Accurate autocorrelation modeling substantially improves fMRI reliability

Wiktor Olszowy[a], John Aston[b], Catarina Rua[a], Guy B. Williams[a]

[a] *Wolfson Brain Imaging Centre, Department of Clinical Neurosciences, University of Cambridge, Cambridge, United Kingdom*
[b] *Statistical Laboratory, Department of Pure Mathematics and Mathematical Statistics, University of Cambridge, Cambridge, United Kingdom*

Abstract

Given the recent controversies in some neuroimaging statistical methods, we compared the most frequently used functional Magnetic Resonance Imaging (fMRI) analysis packages: AFNI, FSL and SPM, with regard to temporal autocorrelation modeling. This process, sometimes known as pre-whitening, is conducted in virtually all task fMRI studies. We employed eleven datasets containing 980 scans corresponding to different fMRI protocols and subject populations. Though autocorrelation modeling in AFNI was not perfect, its performance was much higher than the performance of autocorrelation modeling 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. Also, we observed very severe problems for scans with short repetition times. The resulting false positives and false negatives can be expected to propagate to the group level, especially if the group analysis is performed with a mixed effects model. Our results show superior performance of SPM's alternative pre-whitening: FAST, over the default SPM's method. The reliability of task fMRI studies would increase with more accurate autocorrelation modeling. Furthermore, reliability could increase if the analysis packages provided diagnostic plots. This way the investigator would be aware of residual autocorrelated noise in the GLM residuals. We provide a MATLAB script for the fMRI researchers to check if their analyses might be affected by imperfect pre-whitening.

Keywords: fMRI, statistics, methods validation, temporal autocorrelation

Introduction

Functional Magnetic Resonance Imaging (fMRI) data is known to be positively autocorrelated in time (Bullmore et al., 1996). It results 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 prolonged to the subsequent time points, which increases the likelihood of obtaining false positives in task studies. As a result, parts of the brain might erroneously appear active during an experiment. The degree of temporal autocorrelation is different across the brain (Worsley et al., 2002). In particular, autocorrelation in gray matter is stronger than in white matter and cerebrospinal fluid, but it also varies within gray 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 modeled voxel-wise. For each voxel, an autoregressive-moving-average ARMA(1,1) model is estimated. The ARMA(1,1) estimates are not spatially smoothed. For FSL, a Tukey taper is used to smooth the spectral density estimates voxel-wise. These smoothed estimates are then additionally smoothed within tissue type. Woolrich et al. (2001) showed the appropriateness of the FSL's method for two fMRI protocols: with repetition time (TR) of 1.5s and of 3s, and with voxel size 4x4x7 mm^3. By default, SPM estimates temporal autocorrelation globally as an autoregressive AR(1) plus white noise process (Purdon and Weisskoff, 1998). SPM has an alternative approach: `FAST`, but we know of only two studies which have used it (Todd et al., 2016; Bollmann et al., 2018). Bollmann et al. (2018) explains `FAST` uses a dictionary of covariance components based on exponential covariance functions.

In Lenoski et al. (2008) several fMRI autocorrelation modeling approaches were compared for one fMRI protocol (TR=3s, voxel size 3.75x3.75x4 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 in Purdon and Weisskoff (1998) for the case when autocorrelation was not accounted for.

In this study we investigated the whitening performance of AFNI, FSL and SPM for a wide variety of fMRI protocols. We analyzed both the default SPM's method and the alternative

Email address: `wo222@cam.ac.uk` (Wiktor Olszowy)

Table 1: Overview of the employed datasets. 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.

Study	**Experiment**	**Place**	**Design**	**No. subjects**	**Field [T]**	**TR [s]**	**Voxel size [mm]**	**No. voxels**	**Time points**
FCP	resting state	Beijing	N/A	198	3	2	3.1x3.1x3.6	64x64x33	225
	resting state	Cambridge, US	N/A	198	3	3	3x3x3	72x72x47	119
NKI	resting state	Orangeburg, US	N/A	30	3	1.4	2x2x2	112x112x64	404
	resting state	Orangeburg, US	N/A	30	3	0.645	3x3x3	74x74x40	900
CRIC	resting state	Cambridge, UK	N/A	73	3	2	3x3x3.8	64x64x32	300
neuRosim	resting state	(simulated)	N/A	100	NA	2	3.1x3.1x3.6	64x64x33	225
NKI	checkerboard	Orangeburg, US	20s off+20s on	30	3	1.4	2x2x2	112x112x64	98
	checkerboard	Orangeburg, US	20s off+20s on	30	3	0.645	3x3x3	74x74x40	240
BMMR	checkerboard	Magdeburg	12s off+12s on	21	7	3	1x1x1	182x140x45	80
CRIC	checkerboard	Cambridge, UK	16s off+16s on	70	3	2	3x3x3.8	64x64x32	160
CamCAN	sensorimotor	Cambridge, UK	event-related	200	3	1.97	3x3x4.44	64x64x32	261

one: `FAST`. Furthermore, we analyzed the resulting specificity-sensitivity trade-offs in first level fMRI results. The main part of the paper compares the pre-whitening approaches from AFNI, FSL and SPM for boxcar experimental designs. Supplementary material includes analysis of an event-related design dataset, as well as a group level comparison of SPM's default method with `FAST`. We 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.

Data

In order to explore a range of parameters that may affect autocorrelation, we investigated 11 fMRI datasets (Table 1). These included resting state and task studies, healthy subjects and a patient population, different TRs, magnetic field strengths and voxel sizes. We 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 anonymized 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 anonymized 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, which can be found in Supplementary material, we used the CamCAN dataset (Cambridge Centre for Ageing and Neuroscience, `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, and the scans have not been made public yet. The study of Cambridge Research into Impaired Consciousness (CRIC) was approved by the Cambridge Local Research Ethics Committee (99/391), and the scans have not been made public yet. 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). Simulation details can be found in Supplementary material.

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 (Fig. 1). This way we limited the influence of possible confounders on the results. In order to investigate whether our results are an artefact of the comparison approach used for assessment, we compared AFNI, FSL and SPM by investigating (1) the power spectra of the GLM residuals, (2) the spatial distribution of significant clusters, (3) the average percentage of significant voxels within the brain mask, and (4) 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, we considered voxels in native space using the same brain mask for AFNI, FSL and SPM. For each voxel, we normalized the time series to have variance 1 and calculated the power spectra as the square of the discrete Fourier transform.

Apart from assuming dummy designs for resting state data as in Eklund et al. (2012, 2015, 2016), we also assumed wrong (dummy) designs for task data, and we used resting state scans

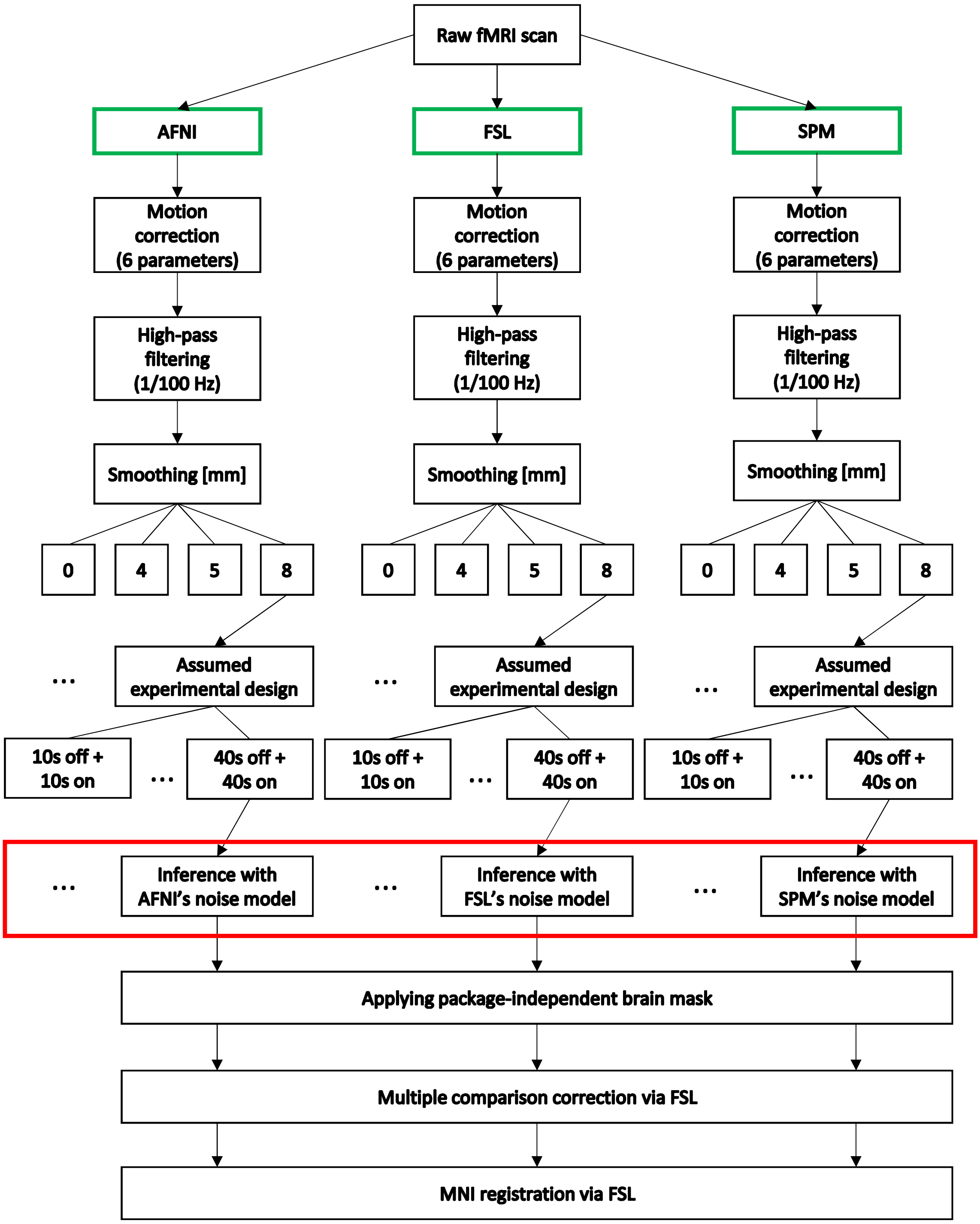
Raw fMRI scan
AFNI
FSL
SPM
Motion correction (6 parameters)
Motion correction (6 parameters)
Motion correction (6 parameters)
High-pass filtering (1/100 Hz)
High-pass filtering (1/100 Hz)
High-pass filtering (1/100 Hz)
Smoothing [mm]
Smoothing [mm]
Smoothing [mm]
0
4
5
8
0
4
5
8
0
4
5
8
...
Assumed experimental design
...
Assumed experimental design
...
Assumed experimental design
10s off + 10s on
...
40s off + 40s on
10s off + 10s on
...
40s off + 40s on
10s off + 10s on
...
40s off + 40s on
...
Inference with AFNI's noise model
...
Inference with FSL's noise model
...
Inference with SPM's noise model
Applying package-independent brain mask
Multiple comparison correction via FSL
MNI registration via FSL

Figure 1: The employed analyses pipelines. The noise models used by AFNI, FSL and SPM were the only relevant difference (marked in a red box).

simulated using the neuRosim package in R (Welvaert et al., 2011). We treated such data as null data. For null data, the positive rate is the familywise error rate, which was employed in Eklund et al. (2012, 2015, 2016). We use the term "significant voxel" to denote a voxel that is covered by one of the clusters returned by the multiple comparison correction.

All the processing scripts needed to fully replicate our study are at https://github.com/wiktorolszowy/fMRI_temporal_autocorrelation. We used AFNI 16.2.02, FSL 5.0.10 and SPM 12 (v7219).

Preprocessing

Slice timing correction was not performed, as for some datasets the slice timing information was not available. In each of the three packages we performed motion correction, which resulted in 6 parameters that we considered as confounders in the consecutive statistical analysis. Furthermore, in each of the three packages we conducted high-pass filtering with frequency cut-off of 1/100 Hz. We 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.

Statistical analysis

For analyses in each package, we 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, we used in the three packages the first derivative of the double gamma model, also known as the temporal derivative. We did not incorporate physiological recordings to the analysis pipeline, as these were not available for most of the datasets used.

We 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. We 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. We masked the statistic maps from AFNI, FSL and SPM using the intersected brain masks from FSL and SPM. We did not confine the analyses to a gray matter mask, because autocorrelation is at strongest in gray matter (Worsley et al., 2002). In other words, false positives caused by imperfect pre-whitening can be expected to occur mainly in gray 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, we extracted the degrees of freedom from each analysis output.

Next, we performed multiple comparison correction in FSL for all the analyses, including for those started in AFNI and SPM. First, we estimated the smoothness of the brain-masked 4-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`), we 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. Finally, we applied previously saved MNI transformations to the binary maps which were showing the location of the significant clusters.

Results

Whitening performance of AFNI, FSL and SPM

To investigate the whitening performance resulting from the use of noise models in AFNI, FSL and SPM, we plotted the power spectra of the GLM residuals. Fig. 2 shows the power spectra averaged across all brain voxels and subjects for smoothing of 8 mm and assumed boxcar design of 10s of rest followed by 10s of stimulus presentation. 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 methods. Fig. 2 does not show respiratory spikes which one could expect to see. This is because the figure refers to averages across subjects. We observed respiratory spikes when analyzing power spectra for single subjects (not shown). Importantly, for the "BMMR checkerboard" dataset analyzed both with the default SPM's method and with `FAST`, there was a small peak at frequency 1/24 Hz, which was the true design frequency. For AFNI and FSL, this peak was higher. As the assumed design was a wrong design, a low power spectrum at the true design frequency suggests too strong pre-whitening, during which negative autocorrelations can be introduced.

Resulting specificity-sensitivity trade-offs

In order to investigate the impact of the whitening performance on first level results, we analyzed the spatial distribution of significant clusters in AFNI, FSL and SPM. Fig. 3 shows an exemplary axial slice in the MNI space for 8 mm smoothing. It was made through the imposition of subjects' binarized 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. 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

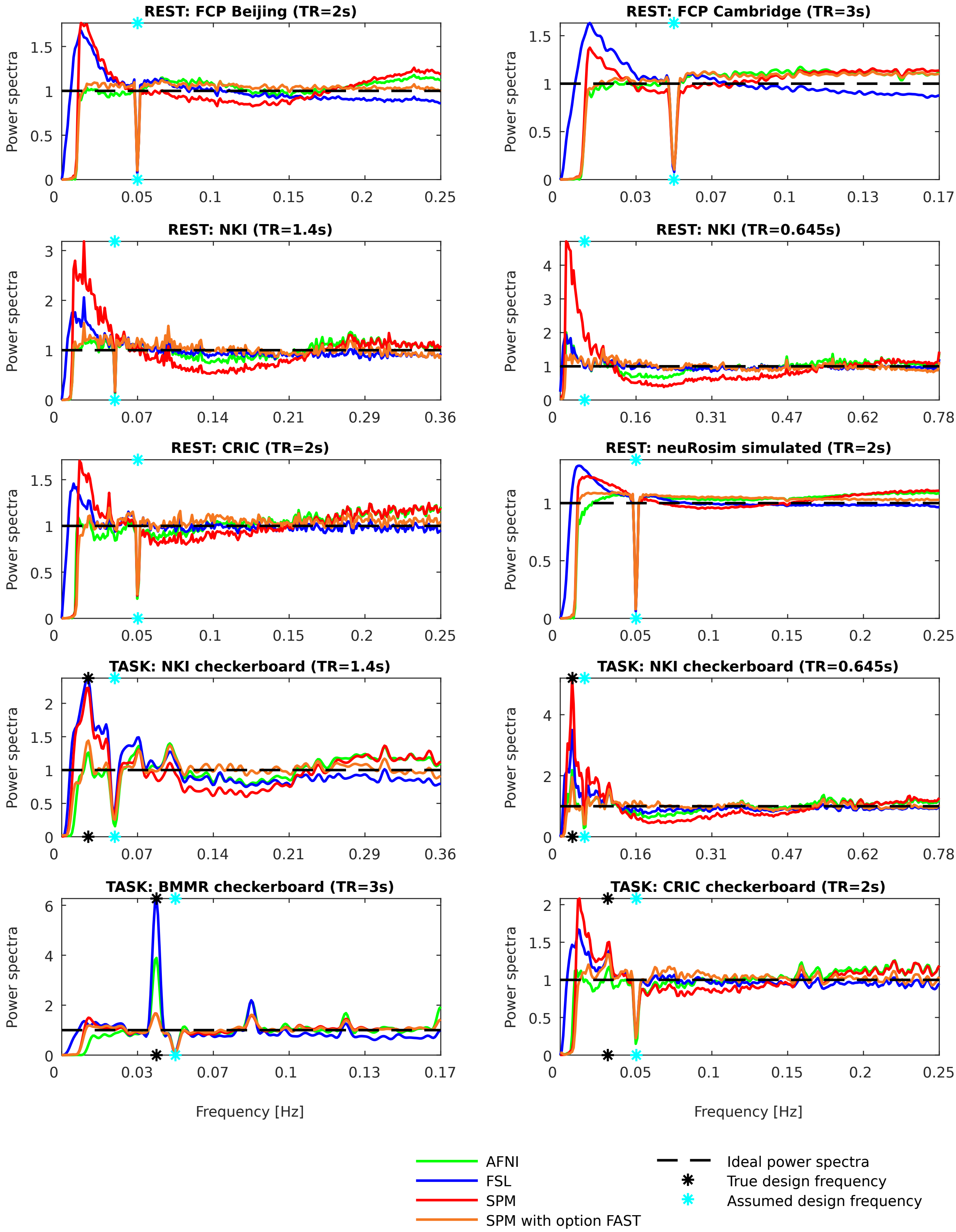

Figure 2: Power spectra of the GLM residuals in native space averaged across brain voxels and across subjects for the assumed boxcar design of 10s of rest followed by 10s of stimulus presentation ("boxcar10"). The dips at 0.05 Hz are due to the assumed design period being 20s (10s + 10s). For some datasets, 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. If after pre-whitening the residuals were white (as it is assumed), the power spectra would be flat. AFNI and SPM's alternative method: FAST, led to best whitening performance (most flat spectra). For FSL and SPM, there was substantial autocorrelated noise left after pre-whitening, particularly at low frequencies.

were scattered primarily within gray 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 gray matter across the brain. False positives in gray 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.

Additional comparison approaches

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, we additionally investigated two other comparison approaches: the percentage of significant voxels and the positive rate. Supplementary material, Fig. S1 shows the average percentage of significant voxels across subjects in 10 datasets for smoothing of 8 mm and for 16 assumed boxcar experimental designs. 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 Fig. 3) 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: 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", FSL and SPM returned 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. For the "BMMR checkerboard" dataset tested with the true design, SPM and FAST resulted in a much lower percentage of significant voxels than AFNI and FSL. The average percentage of significant voxels across subjects for this dataset tested with the true design was 31.5% for AFNI, 36.7% for FSL, 6.7% for SPM and 7.6% for FAST, respectively. This agrees with Fig. 3. For this dataset, the brain mask was limited mainly to the occipital lobe and the percentage relates to the field of view that was used. For this dataset, the power spectrum at the true design was much lower for SPM and FAST than for AFNI and FSL (Fig. 2). This suggests negative autocorrelations were introduced during pre-whitening for the assumed true design. This led to a decrease in perceived activation.

Overall, at an 8 mm smoothing level, AFNI and FAST outperformed FSL and SPM showing a lower average percentage of significant voxels in tests with the wrong design: on average across 10 datasets and across the wrong designs, the average percentage of significant voxels was 0.4% for AFNI, 1% for FSL, 1.9% for SPM and 0.3% for FAST.

As multiple comparison correction depends on the smoothness level of the residual maps, we 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 Gaussian distribution in x-, y-, and z-dimensions across the 10 datasets and across the 16 assumed designs was 10.8 mm for AFNI, 10.3 mm for FSL, 11.7 mm for SPM and 11.5 mm for FAST. Nonetheless, we also 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%. The average percentage across the 10 datasets and across the wrong designs was 0.5% for AFNI, 1.2% for FSL, 2.1% for SPM and 0.4% for FAST.

Supplementary material, Figs. S2-S3 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. 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 the subject's brain. 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.

Since smoothing implicitly affects the voxel size, we considered different smoothing kernel sizes. We 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). With a wider smoothing kernel, the percentage of significant voxels increased (not shown), while the positive rate decreased. Differences between AFNI, FSL, SPM and FAST discussed above for the four comparison approaches and smoothing of 8 mm were consistent across the four smoothing levels.

Discussion

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 significant voxels. Purdon and Weisskoff (1998) showed that this relationship exists when positive autocorrelation is not removed from the data. This phenomenon 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 we

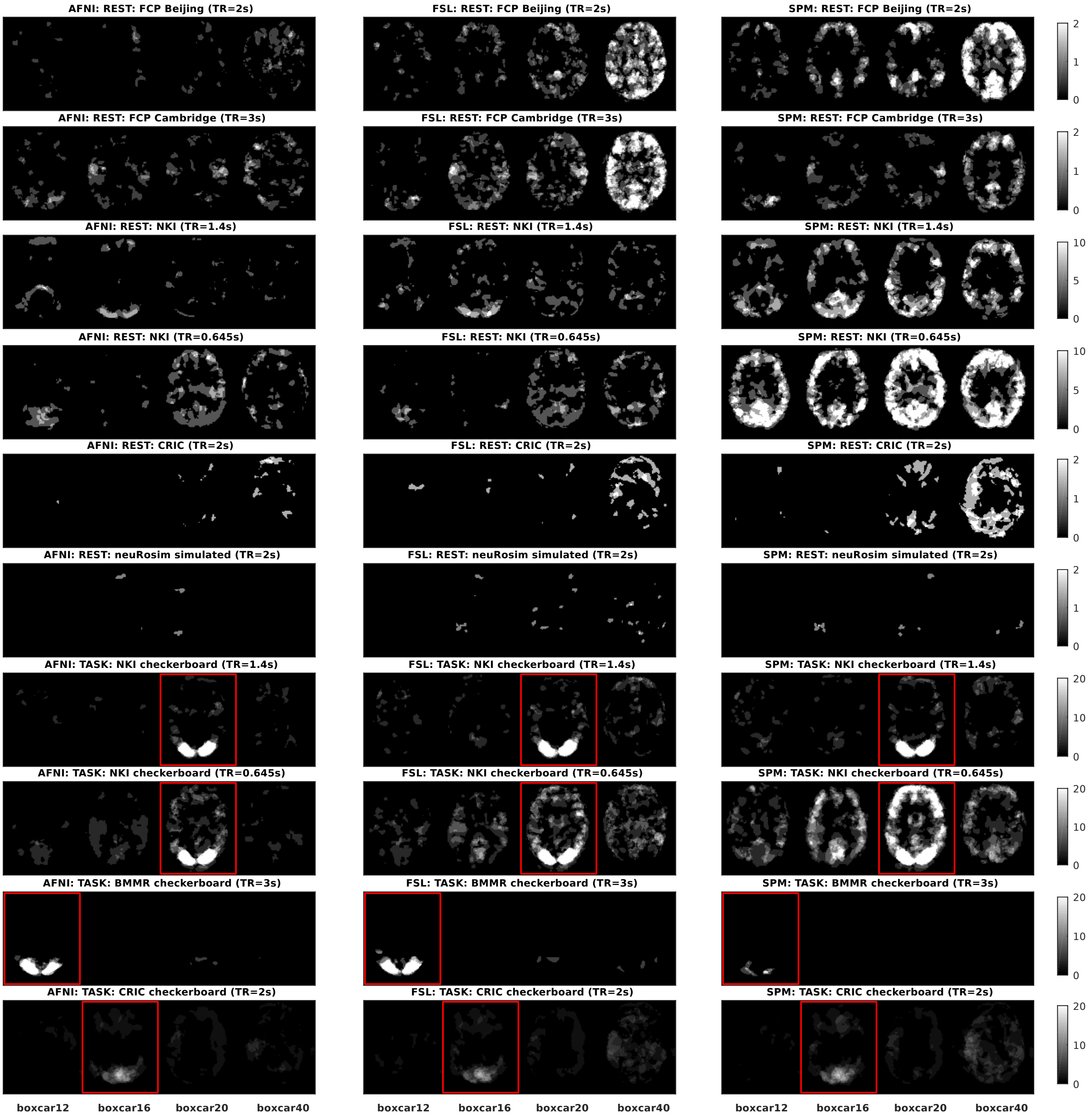
AFNI: REST: FCP Beijing (TR=2s)
FSL: REST: FCP Beijing (TR=2s)
SPM: REST: FCP Beijing (TR=2s)
AFNI: REST: FCP Cambridge (TR=3s)
FSL: REST: FCP Cambridge (TR=3s)
SPM: REST: FCP Cambridge (TR=3s)
AFNI: REST: NKI (TR=1.4s)
FSL: REST: NKI (TR=1.4s)
SPM: REST: NKI (TR=1.4s)
AFNI: REST: NKI (TR=0.645s)
FSL: REST: NKI (TR=0.645s)
SPM: REST: NKI (TR=0.645s)
AFNI: REST: CRIC (TR=2s)
FSL: REST: CRIC (TR=2s)
SPM: REST: CRIC (TR=2s)
AFNI: REST: neuRosim simulated (TR=2s)
FSL: REST: neuRosim simulated (TR=2s)
SPM: REST: neuRosim simulated (TR=2s)
AFNI: TASK: NKI checkerboard (TR=1.4s)
FSL: TASK: NKI checkerboard (TR=1.4s)
SPM: TASK: NKI checkerboard (TR=1.4s)
AFNI: TASK: NKI checkerboard (TR=0.645s)
FSL: TASK: NKI checkerboard (TR=0.645s)
SPM: TASK: NKI checkerboard (TR=0.645s)
AFNI: TASK: BMMR checkerboard (TR=3s)
FSL: TASK: BMMR checkerboard (TR=3s)
SPM: TASK: BMMR checkerboard (TR=3s)
AFNI: TASK: CRIC checkerboard (TR=2s)
FSL: TASK: CRIC checkerboard (TR=2s)
SPM: TASK: CRIC checkerboard (TR=2s)
2 1 0
2 1 0
10 5 0
10 5 0
2 1 0
2 1 0
20 10 0
20 10 0
20 10 0
20 10 0
boxcar12 boxcar16 boxcar20 boxcar40
boxcar12 boxcar16 boxcar20 boxcar40
boxcar12 boxcar16 boxcar20 boxcar40

Figure 3: Spatial distribution of significant clusters in AFNI (left), FSL (middle) and SPM (right) for different assumed experimental designs. 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). Resting state data was used as null data. Thus, low numbers of significant voxels were a desirable outcome, as 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. The clearest cut between the true and the wrong/dummy designs was obtained with AFNI's noise model. FAST performed similarly to AFNI's noise model (not shown).

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.

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 modeling in AFNI and in `FAST`. In our 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 in Eklund et al. (2015). In our 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 modeling of many subjects.

The highly significant responses for the NKI datasets are in line with 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 we considered had TR of 0.645s and 1.4s, 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 study of Eklund et al. (2012) only referred to SPM. In addition to the previous study, we 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=2s ("FCP Beijing" and "CRIC RS"). FSL models autocorrelation more flexibly than SPM, which seems to be confirmed by our study. For short TRs, AFNI's performance deteriorated too, as autocorrelation spans more than one TR (Bollmann et al., 2018) and an ARMA(1,1) noise model can only partially capture it.

Apart from the different TRs, we analyzed the impact of spatial smoothing. If more smoothing is applied, the signal from gray matter will be often mixed with the signal from white matter. As autocorrelation in white matter is lower than in gray matter (Worsley et al., 2002), autocorrelation in a primarily gray 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 shown in Eklund et al. (2012, 2015). More about the impact of smoothing and spatial resolution can be found in Geissler et al. (2005); Weibull et al. (2008); Mueller et al. (2017). We considered smoothing only as a confounder. Importantly, for all four levels of smoothing, AFNI and `FAST` outperformed FSL and SPM.

Compared to FSL, the use of SPM resulted in a lower percentage of significant voxels for the "FCP Cambridge" and "BMMR checkerboard" datasets. These were the only datasets with TR of more than 2 seconds. For the "FCP Cambridge" dataset, a lower percentage of significant voxels was a desirable result, as the dataset was used as null data. However, compared to AFNI and FSL, SPM was less sensitive in detecting activation in the primary visual cortex for the "BMMR checkerboard" dataset. Because the autocorrelation modeling approach in SPM has little flexibility, in case of long TR, where the correlations between adjacent time points become smaller, SPM might introduce negative autocorrelations during pre-whitening. For boxcar designs, this lowers the statistics and increases false negative rates (Lenoski et al., 2008). Surprisingly, compared to AFNI and FSL, for the "BMMR checkerboard" dataset tested with the true design, the use of `FAST` also led to a lower percentage of significant voxels.

Our results confirm Lenoski et al. (2008) insofar as our study also showed best performance of a method that did not involve spatial smoothing of the autocorrelation parameters. Interestingly, in Eklund et al. (2015) AFNI, FSL and SPM were already compared 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. We also observed lowest false positive rates for AFNI. Opposed to Eklund et al. (2015), which compared the packages in their entirety, we 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.

We did not perform slice timing correction, but to account for different slice acquisition times we employed the temporal derivative. The differences in first level results between AFNI, FSL and SPM which we observed could have been smaller if physiological recordings had been modeled. The modeling of physiological noise is known to improve whitening performance, particularly for short TRs (Lund et al., 2006; Bollmann 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.

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

FSL provided 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 1.5s, while for the other protocol TR was 3s. For both protocols, the voxel size was 4x4x7 mm^3. These were large voxels. We suspect that the FSL's pre-whitening approach could have been overfitted to this data.

Friston et al. (2000) and Lenoski et al. (2008) showed that pre-whitening with a global noise model can result in profound bias. SPM's default is a 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).

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 particularly relevant for event-related designs (see Supplementary material). Bias from confounded coefficient maps can be expected to propagate to the group level. In Supplementary material we 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., 2004). These approaches additionally employ standard error maps, which are also 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). We postulate that more accurate autocorrelation modeling at the subject level can substantially improve fMRI reliability both at the subject level and at the group level.

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

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 appear mainly in the posterior cingulate cortex, even though the assumed design for that analysis was a randomized event-related design. The rest activity in these regions can occur at different frequencies and can underlie different patterns (Stark and Squire, 2001). Thus, resting state data is 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. Task fMRI data is 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-dependent low-frequency drifts and head motion. Some of the current simulation toolboxes (Welvaert and Rosseel, 2014) enable the modeling 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. In our study, results for simulated resting state data 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. Because there is no perfect fMRI null data, we used both resting state data with assumed dummy designs and task data with assumed wrong designs. Results for both approaches coincided.

Conclusions

Using data corresponding to a wide variety of fMRI protocols, we 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, all considered fMRI protocols were affected. We 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. The differences between AFNI, FSL and SPM were large and consistent across four different comparison approaches and across 11 datasets. The resulting false positives and false negatives can be expected to propagate to the group level, especially if the group analysis is performed with a mixed 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, our findings support more widespread use of the `FAST` method. 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. We 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. We provide a MATLAB script (GitHub: `plot_power_spectra_of_GLM_residuals.m`) for the fMRI researchers to check if their analyses might be affected by imperfect pre-whitening.

Acknowledgments

We would like to thank Paul Browne, Anders Eklund, Thomas Nichols, Karl Friston, Richard Reynolds, Paola Finoia, Adrian Carpenter, Alison Sleigh, Gang Chen, Guillaume Flandin, and Michał Kosicki for much valuable advice. Furthermore, we would like to thank the James S. McDonnell Foundation for funding the image acquisitions of the Cambridge Research into Impaired Consciousness (CRIC) group, and the CRIC group for sharing their data. Oliver Speck, Michael Hoffmann and Aini Ismafairus from the Otto von Guericke University provided us with the 7T data. We also thank the Neuroimaging Informatics Tools and Resources Clearinghouse and all of the researchers who have contributed with data to the 1,000 Functional Connectomes Project and to the Enhanced Nathan Kline Institute - Rockland Sample. W.O. was in receipt of scholarships from the Cambridge Trust and from the Mateusz B. Grabowski Fund.

References

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

Biswal, B. B., Mennes, M., Zuo, X.-N., Gohel, S., Kelly, C., Smith, S. M., Beckmann, C. F., Adelstein, J. S., Buckner, R. L., Colcombe, S., et al., 2010. Toward discovery science of human brain function. Proceedings of the National Academy of Sciences 107 (10), 4734–4739.

Bollmann, S., Puckett, A. M., Cunnington, R., Barth, M., 2018. Serial correlations in single-subject fMRI with sub-second TR. NeuroImage 166, 152 – 166.

Bullmore, E., Brammer, M., Williams, S. C., Rabe-Hesketh, S., Janot, N., David, A., Mellers, J., Howard, R., Sham, P., 1996. Statistical methods of estimation and inference for functional MR image analysis. Magnetic Resonance in Medicine 35 (2), 261–277.

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

Chen, G., Saad, Z. S., Nath, A. R., Beauchamp, M. S., Cox, R. W., 2012. FMRI group analysis combining effect estimates and their variances. NeuroImage 60 (1), 747–765.

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

Eklund, A., Andersson, M., Josephson, C., Johannesson, M., Knutsson, H., 2012. Does parametric fMRI analysis with SPM yield valid results? – An empirical study of 1484 rest datasets. NeuroImage 61 (3), 565–578.

Eklund, A., Nichols, T., Andersson, M., Knutsson, H., 2015. 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. IEEE, pp. 1376–1380.

Eklund, A., Nichols, T. E., Knutsson, H., 2016. Cluster failure: Why fMRI inferences for spatial extent have inflated false-positive rates. Proceedings of the National Academy of Sciences, 201602413.

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

Friston, K. J., Glaser, D. E., Henson, R. N., Kiebel, S., Phillips, C., Ashburner, J., 2002. Classical and Bayesian inference in neuroimaging: applications. NeuroImage 16 (2), 484–512.

Friston, K. J., Stephan, K. E., Lund, T. E., Morcom, A., Kiebel, S., 2005. Mixed-effects and fMRI studies. NeuroImage 24 (1), 244–252.

Geissler, A., Lanzenberger, R., Barth, M., Tahamtan, A. R., Milakara, D., Gartus, A., Beisteiner, R., 2005. Influence of fMRI smoothing procedures on replicability of fine scale motor localization. NeuroImage 24 (2), 323–331.

Hamid, A. I. A., Speck, O., Hoffmann, M. B., 2015. Quantitative assessment of visual cortex function with fMRI at 7 Tesla–test-retest variability. Frontiers in Human Neuroscience 9.

Jenkinson, M., Beckmann, C. F., Behrens, T. E., Woolrich, M. W., Smith, S. M., 2012. FSL. NeuroImage 62 (2), 782–790.

Larkman, D. J., Hajnal, J. V., Herlihy, A. H., Coutts, G. A., Young, I. R., Ehnholm, G., 2001. Use of multicoil arrays for separation of signal from multiple slices simultaneously excited. Journal of Magnetic Resonance Imaging 13 (2), 313–317.

Lenoski, B., Baxter, L. C., Karam, L. J., Maisog, J., Debbins, J., 2008. On the performance of autocorrelation estimation algorithms for fMRI analysis. IEEE Journal of Selected Topics in Signal Processing 2 (6), 828–838.

Lund, T. E., Madsen, K. H., Sidaros, K., Luo, W.-L., Nichols, T. E., 2006. Non-white noise in fMRI: does modelling have an impact? NeuroImage 29 (1), 54–66.

Luo, W.-L., Nichols, T. E., 2003. Diagnosis and exploration of massively univariate neuroimaging models. NeuroImage 19 (3), 1014–1032.

Mueller, K., Lepsien, J., Möller, H. E., Lohmann, G., 2017. Commentary: Cluster failure: Why fMRI inferences for spatial extent have inflated false-positive rates. Frontiers in Human Neuroscience 11, 345.

Nooner, K. B., Colcombe, S. J., Tobe, R. H., Mennes, M., Benedict, M. M., Moreno, A. L., Panek, L. J., Brown, S., Zavitz, S. T., Li, Q., et al., 2012. The NKI-Rockland sample: a model for accelerating the pace of discovery science in psychiatry. Frontiers in Neuroscience 6.

Penny, W. D., Friston, K. J., Ashburner, J. T., Kiebel, S. J., Nichols, T. E., 2011. Statistical parametric mapping: the analysis of functional brain images. Academic press.

Polimeni, J. R., Renvall, V., Zaretskaya, N., Fischl, B., 2017. Analysis strategies for high-resolution UHF-fMRI data. NeuroImage.

Purdon, P. L., Weisskoff, R. M., 1998. 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.

Raichle, M. E., MacLeod, A. M., Snyder, A. Z., Powers, W. J., Gusnard, D. A., Shulman, G. L., 2001. A default mode of brain function. Proceedings of the National Academy of Sciences 98 (2), 676–682.

Shafto, M. A., Tyler, L. K., Dixon, M., Taylor, J. R., Rowe, J. B., Cusack, R., Calder, A. J., Marslen-Wilson, W. D., Duncan, J., Dalgleish, T., et al., 2014. 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.

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

Stark, C. E., Squire, L. R., 2001. When zero is not zero: the problem of ambiguous baseline conditions in fMRI. Proceedings of the National Academy of Sciences 98 (22), 12760–12766.

Todd, N., Moeller, S., Auerbach, E. J., Yacoub, E., Flandin, G., Weiskopf, N., 2016. 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.

Walter, M., Stadler, J., Tempelmann, C., Speck, O., Northoff, G., 2008. 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.

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

Welvaert, M., Durnez, J., Moerkerke, B., Verdoolaege, G., Rosseel, Y., 2011. neuRosim: An R package for generating fMRI data. Journal of Statistical Software 44 (10), 1–18.

Welvaert, M., Rosseel, Y., 2014. A review of fMRI simulation studies. PLOS ONE 9 (7), e101953.

Woolrich, M. W., Behrens, T. E., Beckmann, C. F., Jenkinson, M., Smith, S. M., 2004. Multilevel linear modelling for FMRI group analysis using Bayesian inference. NeuroImage 21 (4), 1732–1747.

Woolrich, M. W., Ripley, B. D., Brady, M., Smith, S. M., 2001. Temporal autocorrelation in univariate linear modeling of FMRI data. NeuroImage 14 (6), 1370–1386.

Worsley, K. J., Liao, C., Aston, J., Petre, V., Duncan, G., Morales, F., Evans, A., 2002. A general statistical analysis for fMRI data. NeuroImage 15 (1), 1–15.

Supplementary material

Simulation

We used the `neuRosim` package 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. We 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, we 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, i.e. with the same parameters, is used for each voxel. We 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. We 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 1. The number of time points was also chosen as in "FCP Beijing". For the real "FCP Beijing" scan, we arbitrarily chose a cuboidal region of interest, where we 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 we found a parameter for the `neuRosim`'s AR(1) model so that the resulting average AR(1) parameter in the simulated scans in the same cuboidal region of interest was very similar.

Impact on event-related design studies

In order to check if differences in autocorrelation modeling in AFNI, FSL and SPM lead to different first level results for event-related design studies, we analyzed the CamCAN dataset. The task was a sensorimotor one with visual and audio stimuli. The design included the stimulus m-sequence described in Buračas and Boynton (2002). Supplementary material, Fig. S4 shows (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"). 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 3s and 6s. The stimulus duration times were 0.1s.

For the assumed low-frequency design ("B2"), AFNI's autocorrelation modeling 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, we observed a lower percentage of significant voxels compared to SPM tested with 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 we do not know the ground truth, we 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 levels for the truly but little active voxels.

Impact on group studies with a random effects model

To investigate the impact of pre-whitening on the group level, we performed in SPM random effects analyses for a one-sample t-test. We considered only the 8 mm smoothing level and results corresponding to SPM and `FAST`. As there were 10 datasets and 16 assumed designs, for each pre-whitening we ran 160 group analyses. Four of these group analyses were for task data with assumed true design. The rest were analyses on null data. For null data, we found significant clusters in 14 analyses for SPM and in 16 analyses for `FAST`. This corresponded to a familywise error rate of 9% for SPM and 10.3% for `FAST`. For task datasets tested with the true design, the use of `FAST` resulted in a lower percentage of significant voxels than the use of the default method. For the NKI dataset at TR=1.4s, 6.5% of the brain was significant for SPM and 6.2% was significant for `FAST`. For the NKI dataset at TR=0.645s, SPM and `FAST` led to 7.1% and 6.4%, respectively. For the BMMR dataset, 10.8% and 10.7% of the brain was significant following the use of the default noise model of SPM and the use of `FAST`. For the "CRIC checkerboard" dataset, no significant clusters were found at the group level, as several of the subjects had deformed brains and the resulting group brain mask in MNI space did not cover the primary visual cortex.

Furthermore, we performed group analyses for the event-related design dataset: "CamCAN sensorimotor". For the assumed true design, the use of `FAST` led to a higher percentage of significant voxels: 45.6% compared to 42.9% for the SPM's default method. We observed the same relationship at the single subject level (Supplementary material, Fig. S4). While a high percentage of significant voxels might be surprising, the experiment included both visual and audio stimuli, and the dataset consisted of 200 subjects. A large number of subjects makes it easier to find significant activation if the effect size is negligible.

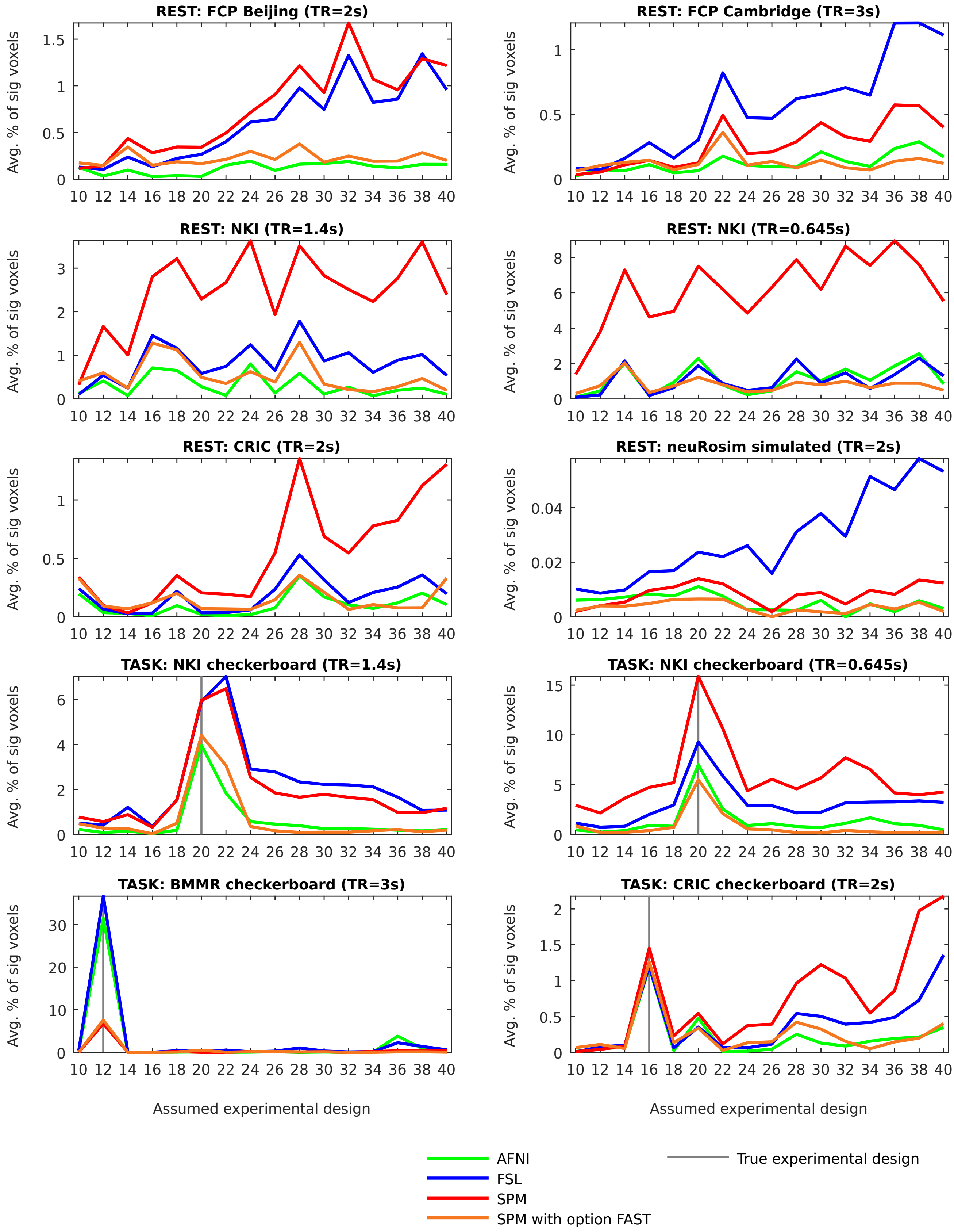

Figure S1: Average percentage of significant voxels across subjects for different packages. x-axis shows the assumed designs, e.g. “10” refers to the boxcar design of 10s of rest followed by 10s of stimulus presentation. Scans were spatially smoothed with FWHM of **8 mm**. Resting state data was used as null data. Thus, a low percentage of significant voxels was a desirable outcome, as 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.

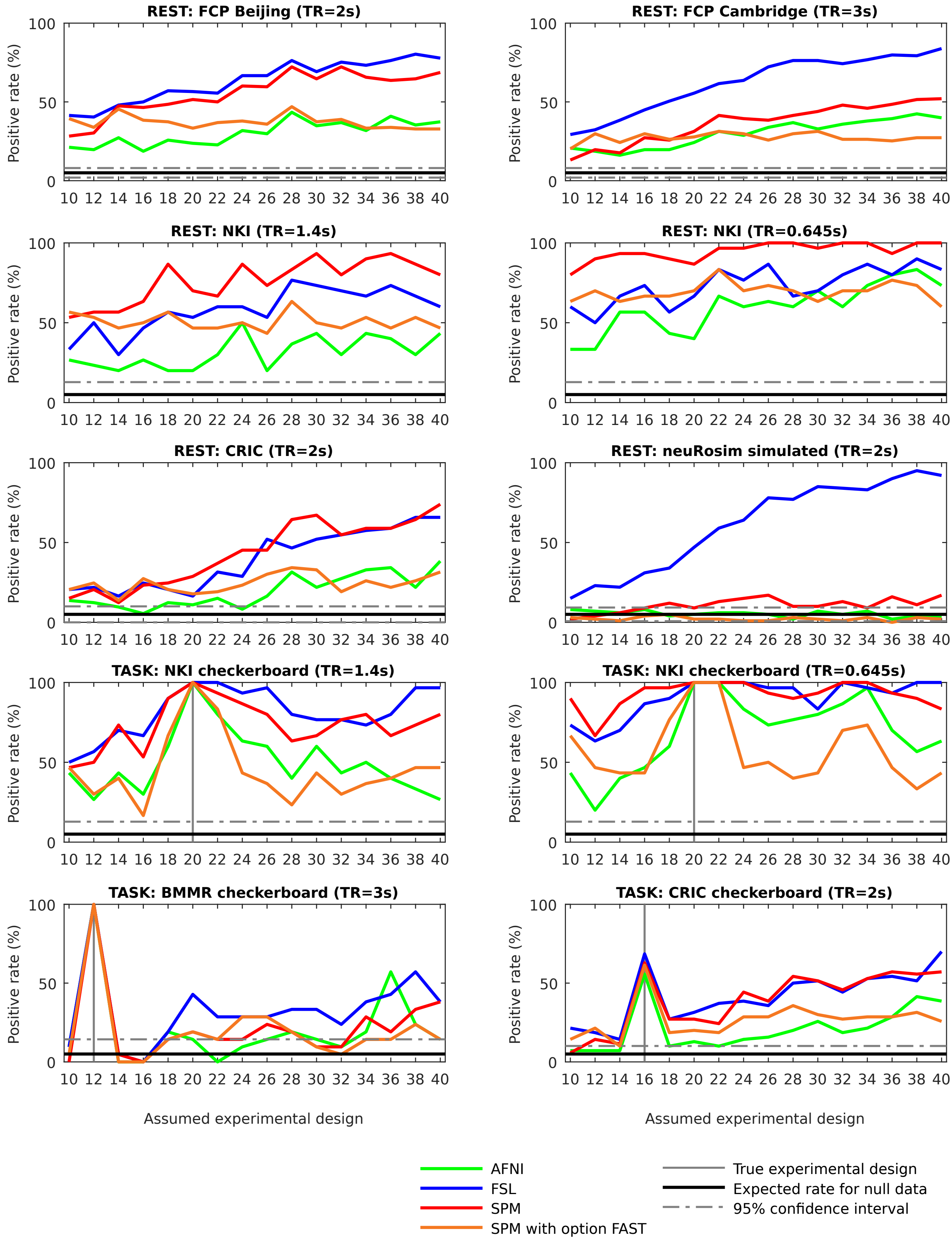

Figure S2: Positive rate for different packages. x-axis shows the assumed designs, e.g. "10" refers to the boxcar design of 10s of rest followed by 10s of stimulus presentation. Scans were spatially smoothed with FWHM of **4 mm**. For null data, the positive rate is the familywise error rate. AFNI and FAST had the highest specificity.

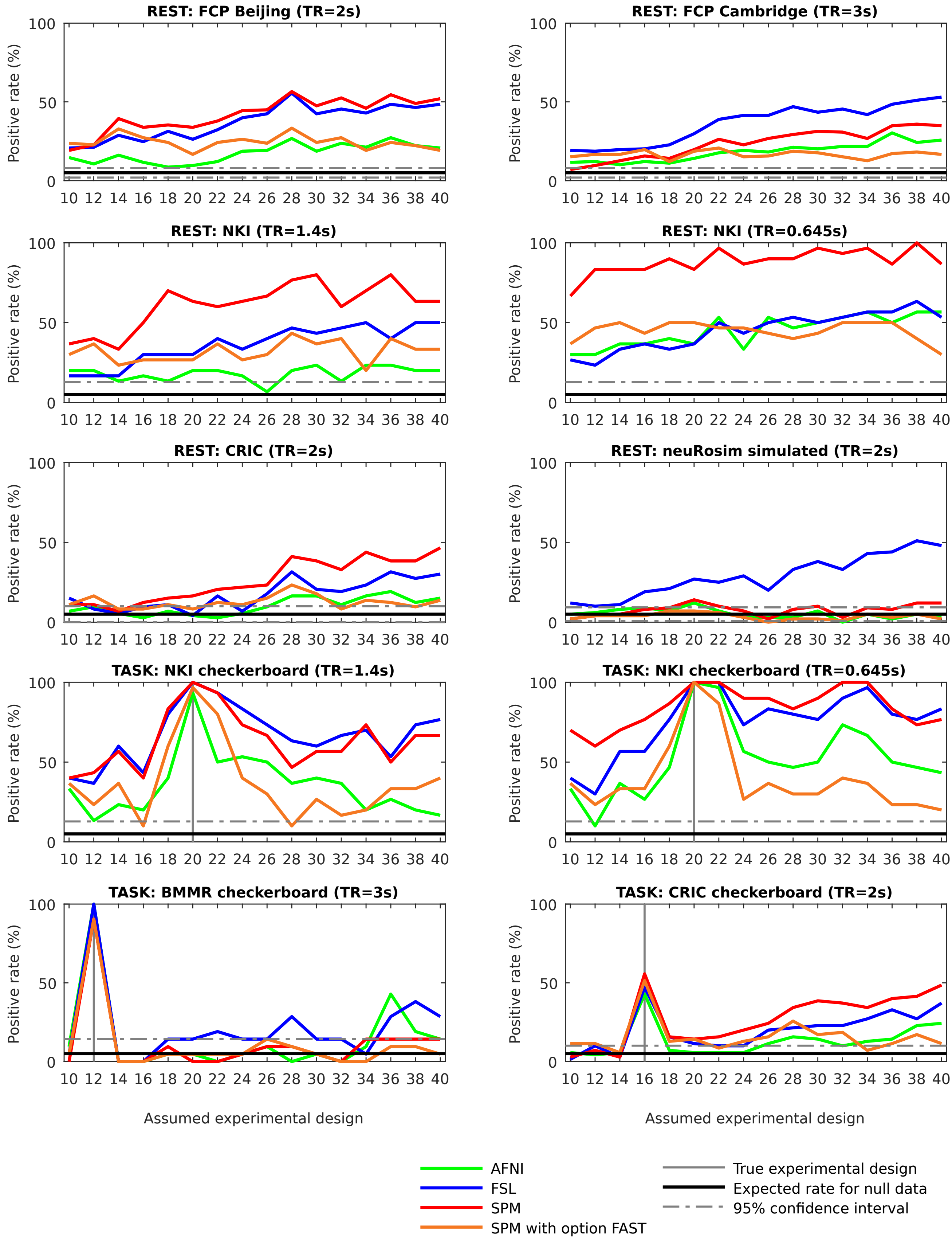

Figure S3: Positive rate for different packages. x-axis shows the assumed designs, e.g. "10" refers to the boxcar design of 10s of rest followed by 10s of stimulus presentation. Scans were spatially smoothed with FWHM of **8 mm**. For null data, the positive rate is the familywise error rate. AFNI and FAST had the highest specificity.

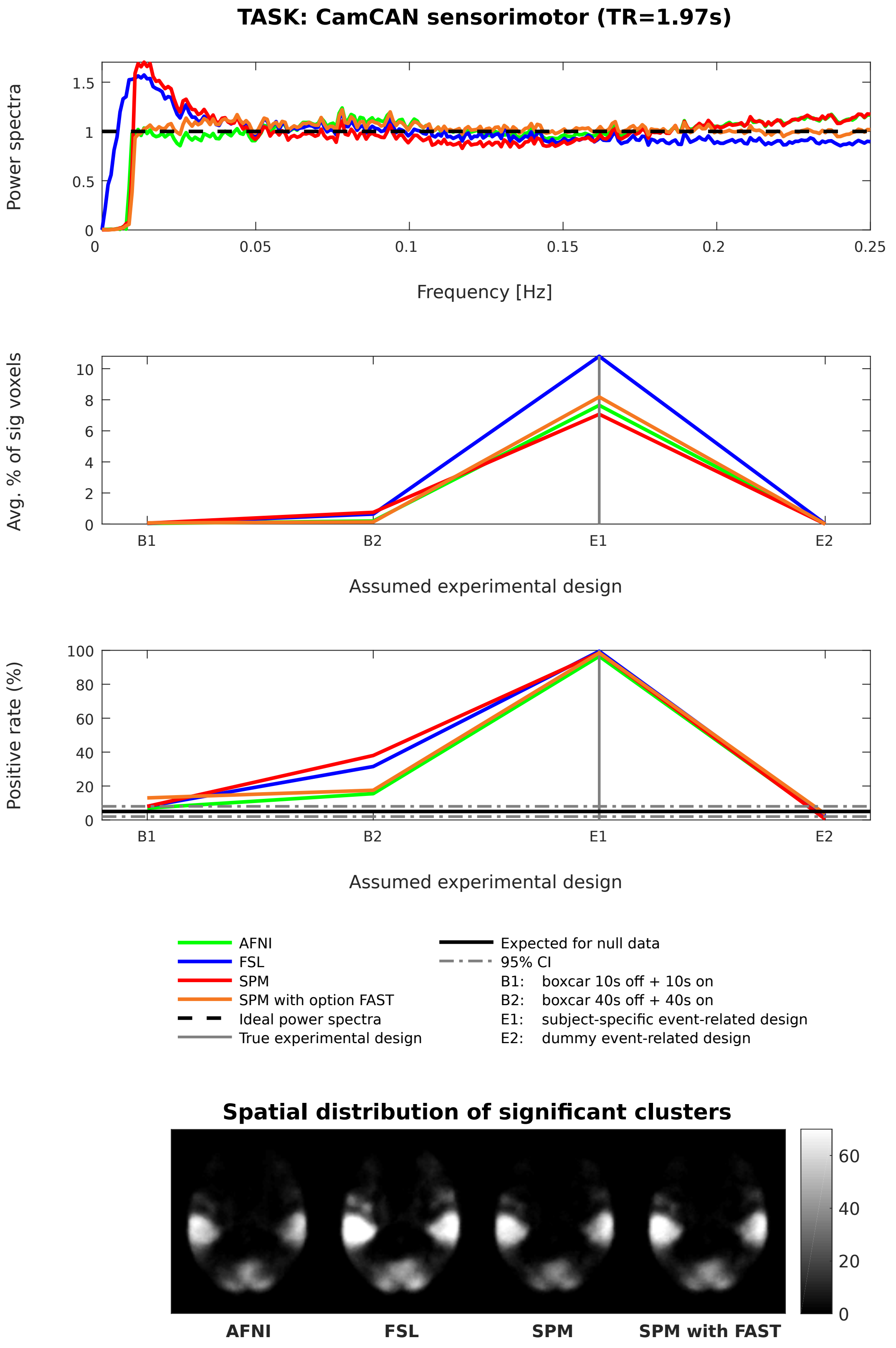

Figure S4: Differences between AFNI, FSL and SPM for a task dataset where the design was an event-related design ("CamCAN sensorimotor"). From top to bottom: (1) power spectra of the GLM residuals in native space averaged across brain voxels and across subjects for the assumed true design ("E1"), (2) 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. Scans were spatially smoothed with FWHM of **8 mm**.