Noise removal in resting-state and task fMRI: functional connectivity and activation maps

IOP Publishing Journal Title
Journal **XX** (XXXX) XXXXXX https://doi.org/XXXX/XXXX

Noise Removal in Resting-state and Task fMRI: Functional Connectivity and Activation Maps

Bianca De Blasi[1], Lorenzo Caciagli[2,3], Silvia Francesca Storti[4], Marian Galovic[2,3,5], Matthias Koepp[2,3], Gloria Menegaz[4], Anna Barnes[6], Ilaria Boscolo Galazzo[4]

[1] Department of Medical Physics and Bioengineering, University College London, London, UK
[2] Department of Clinical and Experimental Epilepsy, Inst. Neurology, University College London, London, UK
[3]MRI Unit, Epilepsy Society, Chalfont St. Peter, Buckinghamshire, UK
[4] Department of Computer Science, University of Verona, Verona, Italy
[5] Department of Neurology, University Hospital Zurich, Zurich, Switzerland
[6] Centre for Medical Imaging, University College London, London, UK

E-mail: deblasib11@gmail.com

Received xxxxxx
Accepted for publication xxxxxx
Published xxxxxx

Abstract

Objective: BOLD-based fMRI is a widely used non-invasive tool for mapping brain function and connectivity. However, the BOLD signal is highly affected by non-neuronal contributions arising from head motion, physiological noise and scanner artefacts. Therefore, it is necessary to recover the signal of interest from the other noise-related fluctuations to obtain reliable functional connectivity results. Several pre-processing pipelines have been developed, mainly based on nuisance regression and ICA. The aim of this work was to investigate the impact of seven widely used denoising methods on both resting-state and task fMRI.
Approach: Task-fMRI can provide some ground truth given that the task administered has well established brain activations. The resulting cleaned data were compared using a wide range of measures: motion evaluation and data quality, resting-state networks and task activations, functional connectivity.
Results: Improved signal quality and reduced motion artefacts were obtained with all advanced pipelines, compared to the minimally pre-processed data. Larger variability was observed in the case of brain activation and functional connectivity estimates, with ICA-based pipelines generally achieving more reliable and accurate results.
Significance: This work provides an evidence-based reference for investigators to choose the most appropriate method for their study and data.

Keywords: pre-processing pipelines, denoising, fMRI, functional connectivity, noise

1. Introduction

Functional magnetic resonance imaging (fMRI), based on the blood-oxygenated-level dependent (BOLD) signal, is a widely used non-invasive tool for mapping brain function and functional connectivity (FC). The latter is defined as the functional coupling of different brain areas usually expressed as correlation between time series (Cole, Smith and Beckmann, 2010; Dipasquale *et al.*, 2017). Common methodologies to assess FC include seed-based analysis and independent component analysis (ICA). The first consists in creating connectivity maps by computing the correlation between the fMRI signal from pairs of regions of interest

(ROIs). The second is a data-driven technique that considers all the voxels at the same time, and uses multivariate statistical analysis to separate the data in distinct networks which are maximally independent and correlated in terms of their fluctuations over time (Fox and Greicius, 2010; Engel *et al.*, 2013).

However, the BOLD signal is generally noisy. Non-neuronal contributions to the BOLD time series arise from several factors including head motion, physiological noise (e.g. cardiac and respiratory) and scanner artefacts (e.g. thermal noise and hardware instability) (Bright, Tench and Murphy, 2017; Caballero-Gaudes and Reynolds, 2017). All these artefacts influence the fMRI signal and can lead to spurious results. Therefore, it is necessary to recover and separate the signal of interest, related to brain function, from the other noise-related fluctuations, so as to obtain reliable estimates, in terms for example of activation/deactivation and connectivity (Caballero-Gaudes and Reynolds, 2017; Dipasquale *et al.*, 2017). To achieve artefact removal, several pre-processing pipelines have been developed which are generally based on nuisance regression or ICA (Pruim, Mennes, Buitelaar, *et al.*, 2015). These pipelines result in cleaned up fMRI time series that more accurately reflect the underlying brain fluctuations of interest and reduce possible bias in post-processing analyses due to noise confounds. In nuisance regression-based pipelines, motion parameters estimated during the realignment procedure are used as regressors of no interest (Caballero-Gaudes and Reynolds, 2017), together with the average time series of white matter (WM) and cerebral spinal fluid (CSF). Often, the expansion of motion terms (e.g. derivatives, squares of derivatives) are included as additional regressors. The second group of pipelines employ ICA, a data-driven method to decompose the fMRI data into signal of interest and structured noise. The classification of these independent components (ICs) into physiological signal or noise is usually carried out manually, resulting in a time consuming and user-dependent procedure. To overcome these limitations, different authors have recently started to propose specific toolboxes for automatically identifying and classifying the ICs, such as AROMA (Pruim, Mennes, van Rooij, *et al.*, 2015) or FIX (Salimi-Khorshidi *et al.*, 2014). Regardless of the basis of the denoising pipelines adopted, their impact on BOLD-fMRI data needs to be assessed in the context of which they are being implemented.

BOLD-fMRI can indeed be acquired during the administration of a task (task-fMRI) or while the subject is resting in the scanner (resting-state fMRI [rs-fMRI]). The latter relies on the BOLD signal to probe neural activity at rest, and it has been shown as stable and reproducible across subjects (Smith *et al.*, 2009; Griffanti *et al.*, 2016). However, by acquiring the data in the absence of any task, *a-priori* information of underlying brain activation is missing (Dipasquale *et al.*, 2017). Therefore, these data are not used to localise brain areas which are activated or deactivated during a specific task, but rather to investigate brain connectivity and network organisation via FC analyses. The effect of different pre-processing methods has been widely investigated in resting-state data. Most of the previous works mainly focused on the ability of different methods to mitigate motion artefacts (Power, Schlaggar and Petersen, 2015), given their significant impact on fMRI time series and hence FC estimates (Parkes *et al.*, 2018). Several benchmarks have been selected to address the impact of motion on the fMRI signal. These include Derivative of root mean square VARiance over voxelS (DVARS), framewise displacement (FD)-DVARS correlations (Muschelli *et al.*, 2014) and QC-FC (quality control/FC) correlations (Power, Schlaggar and Petersen, 2015; Ciric *et al.*, 2017). In general, regression methods including the expansion of motion terms substantially mitigate motion artefacts (Caballero-Gaudes and Reynolds, 2017). Global signal regression (GSR) has been shown to potentially improve motion correction (Satterthwaite *et al.*, 2013; Lydon-Staley *et al.*, 2019). However, GSR itself is still a controversial pre-processing step as it involves regressing an average signal computed across the entire brain (including grey matter [GM], WM and CSF) (Parkes *et al.*, 2018) which might include widespread strong neural fluctuations, removed by the regression. Moreover, GSR tends to introduce negative correlations (anti-correlations) between brain regions which are difficult to interpret and distant-dependence artefacts (Griffanti *et al.*, 2015; Caballero-Gaudes and Reynolds, 2017; Ciric *et al.*, 2017). Therefore, many argue that this removal should be avoided in classical connectivity studies (Cole, Smith and Beckmann, 2010; Griffanti *et al.*, 2015).

In addition to addressing the impact of pre-processing strategies on motion correction, it is also important to assess how different cleaning methods affect the BOLD signal (i.e. using measures like temporal signal to noise ratio [tSNR] or standard deviation [SD]) and derived FC estimates. Pruim et al. (Pruim, Mennes, Buitelaar, *et al.*, 2015) evaluated several regression and ICA-based pipelines by considering their impact on BOLD signal and related FC measures. The authors reported the inability of nuisance regression strategies to fully mitigate the impact of motion on connectivity. On the other hand, ICA-based strategies successfully reduced the effect of head motion and led to higher resting-state networks (RSNs) reproducibility. Similarly, Ciric et al. (Ciric *et al.*, 2017) investigated fourteen pipelines using four benchmark measures including motion related measures and network modularity. The authors emphasised the heterogeneity in terms of pipeline performance across benchmarks, and the importance of identifying the optimal pipeline for a given study.

In the case of rs-fMRI data, there is a lack of ground truth for the BOLD fMRI signal of interest and noise due to multiple sources contributing to the overall signal and to the

unconstrained nature of the paradigm (i.e. the subject is resting in the scanner) (Griffanti *et al.*, 2015). This bottleneck can be naturally overcome by using task-fMRI. Being based on a hypothesis of brain activation/deactivation related to external stimuli over time, a reference is available to distinguish the signal of interest (evoked by the task) from noise, therefore informing the assessment of pipeline performance (Power, Schlaggar and Petersen, 2015). Task-fMRI is affected by similar noise confounds as rs-fMRI. The main advantage of bringing task-fMRI into the loop is that knowing *a-priori* which areas of the brain are active during a give task, provides a convenient reference to identify which pre-processing pipeline may be best suited to reproduce the related activation patterns (Glasser *et al.*, 2018). This can in turn help investigating denoising methods in the context of rs-fMRI, which lacks *a-priori* information about activation or deactivation, but presents a similar underlying signal as task-fMRI.

Unlike rs-fMRI, this topic has been scarcely investigated in the context of task-fMRI. Tierney et al. (Tierney *et al.*, 2016) validated a new pre-processing method (FIACH) via comparison with five other pipelines mainly based on nuisance regression methods, using a language fMRI paradigm in controls. Their pipeline led to more reliable activations in areas expected to be active during the administered language task. Glasser et al. (Glasser *et al.*, 2018) applied temporal ICA to both task and resting-state fMRI datasets, and compared that to GSR, showing usefulness of their method in (a) separating global noise from the global neural signal and (b) selectively removing noise, in both conditions. However, their method would only be suitable for neuroimaging acquisitions similar to those used in the Human Connectome Project, which entail a large number of time points, and high temporal sampling (Glasser *et al.*, 2018). These type of data may not be available in the majority of clinical centres. Finally, Mayer et al. (Mayer *et al.*, 2019) examined the effect of denoising methods (nuisance regression and ICA-based pipelines) on event-related and block-design task-fMRI data. Specifically, they assessed the percentage signal changes in active brain areas versus noisy areas after different denoising strategies and did not focus on connectivity aspects. They showed that FIX and AROMA tend to remove task-related activity, as well as noise when compared to regression of 24 motion parameters. The authors also emphasised the lack of a single appropriate denoising method for all fMRI designs.

In this work, we aimed to investigate the impact of different pre-processing pipelines on both resting-state and task fMRI data. We analysed data from two different groups of healthy controls. The resulting processed data were compared using a wide range of benchmark measures. In particular, we assessed (1) changes in the BOLD signal as expressed by measures of motion evaluation (DVARS) and data quality (tSNR, lost temporal degrees of freedom [tDoF]), (2) modulations of RSNs and task activations, and (3) modulations of FC estimates. These measures have been commonly used in the context of rs-fMRI but are still largely unexplored in task-fMRI.

2. Methods

2.1 Population

Rs-fMRI data were collected for twenty healthy controls (11 males, 38.5 ± 10 y) as part of a larger neuroimaging study approved by the London - Stanmore Research Ethics Committee. Previously collected task-fMRI data were selected from a different group of twenty controls, age- and gender-matched to the rs-fMRI group (11 males, 38 ± 9 y). In this task study, all participants were fluent English speakers and able to understand the instructions for performing a verbal fluency (VF) fMRI task. Recruitment for this study received approval by the London South-East Research Ethics Committee and by the UCL/UCLH Joint Research Office. Written informed consent was obtained for all participants.

2.2 Image Acquisition

The rs-fMRI acquisition was carried out on a 3T Siemens mMR Biograph (Siemens, Erlangen, Germany) PET/MRI scanner equipped with a 16-channel head and neck coil. The subjects were instructed to stay still and relaxed, and to close their eyes without falling asleep. Acquiring rs-fMRI with eyes close was conducted to match the clinical protocols for patients generally scanned in our MRI Unit (mainly epilepsy). Rs-fMRI data were acquired using a 2D echo-planar imaging (EPI) sequence with the following parameters: TR/TE = 2020/30 ms, flip angle = 70°, voxel size = 3 x 3 x 4 mm^3, 36 slices, 260 volumes. High resolution 3D T1-weighted (T1w) anatomical images were also acquired, using an MPRAGE sequence: TR/TE = 2000/2.92 ms, voxel size = 1.1 x 1.1 x 1.1 mm^3, 208 sagittal slices.

The task-fMRI acquisition was carried out on a 3-T Excite HDx scanner (General Electric, Milwaukee, WI, USA), using a standard 8-channel receive head coil and a 2D EPI sequence with the following parameters: TR/TE = 2500/25 ms, flip angle = 70°, voxel size = 3.75 x 3.75 x 2.5 mm^3, 50 slices, 120 volumes. During the acquisition, the participants performed a covert VF task (Wandschneider *et al.*, 2017). The paradigm, lasting 5 min in total, consisted of five 30 s blocks of task alternated with five 30 s resting blocks (crosshair fixation). Participants were instructed to covertly generate words starting with a visually presented letter (A, D, E, S, W). 3D T1w anatomical images were also acquired using a FSPGR sequence: TR/TE = 7.2/2.8 ms, voxel size = 1.1 x 1.1 x 1.1 mm^3, 196 sagittal slices.

2.3 Pre-processing Pipelines

All the pre-processing methods considered in this study were run on both resting-state and task data, using the FSL 5.0.9 software (https://fsl.fmrib.ox.ac.uk/fsl/fslwiki/) and are illustrated hereafter. *Fig.1* summarises the seven pre-processing pipelines and the corresponding analyses used to evaluate these denoising methods. Min was considered as the baseline model (Mayer *et al.*, 2019). Res6 and Res24 were chosen to represent different levels of head motion regression. The former is commonly employed in task-fMRI analyses, but never formally compared with other denoising methods. FIACH was included, as it has been developed specifically for denoising of task-fMRI data. Finally, AROMA, FIX and FIXMC were considered as commonly implemented ICA methods. In the remaining of the manuscript, we will refer to nuisance regression pipelines (Res6, Res24 and FIACH) as regression-based pipelines, while AROMA, FIX and FIXMC will be part of the ICA-based methods. We are aware that ICA-based methods also include regression of the identified noise components, but we considered this distinction, which is also commonly used in literature, given the different modelling of the unwanted signal.

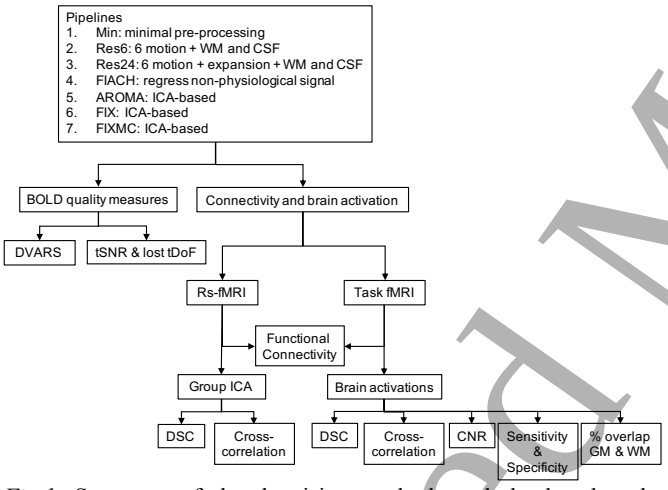

Fig.1: Summary of the denoising methods and the benchmarks adopted in this study. Seven pipelines were considered which were evaluated in terms of BOLD quality measures, connectivity and brain activations. The latter were derived and evaluated in a different way for the resting-state and task fMRI, as reported in the literature. WM: white matter, CSF: cerebral spinal fluid, BOLD: blood-oxygen-level-dependent, DVARS: derivative of root mean square variance over voxels, tSNR: temporal signal-to-noise ratio, tDoF: temporal degrees of freedom, DSC: dice similarity coefficient, CNR: contrast-to-noise ratio.

2.3.1 Minimal Pre-processing - Min.

Both fMRI datasets were minimally pre-processed (Min) using FSL FEAT (https://fsl.fmrib.ox.ac.uk/fsl/fslwiki/FEAT) with the following steps: realignment (MCFLIRT), slice timing correction, high-pass filtering (0.01 Hz cut-off), brain extraction (BET), spatial smoothing with a 6-mm FWHM kernel, co-registration to structural MRI data (FLIRT, 6 degrees of freedom with BBR cost function) and spatial normalisation to 2-mm MNI standard space (FNIRT, non-linear registration).

2.3.2 Regression WM, CSF and 6 Motion Parameters – Res6.

This pre-processing pipeline (Res6) involved regressing out WM, CSF and 6 motion parameters from the minimally pre-processed data. For each subject, the normalized T1w scans were segmented, leading to WM and CSF probability maps that were thresholded at 0.9, to strictly retain only WM and CSF voxels, and binarised. These masks were then used to extract the WM and CSF average time series. WM and CSF signals, together with the 6 motion parameters (estimated in the realignment step and high-pass filtered with cut-off 0.01 Hz, to match the data), were regressed out from the minimally pre-processed data (*fsl_glm*). The resulting residuals were used for further analyses and pipeline comparison.

2.3.3 Regression WM, CSF and 24 Motion Parameters – Res24.

This pre-processing method (Res24) was based on regression of the average WM and CSF time series but entailed removal of 24 motion parameters instead of the previously used 6. The 24 regressors included the original 6 motion parameters, their square, their derivatives and their derivatives squared. High-pass filtering with cut-off 0.01 Hz was applied to the regressors to match the data. These regressors were removed from the minimally pre-processed data and the residuals retained for the subsequent analyses.

2.3.4 FIACH.

This approach was implemented as in Tierney et al. (Tierney *et al.*, 2016). All the raw data were motion corrected (MCFLIRT) and the pipeline FIACH with all the default parameters was subsequently applied (http://www.homepages.ucl.ac.uk/~ucjttie/FIACH.html). The output data (filtered data) were then processed using FEAT, including: slice timing correction, BET, high-pass filtering, spatial smoothing, co-registration to structural data and spatial normalisation to 2-mm MNI standard space (using the same parameters for Min). Finally, the 6 noise regressors estimated by FIACH were regressed out from the data (*fsl_glm*) and the residuals were kept for further analyses.

2.3.5 AROMA.

AROMA is an ICA-based pre-processing pipeline implemented using the corresponding FSL toolbox (https://github.com/maartenmennes/ICA-AROMA), specifically devised to clean each participant's scans from

motion confounds. FSL FEAT minimal pre-processing was initially applied to the data without temporal filtering as explecitely suggested by the developers. The resulting data were then input in the ICA-AROMA toolbox to carry out MELODIC (automatic dimensionality estimation for the optimal number of components) and to automatically identify and remove motion artefacts (non-aggressive option) (Mayer *et al.*, 2019). The cleaned data were then filtered (high-pass, 0.01 Hz), co-registered to structural data (FLIRT, 6 degrees of freedom with BBR cost function) and spatially normalised to 2-mm MNI standard space (FNIRT, non-linear registration).

2.3.6 FIX and FIXMC.

Two additional ICA-based approaches were implemented in FSL using the FIX toolbox (https://fsl.fmrib.ox.ac.uk/fsl/fslwiki/FIX) to clean each participant's scans from various and heterogeneous types of structured noise. The minimal pre-processing was initially applied to the data, and each pre-processed dataset was then decomposed using MELODIC (automatic dimensionality estimation). FIX is based on a hierarchical classifier and it therefore requires a 'training dataset' which closely matches the data under investigation. In this work, we did not use the training input data provided by the FIX package, as these did not match our datasets in terms of TR and resolution. Two specific classifiers were trained, one for resting-state and the other for task fMRI, by classifying ICs from 10 subjects (for each dataset). Each IC was manually classified and labeled by an expert rater as 'noise' or 'signal', according to its spatial distribution (i.e. network of interest or task activation), the temporal power spectrum (i.e. covering frequencies of interest, 0.01-0.1 Hz) and the time series. Using these labels, the classifier was trained, and a summary training file was created. In addition, a leave-one-out test was run to choose the threshold which balances true positive rate (TPR) and true negative rate (TNR). A threshold of 20 was chosen, guaranteeing a mean TPR of 95.7 (rs-fMRI), 88.1 (task-fMRI) and a mean TNR of 91.6 (rs-fMRI), 82.2 (task-fMRI) (Mayer *et al.*, 2019). The trained classifiers were applied to all subjects' data (resting-state and task, separately). Components automatically classified as artefacts were removed from the data using the non-aggressive option (Griffanti *et al.*, 2014; Mayer *et al.*, 2019). FIX was applied without (FIX) and with motion regression (FIXMC). In the latter case, the full variance of 24 motion parameters was regressed out. All cleaned data were finally spatially normalised to 2-mm MNI standard space.

2.4 Pipeline Performance Metrics

2.4.1 Motion Evaluation – DVARS.

To quantify the ability of each pipeline to remove motion artefacts from the data (*Fig.1*), DVARS (root mean square intensity difference of volumes N and N+1) was calculated for every subject. The mean DVARS values were then computed for each participant and pipeline.

2.4.2 Temporal Signal-to-Noise Ratio and Loss of Temporal Degrees of Freedom.

TSNR can be used to determine the SNR of fMRI time series by considering the mean signal over time. This measure can be considered an indication of pipeline performance, as data pre-processing should decrease signal fluctuations around the mean (*Fig.1*) (Griffanti *et al.*, 2014). For each pipeline and corresponding cleaned data, tSNR was voxel-wise computed for each subject by dividing the mean signal over time by the SD over time (leading to a tSNR map). For the resting-state dataset, the mean tSNR value in the GM was computed, using the tissue segmentations previously estimated at the individual level in MNI space (probability values >= 0.9). For the task dataset, the mean tSNR was computed in eight task-related ROIs. Six activations ROIs were derived from the *Neurosynth* VF template (http://neurosynth.org/analyses/terms/verbal%20fluency/, FDR-corrected p-value < 0.01), obtained from a meta-analysis of studies which employed VF paradigms. In addition, deactivations were also considered (two ROIs), including the deactivated areas over the DMN derived from a well-known RSNs template (Smith *et al.*, 2009). Overall, the task-related activation ROIs were: left inferior frontal gyrus (L_IFG), right inferior frontal gyrus (R_IFG), supplementary motor area (SMA), left parietal (L_Par), left temporal lobe (L_TL), and subcortical ROIs (thalamus plus putamen). In terms of deactivations, the medial frontal (paracingulate gyrus) and posterior (precuneous cortex) areas of the default mode network were chosen (DMN_Front and DMN_Post, respectively). The mean tSNR value in these ROIs was finally computed for each subject and pipeline.

The loss of tDoF was used as an additional measure in conjunction with the tSNR to better assess the impact of the pre-processing methods on the statistical power and reliability of the different imaging estimates (*Fig.1*) (Pruim, Mennes, Buitelaar, *et al.*, 2015; Pruim, Mennes, van Rooij, *et al.*, 2015). The total number of fMRI volumes was considered as the total number of available tDoF (i.e. 260 for the resting-state, and 120 for the task data). Each regressor or IC removed from the data was considered as a lost tDoF. The total loss of tDoF was expressed, for each pre-processing method, as a percentage of the initial number of available tDoF. Of note, in the regression-based methods, where the number of regressors was fixed in the model (Res6: 8 regressors; Res24: 26 regressors and FIACH: 6 regressors), a constant value of lost tDoF was expected across subjects.

2.4.3 Statistical Analysis

A one-way repeated-measure analysis of variance (ANOVA) was performed on the DVARS values, along with tSNR for rs-fMRI only (GM values) to test for significant differences across the different pipelines. A *post-hoc* paired sample two-tailed t-test was then applied (p-value < 0.05), which was corrected for multiple comparison using Bonferroni correction. In the case of tSNR in task-fMRI, a two-way repeated measures ANOVA was carried out (factor1 = pipelines, factor2 = ROIs). This was followed by *post-hoc* pairwise comparisons using *multcompare* in MATLAB (p-values < 0.05), which was Bonferroni-corrected for multiple comparison.

2.5 Connectivity Analyses and Brain Activations

2.5.1 Rs-fMRI: Group ICA.

For the resting-state dataset, we tested the ability of the different pre-processing methods to recover RSNs using ICA (*Fig.1*). A group-based ICA (MELODIC) was run for every pipeline, setting the number of ICs to 30 (Griffanti *et al.*, 2014). The most common networks were retained for further analyses. For each pipeline and each RSN of interest, the spatial cross-correlation (CC) and overlap between the group IC map and the corresponding BOLD template (both thresholded at z > 3) were evaluated (Smith *et al.*, 2009). CC was computed using FSL function (*fslcc*), while the spatial overlap was assessed using the Dice Similarity Coefficient (DSC). This index quantifies the cardinality of the intersection of the thresholded maps divided by the average of the cardinality of each thresholded map (Bowring, Maumet and Nichols, 2019). Our assumption was that a higher similarity and overlap between the group ICA maps and the corresponding template would point to higher accuracy of the denoising pipeline in identifying the true signal (Smith *et al.*, 2009; Griffanti *et al.*, 2014).

2.5.2 Task-fMRI: Group Activations.

For the task-fMRI data, brain activations (*Fig.1*) were computed using a general linear model (GLM). This was carried out at the single subject level (first-level analysis) and then at the group level (second-level analysis), independently for each pipeline. At the subject level, the task was modelled by convolving the vector of block onset with a canonical hemodynamic response function (double gamma HRF) to create the regressor of interest, which was temporal filtered to match the data under investigation. Contrast images were created for every subject and pipeline for task related activations. At the second level, we explored activations maps during the VF task for each pipeline using a one-sample t-test. Statistic images were thresholded using clusters determined by z > 2.3 and a corrected cluster significant threshold of p-value = 0.05.

Group-level activation maps were initially compared across pipelines in terms of *i)* number of voxels, *ii)* statistical significance and *iii)* location of the clusters. In order to quantify how much of the activations overlapped with the GM, we computed the percentage of voxels which fell into the different tissue types (GM and WM) for each group activation pipeline. From this analysis, we expected (1) to have most of the activations localised to GM, for all the pipelines and that (2) more effective pipelines would lead to higher overlap between activations and GM, with lower involvement of WM voxels. For this purpose, we used the FSL tissue prior maps, thresholded at 100 and binarised, together with the thresholded and binarised group activation maps. We then computed the percentage of each group map included in either GM and WM tissue maps.

In addition, we computed several measures to quantify the ability of each pipeline to recover the true signal (related to task activation), as well as their accuracy in identifying the brain areas expected to be active in a VF paradigm.

Contrast-to-noise ratio (CNR) was defined at the single-subject level using the time course of the voxel with highest z-score as representative activation signal (Shen and Q. Duong, 2011). This was shifted by two time points to account for the HRF (Liang *et al.*, 2013), and the following equation was used:

$$CNR = \frac{mean(Task) - mean(Rest)}{std(Rest)}$$

where the numerator represents the difference between the mean value of the signal across all activation and baseline conditions, respectively, while the denominator is the standard deviation across the baseline periods. A one-way ANOVA for repeated measures was performed on CNR values to test for significant differences across the seven pipelines. *Post-hoc* paired sample two-tailed t-tests were applied (p-value < 0.05, Bonferroni-corrected).

In terms of brain localisation, we computed the spatial correlation (*fslcc*) and overlap (DSC) of each group-level activation map with respect to the *Neurosynth* VF template. Since the *Neurosynth* VF template was only available with FDR correction (p-value < 0.01) we thresholded the corresponding group activations maps with the same threshold to allow appropriate comparison. This threshold was therefore kept for all the comparative analyses involving the use of the *Neurosynth* VF template.

Sensitivity and specificity were also computed, using the VF template as reference. In particular, sensitivity was defined as the ratio of the number of overlapping voxels between one pipeline and the reference ($Voverlap$) over the number of voxels in the reference map ($Vref$). Specificity was defined as the ratio of the number of overlapping voxels between one pipeline and the reference ($Voverlap$) over the number of

voxels in the pipeline activation map ($Vpipeline$) (Storti *et al.*, 2018):

$$Sensitivity = \frac{Voverlap}{Vref} \; ; \; Specificity = \frac{Voverlap}{Vpipeline}$$

2.5.3 Resting-state and Task fMRI: Functional Connectivity Metrics.

In order to assess the impact of the different cleaning methods on ROI-to-ROI FC, we used the Schaefer functional atlas, comprising 100 parcels derived from 17 RSNs (Schaefer *et al.*, 2018), and extracted the average time series in each ROI for each subject and pipeline (in both datasets, *Fig.1*). For each pipeline, a symmetric connectivity matrix was derived by computing the Pearson correlation coefficient for each pair of nodes, at the single-subject level. These connectivity matrices were visually compared, and the mean and SD matrices were computed across subjects. The 2D spatial correlation between each pair of mean matrices was calculated to assess the extent of similarity of connectivity patterns across pipelines. Additionally, we computed for each pipeline the coefficient of variation (CV), defined as the percentage ratio of the group SD and the mean. We expected the CV to be decreased for the more effective pipelines owing to them removing spurious differences in the control population under investigation. Indeed, effective cleaning methods should increase network similarity between subjects in a homogenous group of healthy controls (Griffanti *et al.*, 2014). We finally assessed the number of negative correlations found in each matrix at the single-subject level.

3. Results

3.1 Motion Evaluation – DVARS.

Fig.2 reports the distribution of mean DVARS values across subjects together with the results of the statistical comparison, for each pipeline and dataset.

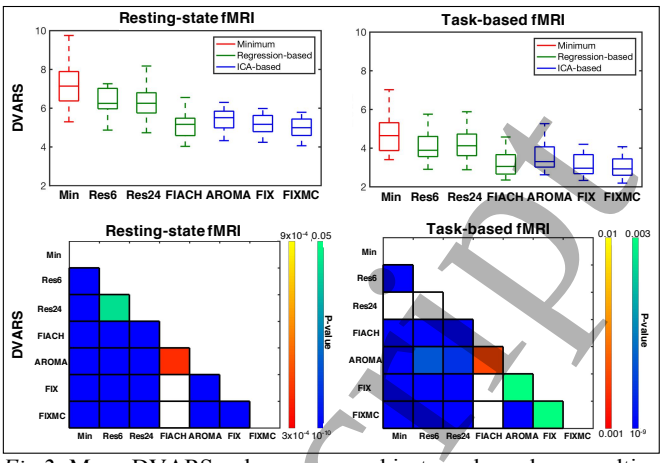

Fig.2: Mean DVARS values across subjects and p-values resulting from the statistical comparisons for the seven pre-processing pipelines. Top: Distribution of mean DVARS values across subjects for every pipeline in the resting-state (left) and task (right) datasets. Bottom: Graphical representation of the p-values resulting from the statistical comparison of mean DVARS values for every pipeline in both datasets. The comparisons which were not statistically significant are reported in white. Each pipeline in the row is compared to each coloumn (i.e. reading row first and then column). The two colour bars report p-values corresponding to positive (hot colour scale) and negative (cold colour scale) t-values.

A similar trend was observed in resting-state and task datasets. One subject appeared as an outlier with regards to DVARS metric in rs-fMRI and was thus discarded from the all the corresponding statistical analyses. Of note, a point was considered outlier if greater than $q_3 + w \times (q_3 - q_1)$ or less than $q_1 - w \times (q_3 - q_1)$, where q_1 and q_3 are the 25th and 75th percentiles of the sample data and $w \times (q_3 - q_1)$ refers to 1.5xIQR (inter quartile range) corresponding to approximately +/–2.7σ. Statistical comparisons revealed significant differences for DVARS values across the seven pipelines, in both resting-state and task fMRI (DVARS: $F_{(6,132)} = 18.05$, p-value < 0.01 in resting-state; $F_{(6,126)} = 12.74$, p-value < 0.01 in task). P*ost-hoc* t-tests between pairs of pipelines revealed significantly lower DVARS values (p-value < 0.05, Bonferroni-corrected) for all the advanced pipelines compared to Min, except the case of Min vs Res24 for DVARS in task which did not achieve the statistical significance. This result points towards a more pronounced removal of motion artefacts in the advanced denoising methods, when compared to Min. All the other pairwise comparisons across advanced pipelines were significant (p-value < 0.05, Bonferroni-corrected) for both datasets, except for few cases reported in white in *Fig.2*. Among the advanced pre-processing methods, FIACH and FIXMC achieved the lowest DVARS values in both resting-state and task fMRI. Overall, regression-based and ICA-based pipelines showed a similar trend for both datasets.

3.2 tSNR and Loss of tDoF.

The tSNR was computed for each subject and pipeline as a measure of signal variation. *Fig.3* reports the distribution across subjects of the mean tSNR values in GM (rs-fMRI) and in a representative active area for task-fMRI (L_IFG) for the seven pipelines. Of note, similar patterns were observed for all the other ROIs for task data, *Fig.S1*.

Increased tSNR was observed for all advanced pipelines when compared to Min, for both resting-state and task data. Similar patterns were observed in both datasets, with increased tSNR values and variability in the task dataset. Of note, the same subject who was an outlier for DVARS also appeared as an outlier for the tSNR metric in rs-fMRI GM and was thus eliminated from all the corresponding statistical analyses. The tSNR values were significantly different across the seven pipelines in both resting-state and task fMRI ($F_{(6,126)} = 11.7$, p-value < 0.01 for GM in resting-state; $F_{(42,1064)} = 3.84$, p-value < 0.01 for L_IFG in task). *Post-hoc* t-tests between pairs of pipelines revealed significantly increased tSNR values (p-value < 0.05, Bonferroni-corrected) for all advanced pipelines when compared to Min (*Fig.3*). Other pairwise t-tests resulted as significant (p-value < 0.05, Bonferroni-corrected) for both datasets, as shown in *Fig.3*. The *post-hoc* t-tests for all the remainder ROIs in the task-fMRI analysis are reported in *Fig.S2*.

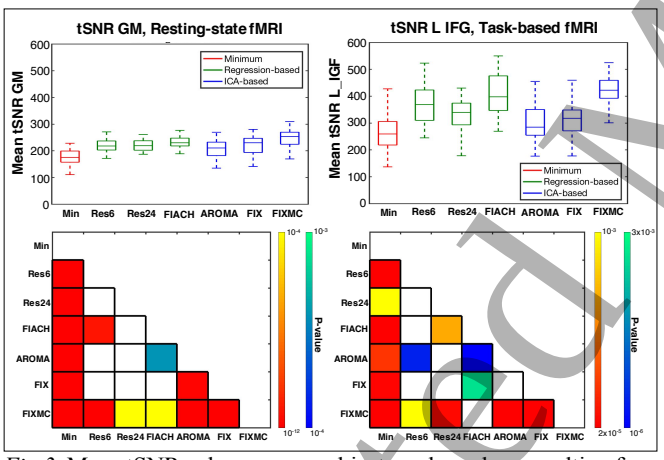

Fig.3: Mean tSNR values across subjects and p-values resulting from the statistical comparisons for the seven pre-processing pipelines. Top: In the case of rs-fMRI, the distribution of the mean tSNR values in GM is reported while, for task-fMRI, we report mean values across a representative area known to be active during a VF paradigm (L_IFG). Bottom: Graphical representation of the p-values resulting from the statistical comparison of mean tSNR values, for every pipeline in resting-state and task fMRI. The comparisons which were not statistically significant are reported in white. Each pipeline in the row is compared to the each coloumn (ref to *Fig.2* for colour bars details).

Being aware of the limitations of tSNR as a quality check metric, we also computed the loss of tDoF to more thoroughly assess the impact of the pre-processing methods on the statistical power and reliability of the different imaging estimates (Pruim, Mennes, Buitelaar, *et al.*, 2015; Pruim, Mennes, van Rooij, *et al.*, 2015). *Fig.4* reports the distribution of lost tDoF for each pipeline in the two datasets, expressed as a percentage of the total tDoF available. As expected, the ICA-based pipelines tended to have a higher loss of tDoF when compared to regression-based methods, with FIXMC having the highest number, as it includes regression of ICs together with the 24 motion parameters.

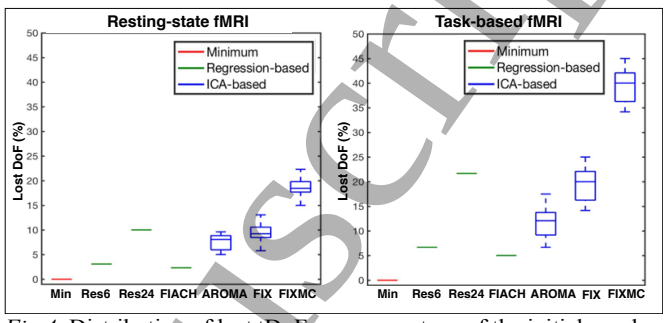

Fig.4: Distribution of lost tDoF as a percentage of the initial number of tDoF. A constant value of lost tDoF was observed for the regression-based methods (fixed number of regressors). For the ICA-based pipelines, the number of lost tDoF was variable and depended on the number of noisy ICs removed from the individual data.

3.3 Rs-fMRI: Group ICA.

A group ICA was carried out to assess how the different pre-processing methods would recover the RSNs. The main RSNs extracted included: visual medial (VISmed), visual occipital (VISocc), visual lateral (VISlat), default mode network (DMN), cerebellum (CER), motor (SMN), auditory (AUD), executive (EXE), fronto-parietal right (FPr) and left (FPl), dorsal attention (DA) and temporal (TEMP), in agreement with previous literature templates (Smith *et al.*, 2009). *Fig.5* reports the resulting maps for DMN, AUD, VISmed and SMN networks, as indicative examples.

When the different RSNs were visually compared, we found that, in the case of the DMN, the ICA-based and FIACH pipelines more accurately recovered the extension of posterior cingulate cortex as compared to Min. All the pipelines accurately recovered the AUD and VISmed networks, however Min, FIX and FIXMC showed noisier maps. The SMN network was recovered in a different pattern by regression-based vs ICA-based pipelines. Indeed, the latter tended to recover the motor network as three separate components (cingulate gyrus and left and right postcentral gyri) while the regression-based methods showed a unique cluster over the cortical motor areas.

Regarding the other networks, the VISocc network was better recovered by the more advanced pipelines, while Min detected less extensive and significant clusters. Regression-based pipelines better recovered the VISlat network, while ICA-based methods did a better job for the CER network. The EXE network was clearly highlighted by all the pipelines with

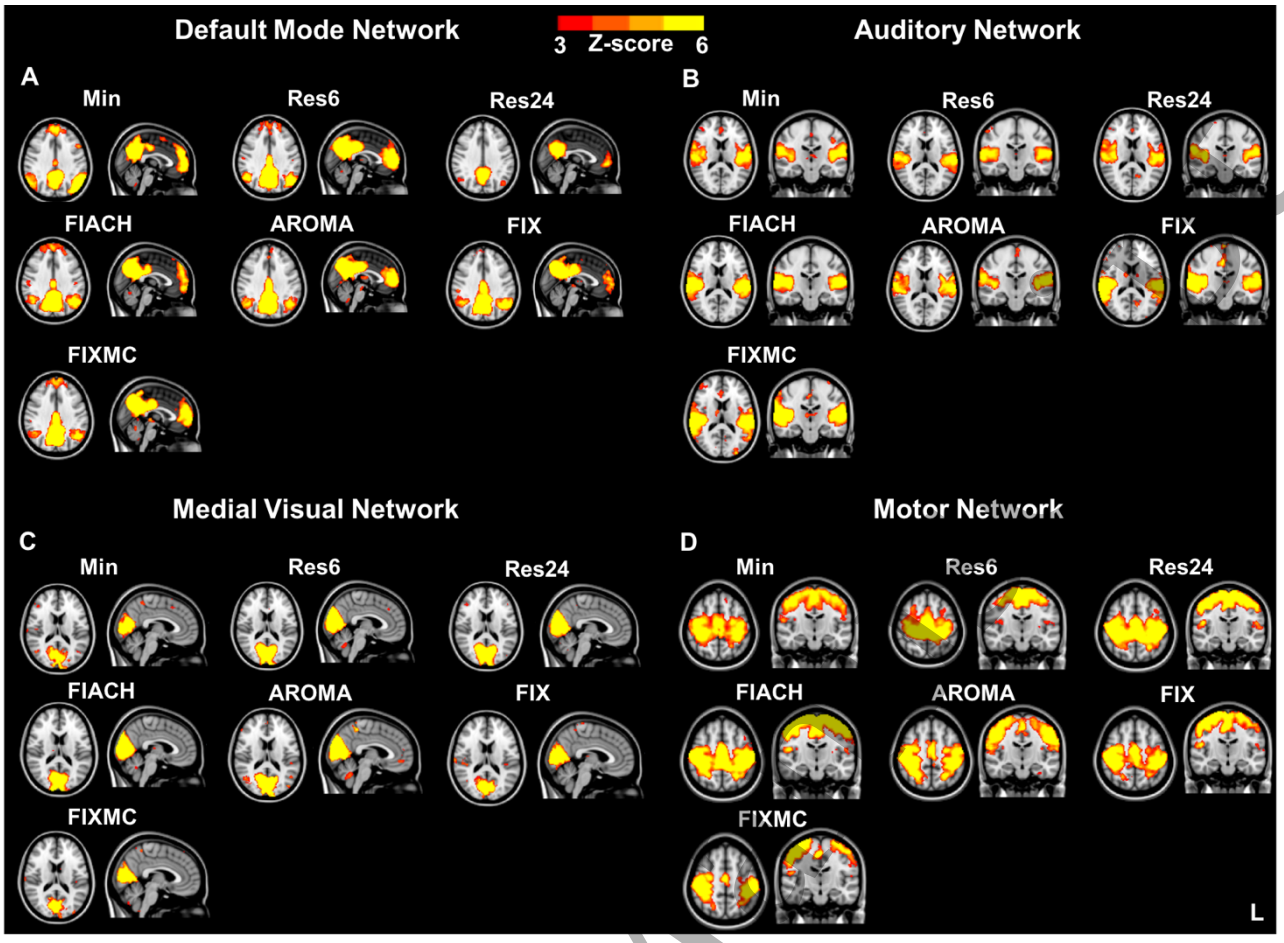

Fig.5: Group ICA maps for four representative networks. In particular, DMN (A), AUD (B), VISmed (C) and SMN (D) networks are reported, after having pre-processed rs-fMRI data with the seven pipelines. The same slices in MNI 2-mm standard space are reported.

the exception of Min, which showed noisy maps. Of note, FIXMC recovered three more distinct clusters (paracingulate and left and right frontal poles/superior frontal gyri) for this network. FPr and FPl networks were reliably recovered by all the pipelines, with the advanced methods generally showing more extensive clusters than Min. In the case of the DA network, Res6 and Res24 more accurately recovered the clusters in the left and right superior frontal gyri. Finally, the TEMP network was recovered by Min and the other advanced methods, with the only exceptions of Res24 which reported a cluster in the cerebellum, likely to be unrelated signal, and FIX/FIXMC which poorly recovered this network.

When considering quantitative measures in terms of spatial similarity and overlap (CC and DSC), these indices confirmed the preliminary qualitative evaluation, as reported in *Table 1*.

Table 1: Cross-correlation (CC) and Dice Similarity Coefficient (DSC) values calculated between each group ICA RSN and the corresponding RSN from Smith et al. template. For a given network and pipeline, the top row reports the CC value and the bottom row the DSC value. Generally high correlation values and high degree of overlap were found for all pipelines. The highest value for each measure and network is reported in bold. Visual medial (VISmed), visual occipital (VISocc), visual lateral (VISlat), default mode network (DMN), cerebellum (CER), motor (SMN), auditory (AUD), executive (EXE), frontoparietal right (FPr) and left (FPl), dorsal attention (DA) and temporal (TEMP).

Cross-correlation and Dice Similarity Coefficient

	Min	Res6	Res24	FIACH	AROMA	FIX	FIXMC
VISMed	0.78	0.84	**0.88**	0.78	0.84	0.76	0.74
	0.61	0.71	**0.76**	0.66	0.69	0.60	0.56
VISocc	0.60	0.87	**0.88**	0.85	0.75	0.84	0.81
	0.48	0.75	**0.78**	0.72	0.68	0.72	0.59
VISlat	0.60	0.58	**0.67**	0.52	0.66	0.58	0.58
	0.54	0.58	**0.62**	0.54	0.60	0.57	0.53
DMN	0.81	**0.86**	0.81	0.77	0.83	0.76	0.78
	0.69	0.63	0.62	0.55	0.63	0.58	0.57
CER	0.60	0.63	0.59	**0.66**	0.54	0.54	0.45
	0.49	0.55	0.51	**0.57**	0.50	0.52	0.45
SMN	**0.76**	0.52	0.65	0.64	0.66	0.67	0.60
	0.71	0.52	0.60	0.59	0.54	0.62	0.51
AUD	0.61	0.67	0.65	0.71	0.67	0.67	**0.72**
	0.55	0.60	0.58	**0.65**	0.60	0.56	0.56
EXE	0.61	0.61	**0.67**	0.55	0.63	0.57	0.59
	0.52	0.52	**0.56**	0.41	0.5	0.46	0.46
FPr	0.66	0.69	0.68	**0.70**	0.65	0.65	0.66
	0.57	0.59	0.59	**0.60**	0.55	0.57	0.58

FPl	**0.80**	0.73	0.79	0.63	0.76	0.77	0.77
	0.70	0.61	0.62	0.53	0.60	0.60	0.59
DA	0.60	0.41	0.55	0.55	**0.69**	0.60	0.61
	0.51	0.37	0.49	0.49	**0.51**	0.49	0.49
TEMP	0.60	0.65	0.47	**0.74**	0.71	0.28	0.36
	0.44	0.51	0.40	**0.62**	0.52	0.20	0.20

Our results point to a general high correlation for all RSNs and pipelines with CC values >=0.25, generally considered as a good cut-off value for classifying an IC from BOLD fMRI (Bright and Murphy, 2015). All correlation values were well above the suggested threshold for all networks and pipelines, except for one case in FIX, which recovered the TEMP with a lower CC value (0.28) though still above the reference cut-off value. In the case of DSC, a value above 0.3 is generally considered as representative of good overlap (Zhu *et al.*, 2013). We found DSC > 0.3 for the large majority of the cases. Lower DSC values, below the suggested optimal threshold, were only found for the TEMP network recovered after using FIX and FIXMC (DSC = 0.2).

Taking the DMN as an example, the highest correlation and spatial overlap were achieved by the Res6 (CC: 0.86) and Min (DSC: 0.69), while the lowest values were achieved via FIX (CC: 0.76) and FIACH (DSC: 0.55), albeit being well above the recommended thresholds. More complex RSNs, e.g. TEMP and DA, were recovered by all the pipelines with higher template correlation/overlap in the case of AROMA (TEMP: 0.71/0.52; DA: 0.69/0.51) and FIACH (TEMP: 0.74/0.62; DA: 0.55/0.49). CC and DSC values generally showed the same trend, with higher correlation values corresponding to higher DSC values.

3.4 Task fMRI: GLM Activation.

The group GLM maps are reported in *Fig.6*, while *Table 2* summarises the main cluster information (location, extension and statistics) of each group-level activation map. The more advanced strategies resulted in more localised activations when compared to Min, which showed more extensive and noisy activations. Similar activation extent and *z*-score values were found for the ICA-based pipeline. In the case of the regression-based pipelines, FIACH and Res24 reported lower *z*-score values, despite showing clusters localised in the expected areas of activations.

Table 3 reports the percentage of overlap between each group activation map and the GM and WM tissue priors. As expected, a high percentage of overlap was reported for GM in all the pipelines, with highest overlap in the case of FIACH (84.28%) and lowest in the case of Min (68.40%) due to the widespread activations. The overlap with WM was far lower than GM with values ranging between 21.07% (Res24) and 35.90 % (AROMA).

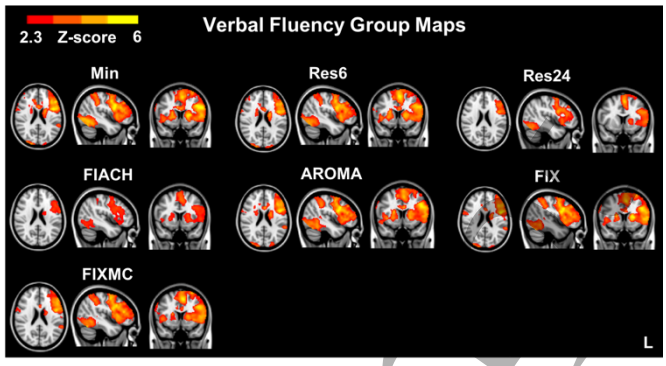

Fig.6: Group GLM activations for task-fMRI with a VF paradigm. Results are displayed in MNI 2-mm standard space with corrected z > 2.3.

Table 2: Summary of the GLM group activations. For each pipeline, the cluster location, cluster extent (number of voxels), cluster significance (p-value), maximum *z*-value and location (mm) in MNI 2-mm space are reported. Of note, the clusters of activations were corrected using cluster determined by z > 2.3 and a (corrected) cluster significant threshold of p-value = 0.05. Left (L), right (R), lateral (lat), superior (sup), inferior frontal gyrus (IFG), precentral gyrus (PreCG), occipital cortex (occ cortex), superior frontal (SF), paracingulate (PAC), middle frontal (MF), supplementary motor cortex (SMA), fusiform gyrus (FFG), insular (INS), supra-marginal gyrus (SMG), temporal (Temp).

	Location in MNI	**Voxels**	**P**	**Zmax**	**Zmax [x,y,z]**
Min	L IFG/Precentral	33533	9E-40	6.8	-46,8,28
	L lat occ cortex	3668	2E-8	4.98	-32,-60,42
	R PreCG	1421	0.0006	4.78	62,6,38
Res6	L SF/PAC gyrus	31296	0	6.54	-4,12,54
	L lat occ cortex	2417	2E-7	4.41	-26,-60,46
	R PreCG	709	0.0106	4.45	58,2,44
Res24	L SF/PAC gyrus	9581	6E-14	5.65	-4,12,54
	L occ cortex	1225	0.0067	4.38	-44,-66,-20
FIACH	L MF/IFG	8934	5E-17	4.64	-54,30,26
	L SMA	1926	1.7E-5	4.44	-4,4,54
	L Occ FFG	960	0.0043	3.23	-38,-68,-8
	R INS cortex	729	0.0133	3.24	32,22,-2
	L lat occ cortex	659	0.034	3.13	-40,-60,68
AROMA	L precentral/IFG	32386	9E-44	6.6	-46,6,24
	L postcentral /SMG	2770	1.2E-7	4.73	-38,-38,42
	R precentral /MFG	598	0.0431	4.92	60,0,44
FIX	L precentral/IFG	25158	8E-37	6.56	-46,4,24
	L lat occ cortex	3281	1E-8	5.29	-26,-70,42
	R caudate	1405	0.0002	5.22	20,12,12
	R precentral gyrus	902	0.0045	4.92	62,6,40
	L SMG	628	0.0339	3.85	-62,-42,22
FIXMC	L IFG	16985	2E-30	6.11	-48,30,16
	L occ FFG	3108	4.7E-9	5.49	-46,-64,-22
	L lat occ cortex	2848	1.1E-8	4.84	-32,-60,44
	R frontal operculum	2398	1.8E-7	5.2	40,22,0
	R occ FFG	1336	0.0001	5.9	36,-66,-20
	L sup temp gyrus	598	0.0272	3.92	-60,-30,4

Table 3: Percentage of overlap between group activation of each pipeline and tissue types (gray [GM] and white [WM] matter). The highest values are reported in bold.

	GM (%)	WM (%)
Min	68.4	34.18
Res6	69.98	29.98
Res24	76.56	21.07
FIACH	**84.28**	27.19
AROMA	71.68	**35.9**
FIX	73.18	30.5
FIXMC	75.45	26.43

These results were further confirmed by extracting additional information from the activations clusters. In terms of CNR, higher values were observed when Min, AROMA and FIX pipelines were used, while regression-based pipelines and FIXMC showed a similar pattern with relatively lower values (*Fig.7*). When statistically compared, the CNR values were significantly different across the seven pipelines ($F_{(6,133)}$ = 33.41, p-value < 0.01). Post-hoc t-tests revealed significant differences (p-value < 0.05, Bonferroni-corrected) for all pairs of pipelines, except for the cases reported in white in *Fig.7*.

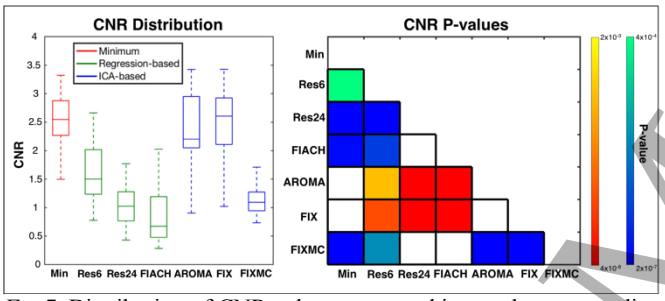

Fig.7: Distribution of CNR values across subject and corresponding p-values from the statistical comparison of every pre-processing method. Left panel: higher CNR values were found in the case of Min, AROMA and FIX. Right panel: graphical representation of the p-values resulting from the statistical comparison of mean CNR values, for every pipeline in task-fMRI . The comparisons which were not statistically significant are reported in white. Each pipeline in the row is compared to the each coloumn (ref to *Fig.2* for colour bar details).

Regarding the spatial comparisons, in order to match the *Neurosynth* VF reference template, we adopted the same threshold for the group activation maps (p-value < 0.01, FDR-corrected) for these subsequent analyses. In the case of FIACH, this threshold left no significant clusters, therefore measures from this pipeline could not be assessed further. CC, DSC, sensitivity and specificity values are reported for the other six pre-processing strategies in *Table 4*. The highest CC and DSC were found for FIXMC, with values of 0.52 and 0.44, respectively, while the lowest values were found in the case of Res24 (CC: 0.38, DSC: 0.24). Overall, ICA-based pipelines showed a good balance between these two measures, together with Min. On the other hand, Res24 was highly specific (0.63), but showed low sensitivity (0.15).

Table 4: Cross-correlation (CC), Dice Similarity Coefficient (DSC), sensitivity (Sens) and specificity (Spec) values between each group GLM map and the reference VF template. The highest values for each measure are reported in bold. NA: not available.

	CC	DSC	Sens	Spec
Min	0.45	0.35	**0.63**	0.25
Res6	0.48	0.39	0.61	0.29
Res24	0.38	0.24	0.15	**0.63**
FIACH	NA	NA	NA	NA
AROMA	0.45	0.36	0.57	0.26
FIX	0.5	0.42	0.58	0.33
FIXMC	**0.52**	**0.44**	0.57	0.35

3.5 Resting-state and Task fMRI: Functional Connectivity Metrics.

The mean connectivity matrices across subjects and CV for every pipeline are reported in *Fig.8*. In the resting-state dataset, higher mean connectivity was found for Min and the ICA-based pipelines. The highest mean correlation value (across ROIs) was obtained by AROMA (0.48), while the lowest mean SD was achieved by Res6 (0.16). Task-fMRI connectivity matrices showed generally higher values compared to resting-state. This was particularly evident in the case of Res24. Min showed the highest mean (0.56) and lowest mean SD (0.19). Lower CV values were observed for Min and AROMA in both resting-state and task fMRI. Overall, matrices were highly correlated with each other, with values ranging between 0.83-0.99 in resting-state and 0.82-0.98 in task. The information provided by these mean matrices was further summarised in *Figs.S3-S4*, where each matrix from a specific pipeline was expressed in terms of mean values for within/between-network connectivity. This would allow to immediately compare the connectivity patterns across pipelines. For example, the mean connectivity inside VIS reached high values in all cases, while its connections with the other networks revealed widely variable patterns with different trends for ICA-based and nuisance-based methods. Moreover, it can be appreciated that all the between-network connections of the limbic systems appeared to be among the lowest ones for all pipelines, in both rs-fMRI and task-fMRI.

With regards to the number of negative correlations, Min showed the lowest number of negative values (2.3 ± 2.4 % resting-state; 2.0 ± 3.9 % task), while the highest number was found in the case of Res6 (37.1 ± 6.2 % resting-state; 42 ± 4.7 % task). The ICA-based pipelines overall showed fewer negative correlations than the regression-based ones.

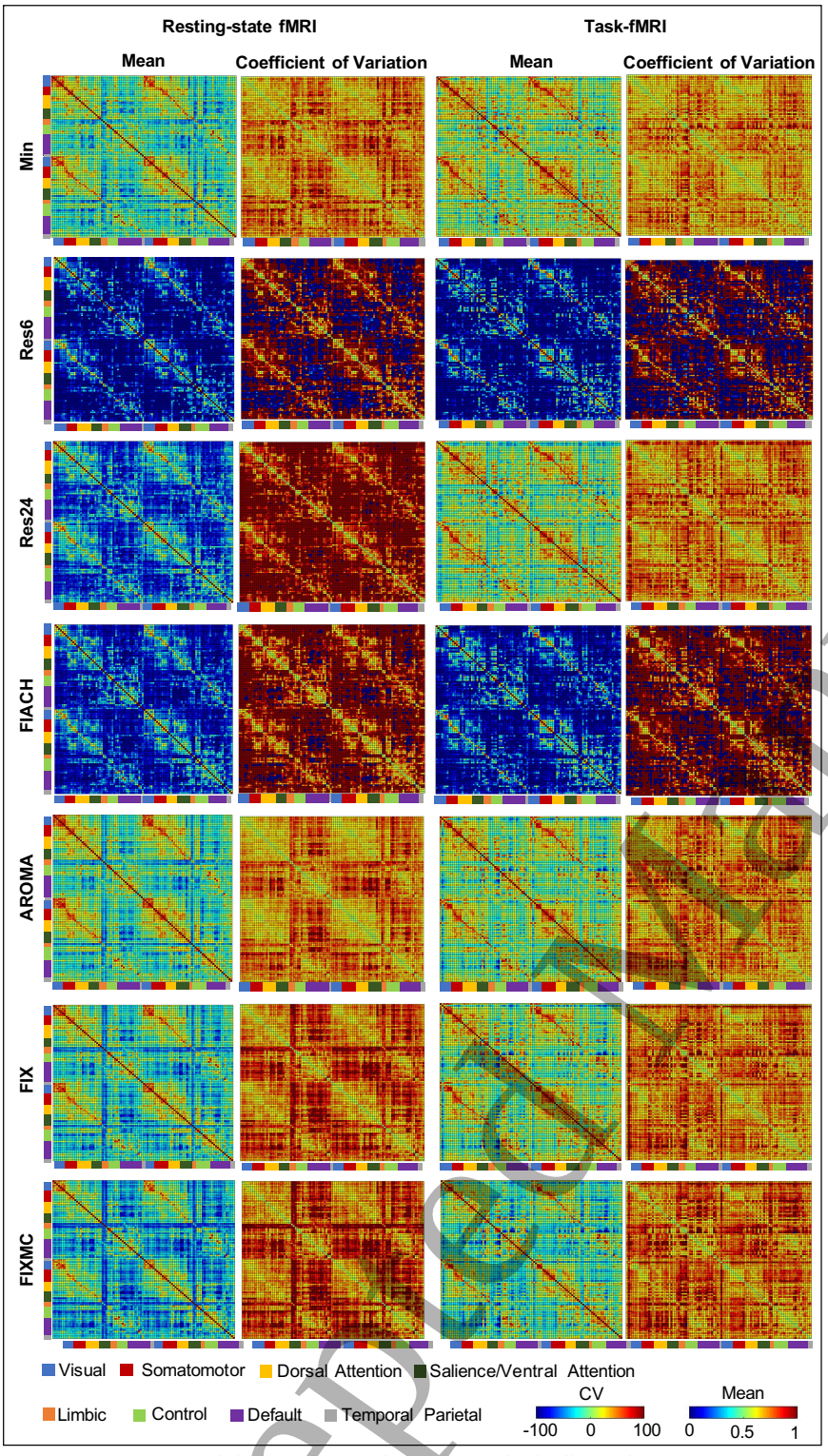

Fig.8: Mean connectivity matrices and CV for each pipeline. The mean matrices across subjects and CV are reported for each pipeline from rs-fMRI (left) and task-fMRI (right) data.

4. Discussion

The issue of data pre-processing for fMRI analysis has been widely investigated, owing to its relevance for any subsequent data processing, and in light of the myriad of different noise removal methods developed across centres (Griffanti *et al.*, 2015; Ciric *et al.*, 2017; Parkes *et al.*, 2018). Non-neuronal contributions to the BOLD signal are many and variable, and often correlated or co-linear to the signal of interest (Caballero-Gaudes and Reynolds, 2017). With this work, we aimed to substantially expand the current literature of fMRI denoising, by assessing the impact of different pre-processing methods on both resting-state and task fMRI datasets. Taking into account the lack of a widely accept "gold standard" or "ground truth" in this context, we decided to benchmark performance by assessing the influence exerted on several fMRI metrics used to investigate brain function: spatially and temporally covarying networks (Griffanti *et al.*, 2014; Pruim, Mennes, Buitelaar, *et al.*, 2015; Pruim, Mennes, van Rooij, *et al.*, 2015), connectivity (Griffanti *et al.*, 2016; Dipasquale *et al.*, 2017; Glasser *et al.*, 2018) and statistical parametric maps (Tierney *et al.*, 2016). Moreover, as previously highlighted (Power, Schlaggar and Petersen, 2015; Glasser *et al.*, 2018), the use of the widely clinically adopted VF task-fMRI brings about some *a priori* expectations, given the associated well characterised patterns of brain activation. Therefore, this work provides extensive guidelines on the impact of each denoising method on the signal of interest and associated connectivity measures for both resting-state and task fMRI, something that to our knowledge has never been previously investigated with such an extensive set of metrics.

The pre-processing pipelines investigated in this work fall into two broad categories: (A) statistical modelling and (B) data-driven methods. The former usually takes the form of a regression analysis explicitly removing nuisance variables from the signal of interest, whilst the latter is based on the theory of ICA. This investigation did not attempt to find the best performing pipeline to pre-process functional imaging data, but rather to provide guidelines on the impact and consequences of using the most common ones. In fact, our results show that there is no clear "winner", as measured by the metrics we implemented. Regression methods are excellent at detecting the artefactual signals associated with motion, but at the same time will necessarily

lead to a reduction of overall connectivity values compared to ICA methods (Pruim, Mennes, Buitelaar, *et al.*, 2015; Parkes *et al.*, 2018). Our results represent a substantial expansion of the state of the art, whilst agreeing with it in several key aspects. The performance of each pre-processing pipeline across benchmarks is summarised in *Table 5*. We are aware that more pipelines have been developed and reported in the literature, and that some of these pipelines we utilised might be considered as relatively "simple" in the context of rs-fMRI. However, assessing the impact of denoising methods in task-fMRI data has scarcely been investigated. Thus, we decided to include a reference baseline (e.g. Min) and relatively simple pipelines (e.g. Res6), to provide an initial investigation into this complex topic, similarly to what other authors proposed (Mayer *et al.*, 2019). In addition, we decided to exclude GSR-based pipelines from the scope of this investigation. GSR remains a controversial step in the pre-processing of fMRI, despite having shown effective results in resting-state data (Ciric *et al.*, 2017; Parkes *et al.*, 2018). In task-fMRI, the global brain signal might be correlated to the administered paradigm (Mayer *et al.*, 2019). We therefore believe that it is important to assess the impact of simplier denoising methods on task data, before benchmarking more complex pipelines which are still controversial in the context of the widely investigated rs-fMRI. A recent publication (Mayer *et al.*, 2019) assessed the impact of similar pipelines on task-fMRI designs (blocked and event-related). Our work provides additional information in this context, by comparing resting-state and task fMRI, with respect to BOLD measures and brain connectivity.

In this work, we chose the following metrics: DVARS, tSNR, tDoF, RSNs/activation maps, CNR and connectivity matrices, to understand the consequences of using one method over another, relative to a minimal amount of processing.

DVARS is used to measure the degree to which each pipeline can remove motion within the time series, which is considered as a major contributor of spurious signal in BOLD-fMRI. All pipelines achieved a decrease in DVARS compared to Min, highlighting more pronounced removal of motion artefacts. FIACH was able to accurately estimate motion time series and achieved the lowest DVARS values across all pipelines. Naturally, a potential risk of using a pipeline that is really good at removing any signal change associated with motion, is that it may also remove relevant signal that happens to be co-linear with the signal of interest. A common problem, particularly in the case of task-fMRI data, is that a stimulus of interest itself may trigger physical movement in human subjects, like nodding or a swallowing (Mayer *et al.*, 2019). For rs-fMRI, the consequences are more complicated, as much of the signal related to pulsatile motion arising from breathing or heartbeat is intrinsic to the spatiotemporal relationship of groups of voxels highlighted by ICA methods. In the case of regression, the use of nonlinear expansion of the motion terms has been proposed to account for spin related contributions of motion related artefacts (Friston *et al.*, 1996; Caballero-Gaudes and Reynolds, 2017). Indeed, we reported a significantly decreased DVARS when Res24 was used as compared to Res6. However, the use of additional motion-related regression, whilst accounting for additional variance, might also lead to increased loss of tDoF and therefore less reliable results in post-processing connectivity analyses (Caballero-Gaudes and Reynolds, 2017).

TSNR can be conceptualised as a measure of variability in the signal of interest, and a reduction of this can be interpreted as a proxy of how well a pre-processing pipeline removes unwanted noise. However, the tSNR taken alone might not be indicative of the goodness of a pipeline and it needs to be accompanied by other measures, such as the loss of tDoF (Boscolo Galazzo *et al.*, 2019). For instance, in task-fMRI, a decreased tSNR might be due to increased mean signal by a small amount (~1%) and increased SD by a higher amount, which is not related to how well a denoising method is actually performing. In ICA-based pipelines, increased tSNR can be found even when a significant number of meaningful components are removed from the data (leading to decreased SD) (Boscolo Galazzo *et al.*, 2019). For both resting state and task fMRI data, we showed an expected decrease in signal variability across time, with an increase in signal mean for all processing pipelines; with both regression and data-driven methods performing well for the rs-fMRI data. Evidence in relation to task-fMRI data, however, is more complex. While investigating the overall mean tSNR of the GM was acceptable for rs-fMRI, the same could not be attained for task-fMRI, because of the specific influence of task-associated stimulation on signal changes in very specific areas. We instead report tSNR values only for ROIs known to be reliably activated during this task, namely the L_IFG (*Fig.3*) and other key fronto-temporo-parietal language-relevant areas, *Fig.S1*. In each region, we found an overall increase in the mean signal for both regression and data-driven methods, though regression methods did recover a higher mean signal. Signal variability, when compared to the minimally processed data, was generally increased both for regression and the data-driven methods. This is in accordance with previous publications on the topic (Griffanti *et al.*, 2016; Dipasquale *et al.*, 2017; De Blasi *et al.*, 2018). For task-based fMRI, a reader might find the effect that the pre-processing methods have on the activated regions more intuitive to understand, which is conveyed via analysis of the CNR, discussed below.

To compensate for the limitations of the tSNR highlighted above, we also considered the loss in tDoF as an additional quality measure. *tDoF* can be lost as a result of the noise removal strategy employed, and can substantially affect the power of the statistical tests implemented to uncover neurobiological effects of interest. Confound regressors and

censoring both reduce the tDoF in data (Ciric *et al.*, 2017), and tDoF loss may introduce a bias if it varies across subjects. It is assumed that each time series regressed out constitutes a single tDoF. ICA-based methods showed higher loss of tDoF, as expected and already reported in the literature (Dipasquale *et al.*, 2017). Indeed, Min, FIACH and Res6 were associated with a lower loss of tDoF which was constant across participants, being pre-specified in the model (Parkes *et al.*, 2018). AROMA showed a moderate loss of tDoF, which varied among subjects but was generally similar to Res24 (Pruim, Mennes, Buitelaar, *et al.*, 2015; Pruim, Mennes, van Rooij, *et al.*, 2015; Ciric *et al.*, 2017). The highest loss of tDoF occurred with FIXMC in both datasets, which may in turn relate to a higher risk of reduced sensitivity to the underlying signal of interest (Pruim, Mennes, van Rooij, *et al.*, 2015). However, as pointed out by Ciric et al. 2017 (Ciric *et al.*, 2017), the loss of tDoF should be interpreted with caution and in conjunction with other measures, such as accuracy of network recovery and connectivity analyses, as the removed tDoF might correspond to artefacts rather than signal of interest (Ciric *et al.*, 2017).

Consequently, *RSNs and Task Activation* can be used to assess accuracy and reproducibility of the expected results in each pre-processing pipeline. In both resting-state and task fMRI, we computed the CC and DSC to assess how well the recovered areas of activation overlap with the expected activation maps. The latter were defined as (a) the RSN templates according to Smith et al. 2009 (Smith *et al.*, 2009) for resting-state data, and (b) the VF activation template for task-fMRI data. For the former (a), the Smith et al. template, considered as gold standard, allowed us to enrich our comprehensive overview, by providing quantitative information on all the considered approaches and allow comparison between them, representing a fixed point to measure the distance from for all the seven pipelines. However, we have to keep in mind that the different approach employed to derive the template could lower down the similarity values when a different processing pipeline is employed to recover the RSNs. We also acknowledge that our eyes-closed acquisition differs from how the subjects were scanned to derive the Smith et al. template (eyes-open). However, we believe that the influence of the acquisition mode may be minor and unlikely to systematically bias our comparison of denoising strategies (Patriat *et al.*, 2013). For the latter (b), we additionally computed sensitivity and specificity values in relation to the *Neurosynth* VF template. Percentage overlap of the activated areas in relation to GM and WM voxels was also computed, to assess the biological relevance of the activations, and specifically whether they would mostly be localised in GM areas. For the resting-state dataset, all advanced pipelines were able to accurately recover the main RSNs and showed similar CC and DSC values. As for task-fMRI, all pipelines related to an improvement of CC and DSC with the *Neurosynth* VF template when compared to Min, with the only exclusion of Res24. Regression-based pipelines performed poorly when compared to ICA-based ones for group GLM activation, which may appear surprising, as the former represent the most used pipelines in the context of task-fMRI literature. Res24 showed suboptimal recovering of RSNs and reduced activations in task-fMRI, which may relate to the decrease in tDoF resulting in a smaller number of voxels to be activated, despite preservation of the central location of the main activation clusters. FIACH accurately recovered all RSNs, with CC values above 0.5 and DSC greater than 0.4. In the case of group-level task activation maps, FIACH resulted in the activation of expected task-related areas, and clusters exhibited the highest overlap with GM voxels (84%) among all pipelines; their extension and z-scores, however, were lower. It is important to note that, as different from previous literature, the FIACH pipeline output was implemented in our analysis via regression of the FIACH regressors from the time series and saving of the residuals as denoised signal, rather than via inclusion of those regressors in a GLM (Tierney *et al.*, 2016; Caballero-Gaudes and Reynolds, 2017; Kronbichler *et al.*, 2018). In our work, this choice was motivated by the need of (a) obtaining a processed time series to be used for further connectivity analyses and (b) keeping the same analytical implementation for all the regression-based pipelines. We acknowledge, however, that this step might have had a differential impact on the FIACH-processed data, leading to a reduced CNR and number of activated voxels. In terms of sensitivity and specificity, Min and ICA-based pipelines showed a good balance, Res6/Res24 were associated with high sensitivity/specificity and relatively low specificity/sensitivity, respectively, which needs to be considered for appropriate assessment of activation maps after either of these pipelines is used. Finally, all advanced pipelines improved the overlap of group activation maps with GM locations when compared to Min, which can be interpreted as an indicator of recovered signal within biologically plausible locations.

In task-based fMRI, *CNR* describes the difference in the means between the task-on condition and the task-off condition (being represented here by VF and baseline, respectively) normalised by the standard deviation in the baseline condition. The CNR is more commonly represented as a statistical parametric map of voxel-wise z-scores (*Fig.6*). Results in the literature are commonly displayed with a reported set threshold of peak significance of a particular number of voxels above this threshold (Shen and Q. Duong, 2011; Molloy, Meyerand and Birn, 2014). For the purposes of this investigation, we use the time series with the maximum z-score (Storti *et al.*, 2018) even if this might lead to different brain areas in different subjects or pipelines. This approach has been previously used in the literature (Storti *et al.*, 2018) and it was chosen to provide a general metric to assess pipeline

performance. With reference to minimal processing, the pipelines recovering the highest signal of interest were both data-driven methods, FIX and AROMA, with the lowest yield attained by FIACH.

Similarity of mean connectivity matrices measured via pairwise correlations, and the *CV of the correlation* between regions, assessed the impact of denoising methods on the subsequently obtained FC matrices. We generally found higher correlations for the task-based datasets, which may be ascribed to the effect of task-related activity, leading to more homogeneous and generally higher brain connectivity estimates. The specific effects of task-based activity on FC values are not yet fully elucidated, and may be highly dependent on the task performed. For example, in Cole et al. 2014, connectivity matrices obtained from different tasks showed different structure, thought being highly correlated with one another and with the matrices obtained from resting-state data (Cole *et al.*, 2014). ICA-based pipelines led to similar mean correlation values and comparable CV matrices across pipelines and datasets. Reduced correlations and higher number of negative correlations were found in the case of Res6 and FIACH for both datasets. The results obtained for Res24 were somehow unexpected in the case of task. Indeed, we expected to find a reduced connectivity as in the case of resting-state data and in line with the other regression-based pipelines. However, we reported correlation values and variability in range with Min and ICA-based pipelines, as opposed to the resting-state dataset where Res24 performance was in line with the other regression-based pipelines.

This study has limitations. As a general caveat of all investigations in this field, we ought to emphasise the lack of an absolute noise-free gold standard, which arguably complicates comparisons among different denoising strategies (Ciric *et al.*, 2017; Dipasquale *et al.*, 2017; Parkes *et al.*, 2018). To overcome this limitation, we implemented task-fMRI dataset relating to an extensively used paradigm, which may provide higher level of generalisability and 'ground truth' as the expected areas of activations are known *a priori*. A further limitation is that the subjects were not the same for resting-state and task fMRI, owing to the unavailability of

Table 5: Summary of the performance of each pre-processing method with respect to benchmarks, considered in this study. For a given method and benchmark, the top/bottom row reports the performance in the resting-state/task fMRI dataset. The +/- symbol indicates good/bad performance for a given metric and pipeline. In the case of DVARS, tSNR, lost DoF, CNR and overlap GM/WM, the results are reported with respect to Min, so that +/- refers to improvement/worsening when compared to minimum pre-processing (Min). For the other benchmarks a general summary is reported. The best/worst performance is indicated in green/red. Overall, we found heterogeneous performance of each method, across benchmarks. This summary is intended for the reader to identify which pipeline works best for a given study. NA: not available.

	DVARS	tSNR	Lost DoF	R	DSC	Sens/ Spec	CNR	Overlap GM	Overlap WM	Connectivity Matrices
Min	//	//	//	>0.25 all	>0.3 all	NA	NA	NA	NA	~ICAbased,slightly flatter/noisier
	//	//	//	~Res6/ AROMA	~Res6/ AROMA	Sens, Spec	//	//	//	Highest mean and lowest SD
Res6	+	+	~FIACH	>0.25 all	>0.3 all	NA	NA	NA	NA	Lowest SD, low/negative values
	+	+	~FIACH	~Min/ AROMA	>Min/ AROMA	High, High	-	+	+	Low/negative values
Res24	+	+	~AROMA/ FIX	>0.25 all	>0.3 all	NA	NA	NA	NA	Low/negative values
	-	+	~AROMA/ FIX	-	-	Sens, Spec	-	+	+	~ICA-based
FIACH	+	+	~Res6	>0.25 all	>0.3 all	NA	NA	NA	NA	Low/negative values
	+	+	~Res6	NA	NA	NA	-	+	+	Low/negative values
AROMA	+	+	~Res24/ FIX	>0.25 all	>0.3 all	NA	NA	NA	NA	Highest mean, low SD
	+	+	~Res24/ FIX	~Min/Res6	~Min/Res6	High, High	~Min	+	-	~FIX
FIX	+	+	~Res24/ AROMA	>0.25 all	>0.3 all, but TEMP	NA	NA	NA	NA	~other ICA-based
	+	+	~Res24/ AROMA	~FIXMC	~FIXMC	High, High	~Min	+	+	~AROMA
FIXMC	+	+	-	>0.25 all	>0.3 all, but TEMP	NA	NA	NA	NA	~other ICA-based (more modules)
	+	+	-	+	+	High, High	-	+	+	~other ICA-based (slightly noisier)

subjects with both acquisitions. Whilst the use of two separate datasets may add variability to the analysis, we found that the effects of noise removal via different pipelines were generally consistent between resting-state and task fMRI data. This finding points to a higher level of generalisability of our results across fMRI acquisitions. These findings are also in line with results by Pruim et al., 2015 (Pruim, Mennes, van Rooij, *et al.*, 2015), who found comparable results between resting-state and task fMRI when assessing those to validate the AROMA pipeline. We acknowledge that our sample size was generally lower than previous studies assessing the impact of denoising methods on rs-fMRI data (Pruim, Mennes, Buitelaar, *et al.*, 2015; Ciric *et al.*, 2017; Parkes *et al.*, 2018). However, our study is the first to convey an extensive comparison of pre-processing methods in both resting-state and task fMRI data, acquired using widely clinically employed acquisitions. Investigating how the fMRI signal and derived FC measures may change as a result of different denoising strategies is timely and important to inform experimental design and post-processing analyses. Therefore, this work provides relevant novel evidence and extensive guidelines on how to pick the best pre-processing depending on the objective of the study.

5. Conclusions

We described a range of commonly used noise removal pipelines for BOLD-fMRI time series and illustrated their application to both resting-state and task fMRI dataset experiments. We also highlighted the heterogeneity in performance of pipelines across benchmarks, especially with respect to FC results. We envision this work as a "brochure" for the users to choose the most appropriate method for their data (*Table 5*), rather than as a performance indicator of any one pipeline in particular.

Acknowledgements

This work was supported by the EPSRC-funded UCL CDT in Medical Imaging (EP/L016478/1), the Department of Health's NIHR-funded BRC at UCLH and by the MRC. This work was carried out in collaboration with the Dept. Computer Science in the University of Verona (Italy). LC was supported by Brain Research UK and acknowledges support from a Berkeley Fellowship by UCL and Gonville and Caius College, Cambridge.

References

De Blasi, B. *et al.* (2018) 'Pipeline comparison for the pre-processing of resting-state data in epilepsy', *26th European Signal Processing Conference, IEEE*, pp. 1137–1141.

Boscolo Galazzo, I. *et al.* (2019) 'Arterial spin labeling reveals disrupted brain networks and functional connectivity in drug-resistant temporal epilepsy', *Frontiers in Neuroinformatics*, 12(101), pp. 1–18. doi: 10.3389/fninf.2018.00101.

Bowring, A., Maumet, C. and Nichols, T. E. (2019) 'Exploring the impact of analysis software on task fMRI results', *Human Brain Mapping*, 40, pp. 3362–3384. doi: https://doi.org/10.1002/hbm.24603.

Bright, M. G. and Murphy, K. (2015) 'Is fMRI "noise" really noise? Resting state nuisance regressors remove variance with network structure', *Neuroimage*. Elsevier B.V., 114, pp. 158–169. doi: 10.1016/j.neuroimage.2015.03.070.

Bright, M. G., Tench, C. R. and Murphy, K. (2017) 'Potential pitfalls when denoising resting state fMRI data using nuisance regression', *NeuroImage*. Elsevier, 154, pp. 159–168. doi: https://doi.org/10.1016/j.neuroimage.2016.12.027.

Caballero-Gaudes, C. and Reynolds, R. C. (2017) 'Methods for cleaning the BOLD fMRI signal', *Neuroimage*. Elsevier, 154, pp. 128–149. doi: 10.1016/j.neuroimage.2016.12.018.

Ciric, R. *et al.* (2017) 'Benchmarking of participant-level confound regression strategies for the control of motion artifact in studies of functional connectivity', *Neuroimage*, 154, pp. 174–187. doi: 10.1016/j.neuroimage.2017.03.020.

Cole, D. M., Smith, S. M. and Beckmann, C. F. (2010) 'Advances and pitfalls in the analysis and interpretation of resting-state FMRI data', *Frontiers in Systems Neuroscience*, 4(8), pp. 1–15. doi: 10.3389/fnsys.2010.00008.

Cole, M. W. *et al.* (2014) 'Intrinsic and task-evoked network architectures of the human brain', *Neuron*. Elsevier Inc., 83, pp. 238–251. doi: 10.1016/j.neuron.2014.05.014.

Dipasquale, O. *et al.* (2017) 'Comparing resting state fMRI de-noising approaches using multi- and single-echo acquisitions', *PLoS ONE*. Edited by A. Gozzi, 12(3), pp. 1–25. doi: 10.1371/journal.pone.0173289.

Engel, J. J. *et al.* (2013) 'Connectomics and epilepsy', *Current Opinion in Neurology*, 26(2), pp. 186–194. doi: 10.1097/WCO.0b013e32835ee5b8.

Fox, M. D. and Greicius, M. (2010) 'Clinical applications of resting state functional connectivity.', *Frontiers in Systems Neuroscience*, 4(19), pp. 1–13. doi: 10.3389/fnsys.2010.00019.

Friston, K. J. *et al.* (1996) 'Movement-related effects in fMRI time-series', *Magnetic Resonance in Medicine*, 35, pp. 346–355.

Glasser, M. F. *et al.* (2018) 'Using temporal ICA to selectively remove global noise while preserving global signal in functional MRI data', *NeuroImage*. Elsevier Inc., 181, pp. 692–717. doi: https://doi.org/10.1016/j.neuroimage.2018.04.076.

Griffanti, L. *et al.* (2014) 'ICA-based artefact removal and accelerated fMRI acquisition for improved resting state network imaging', *NeuroImage*, 95, pp. 232–247. doi: 10.1016/j.neuroimage.2014.03.034.

Griffanti, L. *et al.* (2015) 'Effective artifact removal in resting state fMRI data improves detection of DMN functional connectivity alteration in Alzheimer's disease', *Frontiers in Human Neuroscience*, 9(449), pp. 1–11. doi: 10.3389/fnhum.2015.00449.

Griffanti, L. *et al.* (2016) 'Challenges in the reproducibility of clinical studies with resting state fMRI: an example in early Parkinson's disease', *NeuroImage*. The Authors, 124, pp. 704–713. doi: 10.1016/j.neuroimage.2015.09.021.

Kronbichler, L. *et al.* (2018) 'Schizophrenia and category-selectivity in the brain: Normal for faces but abnormal for houses', *Frontiers in Psychiatry*, 9(47), pp. 1–12. doi: 10.3389/fpsyt.2018.00047.

Liang, X. *et al.* (2013) 'Coupling of functional connectivity and regional cerebral blood flow reveals a physiological basis for network hubs of the human brain', *Proceedings of the National Academy of Sciences*, 110(5), pp. 1929–1934. doi: 10.1073/pnas.1214900110.

Lydon-Staley, D. M. *et al.* (2019) 'Evaluation of confound

regression strategies for the mitigation of micromovement artifact in studies of dynamic resting-state functional connectivity and multilayer network modularity', *Network Neuroscience*, 3(2), pp. 427–454. doi: https://doi.org/10.1162/netn_a_00071.

Mayer, A. R. *et al.* (2019) 'A comparison of denoising pipelines in high temporal resolution task-based functional magnetic resonance imaging data', *Human Brain Mapping*, 40(13), pp. 3843–3859. doi: 10.1002/hbm.24635.

Molloy, E. K., Meyerand, M. E. and Birn, R. M. (2014) 'The influence of spatial resolution and smoothing on the detectability of resting-state and task fMRI', *Neuroimage*. Elsevier B.V., 86, pp. 221–230. doi: 10.1016/j.neuroimage.2013.09.001.

Muschelli, J. *et al.* (2014) 'Reduction of motion-related artifacts in resting state fMRI using aCompCor', *NeuroImage*. Elsevier Inc., 96, pp. 22–35. doi: 10.1016/j.neuroimage.2014.03.028.

Parkes, L. *et al.* (2018) 'An evaluation of the efficacy, reliability, and sensitivity of motion correction strategies for resting-state functional MRI', *Neuroimage*, 171, pp. 415–436. doi: https://doi.org/10.1016/j.neuroimage.2017.12.073.

Patriat, R. *et al.* (2013) 'The effect of resting condition on resting-state fMRI reliability and consistency: A comparison between resting with eyes open, closed, and fixated', *NeuroImage*. Elsevier Inc., 78, pp. 463–473. doi: 10.1016/j.neuroimage.2013.04.013.

Power, J. D., Schlaggar, B. L. and Petersen, S. E. (2015) 'Recent progress and outstanding issues in motion correction in resting state fMRI', *NeuroImage*. Elsevier Inc., 105, pp. 536–551. doi: 10.1016/j.neuroimage.2014.10.044.

Pruim, R. H. R., Mennes, M., Buitelaar, J. K., *et al.* (2015) 'Evaluation of ICA-AROMA and alternative strategies for motion artifact removal in resting state fMRI', *NeuroImage*. Elsevier Inc., 112, pp. 278–287. doi: 10.1016/j.neuroimage.2015.02.063.

Pruim, R. H. R., Mennes, M., van Rooij, D., *et al.* (2015) 'ICA-AROMA: a robust ICA-based strategy for removing motion artifacts from fMRI data', *NeuroImage*. Elsevier Inc., 112, pp. 267–277. doi: 10.1016/j.neuroimage.2015.02.064.

Salimi-Khorshidi, G. *et al.* (2014) 'Automatic denoising of functional MRI data: combining independent component analysis and hierarchical fusion of classifiers', *NeuroImage*. Elsevier B.V., 90, pp. 449–468. doi: 10.1016/j.neuroimage.2013.11.046.

Satterthwaite, T. D. *et al.* (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, pp. 1–39. doi: 10.1016/j.neuroimage.2012.08.052.

Schaefer, A. *et al.* (2018) 'Local-global parcellation of the human cerebral cortex from intrinsic functional connectivity MRI', *Cerebral Cortex*, 28, pp. 3095–3114. doi: 10.1093/cercor/bhx179.

Shen, Q. and Q. Duong, T. (2011) 'Background suppression in arterial spin labeling MRI with a separate neck labeling coil', *NMR Biomed*, 24(9), pp. 1111–1118. doi: 10.1002/nbm.1666.Background.

Smith, S. M. *et al.* (2009) 'Correspondence of the brain's functional architecture during activation and rest', *Proceedings of the National Academy of Sciences*, 106(31), pp. 13040–13045. doi: 10.1073/pnas.0905267106.

Storti, S. F. *et al.* (2018) 'Dual-echo ASL based assessment of motor networks: a feasibility study', *Journal of Neural Engineering*. IOP Publishing, 15, pp. 1–19. doi: 10.1088/1741-2552/aa8b27.

Tierney, T. M. *et al.* (2016) 'FIACH: A biophysical model for automatic retrospective noise control in fMRI', *Neuroimage*. The Authors, 124, pp. 1009–1020. doi: 10.1016/j.neuroimage.2015.09.034.

Wandschneider, B. *et al.* (2017) 'Effect of topiramate and zonisamide on fMRI cognitive networks', *Neurology*, 88(12), pp. 1165–1171. doi: 10.1212/WNL.0000000000003736.

Zhu, S. *et al.* (2013) 'Resting state brain function analysis using concurrent BOLD in ASL perfusion fMRI', *PLoS ONE*, 8(6), pp. 1–9. doi: 10.1371/journal.pone.0065884.