A multidimensional artefact-reduction approach to increase robustness of first-level fMRI analyses: censoring vs. interpolating

Marko **Wilke** [1,2] & Torsten **Baldeweg** [3,4]

[1] Department of Pediatric Neurology and Developmental Medicine, Children's Hospital, and [2] Experimental Pediatric Neuroimaging, Children's Hospital and Department of Neuroradiology, University Hospital Tübingen, Germany

[3] Developmental Neurosciences Programme, UCL Great Ormond Street Institute of Child Health, and [4] Great Ormond Street Hospital NHS Trust, London, United Kingdom.

Running head: Optimizing artefact reduction

Author preprint of the manuscript accepted for publication in the Journal of Neuroscience Methods, the final version is available at https://doi.org/10.1016/j.jneumeth.2019.02.008

Address correspondence to: Marko Wilke, MD, PhD
University Children's Hospital
Dept. III (Pediatric Neurology)
Hoppe-Seyler-Str. 1
D - 72076 Tübingen, Germany
Tel. +49 7071 29-83416
Fax +49 7071 29-5473
Email: marko.wilke@med.uni-tuebingen.de

Abstract

Background: This manuscript describes a new, multidimensional and data-driven approach to identify outlying datapoints from a first-level fMRI dataset.

New method: Using three different indicators of data corruption (the fast variance component of DVARS [Δ%D-var], scan-to-scan total displacement [STS], and each scan's overall explained variance [R^2]), it identifies outlying datapoints while being balanced using Akaike'c corrected criterion (AIC_C) to avoid overcorrection. We then explore the impact of censoring, interpolating, or both, to remove a bad scan's contribution to the final timeseries.

Results and comparison with existing methods: Our results (using three real-life datasets and extensive simulations) show that motion-corrupted datapoints as well as non-motion related image artefacts are detected reliably. Using several indicators is shown to be an advantage over existing single-indicator solutions in different settings. As a result of using our algorithm, stronger activation (as detected by both T-value and number of activated voxels) and an increase in the temporal signal-to-noise ratio can be seen. The effects of censoring and interpolation are distinct and complex.

Conclusions: The multidimensional approach described here is able to identify outlying datapoints in fMRI timeseries, with demonstrable positive effects on several outcome measures. While censoring datapoints may be preferable in many settings, the ultimate choice on which approach to choose may depend on the data at hand. Recommendations are provided for different scenarios.

Highlights

- a new, multidimensional and data-driven approach to identify outlying datapoints in fMRI timeseries is described
- outlier removal is driven by three parameters (Δ%D-var, STS, and R^2) while being balanced by Akaike's corrected information criterion
- the effect of censoring datapoints in the design matrix vs. interpolating them on the raw data level is assessed and compared in different datasets
- stronger activation and a higher signal to noise ratio is seen as an effect of both censoring and interpolation
- some recommendations are provided but the optimal choice of approach may depend on the data at hand

Key words

Functional MRI; outlier detection; artefact reduction; data censoring; data interpolation

Competing interests

Both authors have no competing interests to disclose.

Introduction

Functional MRI has firmly established itself as a prime neuroscience research tool over the last decades. However, it also continues to be a challenging technique which is plagued by a low signal-to-noise ratio (SNR), and consequently is vulnerable to the influence of artefacts (Afyouni & Nichols, 2018; Caballero-Gaudes & Reynolds, 2016; Chen & Glover, 2015; Liu, 2016). Such technical or physiological artefacts may add (random or systematic) noise to a session, rendering a given study harder (or impossible) to interpret (Dipasquale *et al.*, 2017; Liu, 2016; Murphy *et al.*, 2013). Further, the detrimental effects of subject motion are substantial. While it was shown already a long time ago that a large portion of the variance in an fMRI time series is attributable to motion (Friston *et al.*, 1996), the full impact (on resting state studies in particular) has only become clear in recent years (Havsteen *et al.*, 2017; Power *et al.*, 2018; Wilke, 2012b). Consequently, ever more strict guidelines have been suggested with regard to what amount of motion is acceptable, and subjects failing these criteria are often removed from a group study (Afyouni & Nichols, 2018; Power *et al.*, 2012, 2015).

However, there are scenarios in which the decision to discard a subject's dataset is not an easy one, particularly in the presurgical application of fMRI. Termed "clinical functional MRI" early-on (Thulborn *et al.*, 1996), such exams are usually performed in the pre-operative context in subjects with tumorous brain lesions, or structural epilepsy (Benjamin *et al.*, 2017; Krings *et al.*, 2001; Lorenzen *et al.*, 2018; Szaflarski *et al.*, 2017, Wilke *et al.*, 2018). In this setting, much depends on the outcome of a given scanning session, and in the case of failure it may not always be possible to repeat or redo the scan. Here, both sensitivity and specificity are important: "real" activation must not be missed in order to provide correct

information about “eloquent cortex” to the neurosurgeon; this commonly leads to the exploration of several, and lower thresholds to maximize sensitivity (Tyndall *et al.*, 2017; Vlieger *et al.*, 2004; Zsoter *et al.*, 2012). However, false positive (spurious) foci of activation are also a concern (Juenger *et al.*, 2009) as their misinterpretation may lead to a less-aggressive surgical procedure than might otherwise have been considered. The challenge here is to get reliable results out of a given individual dataset of sub-optimal quality, even if it means investing some time and effort. In such a situation, different solutions as compared to a pure research setting are required (Chong, 2017; Wilke *et al.*, 2018).

One of the mainstays to achieving this aim are approaches that aim to “clean” the fMRI dataset of outlying datapoints (Caballero-Gaudes & Reynolds, 2017). One common approach to reduce the impact of the artefacts on ensuing analyses is to include censoring regressors in the statistical design, with the aim to explicitly model the outlier’s undesired variance. To achieve this, the contribution of a given datapoint to the final parameter estimates is down-weighted by introducing a new, binary regressor that contains a “1” for this datapoint and “0” for all others (Lemieux *et al.*, 2007; Siegel *et al.*, 2014). Consequently, unique contributions of this datapoint are very effectively removed (or censored) from the resulting statistical maps, allowing to account for, e.g., fast motion spikes exceeding a given threshold. Encouragingly, real-life analyses showed that usually, one or a few successive datapoints are outliers (Satterthwaite *et al.*, 2013), indicating that single datapoint censoring is a valid starting point. An acknowledged downside is that the loss of degrees of freedom (due to more parameters in the model) leads to an increase in statistical thresholds (Caballero-Gaudes & Reynolds, 2017; Liu *et al.*, 2001; Wilke, 2012b).

Another approach is interpolation: as compared to censoring, interpolation happens at the raw data stage. Here, "bad scans" are removed by directly interpolating the affected datapoint (Caballero-Gaudes & Reynolds, 2017; Mazaika *et al.*, 2009), using information from the unaffected neighboring volumes. To this effect, popular toolboxes are available (Artrepair, 2018). The general idea of this approach is not to explicitly model more variance on the statistical level, but instead to remove unwanted datapoints (and hence, their unwanted variance) from the data before proceeding to statistics.

While the approaches are different, there are similar issues. For one, as datapoints are removed from the analysis, less data is available to fit the model to. This corresponds to a loss of temporal power (Liu *et al.*, 2001). Further, issues pertaining to the identification of which points to censor or interpolate are similar. As motion is the main offender, the most common approach is to apply a motion threshold to identify outlying datapoints (Caballero-Gaudes & Reynolds, 2017), ideally making sure all parameters are meaningfully combined (Wilke, 2014). However, there are at least two distinct disadvantages to this approach. One, the observable motion (usually derived post-hoc by applying a rigid body realignment) may not truly reflect subject motion: for example, a fast moving subject's trajectory in the scanner will not be fully captured when only assessed at each TR, and quick motion only during the acquisition of one volume is not appropriately captured at all (Vaillant *et al.*, 2014). External motion tracking devices could help to overcome this shortcoming (Todd *et al.*, 2015) but are not yet in widespread use. And two, image imperfections induced by other sources (such as RF artefacts, breathing etc. [Birn, 2012; Campbell-Washburn *et al.*, 2016; Liu, 2016; Murphy *et al.*, 2013]) are not reflected in the realignment parameters at all. For explicitly modelling physiological effects in the data, other approaches exist, but these often require prospective collection of such data (Misaki *et al.*, 2015; Murphy *et al.*, 2013)

which again is not commonly available. As there is a limit to how many processing steps, or statistical adjustments can be applied to a given dataset (Powers 2015), a single approach integrating and combining different data-driven criteria seems preferable to identify outliers.

The aim of this manuscript therefore is twofold: one, develop and test a new, multidimensional and data-driven approach to identify outlying datapoints from a first-level fMRI dataset, assessing the effects of motion as well as other sources of artefacts. Two, assess the impact of censoring, interpolating, or both, on the chosen image quality parameters and the resulting statistical maps.

Materials and methods

General Approach and Implementation

In order to evaluate each individual scan's impact on a given session, three parameters are initially calculated, identifying different aspects of data corruption (such as subject motion or imaging artefacts). To then balance the removal of unwanted variance/scans against the increasingly complex design and/or the resulting loss of degrees of freedom, the corrected Akaike information criterion (AIC_C [Akaike, 1974]) is calculated as a fourth parameter. Individual parameter settings are described and discussed in the conclusion section, below.

Δ%D-var: The concept of DVARS (root mean square variance over the differenced timeseries) was first introduced by Smyser *et al.*, 2010, and then expanded upon later (Power *et al.*, 2012). In our approach, the percent change in the magnitude of the fast variance component of DVARS (Δ%D-var) is computed (Afyouni & Nichols, 2018). This parameter allows to identify datapoints showing a suspiciously fast change in signal (see Figure 1 for an illustration). Most elegantly, significance can be calculated on this parameter using a robust estimator of variance and employing χ^2 statistics, allowing to formally designate a scan as an outlier. In our algorithm, processing is done over all voxels in a slice, yielding one value per slice. Additionally or as an alternative, an excessively high value of Δ%D-var itself can be taken to indicate data corruption, especially if present in several slices.

Subject motion: subject motion is described by the scan-to-scan total displacement (STS, also known as framewise displacement, describing subject motion between two consecutive image volumes; Power *et al.*, 2012; Wilke, 2012, 2014). Motion (and fast motion in particular) has long since been recognized as one of the main sources of unwanted variance in fMRI (Friston *et al.*, 1996), and the identification of such fast-motion datapoints is very

commonly used to "scrub" time series (Caballero-Gaudes & Reynolds, 2017). Tukey's criterion (Bliss *et al.*, 1956; Tukey, 1977) is applied to find outliers (see below). However, an upper limit (above which scans are always considered outliers) will usually be specified. On the other end of the spectrum, minute movements may even be "detected" in the absence of actual movement due to outside influences; hence, a lower threshold (below which scans are never considered outliers) can also be set.

Overall explained variance: each scan's contribution to the variance explained by the whole session (expressed in the squared correlation coefficient, R^2 [Pernet, 2014]) is assessed. To this effect, the explained variance of the original model is calculated. Thereafter, each scan is consecutively removed, and the overall variance explained by this reduced model is related to the original model. In the case of an outlier, the variance explained by the new model increases (as the outlier's influence is explicitly explained). This yields a single value for each scan, reflecting the beneficial (ratio > 1) or detrimental (ratio < 1) effect of removing this datapoint from the session. These values are then assessed using Tukey's outlier criterion (see below). Of note, both Δ%D-var and R^2 will detect outliers irrespective of the underlying reason (subject motion, an image artefact, or both). There are no customizable settings for this parameter.

Model complexity: After an outlier is removed, there is always one datapoint that is the next-outlying; however, removing datapoints cannot go on indefinitely (Wilke, 2012a). To achieve balance, the Akaike information criterion was used (McLaren *et al.*, 2012). Its original implementation can be calculated within the SPM-framework according to

$$AIC = 2 \times k + n \times [\log(ResMS \times DOF / n)]$$

where *k* is the number of regressors in the model, *n* is the number of time points, ResMS are the mean squared residuals and DOF are the model's degrees of freedom (McLaren *et al.*, 2012). Of note, in a setting where the number of datapoints *n* does not far exceed k^2 (which is rarely the case in fMRI), a corrected version should be used (Glatting *et al.*, 2007; Hurwich & Tsai, 1989). This is particularly relevant when comparing models with different numbers of regressors, as done here. The corrected AIC can be computed according to

$$AIC_C = 2 \times k + n \times [\log(ResMS \times DOF / n)] + (2 \times k \times (k + 1) / (n - k - 1))$$

the only difference being the appended term penalizing higher model complexity. In our algorithm, AIC_C is initially calculated on the original model. Following the identification of outliers, these are then removed progressively upon which the calculation is repeated. The AIC_C of each (modified) model is then related to the original AIC_C and the global minimum (reflecting the best compromise) is determined. Rules have to be established that describe how to reconcile potentially conflicting suggestions of AIC_C and the other three parameters (see section on combining parameters, below). Other than that, here are no customizable settings for this parameter.

Censoring vs. interpolation: censoring is achieved by modifying the existing design matrix by adding an additional regressor for each outlying volume ("1" for the outlier, "0" for all other volumes). Interpolation is achieved here by a straightforward linear interpolation between neighboring non-outlying volumes, using the approach available within the popular art_repair toolbox (Artrepair, 2018; used here with kind permission).

Outlier Definition

While formal statistics can be calculated for Δ%D-var and the negative maximum of a simple

ratio is used for AIC_C, a criterion is required to find outliers in STS and R^2. An outlier is defined as a datapoint outside of the normal range "which appears to be inconsistent with the remainder of the dataset" (see Hodge & Austin, 2004, and Cousineau & Chartier, 2010, for a review). A common definition in a normal population is "a datapoint outside of two standard deviations of the mean". However, in the case of non-normally distributed data, the mean as well as the standard deviation are subject to bias (Leys *et al.*, 2013). The standard deviation in particular scales with both the number of outliers and the outlier weight (see upper panel in supplementary Figure 1), making this approach ill-suited in a setting with more than a few, and/or severe outliers (Cousineau & Chartier, 2010). We therefore decided to use Tukey's outlier criterion (Bliss *et al.*, 1956; Tukey, 1977), which is based on the more robust estimators of the third quartile and the interquartile range. Here, the upper limit UL of the normal range is defined per

$$UL = Q3 + F \times IQR$$

with Q3 being the third quantile, F being Tukey's factor, and IQR being the interquartile range. An outlier is usually assumed with F = 1.5 (also used here), while F = 3 signifies a far outlier. As can be shown empirically (see lower panel in supplementary Figure 1), the approach is very robust both to outlier weight and outlier number, but only up to a percentage of outliers of about 25. This effect is immediately clear from the formula given above, as both the third quantile as well as the interquartile range (the distance between the 25^{th} and the 75^{th} percentile) are directly affected if the number of outliers exceeds 25%. Above that threshold, there is an increased vulnerability towards both number of outliers and the outlier weight, severely diminishing the advantage of this robust estimator. Our algorithm therefore will, if more than 25% of datapoints are removed, re-estimate the

outlier range on the remaining datapoints. This will effectively re-establish the original robustness of the criterion with regard to outliers in both number and weight. Additionally, the third quartile and the interquartile range were derived here using a bootstrap estimator, with 10.000 estimates.

Combining parameters

Regarding the combination of the results from the 4 parameters (Δ%D-var, STS, R^2, and AIC_c), the following rules were implemented: by default, each datapoint identified as a definite outlier in either Δ%D-var, STS, or R^2, is marked for removal. If the total number across parameters is below the value suggested by AIC_C, no further action is taken (as AIC_C is only meant to provide an upper limit). If, on the other hand, the combined number exceeds the value suggested by AIC_C, definite outliers will still be removed until a prespecified distance factor for AIC_C is reached (default: 2). For example, the removal of 13 datapoints due to subject motion may conflict with the suggestion to only remove 10 datapoints (based on AIC_C). The distance factor then states that (detrimental) increases in AIC_C are only accepted until a factor of 2 w.r.t. the optimum AIC_C value is reached (i.e., if AIC_C increases to more than twice the optimal value). If this criterion is met, no further datapoints are removed. Two individual analyses are illustrated in Figure 2.

Datasets

For this study, we used three datasets: dataset 1 consists of 38 “scientific” imaging sessions from healthy children performing an acoustically-cued verb generation task (Northam *et al.*, 2012). Dataset 2 consists of 84 “clinical” imaging sessions from 28 pediatric patients

performing a left (n = 41) or right (n = 43) hand motor task as part of their assessment prior to epilepsy surgery. Datasets 1 & 2 were acquired at the Great Ormond Street Institute of Child Health, University College London, UK. Dataset 3 consists of 80 "scientific" imaging sessions from 20 adults performing 3 language tasks (Máté *et al.*, 2016; Fiori *et al.*, 2018) and one motor task designed to induce movement artefacts (active ankle movement). It was acquired at Tuebingen University Hospital, Tuebingen, Germany. Demographic details of the participants and further information about the sequence parameters and the task design can be found in Table 1.

Assessing algorithm performance: impact of subject motion

It is very difficult to realistically model motion artefacts as their effects are so diverse (Liu, 2016; Power *et al.*, 2015; Satterthwaite *et al.*, 2013; Wilke, 2012). As no ground truth exists, it is conceptually difficult to use a simulation to validate our approach. We therefore opted to investigate a "research" dataset (of healthy children performing a covert language task, likely of better quality; dataset 1) vs. a "clinical" dataset (of pediatric patients performing a motor task, likely of worse quality; dataset 2). Total subject motion was quantified by summing the absolute scan-to-scan values over time and related to the number of outlying datapoints as detected by the algorithm. This experiment was used to assess the hypothesis that, in more motion-corrupted series, the algorithm would detect more outliers.

Assessing algorithm performance: impact of data corruption

Non-motion image artefacts may be less relevant when compared with subject motion, but they may also be less easy to detect (Liu, 2016). As a model for image degradation, we

decided to simulate a common image artefact by introducing “stripes”. Such artefacts usually are brought about by radiofrequency noise bursts (Astrakas *et al.*, 2016; Campbell-Washburn *et al.*, 2016). We here chose to introduce a very slight signal intensity change (+3%) to each alternating slice of a randomly selected image, mimicking an effect as it might occur in an interleaved image acquisition sequence. This model has the advantage that the ground truth is known. The scans-to-be-corrupted were randomly selected from each individual time series from dataset 1, mimicking the unpredictability of the artefact. The artefact was introduced before realignment and smoothing, and the number of corrupted scans was systematically increased, affecting none (original analyses) or 10/20/30/40/50% of the images, resulting in 6 datasets of 38 sessions each. This experiment was used to assess the hypothesis that, in more artefact-corrupted series, the algorithm would detect more outliers.

Assessing algorithm performance: adaptability

As laid out above, the algorithm uses three different parameters to identify outliers, with the aim to detect outliers due to different influences in a self-adaptive manner, without having to change settings. This adaptability was tested here by assessing its performance in dataset 3, consisting of adults performing three (low-motion) language tasks and one (higher-motion) motor task. All subjects performed all tasks, minimizing between-subject differences. This experiment was used to assess the hypothesis that the different parameters will contribute differently to the final selection of outlying volumes in different settings.

Assessing algorithm performance: impact on resulting statistical map

One issue with declaring a statistical map "better" than another is, again, that no ground truth exists. While this is difficult to test in a cognitive task, the motor task used in dataset 2 must, if performed correctly, induce predictable activation in contralateral primary somatosensory brain regions (Guzzetta *et al.*, 2007). We therefore chose to assess the change in strength of activation (before and after outlier removal) at this cortical location as an indicator of successful outlier removal. To this effect, we manually screened all 84 t-maps from dataset 2 to identify activation in the targeted sensorimotor hand region, in native space. Around the center voxel of this activation, a cubic ROI was defined of ± 1 voxel in each dimension, resulting in a 3 × 3 × 3 voxel ROI. We then assessed the voxelwise t-values within the ROI, as well as the sum of activated voxels (after surviving an initial threshold of $p \leq 0.001$, uncorrected), before and after running our algorithm. This experiment was used to assess the hypothesis that stronger activation would be seen after outlier removal.

Assessing algorithm performance: impact on quality indices (Δ%D-var and STS)

An obvious choice for assessing algorithm performance is the change in the parameters we used to find outliers in the first place. Hence, we assessed in how far our selected data quality indicators changed as a function of progressively removing datapoints identified as outliers, in dataset 1 and dataset 2. As both likely differ in data quality, this also allows observing in how far the original data quality impacts these parameters. To this effect, we assessed the percent change in these parameters following removal of outlying volumes in 5%-steps. Of note, only the first two parameters (Δ%D-var and STS) can be assessed independently of the approach (censoring, interpolation, or both). The effect on R^2 and AIC_C

will vary depending on that approach, they are therefore assessed in a following experiment.

Assessing methodological approach: impact on quality indices (R^2 and AIC_C)

The variance explained by the model (R^2) as well as its complexity (AIC_C) will depend on the chosen approach (censoring, interpolation, and both). To this effect, results from the previous analyses (progressively removing outliers in steps of 5% in dataset 1 and dataset 2) were additionally assessed as a function of approach.

Assessing methodological approach: impact on detection power

Either censoring or interpolation can be pursued following the initial step of identifying the outlying volumes (and they can of course also be combined); however, it is unclear which approach is preferable and both are widely used (Siegel *et al.*, 2014; Mazaika *et al.*, 2009). A downside of censoring is that more complex models have fewer degrees of freedom, resulting in a loss of statistical power (Caballero-Gaudes & Reynolds, 2017; Liu *et al.*, 2001; Wilke, 2012b). An obvious drawback inherent in both approaches is the loss of temporal power, in that less data is available to fit the model to. We therefore investigated this effect here by randomly censoring or interpolating an increasing number of datapoints and assessed the impact on the original statistical map, in the form of counting the suprathreshold voxels at an $p \leq 0.001$, uncorrected. This was performed in dataset 1, in steps of 5, from 5 to 60 scans (i.e., up to a maximum of 50% of scans in this dataset). The procedure was repeated 100 times for each subject. Results were expressed as percent of suprathreshold voxels, as compared to the original images before manipulation (taken here

as the ground truth). The effect of an increase in statistical threshold and the decrease in temporal power were also calculated for comparison.

Assessing methodological approach: impact on temporal signal to noise

To further assess the influence of censoring or interpolation on the resulting timeseries, their temporal signal to noise (tSNR) was calculated, using the simple approach of dividing the mean of a timeseries by its standard deviation (Curtis & Menon, 2014; Welvaert & Rosseel, 2013). To assess the effect of censoring, tSNR was calculated for the reduced (leaving out the censored timepoints) as well as the interpolated timeseries. All results were related to the corresponding original timeseries in a voxelwise fashion. This was done for dataset 1 and dataset 2. Of note, this analysis cannot be conducted on the combined approach.

Image data processing

All processing steps and analyses were conducted within the SPM12 software environment (Wellcome Department of Imaging Neuroscience, UCL, London, UK), running in Matlab (The Mathworks, Natick, MA, USA), partly using custom scripts and functions. Functional MRI data preprocessing was minimal in that only realignment, reslicing, and spatial smoothing (FWHM = 6 mm) were performed. Single-subject, first level analyses were performed in native space using the general linear model (Friston *et al.*, 1995), contrasting the active periods with the intermittent rest periods. This yields native-space T-maps reflecting the active > control contrast.

Statistics

Normality assumptions in the data were assessed using an initial Kolmogorov-Smirnov-Liliefors-Test. Results are presented as means ± standard deviation in the case of normally-distributed data, and as median [standard error of the mean] in the case of non-normally-distributed data. If normality was demonstrated, differences in the mean were assessed using Student's T-test, while correlations were assessed using Pearsons's correlation coefficient. Otherwise, non-parametrical statistical tests were used instead. Specifically, differences in the median were assessed using the Mann-Whitney-U-Test, while correlations were assessed using Spearman's rank correlation. Significance was assumed at $p \leq 0.05$, Bonferroni-corrected for multiple comparisons where appropriate.

Results

Algorithm performance: impact of subject motion

Results of this experiment are shown in Figure 3. Subjects in dataset 2 showed more subject motion (left panel), but the difference did not reach significance (Mann-Whitney, $p > 0.05$). The algorithm identified and removed significantly more datapoints in dataset 2 vs. dataset 1 (middle panel; Mann-Whitney, $p < 0.001$), and there is a significant positive correlation between subject motion and number of removed datapoints (right panel) in both dataset 1 (Spearman, $p < 0.001$) and dataset 2 (Spearman, $p = 0.0024$).

Algorithm performance: impact of data corruption

Results of this experiment are shown in Figure 4. There is a clear increase in the number of removed datapoints as a function of increasing data corruption, up to a level of corruption of 30%. This increase is significant until then (no corruption vs. 10%, T-test, $p < 0.001$; 10% vs. 20%, Mann-Whitney, $p < 0.001$; 20% vs. 30%, Mann-Whitney, $p < 0.001$). At 40% and 50% corruption, the median decreases (although non-significantly) and the spread of results becomes much wider, indicating decreasing stability of outlier detection.

Assessing algorithm performance: adaptability

Results of this experiment are shown in Figure 5. In the three language tasks of dataset 3, the algorithm removes an average of 4-7 outliers out of 100 images, with remarkably stable contributions from the different parameters (mainly based on R^2; Figure 5, left upper and lower panels). In the motor task, the picture changes in that not only more (on average 22)

outliers are detected but also, their vast majority is now identified by Δ%D-var (Figure 5, right upper and lower panel), demonstrating that the algorithm displays a different (adaptive) behavior in different settings.

Algorithm performance: impact on resulting statistical map

Results of this experiment are shown in Figure 6, and some individual examples are provided in the supplementary Figure 2. Regarding activation strength as assessed by the T-value in the target ROI, there is a significant difference only for the original vs. censoring (Mann-Whitney, $p < 0.001$) and the original vs. both approach (Mann-Whitney, $p < 0.001$). The original vs. interpolation difference does not reach significance (Mann-Whitney, $p > 0.05$). Regarding activation strength as assessed by the number of activated voxels in the target ROI, the pattern is similar with a significant difference only for the original vs. censoring (Mann-Whitney, $p < 0.05$) and the original vs. both approach (Mann-Whitney, $p < 0.05$). The original vs. interpolation difference again does not reach significance (Mann-Whitney, $p > 0.05$).

Algorithm performance: impact on quality indices (Δ%D-var and STS)

Results of this experiment are shown in Figure 7. As expected, the progressive removal of outlying datapoints leads to a consistent and substantial reduction in both Δ%D-var and STS. The reduction was significantly more pronounced in dataset 2, for both indices (T-test, each $p < 0.05$).

Methodological approach: impact on quality indices (R^2 and AIC_C)

Results of this experiment are shown in Figure 8. There is a strong impact of approach on R^2, with the approaches involving censoring leading to a substantial increase of explained variance. In contrast to this, R^2 is almost unchanged in the interpolation approach. While there is an appreciable effect of dataset, neither difference reached significance (T-test, each $p > 0.05$). The effect of outlier removal on AIC_C is more complex: there is an early increase of AIC_C in the higher-quality dataset 1 in the censoring approaches, in stark contrast to a consistent decrease in the lower-quality dataset 2. This difference was highly significant for both approaches (T-test, each $p < 0.001$). For the interpolation approach, a consistent decrease was seen, with no differences between the datasets (T-test, $p > 0.05$).

Methodological approach: impact on detection power

Results of this experiment are shown in Figure 9. When progressively and randomly removing datapoints, there is an expected decrease in suprathreshold voxels when more datapoints are removed. This is true for all approaches, but is differently pronounced, in that the interpolation approach shows a lower loss of detection power then both censoring approaches. When assessing the known contributors (higher threshold due to fewer degrees of freedom and lower temporal power due to fewer datapoints), the effect of threshold is only minimal (when removing 50 datapoints, the T-threshold is only 2.84% higher) while temporal power naturally scales linearly with the number of datapoints.

Methodological approach: impact on temporal signal to noise

In dataset 1, there were slight increases in tSNR for both approaches (median 107.8% [SEM 2.0] for censoring and 108.6% [SEM 2.1] for interpolation), but neither difference reached significance, nor were they significantly different from each other. In dataset 2, there were substantial and significant increases in tSNR (median 165.0% [SEM 7.9] for censoring and 171.9% [SEM 8.0] for interpolation) for both approaches (T-test, each $p < 0.001$). Again, there was no significant difference between the approaches.

Discussion

This manuscript was aimed at (1) developing a new multidimensional approach to artefact reduction in first-level functional MRI sessions, and (2) assessing the impact of censoring versus interpolation. After evaluating the impact of subject motion and data corruption, we assessed the impact on the resulting statistical maps and on the four selected quality indices (Δ%D-var, STS, R^2 and AIC_C). Finally, the impact on detection power and temporal signal to noise was evaluated. These results shall now be discussed in more detail.

Impact of subject motion: While seemingly trivial and somewhat expected, ascertaining that the algorithm removes more datapoints in subjects that show more subject motion (Figure 3) seemed prudent as an initial step of assessing algorithm performance. Supporting the notion that the observable motion (described by the realignment parameters) does not represent the full extent of the ensuing data corruption (Friston *et al.*, 1996; Siegel *et al.*, 2014; Smyser *et al.*, 2010; Wilke, 2012b, 2014), the difference in subject motion did not reach significance while the number of removed datapoints did. Importantly, the correlation between subject motion and number of removed datapoints was significant in both datasets, confirming that the effect is not driven by the few high-motion subjects in dataset 2.

Impact of data corruption: Somewhat less trivial was the exploration of the algorithm's behavior in the context of an increasing (and known) amount of realistic data corruption (Figure 4). These results show that the algorithm is well able to detect datapoints contaminated by our synthetically-generated, but realistic non-motion image artefact

(Astrakas *et al.*, 2016; Campbell-Washburn *et al.*, 2016), with significantly increasing numbers removed up to a level of contamination of 30%. This demonstrates that the three indicators not relying on subject motion (Δ%D-var, R^2 and AIC_C) in fact do enable the algorithm to identify outliers not due to subject motion. Beyond a contamination level of 30%, this part of the algorithm becomes unstable, in that no more datapoints are removed despite more severe data corruption. While this could be considered algorithm failure, it should be noted that non-motion artifacts contaminating more than 30% of a dataset must be considered an extremely unlikely scenario. Further, this instability underlines the well-known and obvious fact (Hodge & Austin, 2004; Cousineau & Chartier, 2010) that the identification of an outlier requires the presence of a sufficient amount of normal datapoints (even in a robust implementation [Bliss *et al.*, 1956; Tukey, 1977], as also illustrated in supplementary Figure 1). Hence, attention should always be paid to the amount of datapoints removed during any censoring or interpolation procedure as, if a certain number is exceeded, this may indicate that the dataset is beyond repair (therefore, the final toolbox will generate a warning if more than 40% of datapoints are flagged as outliers). Also, this instability is only present in the case of non-motion induced image artefacts: due to the option to specify an absolute (lower and/or upper) motion threshold, datapoints contaminated by unacceptable subject motion will always be flagged (irrespective of their number) if they exceed a prespecified threshold.

Adaptability: In a real-world setting (and in the clinical application of fMRI in particular; Siegel *et al.*, 2014; Wilke *et al.*, 2018), contaminating influences of all sorts must be expected. Hence, the algorithm should ideally be equally-suitable for different scenarios,

requiring a high degree of adaptability. The results presented in Figure 5 illustrate that this actually is the case: in high-quality, very low-motion (language task) sessions, a low number of datapoints is removed (based mainly on R^2). The pattern changes substantially in that almost four times as many outliers (22 vs. ~6) are identified in the higher-motion (motor) task. It is interesting to note that STS does not dominate the motor task's outlying volume, arguing in favor of the more sophisticated Δ%D-var approach (Afyouni & Nichols, 2018) being more sensitive (although it must be admitted that this may also be an effect of the chosen STS threshold). This pattern demonstrates that the algorithm does not rely on one parameter only and underlines the importance of using a combined outlier detection approach. Further, the excellent performance of the same algorithm with the same settings not only in this "very low vs. low motion" (Figure 5) but also in a "medium vs. high motion" (Figure 3) and in a "low vs. -high artefact" setting (Figure 4) underscores its adaptability. On the single-subject level, this can also be seen in the "best" and "worst" cases: here, some datapoints clearly identified as outliers using Δ%D-var and/or R^2 did not show excessive subject motion (red and green rectangles in Figure 2) which is why it is all the more important to not *a priori* focus on one aspect only.

Impact on resulting statistical maps: The results of outlier removal on the strength of activation in dataset 2 (Figure 6) show that both censoring and interpolation of outlying volumes lead to an observable increase in activation. Some illustrative cases are shown in the supplementary Figure 2. However, in this dataset the difference is only significant in the censoring approaches. While the overall effects are similar and while therefore, this lack of significance should not be over-interpreted, it is based on 84 imaging sessions and the

pattern is similar for both the absolute T-scores as well as for the activated voxels. Results from this experiment therefore indicate that there may be an advantage of the approaches relying on censoring regressors in a real-life scenario, in this respect.

Impact on quality indices (Δ%D-var and STS): Again somewhat as expected, the impact of removing outlying datapoints on Δ%D-var and subject motion is very clear (Figure 7). Both parameters show a substantial and consistent decrease upon removing more "outlying" datasets. Of note, the initial reduction even in the better-quality dataset 1 is already on the magnitude of about 20% when only removing 5% of datapoints. Also, there is a clear difference between the two datasets, with a substantially (and significantly) stronger reduction of both parameters in the worse-quality dataset 2. This clearly indicates that the effect of removing a similar number of outlying datapoints is stronger in worse datasets.

Impact on quality indices (R^2 and AIC_C): When assessing the impact of removing outlying datapoints on R^2 and AIC_C, a more complex pattern emerges (Figure 8), both as a function of approach and of data quality. When assessing R^2, the impact of the methodological approach is very obvious: there is a clear and substantial increase in the total variance explained by the model including censoring regressors (top and bottom left panels in Figure 8). The impact of data quality is visually suggestive in that relatively more variance is explained for the worse-quality dataset 2, but this difference does not reach significance. In stark contrast to this, the interpolation approach has only a negligible (negative) influence on the explained variance (middle left panel in Figure 8). Here, there is no perceptible difference between the datasets. The picture is different for AIC_C, where the influence of

dataset predominates for the censoring approaches. Here, an early increase in the better-quality dataset 1 clearly indicates that the removal of further datapoints does not lead to a better balance between explained variance and model complexity. In contrast to this, AIC_C continues to decrease in the worse-quality dataset 2, showing an additional benefit of removing more datapoints (top and bottom right panels in Figure 8). This behavior clearly illustrates that the criterion's intended effect as a boundary condition (indicating when to stop removing datapoints) is achieved: while in both datasets, more variance is explained by censoring more outlying datapoints, AIC_C ensures that a compromise is enforced between model complexity and explained variance (upper left and right panel in Figure 8). This compromise leads to an earlier termination of datapoint removal in the better-quality dataset than in the worse-quality dataset. Again, the picture is different for the interpolation approach, where model complexity does not change: the model itself remains unchanged as interpolation occurs on the raw data level. However, due to the interpolation and removal of outlying datapoints, the unexplained residual variance is reduced, which in turn leads to a lower AIC_C (middle right panel in Figure 8). Again, the effect of dataset is not significant here.

Impact on detection power: In the next experiment (Figure 9), the influence of methodological approach on detection power (proxied here using the activation strength) is assessed. As expected (Liu *et al.*, 2001; Wilke, 2012b), detection power decreases across all approaches (censoring, interpolation, or both) upon increasingly removing datapoints. The effect of fewer degrees of freedom (open circles in Figure 9, only relevant for the censoring approaches), however, is hardly relevant; while there is an increase in the uncorrected

statistical cutoff, it is below 3% even when censoring 50 datapoints. Compared with this, the loss of datapoints contributing to the model (i.e., temporal power) is much more relevant (solid circles in Figure 9). Interestingly, there is a clear effect of approach in so far that the interpolation approach shows a less-pronounced decrease in detection power than the censoring approaches, becoming more obvious the more datapoints are removed. Between the two approaches including censoring, there is no appreciable difference. This actually, on a side note, confirms the effective removal of all variance explained by a datapoint if a censoring regressor is included (Siegel *et al.*, 2014), as no (positive or negative) effect of interpolation remains if the datapoint is also censored.

Impact on temporal signal to noise: In our final experiment, the influence of methodological approach on temporal signal to noise (tSNR; Curtis & Menon, 2014; Welvaert & Rosseel, 2013) was calculated. Again, the influence of dataset quality is substantial, with the worse-quality dataset 2 benefitting significantly from either censoring or interpolation. In contrast to this, the slight improvement observable in the better-quality dataset 1 was not significant, for either approach. Of note and in both datasets, there was no significant difference w.r.t. the change in tSNR between the two approaches.

Limitations

The question of how many datapoints can meaningfully be removed using such an approach is currently unanswered: if 50% of datapoints are interpolated, on average only one datapoint remains for each datapoint removed, raising concerns not only about the remaining temporal power (a similar issue for all approaches, cf. Figure 9;

Welvaert & Roseel, 20013) but also about the robustness of the resulting interpolation (an issue for interpolation only). In this context, there are also different concerns regarding the calculation of the temporal autocorrelation in fMRI time series between the approaches (Caballero-Gaudes & Reynolds, 2017). As the exact implementation of interpolation is an important step, we also explored other approaches and implemented a hole-filling interpolation ("inpainting") approach, employing partial differential equations (D'Errico, 2018). Here, a smooth interpolant based on the neighboring values is achieved by minimizing the sum of squares of the second derivative at each node, such that the whole dataset (including all to-be-removed datapoints) can be assessed as one single problem. However, while theoretically appealing (and computationally much more demanding), there was no appreciable difference when compared with the much more simple linear interpolation approach (Mazaika, 2009) when a low number of datapoints was interpolated. Surprisingly, the rate of false positive results actually seemed to increase when more datapoints were interpolated, which could be due to the higher interpolation-induced temporal smoothness of the data which may be counterproductive. Finally, the question of whether scattered outlying datapoints are of less concern for interpolation then tightly-clustered ones was also not explored here. It should also be noted that we only investigated block designs: the effect of censoring and/or interpolation on event-related studies was not assessed here. However, given that block designs will constitute the majority of sessions acquired in a clinical context (Wilke *et al.*, 2018), we considered this an acceptable limitation.

Possible Extensions

Following identification of outliers, the original model is re-estimated with an increasingly

larger number of censoring regressors. As a byproduct, the overlap between the statistical maps from these reduced analyses can be calculated, which is in some aspects similar to a previously-described approach to assess reliability on the fMRI group level (Wilke, 2012a). In analogy to this approach, overlapping activations could be considered more reliable if they are present in more of the reduced analyses, thus serving as an additional indicator for the robustness of observed foci of activation.

The impact of our algorithm on second-level statistics was not assessed here as, when analyzing single subjects in a clinical context, the number of censored datapoints may entirely be oriented on individual factors (Siegel *et al.*, 2014). However, this is not the case anymore if such analyses should be performed on subjects prior to entering a group-level statistics. While the impact of fewer degrees of freedom on the first level is not very strong on the second level (Wilke, 2012b), an uneven distribution of censored images between groups still constitutes a bias. For such a scenario (for example when studying rare diseases where every single subject, contributing to a small group, is precious), it is therefore possible to zeropad the final model by adding dummy regressors. This ensures that a constant number of regressors (and hence, degrees of freedom) is achieved across subjects to avoid an undue bias in the ensuing analyses.

Summary

In summary and in response to our first aim, the algorithm suggested here is well able to detect not only motion (Figure 3), but also imaging artefacts of other sources (Figure 4), in first-level fMRI analyses. To this effect, it assesses different dimensions of dataset quality in a data-driven (Figure 1) and statistically robust (supplementary Figure 1) way. Motion-

corrupted datapoints can be flagged using different thresholds, and even sub-threshold motion effects on the data may be captured by the other indicators (Figure 2). Non-motion related image artefacts are also detected reliably (Figure 4); while this behavior is only stable until a contamination of about 40%, heavier contamination by non-motion artefacts is an extremely unlikely scenario. The approach is adaptive in that different parameters predominate for different scenarios (Figure 5). As a result of applying the algorithm, stronger activation as detected by both T-value and number of activated voxels can be seen (Figure 6). The multidimensional approach is effectively controlled by a model complexity parameter (Figures 6 & 7), balancing variance vs. complexity and reliably ensuring that fewer datapoints are removed in less-corrupted datasets. Its application leads to increases in tSNR, significantly more so in lower-quality datasets.

With regard to censoring vs. interpolation, our results do not allow to formulate a clear and unequivocal conclusion as to which approach is superior. As evident from the analyses assessing explained variance and AIC_C (Figure 8), both approaches have a fundamentally different mechanism of action: while "bad" variance is explained using the censoring approach, it is removed using the interpolation approach. While the analyses assessing activation strength (Figure 6) seem to favor censoring, the results assessing detection power (Figure 9) seem to favor interpolation. A similarly mixed picture emerges when assessing the combined approach: as can be seen from the simulations on removing datapoints (Figure 9), censoring so effectively removes variance associated with a given datapoint that an additional interpolation does not seem to make much of a difference. Then again, there is a clear difference in explained variance and AIC_C if interpolated datapoints are additionally

censored (Figure 8), with an intermediate pattern apparent for both parameters in the combined approach. From a conceptual point of view, we believe that it is preferable to not disrupt the timeseries: as the effect of motion or other artefacts on its temporal consistency is not wholly understood (Liu, 2016; Satterthwaite *et al.*, 2013), interpolation of a removed datapoint may only incompletely remove data corruption but may instead introduce new sources of variance. Therefore, while both approaches (and their combination) may have distinct advantages and disadvantages in different scenarios and while all three approaches seem valid, we believe that censoring will generally be preferable.

Recommended settings

Based on the experiments and simulations described herein, two default parameter settings will be available in the toolbox resulting from this work, one for a task-based and one for a resting-state fMRI scenario. While the ultimate choice of parameter settings will depend on many factors (and all settings can still be customized), some recommendations can be provided as follows (for a summary, see also Table 2).

Settings relevant for Δ%D-var are the prespecified p-value (default for both scenarios: 0.05), the required percentage of slices that need to be affected in order to consider the datapoint corrupt (default for both scenarios: 50%), and the absolute Δ%D-var value that is considered excessive. In Afyouni & Nichols (2018), a Δ%D-var value of 5% was suggested for resting state studies; this, however, may be overly strict for task-based fMRI; for this scenario, we suggest a default value of 15%.

Settings relevant for STS are the prespecified upper and lower motion thresholds. For the task-based studies assessed here, we used default values of 1.5 mm and 0.3 mm,

respectively. These, however will be too liberal for resting state studies, where the effect of motion is critically more important (Afyouni & Nichols, 2018; Power *et al.*, 2012, 2015, 2018). Here, values of 0.3 mm and 0 mm, respectively, will be used. The default for the assumed average cortical distance (required to convert radians into distances; Wilke, 2014) is 65 mm and is used for both scenarios.

With regard to on R^2, it must be borne in mind that the overall explained variance of a given scenario will likely be an important parameter for task-based fMRI studies, but may not be the focus of a resting-state study, where further analyses (such as connectivity or independent component analyses) may follow. Hence, requiring a datapoint to contribute meaningfully to a not meaningful outcome parameter does not make sense. Consequently, while the parameter is central for task-based fMRI (cf. Figure 5), it will by default be disabled for resting state analyses.

Similarly, balancing unwanted variance vs. model complexity using AIC_C also only makes sense if the model itself is a meaningful one. Hence, while the balancing effect of AIC_C is central for task-based fMRI (cf. Figure 8), it will by default be disabled for resting state analyses.

With regard to censoring vs. interpolation, based on our considerations above we suggest to usually use censoring for task-based fMRI analyses. Yet again, however, if the main aim of a resting state study is a follow-up analysis using a “cleaned” dataset, interpolation will be more helpful (as the interpolated dataset is written out by the toolbox and can be used instead of the original one).

Acknowledgments

This study was funded in part by a grant of the H.W. & J. Hector Foundation, Mannheim, Germany (M66), to MW. TB was supported by Great Ormond Street Hospital Children's Charity and Epilepsy Research UK. Both sponsors had no part in study design, interpretation of results, or decision to publish.

The algorithm described in this manuscript is available in the form of a SPM12 toolbox for free download at *http://www.medizin.uni-tuebingen.de/kinder/epn*

Literature

Afyouni S, Nichols TE (2018). *Insight and inference for DVARS*. NeuroImage 172: 291-312. doi: 10.1016/j.neuroimage.2017.12.098

Akaike H (1974). A new look at the statistical model identification. IEEE Trans. Autom Control 19: 716-723

Artrepair Toolbox, available at http://cibsr.stanford.edu/tools/human-brain-project/artrepair-software.html (last accessed July 30, 2018)

Astrakas LG, Kallistis NS, Kalef-Ezra JA (2016). Technical Note: Independent component analysis for quality assurance in functional MRI. Med Phys 43:983-992. doi: 10.1118/1.4940123.

Benjamin CF, Walshaw PD, Hale K, Gaillard WD, Baxter LC, Berl MM, Polczynska M, Noble S, Alkawadri R, Hirsch LJ, Constable RT, Bookheimer SY (2017). Presurgical language fMRI: Mapping of six critical regions. Hum Brain Mapp 38:4239-4255. doi: 10.1002/hbm.23661

Bliss CI, Cochran WG, Tukey JW (1956). A Rejection Criterion Based Upon the Range. Biometrika 43: 418-422

Birn RM (2012). The role of physiological noise in resting-state functional connectivity. Neuroimage 62:864-70. doi: 10.1016/j.neuroimage.2012.01.016

Caballero-Gaudes C, Reynolds RC (2017). Methods for cleaning the BOLD fMRI signal. Neuroimage 154: 128-149. doi: 10.1016/j.neuroimage.2016.12.018

Campbell-Washburn AE, Atkinson D, Nagy Z, Chan RW, Josephs O, Lythgoe MF, Ordidge RJ, Thomas DL (2016). Using the robust principal component analysis algorithm to remove RF spike artifacts from MR images. Magn Reson Med 75:2517-25. doi: 10.1002/mrm.25851

Chen JE, Glover GH (2015). Functional Magnetic Resonance Imaging Methods. Neuropsychol Rev 25: 289-313. doi: 10.1007/s11065-015-9294-9

Chong TT (2017). Voodoo surgery? The distinct challenges of functional neuroimaging in clinical neurology. Brain 140(12):e76. doi: 10.1093/brain/awx283

Cousineau D, Chartier S (2010). Outliers detection and treatment: a review. Int J Psychol Res 3: 58-67.

Curtis AT, Menon RS (2014). Highcor: a novel data-driven regressor identification method for BOLD fMRI. Neuroimage 98: 184-194. doi: 10.1016/j.neuroimage.2014.05.013.

D'Errico J (2018). inpaint_nans: Interpolates (& extrapolates) NaN elements in a 2d array. Available at https://de.mathworks.com/matlabcentral/fileexchange/4551-inpaint_nans, last accessed August 9, 2018

Dipasquale O, Sethi A, Laganà MM, Baglio F, Baselli G, Kundu P, et al. (2017) Comparing resting state fMRI de-noising approaches using multi- and single-echo acquisitions. PLoS ONE 12: e0173289. doi: 10.1371/journal.pone.0173289

Fiori S, Zendler C, Hauser TK, Lidzba K, Wilke M (2018).Assessing motor, visual and language function using a single 5-minute fMRI paradigm: three birds with one stone. Brain Imaging Behav: in press

Friston KJ, Holmes AP, Worsley KJ, Poline JB, Frith CD, Frackowiak RSJ (1995). Statistical Parametric Maps in Functional Imaging: A General Linear Approach. Hum Brain Mapp 2:189-210

Friston KJ, Williams S, Howard R, Frackowiack RSJ, Turner R (1996). Movement-related effects in fMRI time-series. Magn Reson Med 35: 346-355

Glatting G, Kletting P, Reske SN, Hohl K, Ring C (2007). Choosing the optimal fit function: comparison of the Akaike information criterion and the F-test. Med Phys 34:4285-92. doi: 10.1118/1.2794176

Guzzetta A, Staudt M, Petacchi E, Ehlers J, Erb M, Wilke M, Krägeloh-Mann I, Cioni G (2007). Brain representation of active and passive hand movements in children. Pediatr Res 61: 485-490.

Havsteen I, Ohlhues A, Madsen KH, Nybing JD, Christensen H, Christensen A (2017). Are Movement Artifacts in Magnetic Resonance Imaging a Real Problem?-A Narrative Review. Front Neurol 8: 232. doi: 10.3389/fneur.2017.00232

Hodge VJ, Austin J (2004). A Survey of Outlier Detection Methodologies. Artif Intell Rev 22: 85-126

Hurwich CM, Tsai CL (1989) Regression and time series model selection in small samples. Biometrika 76: 297–307

Juenger H, Ressel V, Braun C, Ernemann U, Schuhmann M, Krägeloh-Mann I, Staudt M (2009). Misleading functional magnetic resonance imaging mapping of the cortical hand representation in a 4-year-old boy with an arteriovenous malformation of the central region. J Neurosurg Pediatr 4(4): 333-338. doi: 10.3171/2009.5.PEDS08466

Lemieux L, Salek-Haddadi A, Lund TE, Laufs H, Carmichael D (2007). Modelling large motion events in fMRI studies of patients with epilepsy. Magn Reson Imaging 25: 894-901

Leys C, Klein O, Bernard P, Licata L (2013). Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median. J Exp Soc Psychol 49: 764-776

Liu TT, Frank LR, Wong EC, Buxton RB (2001). Detection power, estimation efficiency, and predictability in event-related fMRI. Neuroimage 13:759-773

Liu TT (2016). Noise contributions to the fMRI signal: An overview. Neuroimage 143:141-151. doi: 10.1016/j.neuroimage.2016.09.008

Lorenzen A, Groeschel S, Ernemann U, Wilke M, Schuhmann MU (2018). Role of presurgical functional MRI and diffusion MR tractography in pediatric low-grade brain tumor surgery: a single-center study. Childs Nerv Syst: in press. doi: 10.1007/s00381-018-3828-4

Máté A, Lidzba K, Hauser TK, Staudt M, Wilke M (2016). A "one size fits all" approach to language fMRI: increasing specificity and applicability by adding a self-paced component. Exp Brain Res 234: 673-684

Mazaika P, Hoeft F, Glover GH, Reiss AL (2009). Methods and Software for fMRI Analysis for Clinical Subjects. Ann Meeting of the Organization for Human Brain Mapping, San Francisco

McLaren DG, Ries ML, Xu G, Johnson SC (2012). A generalized form of context-dependent psychophysiological interactions (gPPI): a comparison to standard approaches. Neuroimage 61:1277-1286. doi: 10.1016/j.neuroimage.2012.03.068

Misaki M, Barzigar N, Zotev V, Phillips R, Cheng S, Bodurka J (2015). Real-time fMRI processing with physiological noise correction - Comparison with off-line analysis. J Neurosci Method 256:117-21. doi: 10.1016/j.jneumeth.2015.08.033

Murphy K, Birn RM, Bandettini PA (2013). Resting-state fMRI confounds and cleanup. Neuroimage 80: 349-59. doi: 10.1016/j.neuroimage.2013.04.001

Northam GB, Liégeois F, Tournier JD, Croft LJ, Johns PN, Chong WK, Wyatt JS, Baldeweg T (2012). Interhemispheric temporal lobe connectivity predicts language impairment in adolescents born preterm. Brain 135: 3781-3798. doi: 10.1093/brain/aws276

Pernet CR (2014). Misconceptions in the use of the General Linear Model applied to functional MRI: a tutorial for junior neuro-imagers. Front Neurosci. 8: 1. doi: 10.3389/fnins.2014.00001

Power JD, Barnes KA, Snyder AZ, Schlaggar BL, Petersen SE (2012). Spurious but systematic correlations in functional connectivity MRI networks arise from subject motion. Neuroimage 59:2142-54. doi: 10.1016/j.neuroimage.2011.10.018

Power JD, Schlaggar BL, Petersen SE (2015). Recent progress and outstanding issues in motion correction in resting state fMRI. Neuroimage 105:536-51. doi: 10.1016/j.neuroimage.2014.10.044

Power JD, Plitt M, Gotts SJ, Kundu P, Voon V, Bandettini PA, Martin A (2018). Ridding fMRI data of motion-related influences: Removal of signals with distinct spatial and physical bases in multiecho data. Proc Natl Acad Sci USA 115: E2105-E2114. doi: 10.1073/pnas.1720985115

Satterthwaite TD, Elliott MA, Gerraty RT, Ruparel K, Loughead J, Calkins ME, Eickhoff SB, Hakonarson H, Gur RC, Gur RE, Wolf DH (2013). An improved framework for confound regression and filtering for control of motion artifact in the preprocessing of resting-state functional connectivity data. Neuroimage 64:240-56. doi: 10.1016/j.neuroimage.2012.08.052

Siegel JS, Power JD, Dubis JW, Vogel AC, Church JA, Schlaggar BL, Petersen SE (2014). Statistical improvements in functional magnetic resonance imaging analyses produced by censoring high-motion data points. Hum Brain Mapp 35:1981-1996. doi: 10.1002/hbm.22307

Smyser CD, Inder TE, Shimony JS, Hill JE, Degnan AJ, Snyder AZ, Neil JJ (2010). Longitudinal analysis of neural network development in preterm infants. Cereb Cortex 20:2852-62. doi: 10.1093/cercor/bhq035

Szaflarski JP, Gloss D, Binder JR, Gaillard WD, Golby AJ, Holland SK, Ojemann J, Spencer DC, Swanson SJ, French JA, Theodore WH (2017). Practice guideline summary: Use of fMRI in the presurgical evaluation of patients with epilepsy: Report of the Guideline Development, Dissemination, and Implementation Subcommittee of the American Academy of Neurology. Neurology 88:395-402. doi: 10.1212/WNL.0000000000003532

Thulborn KR, Davis D, Erb P, Strojwas M, Sweeney JA (1996). Clinical fMRI: implementation and experience. Neuroimage 4: S101-107

Tukey JW (1977). Exploratory data analysis. Addison-Wesely, 1977

Todd N, Josephs O, Callaghan MF, Lutti A, Weiskopf N (2015). Prospective motion correction of 3D echo-planar imaging data for functional MRI using optical tracking. Neuroimage 113:1-12. doi: 10.1016/j.neuroimage.2015.03.013

Tyndall AJ, Reinhardt J, Tronnier V, Mariani L, Stippich C (2017). Presurgical motor, somatosensory and language fMRI: Technical feasibility and limitations in 491 patients over 13 years. Eur Radiol 27: 267-278

Vaillant G, Prieto C, Kolbitsch C, Penney G, Schaeffter T (2014). Retrospective Rigid Motion Correction in k-Space for Segmented Radial MRI. IEEE Trans Med Imaging 33:1-10. doi: 10.1109/TMI.2013.2268898. Epub 2013 Jun 14.

Vlieger EJ, Majoie CB, Leenstra S, Den Heeten GJ (2004). Functional magnetic resonance imaging for neurosurgical planning in neurooncology. Eur Radiol 14: 1143-53

Welvaert M, Rosseel Y (2013). On the definition of signal-to-noise ratio and contrast-to-noise ratio for FMRI data. PLoS One 8(11) :e77089. doi: 10.1371/journal.pone.0077089

Wilke M (2012a). An iterative jackknife approach for assessing reliability and power of FMRI group analyses. PLoS One 7: e35578. doi: 10.1371/journal.pone.0035578

Wilke M (2012b). An alternative approach towards assessing and accounting for individual motion in fMRI timeseries. Neuroimage 59: 2062-72. doi: 10.1016/j.neuroimage.2011.10.043

Wilke M (2014). Isolated assessment of translation or rotation severely underestimates the effects of subject motion in fMRI data. PLoS One 9:e106498. doi: 10.1371/journal.pone.0106498

Wilke M, Groeschel S, Lorenzen A, Rona S, Schuhmann MU, Ernemann U, Krägeloh-Mann I (2018). Clinical application of advanced MR methods in children: points to consider. Ann Clin Transl Neurology 27: 1434-1455 (in press). doi: https://doi.org/10.1002/acn3.658

Zsoter A, Staudt M, Wilke M (2012). Identification of successful clinical fMRI sessions in children: an objective approach. Neuropediatrics 43: 249-257. doi: 10.1055/s-0032-1324731

Tables

	Age [years]	Gender [M/F]	TR [s]	VS [mm]	Datapoints	Task	Design	STS/frame [mm]
Dataset 1 (UCL, n=38, 38 sessions)	14.4 ± 2.6	19/19	3.32	3 × 3 × 4	120	Language task	10 active blocks [6 scans]	0.09 [0.02]
Dataset 2 (UCL, n=28, 84 sessions)	12.7 ± 4.7	18/10	2.16	3.3 × 3.3 × 4	50	Motor task	5 active blocks [5 scans]	0.26 [0.05]
Dataset 3 (UKT, n=20, 80 sessions)	31.7 ± 7	8/12	3.00	3 × 3 × 3	100	Language tasks	5 active blocks [10 scans]	0.06 [0.02]
						Motor task		0.14 [0.09]

Table 1: Demographic information about subjects and datasets; UCL, University College London; UKT, University Clinics Tübingen

Parameter	Setting	task-based fMRI	resting-state fMRI
Δ%D-var	p [1]	0.05	→
	% $_{slices}$ [2]	50%	→
	% $_{excessive}$ [3]	15%	5%
STS	upper threshold [4]	1.5 mm	0.3 mm
	lower threshold [5]	0.3 mm	0 mm
	d $_{avg}$ [6]	65 mm	→
R ^{2} [7]		enabled	disabled
AIC $_c$ [8]		enabled	disabled
Default approach		censoring	interpolation

Table 2: Summary of recommended parameter settings for different scenarios (task-based versus resting-state fMRI). Legend: → same value as for task-based fMRI; [1] p-value required for assuming significance; [2] percentage of slices required to be outliers; [3] absolute value of Δ%D-var considered excessive; [4] motion above this threshold will always be considered an outlier; [5] motion below this threshold will never be considered an outlier; [6] value for average cortical distance; [7] calculation of each scan's contribution to the overall explained variance; [8] calculation of corrected Akaike's information criteria to balance scan removal vs. model complexity

Figure Legends

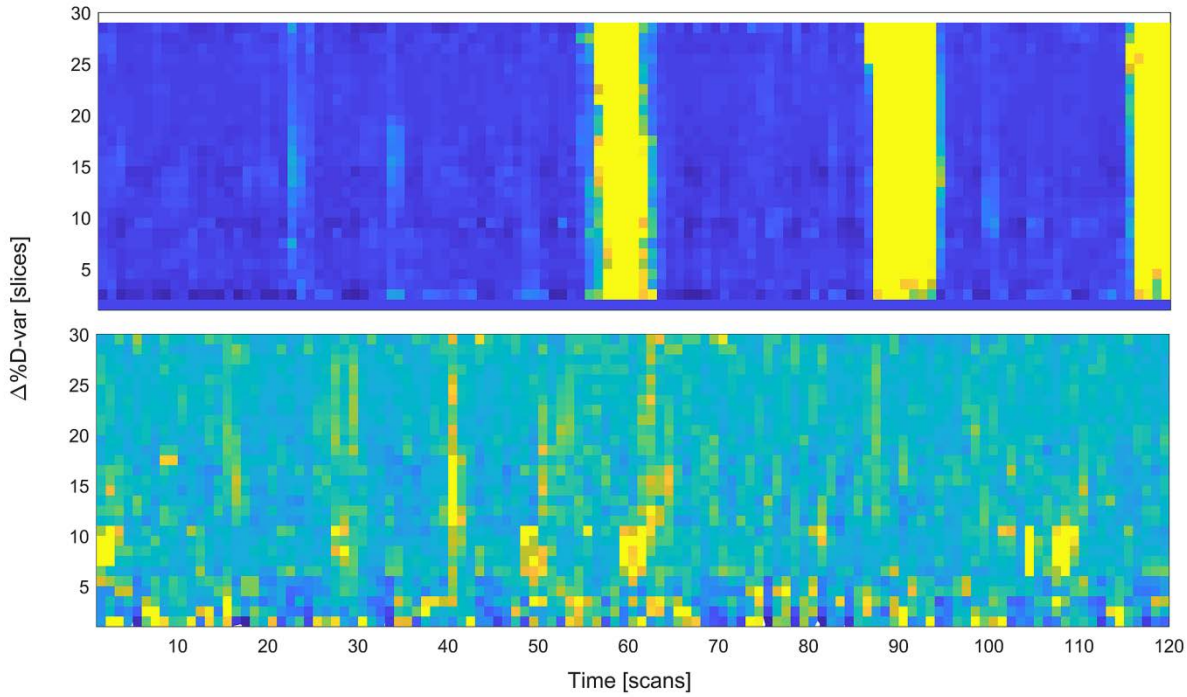


Figure 1: Illustration of the change in the fast variance component of DVARS (Δ%D-var) in the "worst" subject in dataset 1 (top panel) when compared with the "best" subject in dataset 1 (bottom panel). Note three "bursts" of substantially contaminated datapoints in the top panel, corresponding to subject motion (see also Figure 2), and only very low level of data contamination in the lower panel (identical color scaling, 0-20%).

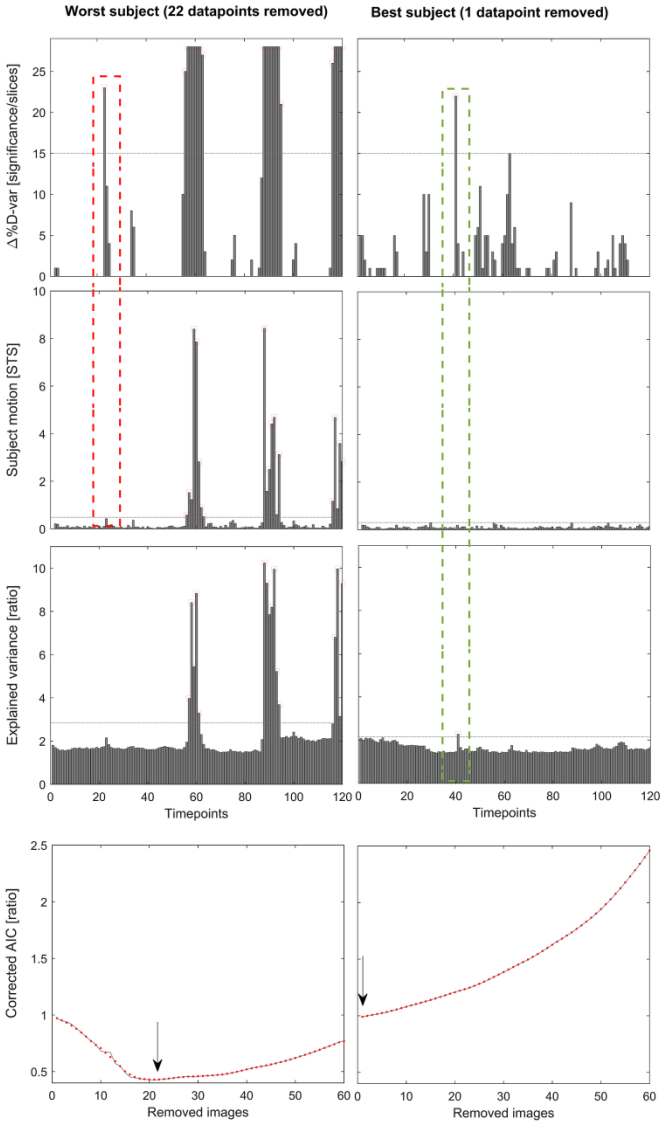


Figure 2: Comparison of all 4 parameters in the "worst" (left panels) vs. the "best" subject (right panels) in dataset 1; cf. also Figure 1. First row: number of slices with significant change in Δ%D-var. Datapoints are removed (indicated by small red squares) if at least half of its slices show a significant change (horizontal dotted line). Second row: subject scan-to-scan motion. Datapoints are removed depending on prespecified thresholds (dotted line). Third row: effect of removing each datapoint on the overall explained variance of the model. Datapoints are removed according to Tukey's outlier criterion (dotted line). Fourth row: corrected Akaike information criterion as a function of progressively removing outlying datapoints (lower values reflect a better compromise between model complexity and explained variance). Note progressive decrease in the artefact-ridden dataset (arrow in lower left panel) vs. no benefit of removing further datapoints in the best dataset (arrow in lower right panel). Red rectangle: removal of one frame detected as an outlier in Δ%D-var despite sub-threshold motion. Green rectangle: identification of the same singular outlying datapoint in the "best" subject by Δ%D-var and R^2 (but not STS).

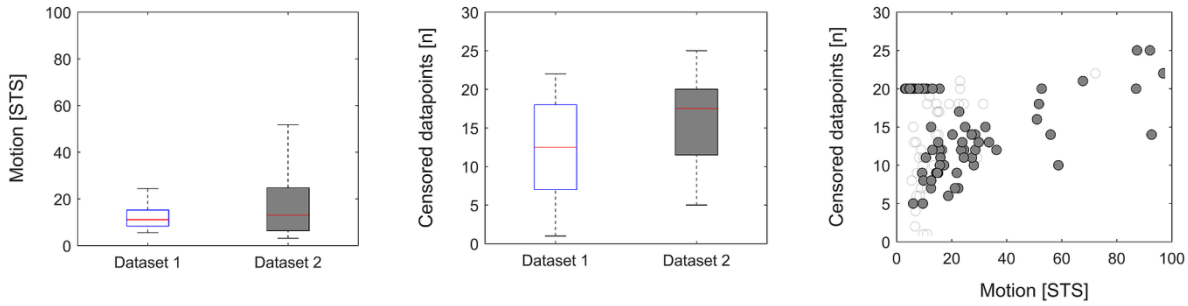


Figure 3: Subject motion versus removed datapoints in datasets 1 (white box/markers) and 2 (gray box/markers). Note substantially higher amount of subject motion in dataset 2 vs. dataset 1 (left panel), leading to a higher number of removed datapoints (middle panel). There is a clear correlation between subject motion and removed datapoints in both datasets (right panel).

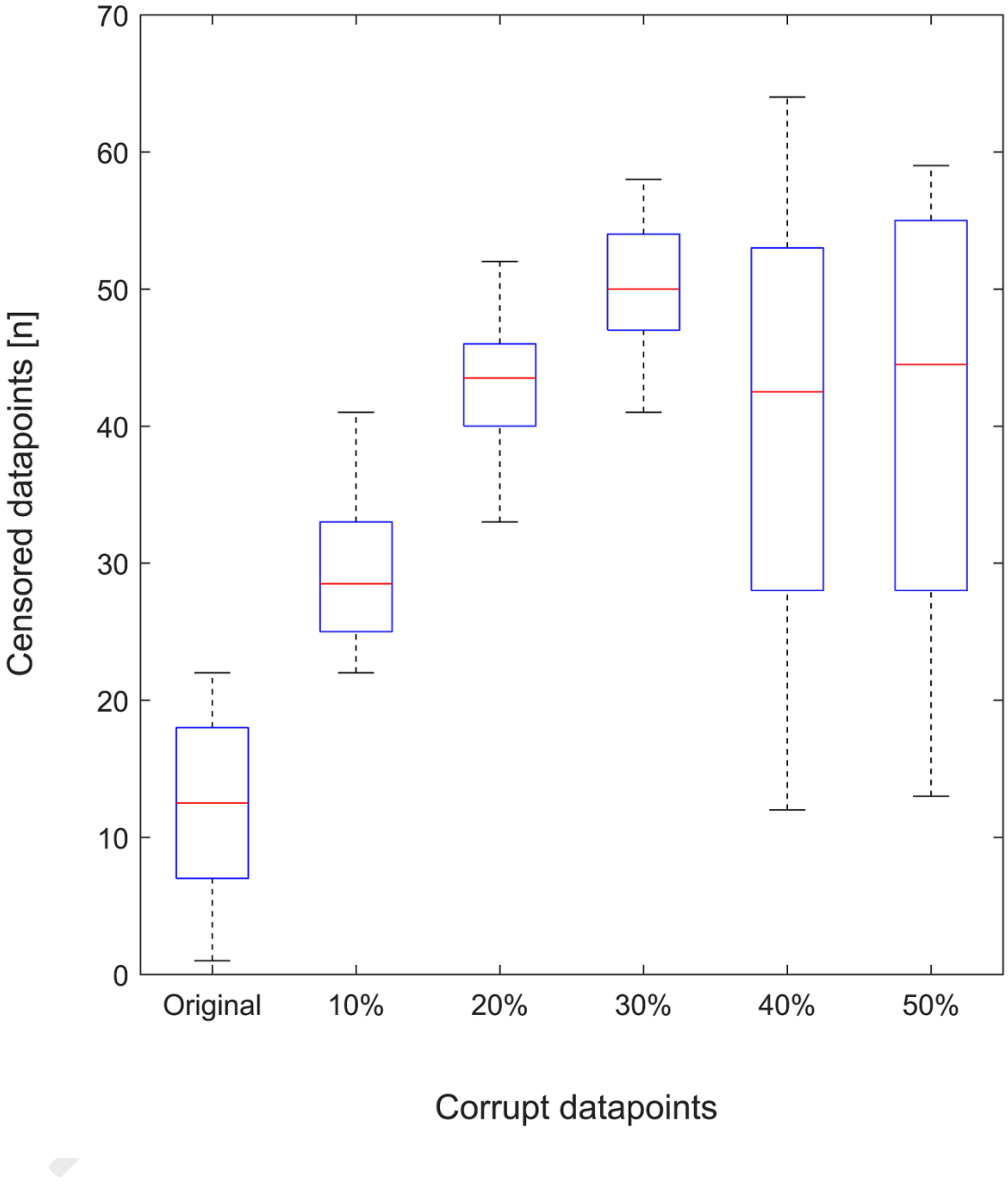


Figure 4: Non-motion data corruption versus removed datapoints in dataset 1. Note steadily increasing number of removed datapoints up to a contamination of 30% of datapoints. Thereafter, no further increase in removed datapoints can be observed.

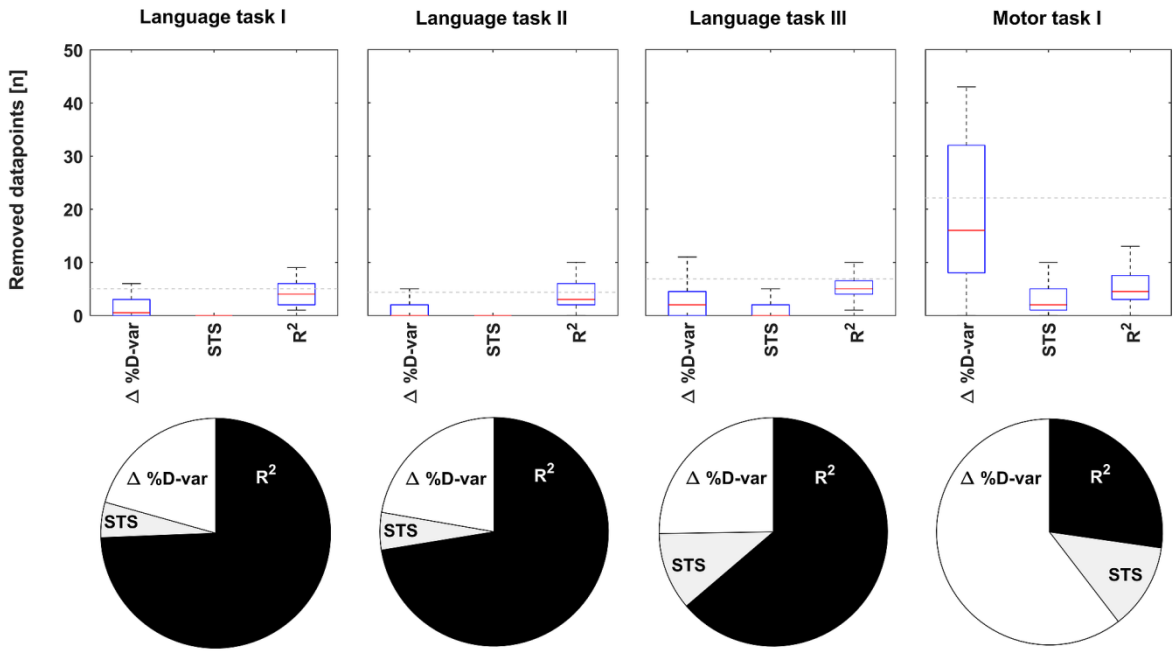


Figure 5: Algorithm performance in dataset 3: language vs. motor task. Note consistent pattern in the language tasks: only few outlying datapoints (grey line in upper panels) are identified, mainly by the R^2 parameter. Substantially more outliers are removed in the motor task (right panel), with now Δ%D-var identifying their majority.

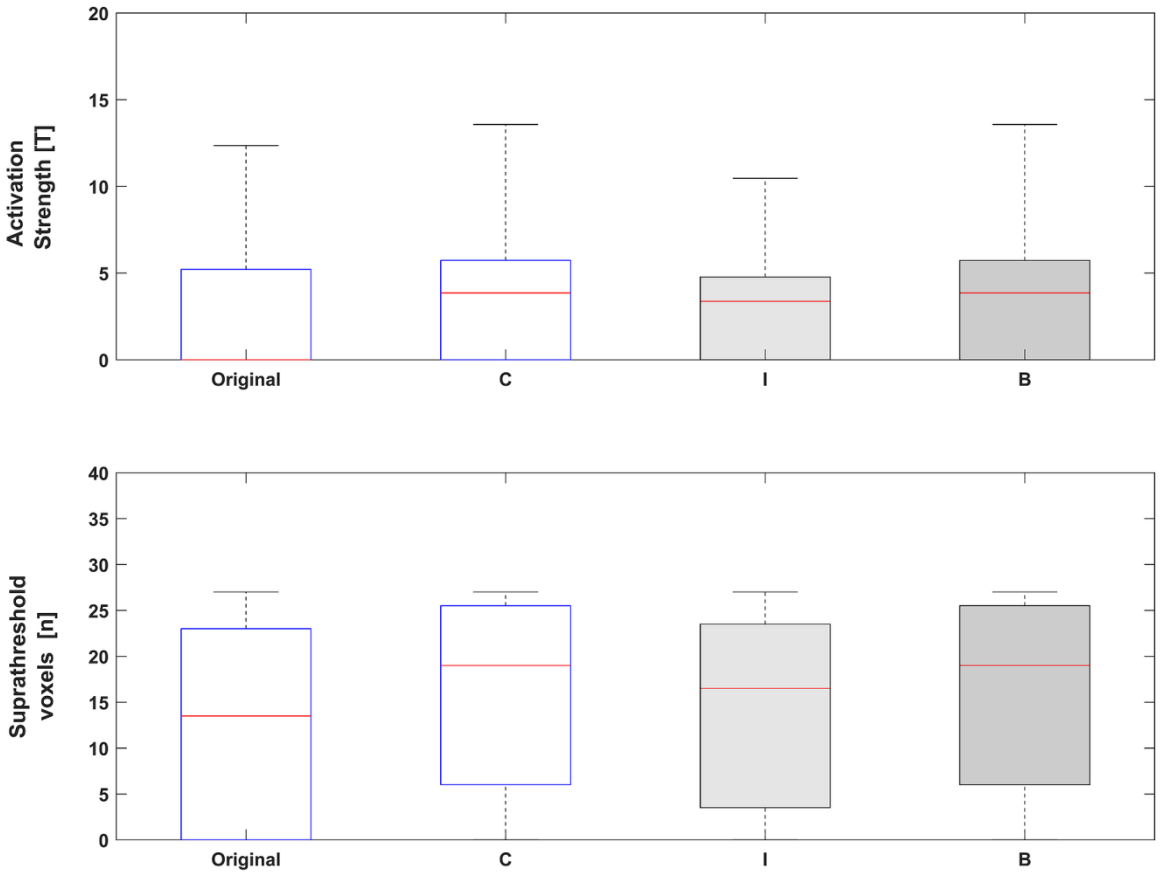


Figure 6: Comparison of activation strength as assessed by the T-value (top panel) versus the number of suprathreshold voxels (bottom panel) in dataset 2. There is a clear effect of removing outliers and a clear effect of approach (censoring [C, white] vs. interpolation [I, light gray] vs. both [B, dark grey]).

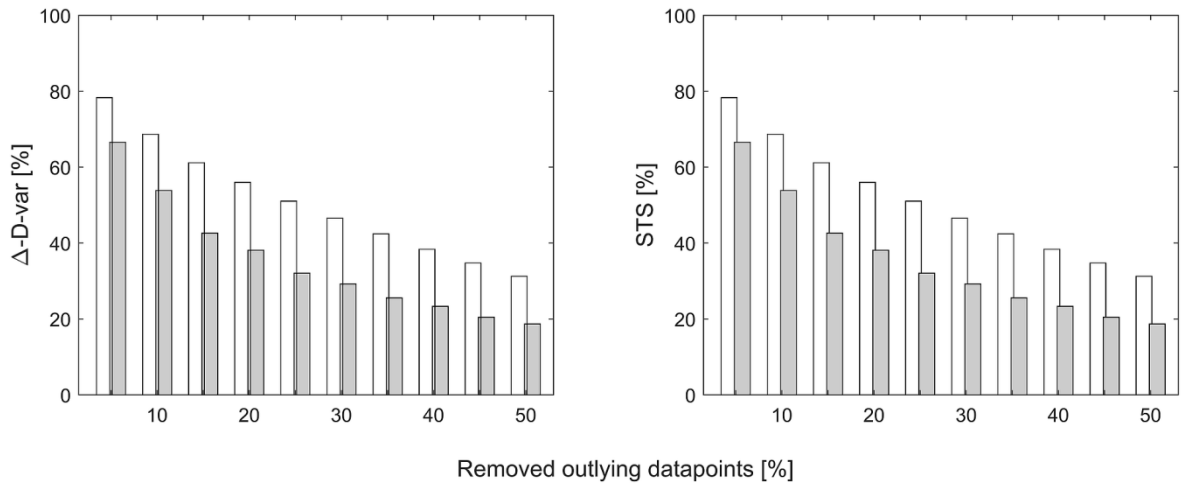


Figure 7: Comparison of Δ%D-var and subject motion (STS) as a function of progressive outlier removal in dataset 1 (white bars) and dataset 2 (gray bars), depicted as percentage w.r.t. respective original dataset. Note linear and progressive reduction of both parameters as a function of progressive outlier removal, but more pronounced reduction in the "worse" dataset 2.

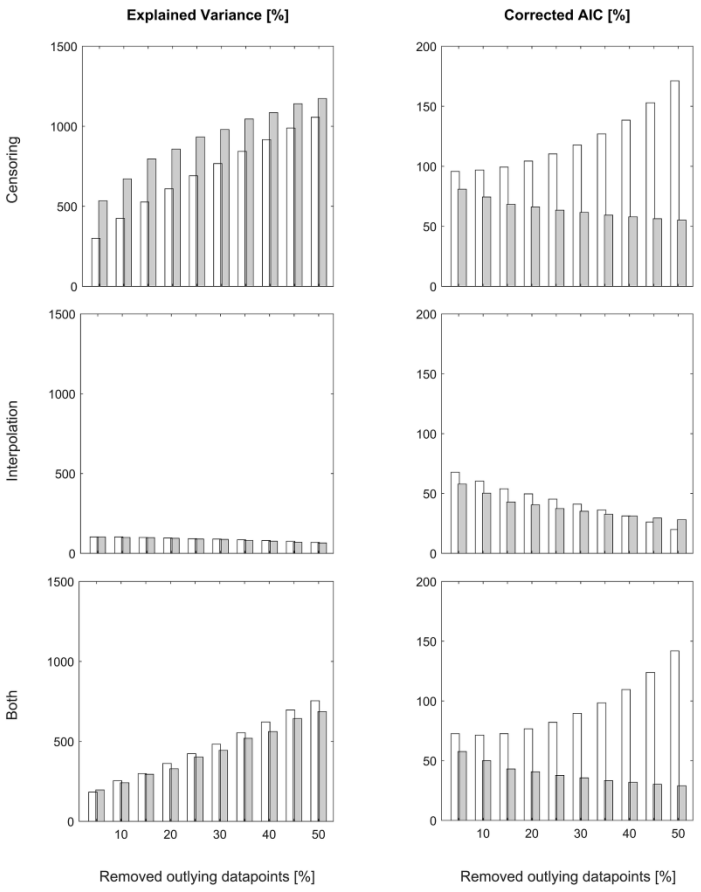


Figure 8: Assessment of explained variance (R^2, higher values indicate more explained variance) and corrected Akaike information criterion (AIC, lower values indicate better overall model fit) as a function of progressive outlier removal in dataset 1 (white bars) and 2 (gray bars), depicted as percentage w.r.t. respective original dataset. **Explained variance** (left panels) strongly increases when censoring (top left panel), but is almost unchanged when interpolating. The combined approach shows an intermediate behavior for both parameters (bottom panel). For **corrected AIC** (right panels), there is a dominant effect of data quality, with AIC_c consistently decreasing independent of approach in the bad-quality dataset 2, but showing an early increase in the good quality dataset 1 in the censoring approaches.

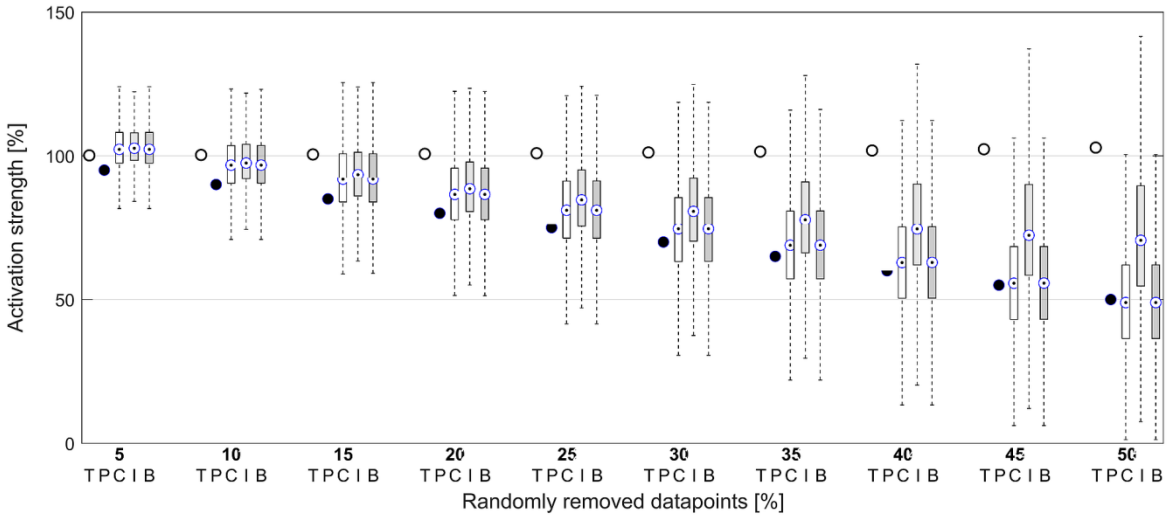
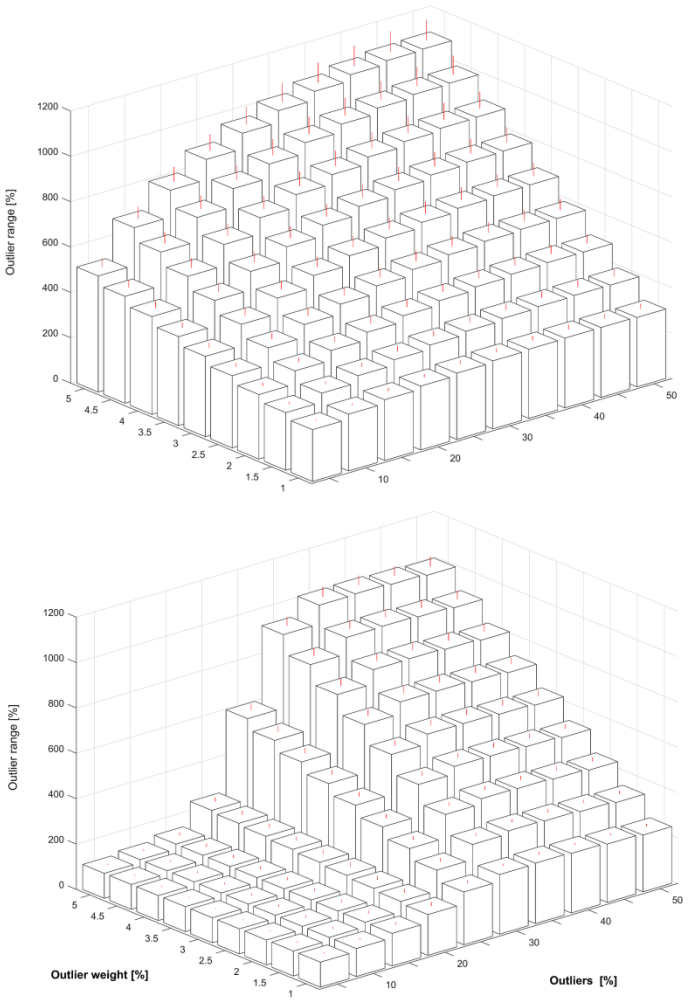
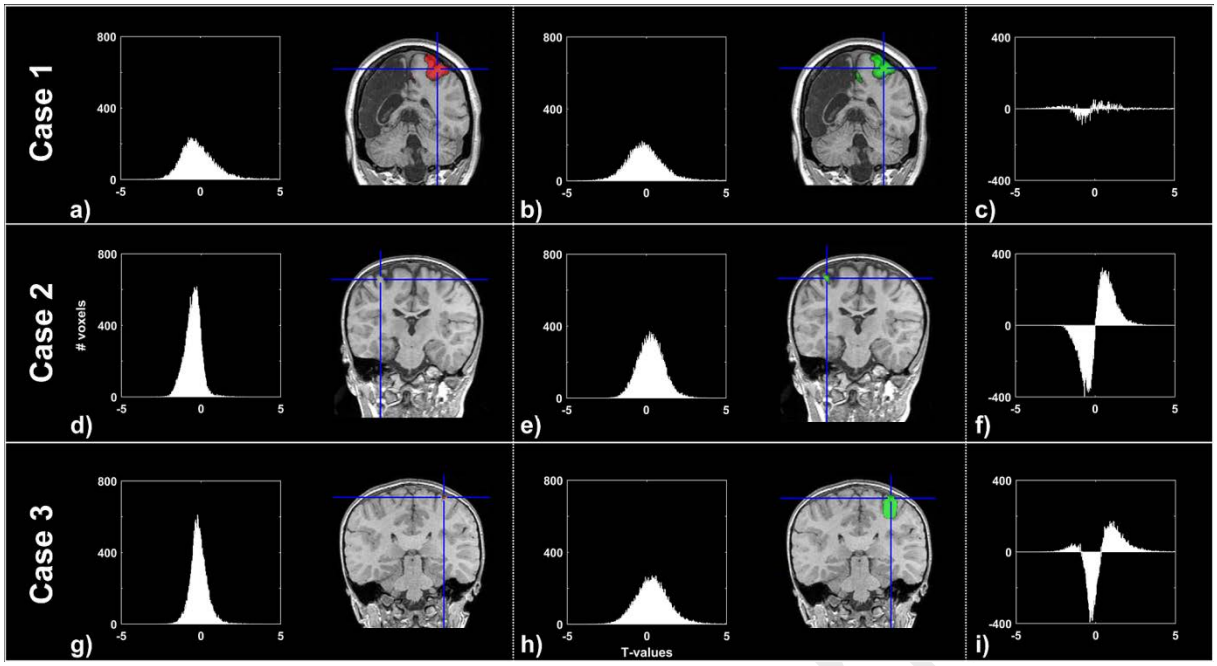


Figure 9: Effect of randomly removing datapoints in dataset 1 (38 subjects, 100 iterations at each step). Note overall clear detrimental effect of removing datapoints for all approaches (censoring [C, white], interpolation [I, light gray] and both [B, dark grey]). The effect is less pronounced for the interpolation approach. Also shown are the effects of decreasing degrees of freedom on the threshold (T, open circles) and the loss of temporal power (P, solid circles).



Supplementary Figure 1: Effect of introducing an increasing number of outlying datapoints with an increasing outlier weight on the standard deviation (upper panel) and the robust estimator used here (Tukey's criterion, lower panel). Outlier weight = 1 was defined as the value of the third quantile plus the interquartile range. Note relatively linear scaling of the standard deviation measure (upper panel) with both increasing number and increasing weight of outliers, and relative stability against both outlier number and outlier weight in the robust measure (lower panel) up to a percentage of outliers of 25%. Thereafter, the robustness of Tukey's criterion is severely impaired and biased by both number and weight of outliers. Results are based on 1000 iterations of a simulated normally distributed random time series (120 datapoints).



Supplementary Figure 2: Effect of running our algorithm in three individual subjects from dataset 2 (all $p < .001$, uncorrected). In case 1 (upper panels), an already strong activation in the right motor cortex (panel a) is not visibly altered by running our algorithm (censoring 5 images; panel b). The respective histograms (number of voxels as a function of T-value) are equally similar, as demonstrated by the difference between them (panel c). In the case of subject 2 (middle panels) with no discernible activation in the original study (panel d), a small but credible activation is seen following censoring 20 images (panel e). The difference histogram (panel f) shows a substantial shift towards higher T-values. In subject 3 (lower panels), a tiny speck of activation in the original study (panel g) is substantially stronger and larger following censoring 7 images only (panel h), again accompanied by a considerable impact on the difference histogram (panel i).