker. Why scientists must share their research code. *Nature News*, 2016.

Robert Becker, Matthias Reinacher, Frank Freyer, Arno Villringer, and Petra Ritter. How ongoing neuronal oscillations account for evoked fMRI variability. *The Journal of Neuroscience*, 31(30):11016–11027, 2011.

Craig M Bennett and Michael B Miller. How reliable are the results from functional magnetic resonance imaging? *Annals of the New York Academy of Sciences*, 1191(1): 133–155, 2010.

M Bianciardi, A Cerasa, F Patria, and GE Hagberg. Evaluation of mixed effects in event-related fMRI studies: impact of first-level design and filtering. *NeuroImage*, 22(3): 1351–1370, 2004.

Bharat B Biswal, Maarten Mennes, Xi-Nian Zuo, Suril Gohel, Clare Kelly, Steve M Smith, Christian F Beckmann, Jonathan S Adelstein, Randy L Buckner, Stan Colcombe, et al.

Toward discovery science of human brain function. *Proceedings of the National Academy of Sciences*, 107(10):4734–4739, 2010.

Saskia Bollmann, Alexander M. Puckett, Ross Cunnington, and Markus Barth. Serial correlations in single-subject fMRI with sub-second TR. *NeuroImage*, 166:152 – 166, 2018. ISSN 1053-8119.

Alexander Bowring, Camille Maumet, and Thomas Nichols. Exploring the Impact of Analysis Software on Task fMRI Results. *bioRxiv*, page 285585, 2018.

Geoffrey M Boynton, Stephen A Engel, Gary H Glover, and David J Heeger. Linear systems analysis of functional magnetic resonance imaging in human V1. *Journal of Neuroscience*, 16(13):4207–4221, 1996.

Randy L Buckner, Abraham Z Snyder, Amy L Sanders, Marcus E Raichle, and John C Morris. Functional brain imaging of young, nondemented, and demented older adults. *Journal of Cognitive Neuroscience*, 12(Supplement 2):24–34, 2000.

Edward Bullmore, Michael Brammer, Steve CR Williams, Sophia Rabe-Hesketh, Nicolas Janot, Anthony David, John Mellers, Robert Howard, and Pak Sham. Statistical methods of estimation and inference for functional MR image analysis. *Magnetic Resonance in Medicine*, 35(2):261–277, 1996.

Giedrius T Buračas and Geoffrey M Boynton. Efficient design of event-related fMRI experiments using M-sequences. *NeuroImage*, 16(3):801–813, 2002.

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

Richard B Buxton, Eric C Wong, and Lawrence R Frank. Dynamics of blood flow and oxygenation changes during brain activation: the balloon model. *Magnetic Resonance in Medicine*, 39(6):855–864, 1998.

Owen Carmichael, Adam J Schwarz, Christopher H Chatham, David Scott, Jessica A Turner, Jaymin Upadhyay, Alexandre Coimbra, James A Goodman, Richard Baumgartner, Brett A English, et al. The role of fMRI in drug development. *Drug discovery today*, 2017.

Joshua Carp. On the plurality of (methodological) worlds: estimating the analytic flexibility of fMRI experiments. *Frontiers in Neuroscience*, 6:149, 2012.

Gang Chen, Ziad S Saad, Audrey R Nath, Michael S Beauchamp, and Robert W Cox. FMRI group analysis combining effect estimates and their variances. *NeuroImage*, 60 (1):747–765, 2012.

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. *Human Brain Mapping*, 2018a.

Gang Chen, Yaqiong Xiao, Paul A Taylor, Justin K Rajendra, Tracy Riggins, Fengji Geng, Elizabeth Redcay, and Robert W Cox. Handling Multiplicity in Neuroimaging through Bayesian Lenses with Multilevel Modeling. *bioRxiv*, page 238998, 2018b.

J Jean Chen and G Bruce Pike. BOLD-specific cerebral blood volume and blood flow changes during neuronal activation in humans. *NMR in Biomedicine*, 22(10):1054–1062, 2009.

Nadège Corbin, Nick Todd, Karl J Friston, and Martina F Callaghan. Accurate modeling of temporal correlations in rapidly sampled fMRI time series. *Human Brain Mapping*, 2018.

Robert W Cox. AFNI: software for analysis and visualization of functional magnetic resonance neuroimages. *Computers and Biomedical Research*, 29(3):162–173, 1996.

Robert W Cox and Andrzej Jesmanowicz. Real-time 3D image registration for functional MRI. *Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine*, 42(6):1014–1018, 1999.

TL Davis, RM Weisskoff, KK Kwong, JL Boxerman, BR Rosen, et al. Temporal aspects of fMRI task activation: Dynamic modeling of oxygen delivery. *Proc. Int. Magnetic Resonance in Medicine*, 2:69, 1994.

JB De Vis, J Hendrikse, A Bhogal, A Adams, LJ Kappelle, and ET Petersen. Age-related changes in brain hemodynamics; A calibrated MRI study. *Human Brain Mapping*, 36 (10):3973–3987, 2015.

Mark D'Esposito, Eric Zarahn, Geoffrey K Aguirre, and Bart Rypma. The effect of normal aging on the coupling of neural activity to the bold hemodynamic response. *NeuroImage*, 10(1):6–14, 1999.

Mark D'Esposito, Leon Y Deouell, and Adam Gazzaley. Alterations in the BOLD fMRI signal with ageing and disease: a challenge for neuroimaging. *Nature Reviews Neuroscience*, 4(11):863, 2003.

Anders Eklund, Mats Andersson, Camilla Josephson, Magnus Johannesson, and Hans Knutsson. Does parametric fMRI analysis with SPM yield valid results? – An empirical study of 1484 rest datasets. *NeuroImage*, 61(3):565–578, 2012.

Anders Eklund, Thomas Nichols, Mats Andersson, and Hans Knutsson. Empirically investigating the statistical validity of SPM, FSL and AFNI for single subject fMRI analysis. In *Biomedical Imaging (ISBI), 2015 IEEE 12th International Symposium on*, pages 1376–1380. IEEE, 2015.

Anders Eklund, Thomas E Nichols, and Hans Knutsson. Cluster failure: Why fMRI inferences for spatial extent have inflated false-positive rates. *Proceedings of the National Academy of Sciences*, page 201602413, 2016.

Anders Eklund, Thomas E Nichols, and Hans Knutsson. Reply to Brown and Behrmann, Cox, et al., and Kessler et al.: Data and code sharing is the way forward for fMRI. *Proceedings of the National Academy of Sciences*, 114(17):E3374–E3375, 2017.

Anders Eklund, Hans Knutsson, and Thomas E Nichols. Cluster failure revisited: Impact of first level design and physiological noise on cluster false positive rates. *Human Brain Mapping*, 2018.

Allen D Elster and Jonathan H Burdette. Questions and Answers in Magnetic Resonance Imaging, 2nd edition. *Mosby, Inc*, 1:1–18, 2001.

Bruce Fischl. FreeSurfer. *NeuroImage*, 62(2):774–781, 2012.

Judith M Ford, Matthew B Johnson, Susan L Whitfield, William O Faustman, and Daniel H Mathalon. Delayed hemodynamic responses in schizophrenia. *NeuroImage*, 26(3):922–931, 2005.

Steven D Forman, Jonathan D Cohen, Mark Fitzgerald, William F Eddy, Mark A Mintun, and Douglas C Noll. Improved assessment of significant activation in functional magnetic resonance imaging (fMRI): use of a cluster-size threshold. *Magnetic Resonance in Medicine*, 33(5):636–647, 1995.

Peter T Fox and Marcus E Raichle. Focal physiological uncoupling of cerebral blood flow and oxidative metabolism during somatosensory stimulation in human subjects. *Proceedings of the National Academy of Sciences*, 83(4):1140–1144, 1986.

Karl J Friston. Bayesian estimation of dynamical systems: an application to fMRI. *NeuroImage*, 16(2):513–530, 2002.

Karl J Friston, Andrew P Holmes, Keith J Worsley, J-P Poline, Chris D Frith, and Richard SJ Frackowiak. Statistical parametric maps in functional imaging: a general linear approach. *Human Brain Mapping*, 2(4):189–210, 1994a.

Karl J Friston, Peter Jezzard, and Robert Turner. Analysis of functional MRI time-series. *Human Brain Mapping*, 1(2):153–171, 1994b.

Karl J Friston, Keith J Worsley, Richard SJ Frackowiak, John C Mazziotta, and Alan C Evans. Assessing the significance of focal activations using their spatial extent. *Human Brain Mapping*, 1(3):210–220, 1994c.

Karl J Friston, John Ashburner, Christopher D Frith, J-B Poline, John D Heather, and Richard SJ Frackowiak. Spatial registration and normalization of images. *Human Brain Mapping*, 3(3):165–189, 1995a.

Karl J Friston, Andrew P Holmes, JB Poline, PJ Grasby, SCR Williams, Richard SJ Frackowiak, and Robert Turner. Analysis of fMRI time-series revisited. *NeuroImage*, 2 (1):45–53, 1995b.

Karl J Friston, P Fletcher, Oliver Josephs, Andrew Holmes, MD Rugg, and Robert Turner. Event-related fMRI: characterizing differential responses. *NeuroImage*, 7(1): 30–40, 1998a.

Karl J Friston, Oliver Josephs, Geraint Rees, and Robert Turner. Nonlinear event-related responses in fMRI. *Magnetic Resonance in Medicine*, 39(1):41–52, 1998b.

Karl J Friston, O Josephs, E Zarahn, AP Holmes, S Rouquette, and J-B Poline. To smooth or not to smooth?: Bias and efficiency in fMRI time-series analysis. *NeuroImage*, 12 (2):196–208, 2000a.

Karl J Friston, Andrea Mechelli, Robert Turner, and Cathy J Price. Nonlinear responses in fMRI: the Balloon model, Volterra kernels, and other hemodynamics. *NeuroImage*, 12(4):466–477, 2000b.

Karl J Friston, Daniel E Glaser, Richard NA Henson, S Kiebel, Christophe Phillips, and John Ashburner. Classical and Bayesian inference in neuroimaging: applications. *NeuroImage*, 16(2):484–512, 2002.

Karl J Friston, Lee Harrison, and Will Penny. Dynamic causal modelling. *NeuroImage*, 19(4):1273–1302, 2003.

Karl J Friston, Klaas Enno Stephan, Torben Ellegaard Lund, Alexa Morcom, and Stefan Kiebel. Mixed-effects and fMRI studies. *NeuroImage*, 24(1):244–252, 2005.

Karl J Friston, Pia Rotshtein, Joy J Geng, Philipp Sterzer, and Rik N Henson. A critique of functional localisers. *NeuroImage*, 30(4):1077–1087, 2006.

Karl J Friston, Katrin H Preller, Chris Mathys, Hayriye Cagnan, Jakob Heinzle, Adeel Razi, and Peter Zeidman. Dynamic causal modelling revisited. *NeuroImage*, 2017.

Claudine J Gauthier, Cécile Madjar, Laurence Desjardins-Crépeau, Pierre Bellec, Louis Bherer, and Richard D Hoge. Age dependence of hemodynamic response characteristics in human functional magnetic resonance imaging. *Neurobiology of Aging*, 34(5):1469–1485, 2013.

Alexander Geissler, Rupert Lanzenberger, Markus Barth, Amir Reza Tahamtan, Denny Milakara, Andreas Gartus, and Roland Beisteiner. Influence of fMRI smoothing procedures on replicability of fine scale motor localization. *NeuroImage*, 24(2):323–331, 2005.

Gary H Glover. Deconvolution of impulse response in event-related BOLD fMRI. *NeuroImage*, 9(4):416–429, 1999.

Gary H Glover, Tie-Qiang Li, and David Ress. Image-based method for retrospective correction of physiological motion effects in fMRI: RETROICOR. *Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine*, 44(1):162–167, 2000.

Rainer Goebel. BrainVoyager – Past, present, future. *NeuroImage*, 62(2):748–756, 2012.

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

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, et al. The brain imaging data structure, a format for organizing and describing outputs of neuroimaging experiments. *Scientific Data*, 3:160044, 2016.

Jack Grinband, Tor D Wager, Martin Lindquist, Vincent P Ferrera, and Joy Hirsch. Detection of time-varying signals in event-related fMRI designs. *NeuroImage*, 43(3): 509–520, 2008.

Jack Grinband, Jason Steffener, Qolamreza R Razlighi, and Yaakov Stern. BOLD neurovascular coupling does not change significantly with normal aging. *Human Brain Mapping*, 38(7):3538–3551, 2017.

Robert L Grubb Jr, Marcus E Raichle, John O Eichling, and Michel M Ter-Pogossian. The effects of changes in $PaCO_2$ on cerebral blood volume, blood flow, and vascular mean transit time. *Stroke*, 5(5):630–639, 1974.

Khalid Hamandi, Afraim Salek-Haddadi, Helmut Laufs, Adam Liston, Karl Friston, David R Fish, John S Duncan, and Louis Lemieux. EEG–fMRI of idiopathic and secondarily generalized epilepsies. *NeuroImage*, 31(4):1700–1710, 2006.

Aini Ismafairus Abd Hamid, Oliver Speck, and Michael B Hoffmann. Quantitative assessment of visual cortex function with fMRI at 7 Tesla–test-retest variability. *Frontiers in Human Neuroscience*, 9, 2015.

Daniel A Handwerker, John M Ollinger, and Mark D'Esposito. Variation of BOLD hemodynamic responses across subjects and brain regions and their effects on statistical analyses. *NeuroImage*, 21(4):1639–1651, 2004.

Daniel A Handwerker, Javier Gonzalez-Castillo, Mark D'esposito, and Peter A Bandettini. The continuing challenge of understanding and modeling hemodynamic variation in fMRI. *NeuroImage*, 62(2):1017–1023, 2012.

Martin Havlicek, Alard Roebroeck, Karl Friston, Anna Gardumi, Dimo Ivanov, and Kamil Uludag. Physiologically informed dynamic causal modeling of fMRI data. *NeuroImage*, 122:355–372, 2015.

Richard Henson and Karl J Friston. Convolution models for fMRI. *Statistical parametric mapping: The analysis of functional brain images*, pages 178–192, 2007.

Richard Henson, Cathy J Price, Michael D Rugg, Robert Turner, and Karl J Friston. Detecting latency differences in event-related BOLD responses: application to words versus nonwords and initial versus repeated face presentations. *NeuroImage*, 15(1): 83–97, 2002.

Paul G Hoel et al. Introduction to mathematical statistics. *Introduction to mathematical statistics*, (2nd Ed), 1954.

Zhenghui Hu and Pengcheng Shi. Sensitivity analysis for biomedical models. *IEEE Transactions on Medical Imaging*, 29(11):1870–1881, 2010.

Jun Hua, Peiying Liu, Tae Kim, Manus Donahue, Swati Rane, J Jean Chen, Qin Qin, and Seong-Gi Kim. MRI techniques to measure arterial and venous cerebral blood volume. *NeuroImage*, 2018.

Myung-Ho In and Oliver Speck. Highly accelerated PSF-mapping for EPI distortion correction with improved fidelity. *Magnetic Resonance Materials in Physics, Biology and Medicine*, 25(3):183–192, 2012.

John PA Ioannidis. Why most published research findings are false. *PLoS medicine*, 2(8): e124, 2005.

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, 2002.

Mark Jenkinson, Christian F Beckmann, Timothy EJ Behrens, Mark W Woolrich, and Stephen M Smith. FSL. *NeuroImage*, 62(2):782–790, 2012.

Robert E Kelly Jr, George S Alexopoulos, Zhishun Wang, Faith M Gunning, Christopher F Murphy, Sarah Shizuko Morimoto, Dora Kanellopoulos, Zhiru Jia, Kelvin O Lim, and Matthew J Hoptman. Visual inspection of independent components: defining a procedure for artifact removal from fMRI data. *Journal of Neuroscience Methods*, 189(2):233–245, 2010.

Nicholas Lange and Scott L Zeger. Non-linear Fourier time series analysis for human brain mapping by functional magnetic resonance imaging. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 46(1):1–29, 1997.

David J Larkman, Joseph V Hajnal, Amy H Herlihy, Glyn A Coutts, Ian R Young, and Gösta Ehnholm. Use of multicoil arrays for separation of signal from multiple slices simultaneously excited. *Journal of Magnetic Resonance Imaging*, 13(2):313–317, 2001.

Helmut Laufs, John L Holt, Robert Elfont, Michael Krams, Joseph S Paul, K Krakow, and A Kleinschmidt. Where the BOLD signal goes when alpha EEG leaves. *NeuroImage*, 31(4):1408–1418, 2006.

Brian Lenoski, Leslie C Baxter, Lina J Karam, José Maisog, and Josef Debbins. On the performance of autocorrelation estimation algorithms for fMRI analysis. *IEEE Journal of Selected Topics in Signal Processing*, 2(6):828–838, 2008.

Pierre LeVan, Louise Tyvaert, Friederike Moeller, and Jean Gotman. Independent component analysis reveals dynamic ictal BOLD responses in EEG-fMRI data from focal epilepsy patients. *NeuroImage*, 49(1):366–378, 2010.

Jonathan M Levin, Blaise deB Frederick, Marjorie H Ross, Jonathan F Fox, Heidi L Von Rosenberg, Marc J Kaufman, Nicholas Lange, Jack H Mendelson, Bruce M Cohen, and Perry F Renshaw. Influence of baseline hematocrit and hemodilution on BOLD fMRI activation. *Magnetic resonance imaging*, 19(8):1055–1062, 2001.

Martin Lindquist, Stephan Geuter, Tor Wager, and Brian Caffo. Modular preprocessing pipelines can reintroduce artifacts into fMRI data. *bioRxiv*, page 407676, 2018.

Martin A Lindquist and Tor D Wager. Validity and power in hemodynamic response modeling: a comparison study and a new approach. *Human Brain Mapping*, 28(8): 764–784, 2007.

Martin A Lindquist, Christian Waugh, and Tor D Wager. Modeling state-related fMRI activity using change-point theory. *NeuroImage*, 35(3):1125–1141, 2007.

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):S187–S198, 2009.

Nikos K Logothetis. What we can do and what we cannot do with fMRI. *Nature*, 453 (7197):869, 2008.

Nikos K Logothetis, Jon Pauls, Mark Augath, Torsten Trinath, and Axel Oeltermann. Neurophysiological investigation of the basis of the fMRI signal. *Nature*, 412(6843):150, 2001.

Ji Meng Loh, Martin A Lindquist, and Tor D Wager. Residual analysis for detecting mis-modeling in fMRI. *Statistica Sinica*, pages 1421–1448, 2008.

Gabriele Lohmann, Johannes Stelzer, Eric Lacosse, Vinod J Kumar, Karsten Mueller, Esther Kuehn, Wolfgang Grodd, and Klaus Scheffler. LISA improves statistical analysis for fMRI. *Nature Communications*, 9(1):4014, 2018.

Hanzhang Lu and Peter CM van Zijl. A review of the development of Vascular-Space-Occupancy (VASO) fMRI. *NeuroImage*, 62(2):736–742, 2012.

Yingli Lu, Andrew P Bagshaw, Christophe Grova, Eliane Kobayashi, François Dubeau, and Jean Gotman. Using voxel-specific hemodynamic response function in EEG-fMRI data analysis. *NeuroImage*, 32(1):238–247, 2006.

Yingli Lu, Christophe Grova, Eliane Kobayashi, François Dubeau, and Jean Gotman. Using voxel-specific hemodynamic response function in EEG-fMRI data analysis: An estimation and detection model. *NeuroImage*, 34(1):195–203, 2007.

Torben E Lund, Kristoffer H Madsen, Karam Sidaros, Wen-Lin Luo, and Thomas E Nichols. Non-white noise in fMRI: does modelling have an impact? *NeuroImage*, 29 (1):54–66, 2006.

Wen-Lin Luo and Thomas E Nichols. Diagnosis and exploration of massively univariate neuroimaging models. *NeuroImage*, 19(3):1014–1032, 2003.

Dov Malonek and Amiram Grinvald. Interactions between electrical activity and cortical microcirculation revealed by imaging spectroscopy: implications for functional brain mapping. *Science*, 272(5261):551–554, 1996.

Jonathan L Marchini and Brian D Ripley. A new statistical approach to detecting significant activation in functional MRI. *NeuroImage*, 12(4):366–380, 2000.

Friederike Moeller, Hartwig R Siebner, Stephan Wolff, Hiltrud Muhle, Rainer Boor, Oliver Granert, Olav Jansen, Ulrich Stephani, and Michael Siniatchkin. Changes in activity of striato–thalamo–cortical network precede generalized spike wave discharges. *NeuroImage*, 39(4):1839–1849, 2008.

Martin M Monti. Statistical analysis of fMRI time-series: a critical review of the GLM approach. *Frontiers in Human Neuroscience*, 5(28), 2011.

Karsten Mueller, Jöran Lepsien, Harald E Möller, and Gabriele Lohmann. Commentary: Cluster failure: Why fMRI inferences for spatial extent have inflated false-positive rates. *Frontiers in Human Neuroscience*, 11:345, 2017.

Jeanette A Mumford and Thomas Nichols. Simple group fMRI modeling and inference. *NeuroImage*, 47(4):1469–1475, 2009.

Christopher FH Nam, John AD Aston, and Adam M Johansen. Quantifying the uncertainty in change points. *Journal of Time Series Analysis*, 33(5):807–823, 2012.

Thomas E Nichols, Samir Das, Simon B Eickhoff, Alan C Evans, Tristan Glatard, Michael Hanke, Nikolaus Kriegeskorte, Michael P Milham, Russell A Poldrack, Jean-Baptiste Poline, et al. Best practices in data analysis and sharing in neuroimaging using MRI. *Nature Neuroscience*, 20(3):299, 2017.

Kate Brody Nooner, Stanley J Colcombe, Russell H Tobe, Maarten Mennes, Melissa M Benedict, Alexis L Moreno, Laura J Panek, Shaquanna Brown, Stephen T Zavitz, Qingyang Li, et al. The NKI-Rockland sample: a model for accelerating the pace of discovery science in psychiatry. *Frontiers in Neuroscience*, 6, 2012.

Terrence R Oakes, Tom Johnstone, KS Ores Walsh, Lawrence L Greischar, Andrew L Alexander, Andrew S Fox, and RJ Davidson. Comparison of fMRI motion correction software tools. *NeuroImage*, 28(3):529–543, 2005.

Seiji Ogawa, Tso-Ming Lee, Alan R Kay, and David W Tank. Brain magnetic resonance imaging with contrast dependent on blood oxygenation. *Proceedings of the National Academy of Sciences*, 87(24):9868–9872, 1990.

Seiji Ogawa, David W Tank, Ravi Menon, Jutta M Ellermann, Seong G Kim, Helmut Merkle, and Kamil Ugurbil. Intrinsic signal changes accompanying sensory stimulation: functional brain mapping with magnetic resonance imaging. *Proceedings of the National Academy of Sciences*, 89(13):5951–5955, 1992.

Wiktor Olszowy, Guy B Williams, and John Aston. fMRI experiments: can frequency-domain methods perform better than standard FSL routines? In *Novel Statistical Methods in Neuroscience workshops, Magdeburg, Germany*, 2016.

Wiktor Olszowy, Guy B Williams, Catarina Rua, and John Aston. Validation of the canonical hemodynamic response function model used in fMRI studies. *European Neuropsychopharmacology*, 28:S58–S59, 2018.

Wiktor Olszowy, John Aston, Catarina Rua, and Guy B Williams. Accurate autocorrelation modeling substantially improves fMRI reliability. *Nature communications*, 10(1): 1220, 2019.

David Parker, Georges Rotival, Andrew Laine, and Qolamreza R Razlighi. Retrospective detection of interleaved slice acquisition parameters from fMRI data. In *Biomedical Imaging (ISBI), 2014 IEEE 11th International Symposium on*, pages 37–40. IEEE, 2014.

Ameera X Patel, Prantik Kundu, Mikail Rubinov, P Simon Jones, Petra E Vértes, Karen D Ersche, John Suckling, and Edward T Bullmore. A wavelet method for modeling and despiking motion artifacts from resting-state fMRI time series. *NeuroImage*, 95:287–304, 2014.

Linus Pauling and Charles D Coryell. The magnetic properties and structure of hemoglobin, oxyhemoglobin and carbonmonoxyhemoglobin. *Proceedings of the National Academy of Sciences*, 22(4):210–216, 1936.

Shin-Lei Peng, Julie A Dumas, Denise C Park, Peiying Liu, Francesca M Filbey, Carrie J McAdams, Amy E Pinkham, Bryon Adinoff, Rong Zhang, and Hanzhang Lu. Age-related increase of resting metabolic rate in the human brain. *NeuroImage*, 98:176–183, 2014.

Shin-lei Peng, Xi Chen, Yang Li, Karen M Rodrigue, Denise C Park, and Hanzhang Lu. Age-related changes in cerebrovascular reactivity and their relationship to cognition: A four-year longitudinal study. *NeuroImage*, 174:257–262, 2018.

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*. Academic press, 2011.

Russell A Poldrack and Krzysztof J Gorgolewski. Making big data open: data sharing in neuroimaging. *Nature Neuroscience*, 17(11):1510, 2014.

Jonathan R Polimeni, Ville Renvall, Natalia Zaretskaya, and Bruce Fischl. Analysis strategies for high-resolution UHF-fMRI data. *NeuroImage*, 2017.

Jean-Baptiste Poline, Stephen C Strother, Ghislaine Dehaene-Lambertz, Gary F Egan, and Jack L Lancaster. Motivation and synthesis of the FIAC experiment: reproducibility of fMRI results across expert analyses. *Human Brain Mapping*, 27(5):351–359, 2006.

Jonathan D Power, Kelly A Barnes, Abraham Z Snyder, Bradley L Schlaggar, and Steven E Petersen. Spurious but systematic correlations in functional connectivity MRI networks arise from subject motion. *NeuroImage*, 59(3):2142–2154, 2012.

Darren Price, Lorraine Komisarjevsky Tyler, R Neto Henriques, KL Campbell, Nitin Williams, MS Treder, JR Taylor, Carol Brayne, Edward T Bullmore, Andrew C Calder, et al. Age-related delay in visual and auditory evoked responses is mediated by white- and grey-matter differences. *Nature Communications*, 8:15671, 2017.

Patrick L Purdon and Robert M Weisskoff. Effect of temporal autocorrelation due to physiological noise and stimulus paradigm on voxel-level false-positive rates in fMRI. *Human Brain Mapping*, 6(4):239–249, 1998.

Marcus E Raichle, Ann Mary MacLeod, Abraham Z Snyder, William J Powers, Debra A Gusnard, and Gordon L Shulman. A default mode of brain function. *Proceedings of the National Academy of Sciences*, 98(2):676–682, 2001.

Afra Ritzl, John C Marshall, Peter H Weiss, Oliver Zafiris, Nadim J Shah, Karl Zilles, and Gereon R Fink. Functional anatomy and differential time courses of neural processing for explicit, inferred, and illusory contours: An event-related fMRI study. *NeuroImage*, 19(4):1567–1577, 2003.

Serge ARB Rombouts, Rutger Goekoop, Cornelis J Stam, Frederik Barkhof, and Philip Scheltens. Delayed rather than decreased BOLD response as a marker for early Alzheimer's disease. *NeuroImage*, 26(4):1078–1085, 2005.

Ziad S Saad, Kristina M Ropella, Robert W Cox, and Edgar A DeYoe. Analysis and use of FMRI response delays. *Human Brain Mapping*, 13(2):74–93, 2001.

René Scheeringa, Ali Mazaheri, Ingo Bojak, David G Norris, and Andreas Kleinschmidt. Modulation of visually evoked cortical FMRI responses by phase of ongoing occipital alpha oscillations. *The Journal of neuroscience*, 31(10):3813–3820, 2011.

Meredith A Shafto, Lorraine K Tyler, Marie Dixon, Jason R Taylor, James B Rowe, Rhodri Cusack, Andrew J Calder, William D Marslen-Wilson, John Duncan, Tim Dalgleish, et al. The Cambridge Centre for Ageing and Neuroscience (Cam-CAN) study protocol: a cross-sectional, lifespan, multidisciplinary examination of healthy cognitive ageing. *BMC Neurology*, 14(1):204, 2014.

Ronald Sladky, Karl J Friston, Jasmin Tröstl, Ross Cunnington, Ewald Moser, and Christian Windischberger. Slice-timing effects and their correction in functional MRI. *NeuroImage*, 58(2):588–594, 2011.

Stephen M Smith. Fast robust automated brain extraction. *Human Brain Mapping*, 17 (3):143–155, 2002.

Stephen M Smith and Thomas E Nichols. Threshold-free cluster enhancement: addressing problems of smoothing, threshold dependence and localisation in cluster inference. *NeuroImage*, 44(1):83–98, 2009.

Craig EL Stark and Larry R Squire. When zero is not zero: the problem of ambiguous baseline conditions in fMRI. *Proceedings of the National Academy of Sciences*, 98(22): 12760–12766, 2001.

Jerzy P Szaflarski, Mark DiFrancesco, Thomas Hirschauer, Christi Banks, Michael D Privitera, Jean Gotman, and Scott K Holland. Cortical and subcortical contributions to absence seizure onset examined with EEG/fMRI. *Epilepsy & Behavior*, 18(4):404–413, 2010.

Jason R Taylor, Nitin Williams, Rhodri Cusack, Tibor Auer, Meredith A Shafto, Marie Dixon, Lorraine K Tyler, Richard N Henson, et al. The Cambridge Centre for Ageing and Neuroscience (Cam-CAN) data repository: structural and functional MRI, MEG, and cognitive data from a cross-sectional adult lifespan sample. *NeuroImage*, 144:262–269, 2017.

Stefan Thesen, Oliver Heid, Edgar Mueller, and Lothar R Schad. Prospective acquisition correction for head motion with image-based tracking for real-time fMRI. *Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine*, 44(3):457–465, 2000.

Nick Todd, Steen Moeller, Edward J Auerbach, Essa Yacoub, Guillaume Flandin, and Nikolaus Weiskopf. Evaluation of 2D multiband EPI imaging for high-resolution, whole-brain, task-based fMRI studies at 3T: Sensitivity and slice leakage artifacts. *NeuroImage*, 124:32–42, 2016.

Kamen A Tsvetanov, Richard NA Henson, Lorraine K Tyler, Simon W Davis, Meredith A Shafto, Jason R Taylor, Nitin Williams, and James B Rowe. The effect of ageing on fMRI: Correction for the confounding effects of vascular reactivity evaluated by joint fMRI and MEG in 335 adults. *Human Brain Mapping*, 36(6):2248–2269, 2015.

Monroe P Turner, Nicholas A Hubbard, Dinesh K Sivakolundu, Lyndahl M Himes, Joanna L Hutchison, John Hart, Jeffrey Spence, Elliot Frohman, Teresa Frohman, Darin Okuda, et al. Preserved canonicality of the BOLD hemodynamic response reflects healthy cognition: Insights into the healthy brain through the window of multiple sclerosis. *NeuroImage*, 2018.

Martin Walter, Joerg Stadler, Claus Tempelmann, Oliver Speck, and Georg Northoff. High resolution fMRI of subcortical regions during visual erotic stimulation at 7 T. *Magnetic Resonance Materials in Physics, Biology and Medicine*, 21(1):103–111, 2008.

B Douglas Ward. Deconvolution analysis of fMRI time series data [software manual]. *Retrieved from: https://afni.nimh.nih.gov/pub/dist/doc/manual/Deconvolvem.pdf*, 1998/2006.

Andreas Weibull, H Gustavsson, Sören Mattsson, and Jonas Svensson. Investigation of spatial resolution, partial volume effects and smoothing in functional MRI using artificial 3D time series. *NeuroImage*, 41(2):346–353, 2008.

Marijke Welvaert and Yves Rosseel. A review of fMRI simulation studies. *PLOS ONE*, 9 (7):e101953, 2014.

Marijke Welvaert, Joke Durnez, Beatrijs Moerkerke, Geert Verdoolaege, and Yves Rosseel. neuRosim: An R package for generating fMRI data. *Journal of Statistical Software*, 44 (10):1–18, 2011.

Kathryn L. West, Mark D. Zuppichini, Monroe P. Turner, Dinesh K. Sivakolundu, Yuguang Zhao, Dema Abdelkarim, Jeffrey S. Spence, and Bart Rypma. BOLD hemodynamic response function changes significantly with healthy aging. *NeuroImage*, 2018.

Mark W Woolrich, Brian D Ripley, Michael Brady, and Stephen M Smith. Temporal autocorrelation in univariate linear modeling of FMRI data. *NeuroImage*, 14(6):1370–1386, 2001.

Mark W Woolrich, Timothy EJ 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, 2004a.

Mark W Woolrich, Timothy EJ Behrens, and Stephen M Smith. Constrained linear basis sets for HRF modelling using Variational Bayes. *NeuroImage*, 21(4):1748–1761, 2004b.

Keith J Worsley and Karl J Friston. Analysis of fMRI time-series revisited again. *NeuroImage*, 2(3):173–181, 1995.

Keith J Worsley, CH Liao, J Aston, V Petre, GH Duncan, F Morales, and AC Evans. A general statistical analysis for fMRI data. *NeuroImage*, 15(1):1–15, 2002.

Melissa Emily Wright and Richard Wise. Can blood oxygenation level dependent functional magnetic resonance imaging be used accurately to compare older and younger populations? A mini literature review. *Frontiers in Aging Neuroscience*, 10:371, 2018.

Renat Yakupov, Juan Lei, Michael B Hoffmann, and Oliver Speck. False fMRI activation after motion correction. *Human Brain Mapping*, 38(9):4497–4510, 2017.

Andy WK Yeung. An updated survey on statistical thresholding and sample size of fMRI studies. *Frontiers in Human Neuroscience*, 12:16, 2018.

Chadia Zayane and Taous Meriem Laleg-Kirati. A sensitivity analysis of fMRI balloon model. *Computational and mathematical methods in medicine*, 2015, 2015.

Glossary

ASL	Arterial Spin Labelling
BOLD	Blood Oxygenation Level Dependent
BMMR	Biomedical Magnetic Resonance
CamCAN	Cambridge Centre for Ageing and Neuroscience
CBF	Cerebral Blood Flow
CBV	Cerebral Blood Volume
$CMRO_2$	Cerebral Metabolic Rate of Oxygen Consumption
CRIC	Cambridge Research into Impaired Consciousness
DCM	Dynamic Causal Modelling
FCP	Functional Connectomes Project
FIR	Finite Impulse Response
fMRI	functional Magnetic Resonance Imaging
FWER	Familywise Error Rate
FWHM	Full Width at Half Maximum
GLM	General Linear Model
HRF	Hemodynamic Response Function
IRB	Institutional Review Board
MEG	Magnetoencephalography
MNI	Montreal Neurological Institute
NKI	Nathan Kline Institute
OEF	Oxygen-Extraction Fraction
SMC	Supplementary Motor Cortex
SNR	Signal to Noise Ratio
STG	Superior Temporal Gyrus
TR	Repetition Time
VASO	Vascular Space Occupancy
VOI	Volume of Interest