NeuroImage 200 (2019) 511–527

Contents lists available at ScienceDirect

NeuroImage

journal homepage: www.elsevier.com/locate/neuroimage

Recommendations for motion correction of infant fNIRS data applicable to multiple data sets and acquisition systems

Renata Di Lorenzo [a,b,*,1], Laura Pirazzoli [c,i,j,1], Anna Blasi [d], Chiara Bulgarelli [c,d], Yoko Hakuno [e,f], Yasuyo Minagawa [f], Sabrina Brigadoi [g,h]

[a] Experimental Psychology, Helmholtz Institute, Utrecht University, Utrecht, the Netherlands
[b] Developmental Psychology, Utrecht University, Utrecht, the Netherlands
[c] Centre for Brain and Cognitive Development, Birkbeck College, London, UK
[d] Department of Medical Physics and Biomedical Engineering, University College London, UK
[e] Research Fellow of Japan Society for the Promotion of Science, 5-3-1 Kojimachi, Chiyoda-ku, Tokyo, 102-0083, Japan
[f] Department of Psychology, Faculty of Letters, Keio University, 2-15-45 Mita, Minato-ku, Tokyo, 108-8345, Japan
[g] Department of Developmental Psychology, University of Padova, Padova, Italy
[h] Department of Information Engineering, University of Padova, Padova, Italy
[i] Laboratories of Cognitive Neuroscience, Division of Developmental Medicine, Department of Medicine, Boston Children's Hospital, Boston, MA, USA
[j] Harvard Medical School, Boston, MA, USA

ARTICLE INFO

Keywords:
fNIRS
Motion correction
Infants
Semi-simulated data

ABSTRACT

Despite motion artifacts are a major source of noise in fNIRS infant data, how to approach motion correction in this population has only recently started to be investigated. Homer2 offers a wide range of motion correction methods and previous work on simulated and adult data suggested the use of Spline interpolation and Wavelet filtering as optimal methods for the recovery of trials affected by motion. However, motion artifacts in infant data differ from those in adults' both in amplitude and frequency of occurrence. Therefore, artifact correction recommendations derived from adult data might not be optimal for infant data. We hypothesized that the combined use of Spline and Wavelet would outperform their individual use on data with complex profiles of motion artifacts. To demonstrate this, we first compared, on infant semi-simulated data, the performance of several motion correction techniques on their own and of the novel combined approach; then, we investigated the performance of Spline and Wavelet alone and in combination on real cognitive data from three datasets collected with infants of different ages (5, 7 and 10 months), with different tasks (auditory, visual and tactile) and with different NIRS systems. To quantitatively estimate and compare the efficacy of these techniques, we adopted four metrics: hemodynamic response recovery error, within-subject standard deviation, between-subjects standard deviation and number of trials that survived each correction method. Our results demonstrated that (i) it is always better correcting for motion artifacts than rejecting the corrupted trials; (ii) Wavelet filtering on its own and in combination with Spline interpolation seems to be the most effective approach in reducing the between- and the within-subject standard deviations. Importantly, the combination of Spline and Wavelet was the approach providing the best performance in semi-simulation both at low and high levels of noise, also recovering most of the trials affected by motion artifacts across all datasets, a crucial result when working with infant data.

1. Introduction

Functional near-infrared spectroscopy (fNIRS) is a neuroimaging technique that has experienced an exponential increase in its application to study the infant brain and cognitive development (for a review see Wilcox and Biondi, 2015). fNIRS has been adopted in several developmental areas, such as social cognition, language, memory, numerosity (for a recent review see Aslin et al., 2015) and it also shows potential for the study of connectivity (e.g., Bulgarelli et al., 2018; Homae et al., 2010; Molavi et al., 2014). fNIRS infers localized brain activation by

* Corresponding author. Developmental and Experimental Psychology, Helmholtz Institute, Utrecht University, Heidelberglaan 1, 3584, CS Utrecht, the Netherlands.
E-mail address: renata.dlorenzo@gmail.com (R. Di Lorenzo).
[1] Renata Di Lorenzo and Laura Pirazzoli contributed equally to this work and therefore share joint first authorship.

https://doi.org/10.1016/j.neuroimage.2019.06.056
Received 3 April 2019; Received in revised form 11 June 2019; Accepted 24 June 2019
Available online 25 June 2019

quantifying changes in blood oxy-hemoglobin (HbO) and deoxy-hemoglobin (HbR) from changes in intensity of near-infrared light, migrating from a source to a detector (Obrig and Villringer, 2003). In developmental research, fNIRS is praised because it measures non-invasively brain activity similarly to fMRI, without some of the limitations of the latter (e.g., allowing the study of awake infants in more naturalistic environments). Whereas in fMRI even small head movements can have a large negative impact on data quality, fNIRS, with a good design and tight fit of the headgear, will allow participants a much wider range of movement. This does not mean that the technique is insensitive to motion artifacts; on the contrary, the presence of significant motion artifacts in infant fNIRS time recordings usually constitutes a considerable challenge for the data analysis and recovery of brain activation. Motion artifacts are typically seen in the form of abrupt signal changes and can be classified based on their amplitude and frequency (e.g., spikes, slow drifts, baseline shifts). Despite the improvements achieved in fNIRS technology and headgear sets (Lloyd-Fox et al., 2010), fNIRS users are still facing the issue of how to best approach motion artifacts that are particularly disrupting in infant fNIRS data.

The presence of motion artifacts in data recordings might affect the shape and validity of the hemodynamic response function (HRF). Indeed, abrupt changes in the signal might mask or inflate the true hemodynamic response, possibly causing, at the level of statistical inference, false negatives or false positives. Furthermore, the undesired modulation of the NIRS signal due to artifacts, as well as improper pre-processing pipelines, might contribute to the variability of hemodynamic responses often reported in infant studies (for further discussion on this issue see Issard and Gervain, 2018). Hence, addressing the issue of motion artifacts is crucial to draw valid neuroscientific conclusions when interpreting fNIRS results; also, identifying the most appropriate pipeline for motion correction might help reduce the reported variability in infant HRFs and to establish a typical infant hemodynamic response.

In the standard fNIRS processing pipeline, the main options for dealing with motion artifacts are: (i) rejection of the trials contaminated by artifacts; or (ii) correction of the corrupted signal. Trial rejection is not desirable and is often not possible in studies involving challenging populations (such as infants and clinical populations), where the total number of trials per session is often limited by the tolerance and compliance of the participants. Therefore, excluding trials will significantly undermine the calculation of a reliable hemodynamic response and/or weaken statistical power. Thus, in these cases motion correction is needed to reduce the number of rejected trials and, ultimately, the number of excluded datasets from a study. Homer2 (Huppert et al., 2009), is an open-source software for fNIRS data pre-processing that offers a wide range of motion correction methods and it is being used by an increasing number of research groups (e.g., Lloyd-Fox et al., 2015; Miguel et al., 2017; Ravicz et al., 2015; Timeo et al., 2017) who are moving away from their in-house fNIRS data processing software options, and are increasingly using motion correction tools over trial rejection.

To date, several motion correction methods have been published and their performances have been compared in previous studies using both simulated and real data (Brigadoi et al., 2014; Cooper et al., 2012). These studies have highlighted that (i) correction is always better than trial rejection; (ii) Spline interpolation (Scholkmann et al., 2010) and Wavelet filtering (Molavi and Dumont, 2012) are the most effective methods in terms of recovering the hemodynamic responses affected by motion artifacts (Brigadoi et al., 2014; Cooper et al., 2012). It is worth noting that all these studies used adult data, both to validate the motion correction technique and to compare the performance of a set of correction techniques.

However, motion artifacts in infant data (and in some clinical adult populations) occur more frequently, sometimes sequentially, and encompass a wider amplitude range compared to what is seen in typical adult datasets. In adult experiments, movement is usually kept to a minimum by simply giving specific instructions to the participants. Therefore, in adult datasets, artifacts are relatively rare and easy to identify. On the contrary, infants cannot be instructed to remain still and, apart from limited cases, motion artifacts typically affect the entire recording. For example, movements occur during bouts of fussiness, boredom or even excitement, which can result in the recording of multiple motion artifacts within a short time window; the variability and the unpredictability of these behaviors, coupled with their high frequency of occurrence during the recording, makes it hard to identify their effects and correct the resulting artifacts that corrupt the data.

Since all motion correction techniques have been validated on adult data and given the substantial difference between the rate of motion artifacts typical in infant compared to adult studies, a worthy question is whether these techniques will show the same performance when applied to datasets with frequent and unpredictable motion artifacts.

To our knowledge, only two studies compared motion correction methods on data from youngsters, one with infant participants (Behrendt et al., 2018) and another one with children (Hu et al., 2015). Compared to previous work on adult and simulated data (Brigadoi et al., 2014; Cooper et al., 2012; Hu et al., 2015), these studies introduce a new approach to motion correction that involves the combined use of two correction techniques previously tested and used separately. This choice is driven by the variability of motion artifacts typically seen in data from young participants: combining two techniques that target different types of motion artifacts should outperform the use of each technique on its own. Specifically, Behrendt and colleagues (2018), compared the performance of Wavelet filtering, targeted Principal Component Analysis (tPCA) and their combination on infant data; their findings indicated that Wavelet performs best across all the datasets (4 datasets), task types (video vs. live stimuli presentation) and age groups involved (five, seven and twelve-month-olds with $N = 20$ in each age group, six-to-eight-month-olds with $N = 10$). Moreover, Hu et al. (2015), demonstrated that for their sample and task (six-to-twelve-year-olds with $N = 12$; language event-related design), the combined use of Wavelet and Moving Average (MA) techniques was preferable to the use of each method on its own, or to other methods (e.g., Spline interpolation, PCA, correlation based signal improvement — CBSI). Although both works, in line with findings on adult data, confirmed that Wavelet filtering performs well on more noisy data, they did not show a clear advantage in using a combination of motion correction techniques on this type of data. Indeed, the combination of Wavelet with tPCA did not outperform Wavelet alone (Behrendt et al., 2018). Further, it should be noted that the MA method used by Hu and colleagues (2015) serves the function of a high pass filter (removes slow drifts in the data) and for this reason it cannot be considered a motion correction method such as Wavelet or tPCA. Slow drifts removal is a processing step commonly implemented also in works that discard trials affected by artifacts (e.g., Grossman et al., 2008; Lloyd-Fox et al., 2009).

While the approach of combining more than one motion correction technique holds great potential for dealing with infant data, it is possible that the optimal combination(s) has not yet been identified. Given the consistently efficient performance of Wavelet across age-groups, it is sensible, moving forward, to test new combinations that include this method. With the present work we aim to take a step in this direction and assess the performance of a still untested combination of motion correction techniques: Wavelet filtering and Spline interpolation. The selection of these two methods was guided by previous recommendations from work on adult, children and simulated data (Brigadoi et al., 2014; Cooper et al., 2012; Hu et al., 2015) that compared most of the correction techniques available in Homer2.

Besides introducing this new combination, this work also tests and compares the independent performance of a wide range of motion correction techniques currently available in Homer2 (and of the proposed combination) on infant semi-simulated data: Spline interpolation, Wavelet filtering, tPCA, Wavelet Kurtosis, and Spline Savitzky-Golay.

In brief, the present work has a two-fold aim: 1) test and compare the independent performance of several motion correction techniques on

their own and of a novel combination (Spline + Wavelet) on infant semi-simulated data; 2) investigate the performance of Spline, Wavelet, and their combinations on three different infant datasets. These datasets were collected at different research labs, with different tasks, age-groups, headgears and NIRS acquisition systems. Regarding the analysis on the semi-simulated data we hypothesize that Wavelet and tPCA will perform similarly (Behrendt et al., 2018) and that the new combination will potentially outperform techniques applied individually. If this hypothesis is verified, we also expect that the improved performance of this new combination of motion correction techniques will be greater in signals containing a higher percentage of motion artifacts. Since Wavelet Kurtosis (Chiarelli et al., 2015) and Spline Savitzky-Golay (Jahani et al., 2018) are more recent techniques that have never been investigated in a comparative work, a-priori hypotheses cannot be advanced on their performance. With regard to the second part of this work, we hypothesize that the efficacy of each correction method will be highly dependent on the amount and nature of motion artifacts, which in turn might be related to task design, age group, and scalp-optode coupling (that could also be associated with the headgear design) and we expect that the new combination will provide improved or at least similar performance than when applying each technique alone.

The ultimate goal of the present study is to contribute to the common aim of the standardization of infant data preprocessing. The results of this work will provide guidelines for the developmental community for the analysis of infant fNIRS data.

2. Materials and methods

2.1. Infant fNIRS datasets

In this work, we analyzed data collected in four, independent and unrelated studies, each study contributing one dataset. For convenience, we refer to each dataset using numbers.

2.1.1. Dataset 1

Data from twelve 11-month-old infants (7 girls, $M_{age} = 347$, $SD_{age} = 10.8$, range = 361-331 days) was retrospectively selected from a group of infants recruited for a longitudinal study at the Centre for Brain and Cognitive Development (CBCD), Birkbeck, University of London (UK) with a resting state protocol while participants were awake. The data was selected from the first time point of the study at 11 months, as it was the closest to the other datasets. Data from seven additional participants was rejected for this analysis because the recording was not long enough for the purpose of this part of the study or there was excessive noise for any of the motion correction techniques to recover any trials (and therefore we would only be measuring noise and not motion correction performance).

The fNIRS data was acquired with the NTS diffuse optical imaging system (Gowerlabs Ltd UK; Everdell et al., 2005). The infants wore a custom-made headgear set consisting of three source–detector arrays located over temporal and frontal regions. 12 source- and 12 detector-optodes were embedded in a custom-built fNIRS headgear, with source-detector separations of 2.5 cm over the temporal regions and 3 cm on the frontal regions.

During data acquisition, the infants sat on their parent's lap in a dimly lit and sound-proofed room in front of a 46-inch plasma screen showing a screensaver video with coloured bubbles accompanied by relaxing music, with no identifiable shapes or social stimuli.

Dataset 1 was used to create semi-simulated data, in order to test the performance of the motion correction techniques alone (i.e., Spline, Wavelet, tPCA, Wavelet Kurtosis, and Spline Savitzky-Golay) and the combinations of Spline and Wavelet on a dataset with a known hemodynamic response. Three different semi-simulated datasets were created starting from the same resting state data, by adding different types of hemodynamic responses. These three sets of data (see Fig. 3) were designed to simulate a wide range of data typically collected in infant studies using different stimuli presentation lengths: (1) a 20 s HRF, corresponding to a stimuli presentation of about 10 s (hereafter referred as "Standard HRF"); (2) a 40 s HRF, to simulate longer stimuli presentation times ("Block design HRF" in the text); (3) a 12 s HRF, simulating shorter stimuli presentation that is typical of also event-related designs ("Short HRF" in the text). Hemodynamic responses were simulated by a linear combination of two gamma-variant functions (Abdelnour and Huppert, 2009), the parameters of which were tuned so as to allow small variations in peak amplitude and latency between trials. For the "Standard HRF" dataset, this led to a peak HRF amplitude of 1.43 ± 0.03 μM for HbO and -0.68 ± 0.01 μM for HbR; for the "Block Design HRF" dataset, to a peak HRF amplitude of 1.30 ± 0.01 μM for HbO and -0.65 ± 0.005 μM for HbR and for the "Short HRF" dataset, to a peak HRF amplitude of 1.53 ± 0.05 μM for HbO and -0.59 ± 0.02 μM for HbR. A minimum of 5 and a maximum of 8 HRFs per participant were added in the "Standard HRF" dataset, a minimum of 3 and a maximum of 6 in the "Block Design HRF" dataset and a minimum of 7 and a maximum of 8 in the "Short HRF" dataset, with a variable inter-trial interval (ITI), always long enough to allow the HRF to come back to baseline. ITIs were selected randomly from a normal distribution with mean (8+α) s and standard deviation 2 s. The α value was the length of the simulated HRF in that dataset.

2.1.2. Dataset 2

Sixteen 4- to 6-month-old infants (6 girls, $M_{age} = 154$, $SD_{age} = 25.7$, range = 115–205 days) were retrospectively selected from a group of infants who participated in a fNIRS study at the Centre for Brain and Cognitive Behaviour (CBCD), Birkbeck, University of London (UK) (Lloyd-Fox et al., 2018). The aim of the original study was to examine early brain responses to social and non-social stimuli in two groups of infants, one with increased familial risk for later development of Autism Spectrum Disorder (ASD) and the other one not. The total number of participants included in the original study was 36, and, for the present study, we selected the 16 infants included in the low-risk group (i.e., infants with no increased risk for later ASD). A further 9 infants participated in the study but were excluded because they did not attend to enough trials. We chose to include only the low risk group to avoid any bias in the interpretation of the results (see paragraph Metrics of Comparison). Recruitment, ethical approval (UK National Health Service National Research Ethics Service London REC 08/H0718/76 and 06/MRE02/73) and informed consent, as well as background data on participating families with high- and low-risk infants, were made available for the ASD study through the BASIS network (http://www.basisnetwork.org). All methods and experimental protocols were approved and carried out in accordance with the NHS and Birkbeck, University of London Ethics Committee guidelines and regulations. Informed consent was obtained from the parent/legal guardian for each participant.

Infants wore custom-built fNIRS headgear consisting of two source–detector arrays, containing a total of 26 channels (source–detector separations: 2 cm; see Fig. 1a), and were tested with the NTS diffuse optical imaging system (Gowerlabs Ltd UK; Everdell et al., 2005), the same instrument used for collection of Dataset 1 with a 10 Hz sampling rate. This system used two continuous wavelengths of source light at 770 and 850 nm. The two fNIRS arrays were placed bilaterally on the participants' head and covered from the inferior frontal to the posterior temporal regions (Lloyd-Fox et al., 2018). During data acquisition, infants sat on their parent's lap facing a 46-inch plasma screen situated about 100 cm away where visual stimuli were displayed. Hidden behind the screen, two speakers played auditory stimuli. The experimental condition trials, displayed for 9–12 s each, alternated one after the other, with a reference trial (silent presentation of visual static non-social images, also displayed for 9–12 s) between each. During the three types of experimental conditions (silent, S; auditory vocal, V; auditory non-vocal, N) the infants were looking at videos displaying a social interaction. In between two consecutive experimental conditions, static images of vehicles were displayed with no sounds. The conditions were presented in the same order across infants until the infants became bored or fussy as

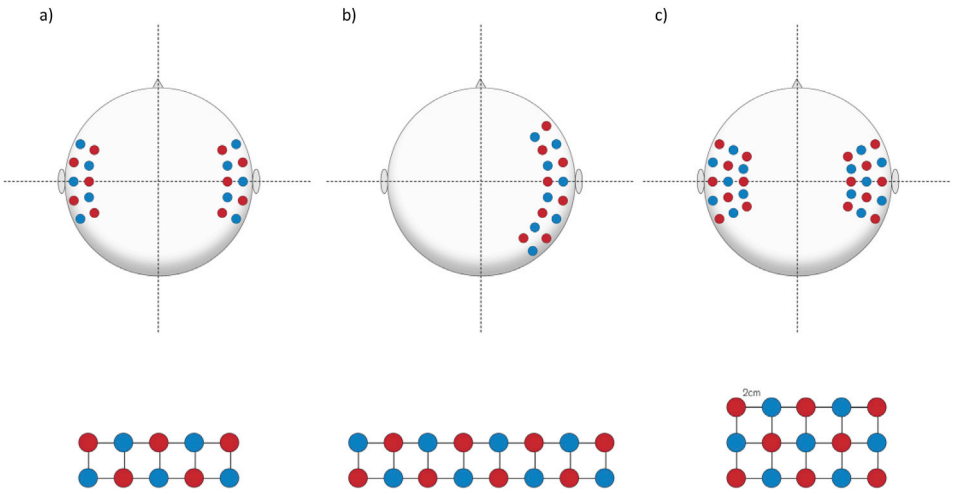

Fig. 1. Probe layout and locations of the channels for a) Dataset 2, b) Dataset 3, and c) Dataset 4. Red dots represent sources and blue dots indicate detectors.

judged by the experimenter who was monitoring their behaviour (see Lloyd-Fox et al., 2018 for more details).

2.1.3. Dataset 3

Seventeen 5-month-olds (9 girls; $M_{age} = 163$, $SD_{age} = 12.7$, range = 127–182 days) took part in the experiment conducted at Utrecht University, The Netherlands; two additional infants were excluded because they viewed less than three trials per condition. Both parents gave informed consent prior to participation. The Medical Ethical Committee of the University Medical Centre of Utrecht approved the study (protocol number: NL50617.041.14, METC 14–526), which was conducted in accordance with the Declaration of Helsinki.

The fNIRS data was acquired with the NTS diffuse optical imaging half-system (NTS2 – Gowerlabs Ltd UK; Everdell et al., 2005), consisting of eight dual-wavelength sources (780 nm, 850 nm) and eight detectors, which form 22 source-detector channels at a separation of 2 cm (see Fig. 1b). Data were sampled at a frequency of 10 Hz. The probe array was placed over the infants' right hemisphere, covering parts of the occipital, temporal and frontal cortices.

During the experiment infants sat on their parent's lap at ~60 cm from a 23-inch computer monitor (refresh rate 60 Hz, 1920 × 1080 resolution), in a dimly lit room. The task consisted in alternating 5-s experimental trials displaying sequences of five female pictures posing either happy or fearful expressions interleaved with ~10 s baseline trials displaying sequences of at least 10 houses (the end of the baseline trial was controlled by the experimenter). Every stimulus was displayed for 800 ms and followed by a 200 ms inter-stimulus interval showing a fixation cross; the order of the stimuli was randomized within trials (see Di Lorenzo et al., 2019 for more details).

2.1.4. Dataset 4

Twenty-two 10-month-old infants (13 girls; $M_{age} = 332$, $SD_{age} = 22.9$, range = 281–370 days) took part in the experiment; eight additional infants took part in the experiment but were excluded from the study owing to fussiness. This experiment was carried out at Keio University, Tokyo, Japan.

Parents gave informed consent in compliance with a protocol approved by the ethic committee of Keio University, faculty of letters (14034-0-2).

The fNIRS data was acquired with the Hitachi ETG-7000 (Hitachi, Tokyo, Japan), consisting of sixteen dual-wavelength sources (780 nm, 830 nm) and fourteen detectors, which formed 44 source-detector channels at a separation of 2 cm (see Fig. 1c). Data were sampled at a frequency of 10 Hz. The probe array was placed over the infants' right and left hemispheres, covering parts of the parietal, temporal and frontal cortices. For the work in this paper, we only analised data from the right emisphere (22 channels).

During the task infants sat on their parents' lap in front of a table. An experimenter sat next to the infant and delivered brushstrokes on the infant's right forearm at two different speeds: 3 cm/s or 30 cm/s. A second experimenter sat in front of the infant and distracted him/her with toys to prevent the infant from directing his attention to the tactile stimulation. Each experimental stimulus lasted 10 s and was followed by a baseline trial, where no touch was applied, with a duration of either 10 or 15 s. The order of the stimuli was fully randomized.

2.1.5. Dataset comparison: contamination by motion artifacts

Since infant data are likely to be contaminated more than adult data by motion artifacts, we expected different performance of the motion correction techniques depending on the amount of artifacts. Therefore, as first step, we evaluated the degree of motion artifact contamination in each dataset. To this end, we quantified for each channel and subject of each of the four datasets the percentage of signal identified as motion artifact (prior to correction) relative to the total duration of the signal. Motion artifactual sections of the signal were identified as described in section 2.2.1.

2.2. Motion correction techniques[2]

The main goal of this work is to investigate on different infant datasets and acquisition systems whether the combination of Spline and Wavelet overcomes the use of these techniques alone and of other techniques currently available in Homer2. We included the two possible combinations of Spline and Wavelet to examine whether the sequential order of application of the functions influenced their performance (hereafter referred to as Spl + Wav, with the opposite-ordered combination referred as Wav + Spl). In particular, we tested on Dataset 1 (semi-simulated data with a known hemodynamic response) the performance of the following methods: trial rejection, Spline, Wavelet, Spl + Wav, Wav + Spl, tPCA, Wavelet Kurtosis, and Spline Savitzky-Golay. As Wavelet requires the user to set specific tuning parameters, we further investigated on Dataset 1 whether different parameters can influence the performance of this motion correction technique. Based on the results of Dataset 1, we chose to evaluate the performance on real data of trial

[2] All analyses presented in this work were run using Homer2. To aid Homer2 users, in the main text we always report the Homer2 abbreviations for functions and parameters. However, it is important to note that the same functions can be implemented in processing streams outside of the Homer2 environment, where different naming conventions are used.

Table 1
Parameters used in hmrMotionArtifactByChannel and hmrMotionArtifact functions for each dataset.

	Dataset 1	Dataset 2	Dataset 3	Dataset 4
tMotion	1	1	1	1
tMask	1	1	1	1
STDEVthresh	13	15	15.5	13.5
AMPthresh	0.4	0.4	0.4	0.4

rejection, Spline, Wavelet, Spl + Wav, Wav + Spl, which were tested using Datasets 2, 3, and 4.

2.2.1. Trial rejection

Trial Rejection consists in the exclusion of trials contaminated by motion artifacts from further analysis. The excluded trials were pre-selected in a channel-by-channel mode by the Homer2 function hmrMotionArtifactByChannel. This function detects the signal exceeding a threshold in change of amplitude (AMPthresh) or/and a threshold in change of standard deviation (STDEVthresh) within a predefined time-window (tMotion) and marks as artifacts the data points around the detected motion (+/− tMask). After the detection of artifacts, a second step rejects the corrupted trials from all channels (not on a channel-base). All parameters for this function are defined by the user. For each dataset, we selected values that identified the majority of spike-like motion artifacts (see Table 1). In particular, values of standard deviation thresholds were decided after visual inspection of each dataset in order to optimize motion detection and avoid both the over-identification of signal as noise and to miss important artifacts. This same approach was used in Cooper et al. (2012).

2.2.2. Spline

Spline interpolation, first proposed by Scholkmann et al. (2010), is a channel-by-channel correction method that acts on previously detected motion artifacts (hmrMotionArtifactByChannel; see description in the Trial Rejection section). After motion detection, the spline function corrects the artifact by performing a cubic spline interpolation of the artifact; the interpolation is then subtracted from the original signal. After this, the signal is baseline corrected to ensure that signal time-course before and after the corrected artifact is continuous. Spline interpolation depends on a parameter (p) that can be set by the user; in this study we used $p = 0.99$, the same value used by previous studies (Brigadoi et al., 2014; Cooper et al., 2012; Scholkmann et al., 2010). The positive aspect of this correction method is that it only corrects the pre-localized artifacts without modifying the other portions of the time-series. However, it depends heavily on the ability of the motion detection step in identifying motion artifacts and hence on the parameters set in the motion detection function (hmrMotionArtifactByChannel).

2.2.3. Wavelet

The wavelet filtering available in Homer2, first proposed by Molavi and Dumont (2012), is a function that detects and corrects artifacts channel-by-channel in a single step. This function decomposes the signal time-course of every channel in a series of wavelet detail coefficients which are characterized by a Gaussian distribution: while the coefficients linked to the physiological components (NIRS signal of interest) will be distributed around zero, the coefficients reflecting motion artifacts can be identified as the outliers of the Gaussian distribution. Then, by setting to zero all detail coefficients identified as outliers of the distribution (<first quartile - α times the interquartile range or > third quartile + α times the interquartile range) and reconstructing the signal with the modified coefficients (with the inverse discrete wavelet transform), we can obtain a version of the original signal with a much reduced presence of motion artifacts. In Homer2 the α threshold can be defined by setting the tuning parameter *iqr*. The advantage of the wavelet method is that it does not require a prior step of motion artifact detection, however the choice of *iqr* value will depend on the timing of the task paradigm and consequently on the characteristics of the evoked HRF. For instance, using a too low *iqr* will reduce and/or filter out the hemodynamic response itself in event-related designs. For Dataset 1, six different *iqr* values were employed (*iqr* = 0.1, 0.5, 0.8, 1.0, 1.2, and 1.5) to evaluate the influence of the choice of *iqr* on motion correction and HRF estimation. For Datasets 2, 3 and 4 we used *iqr* = 0.8, which was defined by visually inspecting the effects of this and the other *iqr* values on the group-averaged HRFs (i.e., 1.2, 1.0, 0.5, 0.1); also, this value yielded a good performance in recovering the hemodynamic response and the number of trials when tested on semi-simulated data.

2.2.4. tPCA

Targeted Principal Component Analysis (tPCA) was first described by Yücel et al. (2014). This motion correction method, like Spline interpolation, acts on previously detected motion artifacts (using the function hmrMotionArtifactByChannel) and is applied to artifactual parts of the signal time-course across all channels. First motion artifacts are detected. Then, segments of data containing motion artifacts are extracted from the original signal of all channels and concatenated together into a new single data matrix. The correction technique (PCA) is applied only to this dataset which mainly consists of epochs of motion ('targeted'). PCA applies an orthogonal transformation to this dataset composed of N measurements (number of channels) to produce N uncorrelated components, ordered by their contribution to the variance of the data. Thus, the first components will account for the largest proportion of the variance of the data. Since motion artifacts should constitute a large proportion of the variance of the data, the first M components should represent the variance caused by the motion artifacts. Removing the first M components from the signal should result in the correction of the motion artifacts (Zhang et al., 2005). The corrected segments are stitched back into the original signal. Analogously to Spline, the signal is baseline corrected to ensure that the time-course before and after the corrected artifact is continuous. This procedure can be reiterated for a number of times (as defined by the user) to allow for correction of any residual motion artifacts not corrected during the first iteration.

In Homer2, tPCA depends on two parameters that can be set by the user: the percentage of variance (nSV) and the maximum number of iterations (maxIter). In this study, we used nSV = 0.97, that is we removed 97% of the total variance, and maxIter = 5, the same values suggested by Yücel et al. (2014). The performance of PCA is directly dependent on the number of measurements available, i.e., on the number of channels. To test this dependency, we run tPCA on Dataset 1 twice, once by using the complete dataset and once by using only the channels made up by half of the available sources and half of the available detectors, to emulate a smaller dataset.

2.2.5. Wavelet Kurtosis

Kurtosis-based Wavelet filtering (Wav Kurt) was first proposed by Chiarelli et al. (2015) with the aim to overcome some of the limitations associated with wavelet filtering (i.e. in cases of high signal-to-noise ratio (SNR) this correction can lead to the reduction of signal amplitude). The novel idea introduced by the authors is that the wavelet coefficients generated from fNIRS signals have sub-Gaussian (kurtosis <3) or Gaussian (kurtosis = 3) distributions. In contrast, data affected by motion artifacts tend to have larger kurtosis values, reflecting the presence of outliers. Unlike the α value of wavelet filtering, the kurtosis value is independent of the SNR of the data.

Analogously to wavelet filtering, kurtosis-based wavelet filtering performs a wavelet transformation of the data and computes the distribution of the wavelet coefficients. The presence of motion artifacts is assessed by computing the kurtosis of this distribution and evaluating whether it is higher than a given threshold. If this is the case, the most extreme coefficients are set to zero, the kurtosis computed again and compared to the given threshold. This procedure iterates until the computed kurtosis is smaller than the given threshold. After this, the

inverse wavelet transformation is applied and the signal reconstructed. In this work, we used a kurtosis threshold of 3.3, as suggested in the original paper (Chiarelli et al., 2015).

2.2.6. Spline Savitzky-Golay

The combination of Spline interpolation with Savitzky-Golay (SG) filtering was recently proposed by Jahani et al. (2018). The authors intended to devise a single approach that could deal with different types of motion artifacts. They proposed the combination of Spline interpolation, which best corrects baseline shifts, followed by SG filtering, a method suited for correction of high frequency spikes. This algorithm, as shown in the original paper, works when the SNR of the data is greater than 3, while the authors suggest to apply only the SG filtering part of the algorithm when SNR <3. In this work, for the Spline part of the algorithm we employed the same parameters defined in sections 2.2.1 and 2.2.2. For the SG part of the algorithm, we set the frame size to 6 s, as in the original work (Jahani et al., 2018).

2.3. Data processing streams and values used in each dataset

Data was processed using the Homer2 package in MATLAB. We created different processing streams, each tailored for a specific correction method: trial rejection, Spline, Wavelet, Spl + Wav, Wav + Spl, tPCA, Wavelet Kurtosis, and Spline Savitsky-Golay (Fig. 2 reports the steps of all processing streams). We also processed the datasets without correcting for motion artifacts (No Correction approach).

For each processing stream, channels showing very high or low optical intensity readings were excluded from further analyses (using the function enPruneChannels). Intensity thresholds were differently chosen for each device according to the company recommendations. After that, the raw intensity data were converted to optical density (OD) changes. Then invalid trials (e.g., non-looking trials) were discarded. From this

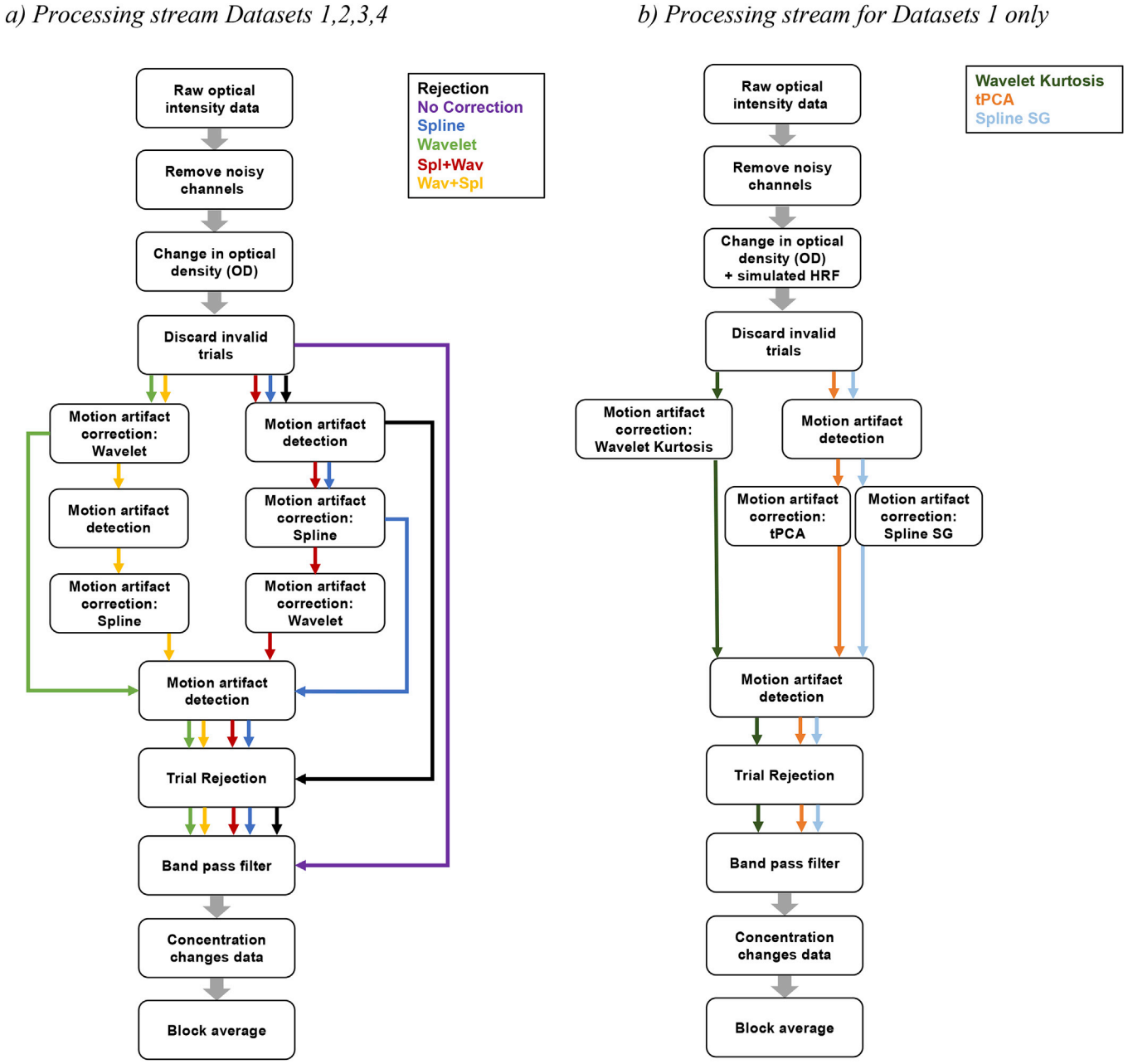

Fig. 2. Processing streams of the techniques applied to all datasets (a) and to Dataset 1 only (b). Each coloured arrow represents the specific workflow for each technique, while thicker grey arrows indicate the common processing steps.

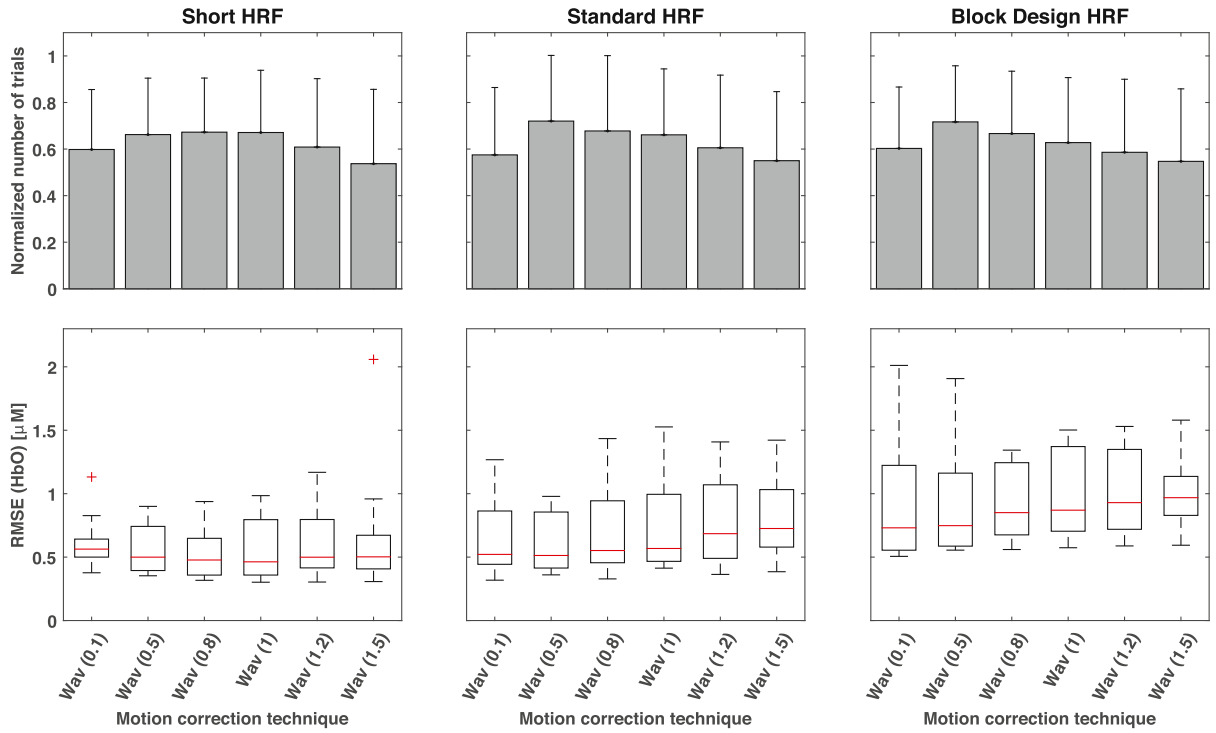

Fig. 3. Normalized number of trials (upper row) and RMSE (bottom row) recovered after applying the Wavelet motion correction technique with different iqr values for each set of data of Dataset 1. Error-bars in the upper panels represent standard error across participants. The red line in the box plots of the bottom panels depicts the median value, the two whiskers denote the first and third quartile and outliers are represented by red crosses.

step on, different streams were applied (see Fig. 2). After any correction method, motion detection, using hmrMotionArtifact, was applied to identify the remaining uncorrected motion artifacts; the parameters chosen for this function were the same as the ones used in hmrMotionArtifactByChannel (see Table 1). Note that motion detection was also part of the trial rejection stream. A subsequent step allowed the rejection of trials affected by motion artifacts, for all processing streams but the No Correction one. After this step, a band-pass filter (third order Butterworth) was applied to reduce slow drifts and high-frequency noise. Note that specific values for the cut-off frequencies of the band-pass filter were set for each dataset to maximize the reduction of noise without removing or corrupting the hemodynamic responses, taking into account the different stimulation timing of each experiment (e.g., Dataset 3, having the shortest stimuli presentation time allowed the use of a higher frequency for the low cut-off). Then the OD data were converted to concentration changes using the modified Beer–Lambert law (Cope and Delpy, 1988; Delpy et al., 1988) with differential pathlength factor of 5.1 (Duncan et al., 1995). Finally, all remaining trials were block-averaged for every condition, channel and participant. Table 2 reports the values of the parameters of the functions used in the processing streams used for each dataset.

2.4. Metrics of Comparison

In order to compare the performance of the motion correction techniques, we used four metrics: the hemodynamic response recovery error, measured with the root mean squared error (RMSE) (applied on Dataset 1), the within-subject standard deviation (applied on Datasets 2, 3, 4), the between-subjects standard deviation (applied on Datasets 2, 3, 4), and the number of trials that survived each correction method (applied on all datasets). As for the number of trials, the within-, and between-subjects SD, every motion correction technique was first compared to the No Correction approach (purple path in Fig. 2), which does not correct or reject trials affected by motion artifacts, and then to each other. Both results for HbO and HbR concentration changes were investigated.

For the semi-simulated Dataset 1, two types of comparisons were performed: the first investigated the impact of changing the *iqr* value in the Wavelet motion correction technique, while the second investigated the performance of all motion correction techniques in semi-simulation. Therefore, for the first comparison, the performance of the Wavelet technique was compared across *iqr* values (i.e., 0.1, 0.5, 0.8, 1.0, 1.2, 1.5), whereas when comparing performance across techniques (and the Wavelet-Spline combinations), only two options for the Wavelet parameter *iqr* were tested: $iqr = 0.5$ and $iqr = 0.8$. These two *iqr* values were chosen because they showed the highest performance in the recovery of the HRF and of the number of trials on the first comparison. Note that an *iqr* of 1.0 also performed well in these two metrics, but with a higher variability. In addition, the infant literature reports studies using *iqr* values of 0.5 (e.g., Ravicz et al., 2015). Therefore, we decided to further test only the two lower *iqr*s, 0.5 and 0.8.

2.4.1. Hemodynamic response recovery error

The RMSE between the true simulated HRF and the HRF recovered by each processing stream was computed for each of the three sets of data of Dataset 1, each participant and channel. The true HRF was computed as the average of the true single-trial HRFs for each participant and channel. Results are reported aggregated across channels. A lower RMSE is an index of improved HRF estimate.

2.4.2. Within-subject standard deviation

The within-subject standard deviation (SD) compares, for every participant, the mean of the standard deviation of the single-trial hemodynamic responses across channels and conditions. The assumption behind this metric is that most of the variability between single-trial hemodynamic responses within the same participant should be due to

Table 2
Parameters and values used in each processing stream for each of the four datasets. Abbreviations for functions and parameters used in Homer2 are in italics.

		Dataset 1			Dataset 2	Dataset 3	Dataset 4
		Standard HRF	Block Design HRF	Short HRF			
Channel rejection *enPruneChannels*	Intensity range *(dRange)*	1.00E-03 1.00E+07	1.00E-03 1.00E+07	1.00E-03 1.00E+07	1.00E-03 1.00E+07	2.00E-03 1.00E+07	9.00E-01 4.00E+05
	SNR threshold	0	0	0	0	0	0
	Source-Detector separation range *(SDrange)*	0.0 45.0	0.0 45.0	0.0 45.0	0.0 45.0	0.0 45.0	0.0 45.0
Spline interpolation *hmrMotionCorrectSpline*	*p*	0.99	0.99	0.99	0.99	0.99	0.99
Wavelet filtering *hmrMotionCorrectWavelet*	*iqr*	0.1/0.5/0.8/1.0/ 1.2/1.5	0.1/0.5/0.8/1.0/ 1.2/1.5	0.1/0.5/0.8/1.0/ 1.2/1.5	0.8	0.8	0.8
Targeted PCA	Variance amount % *(nSV)*	0.97	0.97	0.97	–	–	–
hmrMotionCorrectPCArecurse	Max # of iterations *(maxIter)*	5	5	5	–	–	–
Wavelet Kurtosis *hmrMotionCorrectKurtosisWavelet*	Kurtosis threshold *(kurt)*	3.3	3.3	3.3	–	–	–
Spline Savitzky-Golay	*p*	0.99	0.99	0.99	–	–	–
hmrMotionCorrectSplineSG	*FrameSize_sec*	6	6	6	–	–	–
Trial rejection based on artifact presence *enstimRejection*	time Range *(tRange)*	−2.0 10.0	−2.0 10.0	−2.0 10.0	−2.0 10.0	−2.0 5.0	−2.0 10.0
Bandpass filter	*hpf*	0.01	0.01	0.01	0.01	0.03	0.025
hmrBandPass	*lpf*	1	1	1	1	0.8	1
Block average *hmrBlockAverage*	time Range *(tRange)*	−2.0 20.0	−2.0 40.0	−2.0 12.0	−2.0 20.0	−2.0 15.0	−2.0 20.0

the presence of motion artifacts.

2.4.3. Between-subjects standard deviation

The between-subjects SD was used to investigate the variability between the averaged hemodynamic responses across subjects for every channel and condition. Similarly to the within-subject SD, we assume that most of the variability between hemodynamic responses across subjects is due to the presence of motion artifacts.

2.4.4. Number of recovered trials

We also quantified the number of trials included in the averaged HRFs for every participant and condition, after the different correction methods were applied. The number of trials recovered by each technique for each participant was normalized on the original number of trials available in each participant (i.e. on the number of trials obtained with the No Correction approach).

2.5. Statistical analyses

Statistical analyses were performed using SPSS (IBM Corporation, Armonk, NY, USA).

The mean of the percentages of motion artifacts calculated across channels for each subject and wavelength was submitted to a one way ANOVA with dataset as between-subject factor, to evaluate differences in motion artifact contamination between datasets; independent samples t-tests were used to follow-up the significant difference among the datasets.

To compare the performance of the techniques, separate repeated measure ANOVAs with technique as within-subject factor were computed for the number of recovered trials (for all datasets), the RMSE (for Dataset 1) and the within-subject SD (for datasets 2, 3 and 4) metrics for both chromophores (HbO, HbR). For the RMSE and the within-subject SD, we calculated for every subject a unique value representative of each technique, which consisted in the mean of all values across channels and conditions. Main effects of technique were followed up with two-tailed paired sample t-tests to compare the performances of all the techniques to each other, within each dataset. The False Discovery Rate (FDR) criteria was employed to correct for multiple comparisons (note that all figures display FDR-corrected results).

2.6. Motion correction performance at different percentages of artifact contamination

A further analysis was performed on the semi-simulated data of Dataset 1 to evaluate the motion correction performance (indexed by the RMSE) at different percentages of motion artifact contamination. This analysis was devised to understand whether the different performance of each motion correction technique compared to the other was constant across the different percentages of presence of motion artifacts or diverged as this percentage increased.

For each subject and channel, we computed the percentage of motion artifact contamination as described in section 2.1.5, and for each subject, channel and motion correction method, we used RMSE as metric of its performance. The amount of available data at lower percentages of motion artifact contamination was higher, getting sparser at increased percentages of motion artifacts. Therefore, we fitted a linear model (B-spline with 3 degrees of freedom) on the RMSE values at the different percentages of motion artifacts and we predicted the RMSE values at the missing points, from 0 to 21% of motion artifact contamination, at steps of 0.5% (there were only few samples exceeding 21% of noise, not providing enough power to perform statistical analyses). Within this framework, we evaluated the confidence intervals of the estimate as well. For each pair of motion correction techniques, and at each estimated percentage of motion contamination, a *t*-test was performed to evaluate whether the performance of the two techniques differed statistically. We employed the Welch's *t*-test, which takes into account the unequal variance of the two populations. The standard deviation of the two populations at each percentage of motion artifact was computed knowing the estimated RMSE value and the confidence interval at that point (90% confidence level). The degrees of freedom were computed using the Welch–Satterthwaite equation. FDR was employed to correct for multiple comparisons (separately for each comparison between any two technique). All these analysis were performed with the RStudio software package (RStudio Team, 2016).

We hypothesize that at very low percentages of motion artifact contamination all techniques will perform similarly. Then, as the number of motion artifacts increases, standard techniques used for motion correction in adult datasets (e.g., Wavelet) should outperform the Rejection, No correction and Spline techniques. Further, as the number of

motion artifacts reaches levels compatible with infants datasets, the combination of Spline and Wavelet should outperform the other techniques, thus demonstrating both the importance of evaluating motion correction techniques in infants dataset and that in this type of datasets stacking two techniques could be ideal.

3. Results

3.1. Motion artifact quantification

The assessment of the percentage of motion artifacts per channel and participant revealed that Dataset 2 was the least noisy one, with a maximum of 12.5% of signal identified as motion artifact ($M = 2.21$, $SD = 2.72$, range = 0–12.5%), while the noisiest datasets were Dataset 1 ($M = 7.26$, $SD = 6.76$, range = 0–46.0%), and Dataset 3 ($M = 6.13$, $SD = 7.83$, range = 0–65.7%); Dataset 4 showed an intermediate level of noise ($M = 3.99$, $SD = 3.85$, range = 0–18.1%). Note that the amount of motion artifact contamination in a signal is dependent on the size of the temporal mask employed (tMask).

The one-way ANOVA (with *Dataset* as between subject factor) showed a significant difference between the percentage of artifacts in each dataset, $F(3,70) = 11.05$, $p < .001$, $\eta_p^2 = 0.32$. Post-hoc t-tests confirmed what previously described, that is Dataset 2 was less corrupted by motion artifacts than Dataset 1 ($t(19.6) = -5.98$, $p < .001$), Dataset 3 ($t(17.1) = -4.14$, $p = .002$), and Dataset 4 ($t(27.6) = -3.83$, $p = .002$); whereas a similar percentage of artifacts was observed between Datasets 1 and 3 ($t(34) = 0.91$, $p = .37$), and Datasets 3 and 4 ($t(22.7) = 2.09$, $p = .058$). Also, Dataset 1 had a higher percentage of artifacts compared to Dataset 4 ($t(27.3) = 3.52$, $p = .003$).

3.2. Motion correction performance in semi-simulation (dataset 1)

3.2.1. Tuning parameters of wavelet

Fig. 3 reports the performance of the Wavelet motion correction technique across *iqr* values for both the normalized number of recovered trials and the RMSE metric. Results for each set of data of Dataset 1 are reported in the three panels.

The number of recovered trials was significantly different among *iqr* values in both the "Standard HRF" and the "Short HRF" sets of data ($F(5,55) = 3.21$, $p = .013$ and $F(5,55) = 3.51$, $p = .008$, respectively) but not in the "Block Design HRF" set ($F(5,55) = 1.59$, $p = .177$). In particular, there was a trend towards a higher trial recovery with $iqr = 1$ compared to $iqr = 1.2$ or 1.5 and with $iqr = 0.8$ compared to $iqr = 1.5$ in both the "Standard HRF" and the "Short HRF" sets of data. Furthermore, there was a trend towards a higher trial recovery with $iqr = 0.5$ compared to $iqr = 0.1$ in the "Standard HRF" and compared to $iqr = 1.5$ in the "Short HRF" set of data (min $t = 2.65$, max $p = .082$, corrected, df = 11).

RMSE was significantly different among *iqr* values in the "Standard HRF" set of data ($F(5,55) = 3.18$, $p = .014$), although no significant differences emerged from the t-tests comparing the different techniques. No significant differences between techniques were found in the "Short HRF" and in the "Block Design HRF" sets of data ($F(5,55) = 0.85$, $p = .52$ and $F(5,55) = 0.49$, $p = .78$, respectively).

3.2.2. Performance of correction techniques in simulated HRFs

Fig. 4 reports the number of trials recovered after applying each motion correction processing pipeline normalized to the original number of trials (equivalent to the number obtained with the No Correction approach) for each participant. Results for each set of data of Dataset 1 are reported in the three panels. Note that the results relative to the

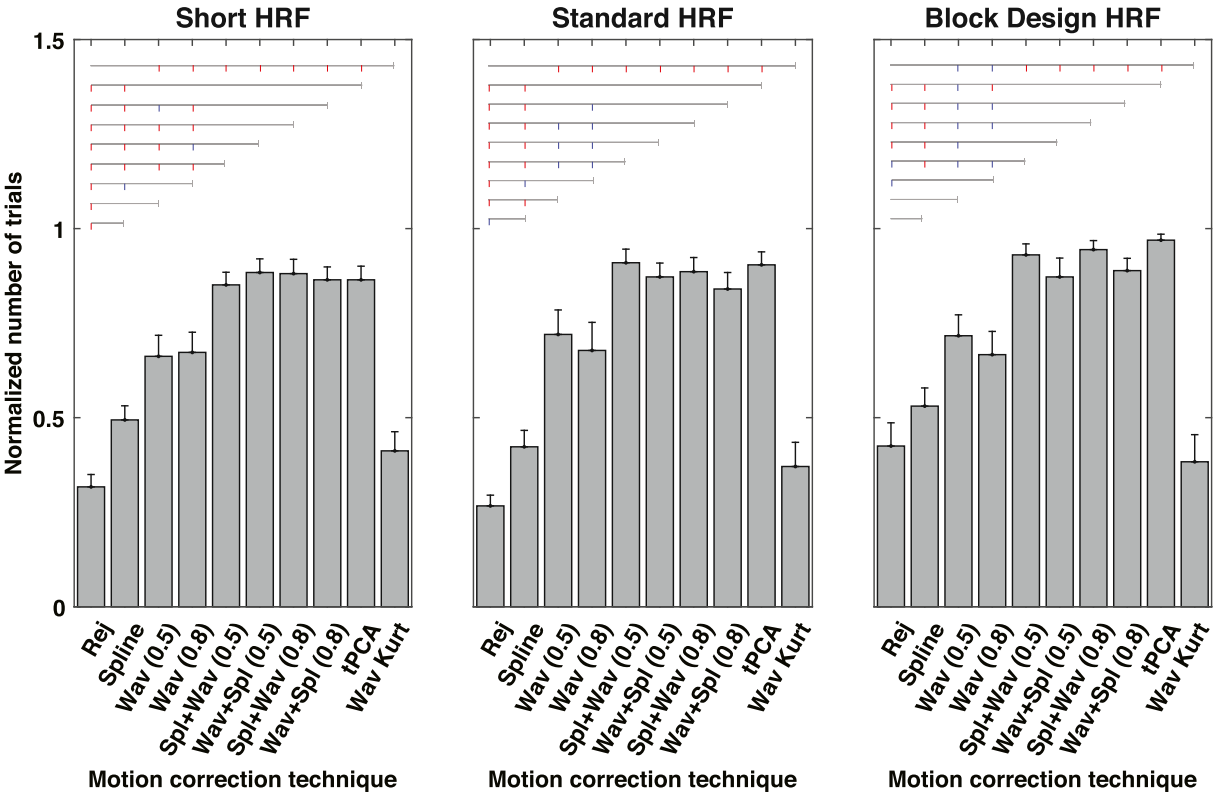

Fig. 4. Normalized recovered number of trials for each processing pipeline and for each of the three sets of data of Dataset 1. Error-bars indicate standard error across participants. The lines above linking the different techniques indicate significant statistical difference: blue lines correspond to $p < .05$ and red lines to $p < .01$. Degrees of freedom = 11.

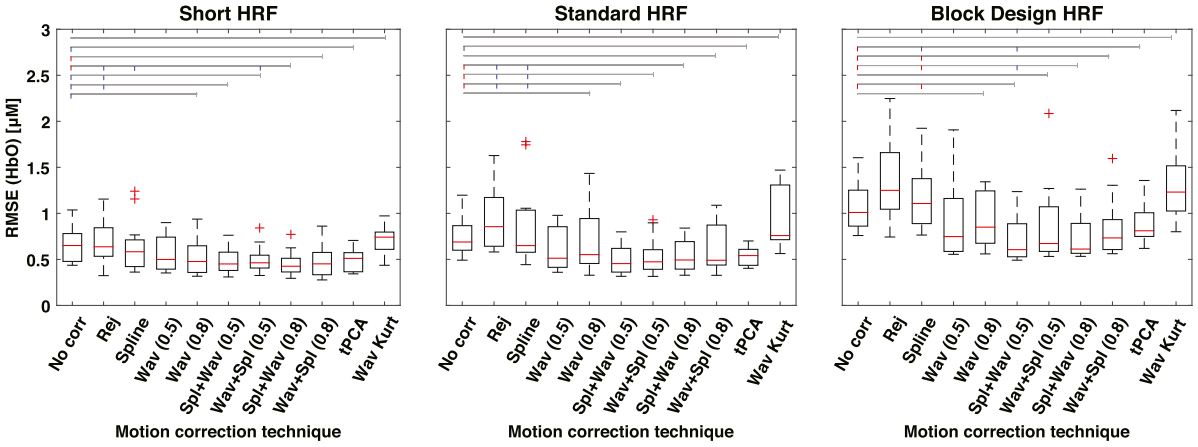

Fig. 5. RMSE for each processing pipeline and for each of the three sets of data of Dataset 1. The red line in the box plots depicts the median value, the two whiskers denote the first and third quartile and outliers are represented by red crosses. The lines above linking the different techniques indicate the significant statistical difference: blue lines correspond to $p < .05$ and red lines to $p < .01$. Degrees of freedom = 11.

performance of Spline Savitzky-Golay are not shown in this and the following figures since the application of this technique had a strong impact on the signal, causing the rejection of all the trials in all subjects.

Regarding the remaining correction methods, the number of recovered trials was significantly different among techniques in all three sets of data (min $F = 15.3$, *all* $ps < .001$). The results of the t-tests between techniques are reported in Fig. 4 and are consistent across sets of data. In general, while Rejection and Wavelet Kurtosis are the techniques that significantly recover less trials, tPCA and the combinations of Spline and Wavelet (regardless of order and *iqr* value) are the approaches recovering almost all trials.

Fig. 5 reports the RMSE obtained after applying each motion correction processing pipeline (see Figure S2 for the HbR results). Results for each set of data of Dataset 1 are reported in the three panels. RMSE was significantly different among techniques in all three sets of data (min $F = 2.03$, max $p = .037$). Pairwise t-tests between techniques are reported in Fig. 5. For all sets of data, trial rejection, No Correction, and Wavelet Kurtosis perform worse. tPCA and Wavelet showed similar performance levels, with Wavelet scoring a slightly lower median RMSE compared to tPCA, although this is not statistically significant. However, tPCA showed a smaller interquartile range compared to Wavelet. The performances of the combinations of Spline and Wavelet were also comparable with those of Wavelet and tPCA, however, the combined approach reported the lowest median RMSE across all HRF types and, for the "Block Design HRF", the hemodynamic response was significantly better recovered by Spl + Wav (0.5) than tPCA.

Further analyses on tPCA, comparing the results of tPCA using all available channels and tPCA using approximately half the number of channels, showed that its performance is dependent on the number of channels available, although this difference was not statistically significant, likely due to the low power of this analysis. We report these analyses in the supplementary materials (S1.3).

Overall, the method scoring the lowest median RMSE for all sets of data was the combination of Spline (performed at first step) and Wavelet (performed at second step), which therefore proved to be a robust method across different types of hemodynamic responses. Interestingly, in the "Short HRF" set of data, the lowest RMSE was scored by this combination when using $iqr = 0.8$, while for the "Standard HRF" and the "Block Design HRF" when using $iqr = 0.5$. This result is in line with the formulation of Wavelet approach: lowering the iqr, indeed, sets to 0 more detail coefficients, increasing the chances to reduce or even remove the actual hemodynamic response. Detail coefficients represent the output of the different high-pass filters performed during wavelet decomposition. In the "Short HRF" set of data, the hemodynamic response has higher frequency components than in the "Standard HRF" and in the "Block Design HRF" sets of data. Therefore, it should be hypothesized that it is more likely that some of the detail coefficients contain information related to the hemodynamic response in the "Short HRF" dataset and therefore a higher threshold for rejection should be selected.

Fig. 6 reports example individual average HRFs for one channel recovered after applying the different motion correction techniques on the three sets of data (see Figure S3 for the HbR responses).

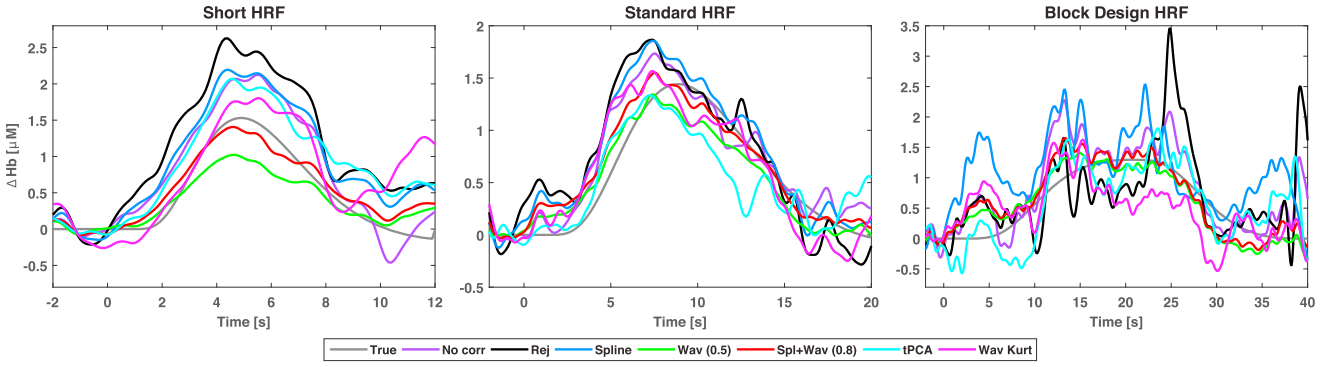

Fig. 6. Examples of individual average HbO responses for one channel recovered after applying different motion correction approaches on the three sets of data of Dataset 1. Only some of the tested techniques are reported for visualization purposes. The missing techniques, Wavelet with $iqr = 0.8$, Spl + Wav with $iqr = 0.5$ and Wav + Spl with both *iqrs* performed similarly to Spl + Wav (0.8), Wavelet (0.5), Wavelet (0.5) and Spl + Wav (0.8), respectively. The true simulated HRF is displayed in grey.

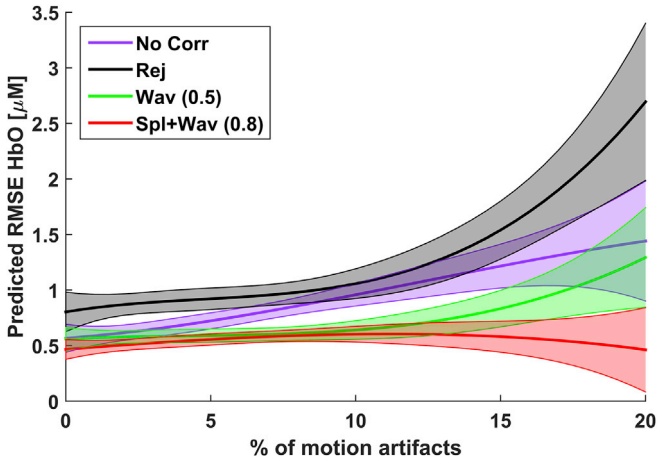

Fig. 7. Estimated RMSE HbO values at different percentages of motion artifact contamination for a selection of motion correction methods. Rejection, No Correction, Wav (0.5) and Spl + Wav (0.8) were selected for visualization purposes. The shaded areas represent the confidence intervals.

3.2.3. *Motion correction performance at different percentages of motion artifact contamination*

The performance of almost all motion correction techniques was dependent on the amount of motion artifacts identified in the signal, with a worsening of the performance at increased percentages of motion artifact contamination (see Fig. 7 reporting HbO results, and Figure S5 for the HbR results). This decline in the performance was smaller for Wavelet, tPCA and above all for the combination of Spline and Wavelet (with Spline applied before Wavelet).

Fig. 8 displays the corrected p values resulting from the Welch's *t*-test at three exemplary percentages of motion artifacts (0%, 5% and 15%) for each comparison between techniques for HbO (see Figure S6 for the HbR results). As expected, when no motion artifacts were detected only few comparisons resulted statistically significant. At increasing percentages of motion artifact contamination, the performance of the different techniques started to diverge, with Wavelet (at $iqr = 0.5$), the stacked techniques and tPCA showing statistically significant better performances compared to the other techniques. Note that 0% corresponds to the percentage of artifacts identified by the detection process rather than the absence of artifacts; it is possible that some subtler artifacts were not captured.

3.3. *Motion correction performance in task-based datasets (datasets 2, 3, 4)*

For reasons of brevity and since the results for HbO and HbR were similar, in the main manuscript we report only the HbO results; HbR results are reported in the supplementary materials (S2). For each dataset, repeated measures ANOVA analysis of within-subject SDs revealed a main effect of *technique* for all datasets (all $ps < .0001$). In general, across all datasets, No Correction, Rejection and Spline performed poorly in reducing the within-subject SDs, while Wavelet and the two possible combinations of Spline and Wavelet reduced more effectively this measure. However, subtle differences can be observed between datasets. For instance, for Datasets 3 and 4 Spl + Wav was more effective in SD reduction, while for Dataset 2 Wav + Spl performed best. Fig. 9 reports the p-values of all the t-tests performed between each technique. Both combinations of Spline and Wavelet recovered most of the trials affected by artifacts in all datasets (see Fig. 10).

3.3.1. *Performance of rejection versus No correction*

The between-subjects SD scatter plots (first column of Fig. 11) show that, for every dataset, rejecting trials performs worse than including noisy trials, increasing the between-subjects SD among the subjects'

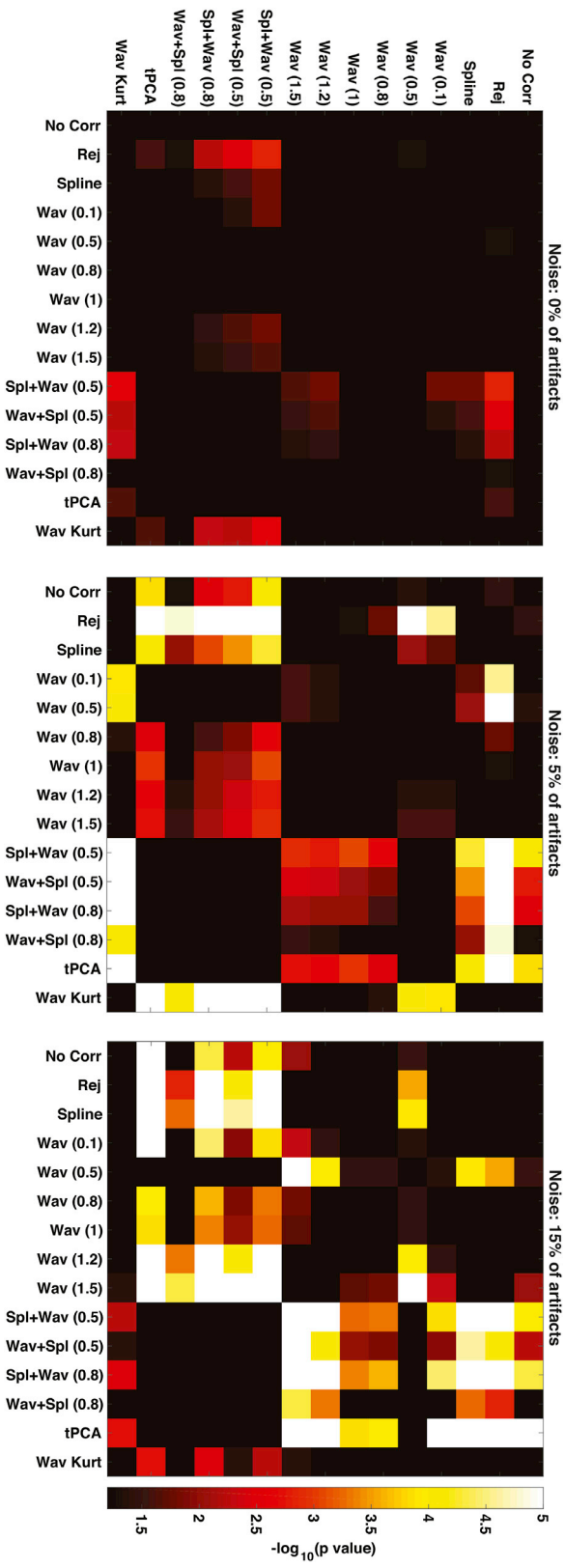

Fig. 8. Corrected p values resulting from the paired t-tests computed between any pair of correction methods on the RMSE HbO values of the "Standard HRF" dataset at 0%, 5%, and 15% of artifact contamination. For visualization purposes, we selected only 3 levels of noise for one HRF type. The colorbar represents -log10(p value) for visualization purposes (the value 1.2, which is the lower limit, corresponds to $p = .05$, higher values correspond to $p < .05$, the higher the more significant).

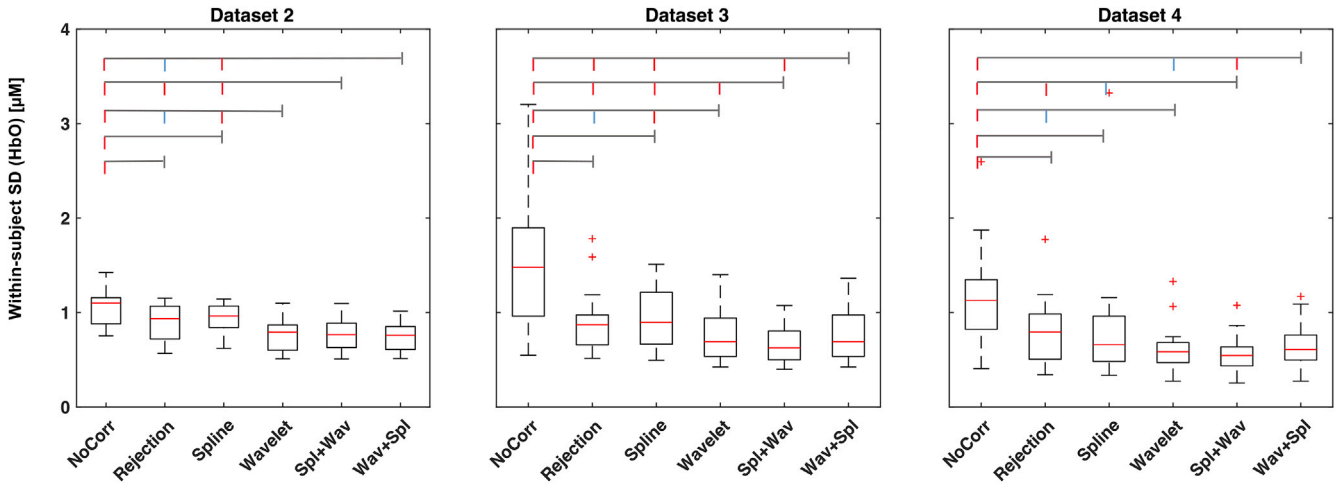

Fig. 9. Box plots of the within-subject SD calculated for HbO of all techniques of Datasets 2, 3 and 4. The red line in the box plots indicates the median, while the two whiskers denote the first and third quartile. Outliers are represented with red crosses. The lines above linking the different techniques indicate significant statistical differences: blue lines correspond to $p < .05$ and red lines to $p < .01$. Degrees of freedom: Dataset 2, min = 13, max = 15; Dataset 3, min = 14, max = 16; Dataset 4, min = 16, max = 21.

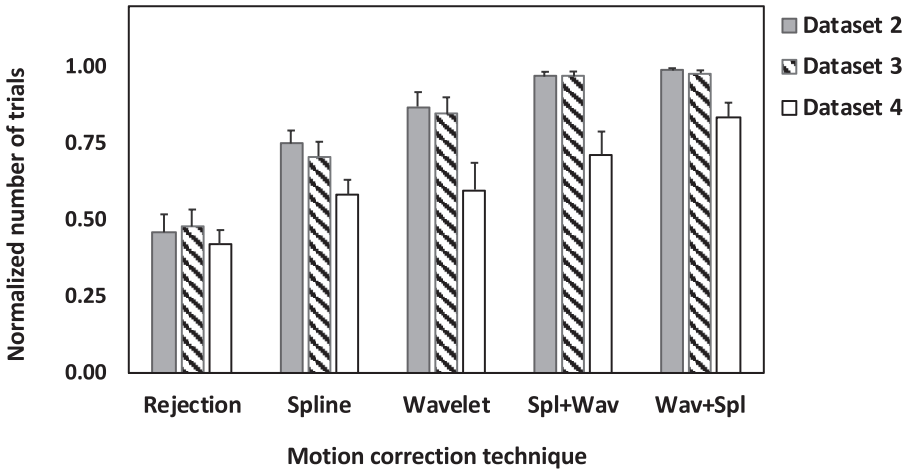

Fig. 10. The bars represent the mean number of trials averaged for each technique and dataset, normalized to the mean number of trials averaged when no motion correction was applied; the error bars indicate standard error (SEM).

mean HRF more than 60% of the time. This outcome could be the result of some very noisy mean HRFs measured from infants that had many trials rejected.

Instead, the rejection method has a better performance reducing the within-subject SD as compared to the No Correction approach (second column of Fig. 11); the rejection of trials affected by motion artifacts is indeed efficient in reducing the variability between trials within each subject. For all datasets, paired t-tests supported this and showed a significant difference between these two techniques (all *ps* < .01; see Fig. 9).

Rejection excluded more than 50% of the total presented trials in each dataset (Fig. 10).

3.3.2. *Performance of motion correction techniques versus No correction*

All the correction methods significantly reduced the within-subject SD across all datasets (Fig. 9; paired t-tests No Correction vs. correction, all *ps* < .01). Specifically, we found that the Wavelet and its combinations with Spline (Wav + Spl and Spl + Wav) performed best at reducing the SD in 97%–100% of cases (Fig. 12 lower panel). Spline alone reduced the within-subject SD in 77%–90% of the cases (Fig. 12 lower panel). Two-tailed paired t-tests between all the correction methods considered in this study revealed that: *(i)* Wavelet significantly reduced the within-subject SD compared to Spline in Datasets 2 and 3 (*ps* < .001), and to the combination of Wav + Spl in Dataset 4 ($p < .03$); *(ii)* Spl + Wav significantly decreased the SD when compared to Spline in all datasets (*ps* < .03), to Wavelet (*ps* < .04) and Wav + Spl (*ps* < .002) in Datasets 3 and 4; *(iii)* Wav + Spl significantly reduced the SD when compared to Spline in Datasets 2 and 3 (*ps* < .007) and to Spl + Wav in Dataset 2 ($p < .01$).

The between-subjects SD was reduced in 100% of cases with Wav + Spl in all datasets, with Spl + Wav performing similarly (Fig. 12 upper panel). These techniques on their own have a lower power in reducing the between-subjects SD, with Spline having the worst performance. Note that all techniques succeeded in reducing the between-subjects SD metric relative to the No Correction approach.

4. Discussion

The primary objective of this work is to provide recommendations for researchers who first approach infant fNIRS data analysis. fNIRS infant recordings are often contaminated by motion artifacts of various shapes and frequencies that are not easy to detect and consequently correct. To our knowledge, only one recent study measured the effects of a selection

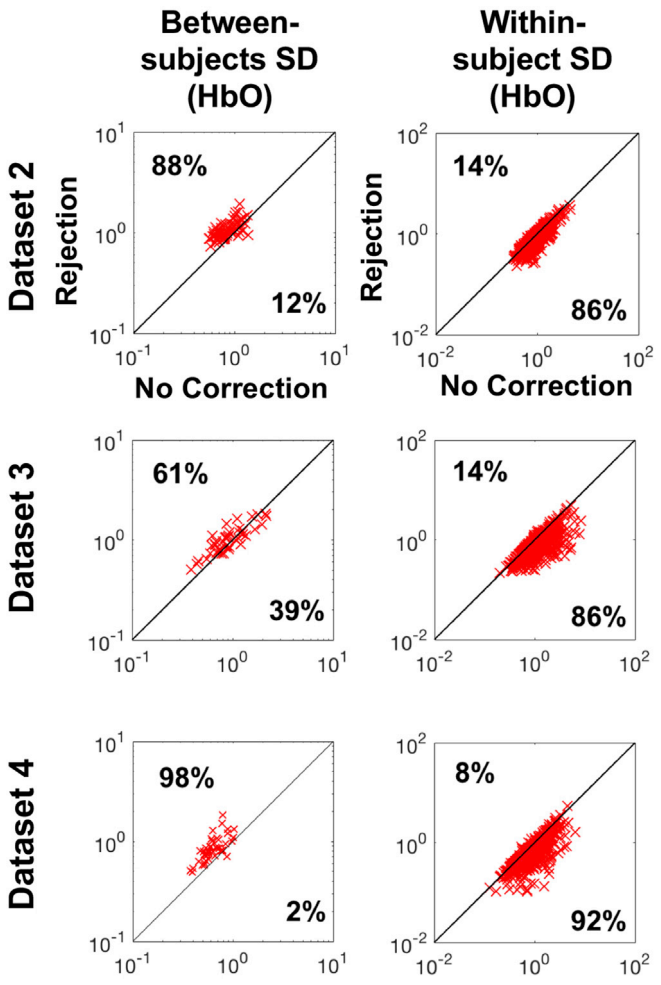

Fig. 11. Scatter plots of the between-subjects (left column) and within-subject (right column) SD calculated for the comparison between Rejection (y axis) and No Correction technique (x axis) for HbO values of datasets 2, 3 and 4 (row 1, 2, 3, respectively). The rejection of trials affected by predefined motion artifacts decreases the between-subjects SD between 39% and 2% of the time for HbO compared to No Correction, across all datasets (left panels); however, when looking at a single dataset the percentage of standard deviation reduction has a smaller range (right panels).

of motion correction techniques on infant data (Behrendt et al., 2018). Our study contributes to the current literature in two ways. First, by comparing the performance of most motion correction techniques available in Homer2 on infant semi-simulated data and by further assessing their performance based on the percentage of motion artifacts present in the signal. Second, by testing the efficacy of a novel combination of motion correction methods, Spline interpolation and Wavelet filtering, on real infant data collected from different sites using different paradigms, headgears and NIRS systems.

4.1. Motion correction performance in semi-simulation

To begin with, we tested the effect of changing the iqr values (i.e., 0.1, 0.5, 0.8, 1.0, 1.2, 1.5) of the Wavelet motion correction technique on three different types of simulated HRFs (previously defined as three sets of data: Short, Block Design, Standard HRFs) that were added to infant resting state recordings. Two iqr values (0.5 and 0.8) tended to best recover the majority of trials and improved the overall estimation of the HRF across all sets of data. Subsequently, we evaluated the performance of all the tested motion correction techniques. Trial Rejection and No Correction were the techniques with the worst performance in recovering trials and HRFs in all three sets of data. This is in line with the findings from previous studies on adult and infant data (Behrendt et al., 2018; Brigadoi et al., 2014), that also discourage the use of rejection of trials affected by motion artifacts as main correction strategy. This approach might be effective only in cases where motion artifacts are not frequent (Brigadoi et al., 2014) and a high number of trials are available, which is not common in infant studies.

Of the motion correction algorithms, Spline SG showed the worst performance as it did not recover any of the trials. One possible reason for this is that the spline part of this method is only applied when SNR >3. Thus, given the low SNR typical of infant data, the SG part was more often applied to our datasets alone than in combination with spline (an average of 30% of channels per subject showed SNR > 3 in this dataset, with 7 infants not showing any channel with SNR > 3). The use of this smoothing filter on its own might not have been sufficient to correct the artifacts or might even alter the signal resulting in larger artifacts. The second worst performing technique was Wavelet Kurtosis. For both metrics (number of recovered trials and RMSE) its performance was similar to that of trial rejection. In light of these findings, we did not test either Spline SG or Wavelet Kurtosis on the task-based datasets and we do not recommend the use of either correction algorithms for the analysis of infant data.

The techniques with the best performance on both metrics were tPCA and the combinations of Spline and Wavelet. A key finding of this first set of analyses is that the combinations of Spline and Wavelet, compared to the same techniques on their own, were able to recover most of the trials regardless of *iqr* (0.5 or 0.8) and order (Spl + Wav or Wav + Spl), with Spl + Wav being the method with the lowest RMSE median and, on average, the highest number of recovered trials. These results suggest that the combination represents a valid option for recovering the HRF across different types of datasets. In their study, Behrendt et al. (2018) tested Wavelet filtering with *iqr*s of 0.1, 0.5, 1.0 and 1.5 on infant data and recommended the use of 0.5. While our results do not entirely diverge from this suggestion, we advise infant researchers to try $iqr = 0.5$ and $iqr = 0.8$, and choose the value that prevents the risk of underestimating the true hemodynamic response. We suggest to use $iqr = 0.5$ when analyzing data collected with stimuli of 10 s or longer, and $iqr = 0.8$ when analyzing short event-related data (i.e., collected with stimuli shorter than 10 s).

One technique that, despite being available in Homer2, was not included in our analysis was CBSI. The reason for this choice reflects the concerns previously raised by Brigadoi et al. (2014), namely that CBSI does not correct HbR starting from the original signal, but creates its surrogate from HbO (therefore HbR does not reflect real measured data). Further, CBSI relies on strict assumptions on the relation between HbO and HbR that are not always met, and when this happens, the performance of this technique is negatively affected (Brigadoi et al., 2014).

The superiority of the combination of Spline and Wavelet has been further confirmed by the analysis computed on data containing from 0% to 21% of artifacts. Namely, we found that when no artifacts were detected, all the techniques had a similar performance in recovering the HRF, whereas at increasing percentages of artifacts Wavelet (at $iqr = 0.5$), tPCA and particularly Spl + Wav were more effective compared to the other techniques. Our findings are in contrast with the work by Behrendt and colleagues (2018), showing that tPCA performs better than Wavelet also with a higher presence of artifacts in the data. The reason of these conflicting results might be related to the choice of input parameters required for tPCA; perhaps the parameters used in the current work were more efficient in identifying the majority of artifacts in Dataset 1 as compared to the parameters used by Behrendt and colleagues (2018) in their dataset. A negative feature of tPCA is its dependence on the number of available channels; when more channels are available this usually leads to a better correction of the signal. However, infant research sometimes uses only few channels (e.g., Kida and Shinohara, 2013; Minagawa et al., 2008). Therefore, researchers should be aware of this tPCA feature when deciding which motion correction technique to apply,

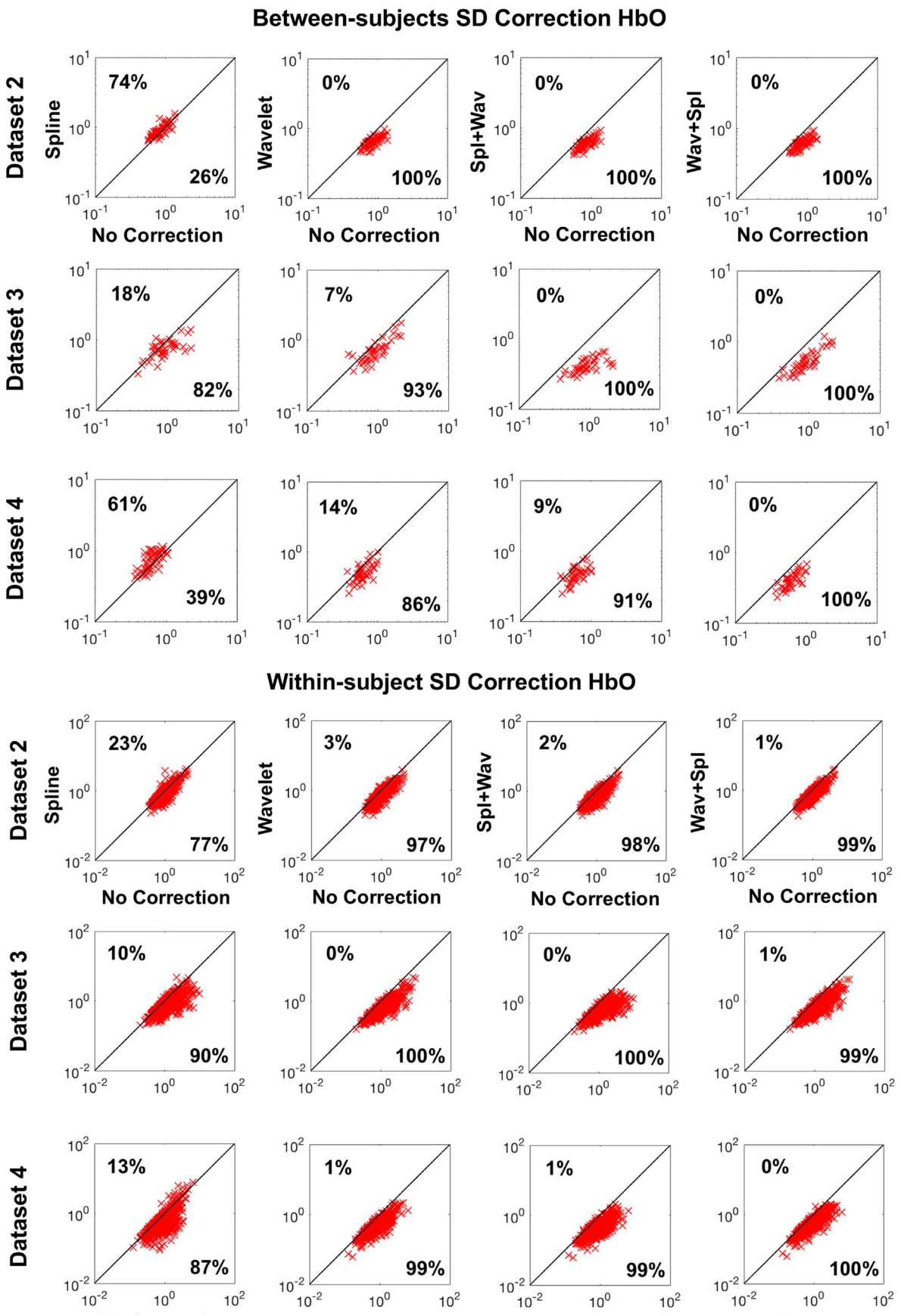

Fig. 12. Scatter plots of the between-subjects (upper panel) and within-subject (lower panel) SD computed for all the motion correction techniques (y axis) versus no motion correction (x axis) for HbO of datasets 2, 3 and 4 (rows 1, 2, and 3 of each panel). Spl + Wav and Wav + Spl have the best performance in reducing the within and between standard deviation across all datasets (i.e., SD reduced by 100%.ca of the cases).

particularly when their dataset has few channels available. In this work, we aimed to evaluate and suggest a new stacking approach that can be used with noisy infant data acquired with all types of probe arrays (i.e., multi- or single-channel arrays); hence, we discarded tPCA in the tests performed on the task-based datasets.

In sum, our findings on Dataset 1 suggest that Spl + Wav might be the optimal correction method for the recovery of the true HRF in infant semi-simulated data, at all percentages of motion artifacts.

4.2. Motion correction performance in the task-based datasets

The results of the comparison between the different correction methods on infant data collected during cognitive/perceptual tasks (i.e., Datasets 2, 3, 4), support the findings of our first analyses on semi-simulated infant data. Also in this case, Rejection and No Correction are the worst methods in reducing artifacts. Spline, when compared to its combination with Wavelet or to Wavelet on its own, had the worst performance in reducing the within- and between-subjects SDs and in recovering a sufficient number of trials in all three datasets. On the other hand, the use of Wavelet alone seems to work well in reducing the within- and between-subjects SDs across all datasets; however, this technique also saves less trials from rejection in comparison to the combined use of Wavelet and Spline. In infant research, including the maximum number of valid trials in the data analysis is crucial for obtaining a reliable hemodynamic response, and, ultimately, for reaching sufficient power in the statistical analyses.

4.3. General conclusions

In the current study we used four different metrics (i.e., RMSE; within- and between-subjects SD; number of recovered trials) to evaluate and compare the performance of different motion correction techniques in infant data. As each metric holds strengths and limitations (e.g., within-subject SD could be affected by habituation) conclusions on the effectiveness of the correction methods can only be drawn by taking into account the results obtained for all metrics. Therefore, in this section we will further discuss evidence coming from tests on all datasets and all metrics.

The combination of Spline and Wavelet was the most effective in (1) recovering the true HRF in infant semi-simulated data, both at low and high percentages of motion artifacts; (2) reducing the within- and between-subjects standard deviations; (3) saving nearly all trials across the three task-based datasets. Therefore, for infant data containing a large number of motion artifacts we suggest the combined use of Spline and Wavelet. Moreover, although we found that the optimal order of these correction methods differs per dataset, we suggest the application of Spline followed by Wavelet since this order showed a slightly better recovery of the true HRF in semi-simulated data (i.e., lowest RMSE median).

Given that the combined use of Spline and Wavelet had an optimal performance on all datasets that varied from one another on a number of levels (percentage of artifacts, age, task, NIRS system, headgear), we can conclude that our findings can be safely generalized to any infant dataset and, possibly, also to datasets acquired from other challenging samples, such as clinical population, under the hypothesis that several unpredictable motion artifacts could be present in these clinical acquisitions, coupled with few available trials. Our findings in semi-simulation show that the stacking recovered the true HRF better than when applying Wavelet alone. This result could apparently be in contrast with the results of Behrendt et al. (2018) who reported that the recovery of the hemodynamic response was not improved by the combination of tPCA and Wavelet compared to when only Wavelet was applied. The difference between the two studies could be identified in the differences between the modus operandi of tPCA and Spline. Similarly to Spline, tPCA works on predefined motion artifacts and performs well in correcting step changes in the signal (Yücel et al., 2014); conversely, tPCA applies the correction to all channels, including those that are not affected by motion, which might result in an undesired modification of the original signal or even a loss of useful signal in multiple channels when many motion artifacts sparsely located across channels are present.

Whereas we suggest the use of Spline and Wavelet in combination, we advise caution when selecting the parameters of the motion detection function (hmrMotionArtifactbyChannel function in Homer2). This step is crucial since the performance of Spline is highly dependent on the detection of the motion artifacts present in the data, which is determined, in turn, by the chosen values. Specifically, we found that the STDEV-thresh was the parameter that needed most careful tuning. In this work, the STDEVthresh value was defined by visually evaluating how efficient the function was in identifying the major motion artifacts present in the data. While this procedure is necessary to correctly select the motion artifacts, it is time consuming and subjective. Future work should address these issues by creating a more automatic and objective way to define these parameters.

On this note, we take the opportunity to also stress the importance of carefully selecting the parameters of another user-dependent function: the band-pass filter (for more in depth discussion on filter selection see Pinti et al., 2018). The choice of appropriate cut-off frequencies is critical to preserve hemodynamic responses evoked by experimental conditions while excluding irrelevant frequencies associated with physiological oscillations. In infant research, it is also important to take into account the age of the participants when setting the filter cut-offs, as the frequency of some of these oscillations (e.g., heart rate) change during development. While we are confident that the slightly different filter cut-offs employed in all datasets tested in the current study did not impact our motion correction results, which are consistent across datasets, we suggest that future studies should investigate which is the most appropriate combination of motion correction technique and band-pass filter for infant fNIRS data.

Despite our findings are in favor of the combined use of Spline and Wavelet, we do not discourage the application of Wavelet as exclusive correction method. Our results confirmed previous findings that Wavelet on its own has a better performance compared to Spline interpolation (Brigadoi et al., 2014). We showed that this technique is a good option with moderately noisy datasets but the combination of Spline and Wavelet should be preferred as the percentage of noise in the signal increases. Therefore, we suggest that the decision of which technique to apply should be driven by a critical evaluation of the number and characteristics of artifacts embedded in the fNIRS time courses. One benefit of using Wavelet filtering alone is that it does not require the detection of motion artifacts, a challenging task when the artifacts affecting the fNIRS recordings are extremely difficult to detect using automatic methods. We do not advise including Spline as sole correction method in the preprocessing of infant data. Specifically, we showed that Spline, compared to Wavelet and to the stacking of Spline and Wavelet, had the worst performance in recovering the true HRF in semi-simulated data (highest RMSE), in lowering the within- and between-subjects SDs in the task-based datasets and in recovering an acceptable number of trials in all four datasets.

To summarize, the convergence of results across all the different datasets tested in our work does not only indicate that the efficacy of the combination of Spline and Wavelet is independent from infants' age, the task design (visual/auditory vs. tactile stimulation), or percentage of noise, but that it is not determined either by the fNIRS system and headgear used.

It is likely that in the near future, thanks to technological advances such as the development of lightweight fibers or portable devices, we will witness an increase in the number of channels used for infants studies and, eventually, full head coverage arrays. While in this work we did not test the performance of our techniques on data collected with large channel arrays we believe that the same recommendations can be applied to such datasets. As long as the headgear is tightly fitted to the subject's head, the data quality should be comparable to the datasets tested in the

present work.

Acknowledgements

The authors would like to thank the infants and their families for their participation.

The authors would like to thank Prof. Livio Finos for the fruitful discussion and advices on statistical analyses.

Dataset 1 was supported by the Leverhulme Trust Research Project Grant (2015-115). Data was collected by Chiara Bulgarelli and Dr. Carina De Klerk. The authors would like to thank Professor Victoria Southgate and Professor Antonia Hamilton as PIs of the project for sharing the data.

Dataset 2 was supported by the UK Medical Research Council (G0701484), the Simons Foundation (no. SFARI201287), the BASIS Funding Consortium Led by Autistica (www.basisnetwork.org) and the Innovative Medicines Initiative Joint Undertaking under grant agreement no. 115300, resources of which are composed of financial contribution from the European Union's Seventh Framework Programme (FP7/2007–2013) and EFPIA companies. The BASIS team in alphabetical order: Baron-Cohen, S., Bedford, R. Bolton, P., Blasi, A., Charman, T., Cheung, H.M., Davies, K., Elsabbagh, M., Fernandes, J., Gammer, I., Gliga, T., Green, J., Guiraud, J., Johnson, M.H., Liew, M., Lloyd-Fox, S., Maris, H., O'Hara, L., Pasco, G., Pickles, A., Ribeiro, H., Salomone, E., Tucker, L., Yemane, F.

Dataset 3 was supported by a grant from the European Community's Horizon 2020 Program under grant agreement n° 642996 (Brainview) (RDL). The authors would like to thank Rianne van Rooijen for helping with the fNIRS data acquisition, Carlijn van den Boomen for lab support and Professor Chantal Kemner as PI of the project for providing lab space and equipment.

Dataset 4 was supported by a grant from a Grant-in-Aid for Scientific Research (A) (15H01691) (YM) and MEXT Supported Program for the Strategic Research Foundation at Private Universities. The authors would like to thank Aika Yasui for the fNIRS data collection.

S.B. was supported by grant "Progetti di Ateneo Bando 2015" C92I1600012005 and by "Progetto STARS Grants 2017" C96C18001930005 both from the University of Padova.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.neuroimage.2019.06.056.

References

Lloyd-Fox, S., Blasi, A., Volein, A., Everdell, N., Elwell, C.E., Johnson, M.H., 2009. Social perception in infancy: a near infrared spectroscopy study. Child Dev. 80 (4), 986–999. http://doi.org/10.1111/j.1467-8624.2009.01312.x.

Abdelnour, A.F., Huppert, T.J., 2009. Real-time imaging of human brain function by near-infrared spectroscopy using an adaptive general linear model. Neuroimage 46 (1), 133–143. http://doi.org/10.1016/j.neuroimage.2009.01.033.

Aslin, R., Shukla, M., Emberson, L.L., 2015. Hemodynamic correlates of cognition in human infants. Annu. Rev. Psychol. 33 (4), 395–401. http://doi.org/10.1038/nbt.3121.

Behrendt, H.F., Firk, C., Nelson, C.A., Perdue, K.L., 2018. Motion correction for infant functional near-infrared spectroscopy with an application to live interaction data. Neurophotonics 5 (01), 1. http://doi.org/10.1117/1.NPh.5.1.015004.

Brigadoi, S., Ceccherini, L., Cutini, S., Scarpa, F., Scatturin, P., Selb, J., et al., 2014. Motion artifacts in functional near-infrared spectroscopy: a comparison of motion correction techniques applied to real cognitive data. Neuroimage 85, 181–191. http://doi.org/10.1016/j.neuroimage.2013.04.082.

Bulgarelli, C., Blasi, A., Arridge, S., Powell, S., de Klerk, C.C., Southgate, V., et al., 2018. Dynamic causal modelling on infant fNIRS data: a validation study on a simultaneously recorded fNIRS-fMRI dataset. Neuroimage 175, 413–424. https://doi.org/10.1016/j.neuroimage.2018.04.022.

Chiarelli, A.M., Maclin, E.L., Fabiani, M., Gratton, G., 2015. A kurtosis-based wavelet algorithm for motion artifact correction of fNIRS data. Neuroimage 112, 128–137. http://doi.org/10.1016/j.neuroimage.2015.02.057.

Cooper, R.J., Selb, J., Gagnon, L., Phillip, D., Schytz, H.W., Iversen, H.K., et al., 2012. A systematic comparison of motion artifact correction techniques for functional near-infrared spectroscopy. Front. Neurosci. 6 (OCT), 1–10. http://doi.org/10.3389/fnins.2012.00147.

Cope, M., Delpy, D.T., 1988. System for long-term measurement of cerebral blood and tissue oxygenation on newborn infants by near infra-red transillumination. Med. Biol. Eng. Comput. 26 (3), 289–294.

Delpy, D.T., Cope, M., van der Zee, P., Arridge, S., Wray, S., Wyatt, J., 1988. Estimation of optical pathlength through tissue from direct time of flight measurement. Phys. Med. Biol. 33 (12), 1433–1442. http://doi.org/10.1088/0031-9155/33/12/008.

Di Lorenzo, R., Blasi, A., Junge, C., Van Den Boomen, C., Van Rooijen, R., Kemner, C., 2019. Brain responses to faces and facial expressions in 5-month-olds: an fNIRS study. Front. Psychol. 10, 1240. https://doi.org/10.3389/fpsyg.2019.01240.

Duncan, a, Meek, J.H., Clemence, M., Elwell, C.E., Tyszczuk, L., Cope, M., Delpy, D.T., 1995. Optical pathlength measurements on adult head, calf and forearm and the head of the newborn infant using phase resolved optical spectroscopy. Phys. Med. Biol. 40 (2), 295–304. http://doi.org/10.1088/0031-9155/40/2/007.

Everdell, N.L., Gibson, A.P., Tullis, I.D.C., Vaithianathan, T., Hebden, J.C., Delpy, D.T., 2005. A frequency multiplexed near-infrared topography system for imaging functional activation in the brain. Rev. Sci. Instrum. 76 (9), 093705. http://doi.org/10.1063/1.2038567.

Grossmann, T., Johnson, M.H., Lloyd-Fox, S., Blasi, A., Deligianni, F., Elwell, C., Csibra, G., 2008. Early cortical specialization for face-to-face communication in human infants. Proc. Biol. Sci. 275 (1653), 2803–2811.

Homae, F., Watanabe, H., Otobe, T., Nakano, T., Go, T., Konishi, Y., Taga, G., 2010. Development of global cortical networks in early infancy. J. Neurosci. 30 (14), 4877–4882. https://doi.org/10.1523/JNEUROSCI.5618-09.2010.

Hu, X.-S., Arredondo, M.M., Gomba, M., Confer, N., DaSilva, A.F., Johnson, T.D., et al., 2015. Comparison of motion correction techniques applied to functional near-infrared spectroscopy data from children. J. Biomed. Opt. 20 (12), 126003. http://doi.org/10.1117/1.JBO.20.12.126003.

Huppert, T.J., Diamond, S.G., Franceschini, M.A., Boas, D.A., 2009. Hom{ER}: a review of time-series analysis methods for near-infrared spectroscopy of the brain. Appl. Opt. 48 (10), D280–D298. http://doi.org/10.1364/AO.48.00D280.

Issard, C., Gervain, J., 2018. Variability of the hemodynamic response in infants: influence of experimental design and stimulus complexity. Dev. Cognitive. Neurosci. 33, 182–193. https://doi.org/10.1016/j.dcn.2018.01.009.

Jahani, S., Setarehdan, S.K., Boas, D.A., Yücel, M.A., 2018. Motion artifact detection and correction in functional near-infrared spectroscopy: a new hybrid method based on spline interpolation method and Savitzky–Golay filtering. Neurophotonics 5 (1), 015003. https://doi.org/10.1117/1.NPh.5.1.015003.

Kida, T., Shinohara, K., 2013. Gentle touch activates the prefrontal cortex in infancy: an NIRS study. Neurosci. Lett. 541, 63–66. http://doi.org/10.1016/j.neulet.2013.01.048.

Lloyd-Fox, S., Blasi, A., Elwell, C.E., 2010. Illuminating the developing brain: the past, present and future of functional near infrared spectroscopy. Neurosci. Biobehav. Rev. 34 (3), 269–284. http://doi.org/10.1016/j.neubiorev.2009.07.008.

Lloyd-Fox, S., Széplaki-Köllod, B., Yin, J., Csibra, G., 2015. Are you talking to me? Neural activations in 6-month-old infants in response to being addressed during natural interactions. Cortex 70, 35–48. http://doi.org/10.1016/j.cortex.2015.02.005.

Lloyd-Fox, S., Blasi, A., Pasco, G., Gliga, T., Jones, E.J.H., Murphy, D.G.M., et al., 2018. Cortical responses before 6 months of life associate with later autism. Eur. J. Neurosci. 47 (6), 736–749. http://doi.org/10.1111/ejn.13757.

Miguel, H.O., Lisboa, I.C., Gonçalves, O.F., Sampaio, A., 2017. Brain mechanisms for processing discriminative and affective touch in 7-month-old infants. Dev. Cognitive. Neurosci. (February), 0–1. http://doi.org/10.1016/j.dcn.2017.10.008.

Minagawa-Kawai, Y., Matsuoka, S., Dan, I., Naoi, N., Nakamura, K., Kojima, S., 2008. Prefrontal activation associated with social attachment: facial-emotion recognition in mothers and infants. Cerebr. Cortex 19 (2), 284–292. https://doi.org/10.1093/cercor/bhn081.

Molavi, B., Dumont, G.A., 2012. Wavelet-based motion artifact removal for functional near-infrared spectroscopy. Physiol. Meas. 33 (2), 259–270. http://doi.org/10.1088/0967-3334/33/2/259.

Molavi, B., May, L., Gervain, J., Carreiras, M., Werker, J.F., Dumont, G.A., 2014. Analyzing the resting state functional connectivity in the human language system using near infrared spectroscopy. Front. Hum. Neurosci. 7 (January), 1–9. http://doi.org/10.3389/fnhum.2013.00921.

Obrig, H., Villringer, A., 2003. Beyond the visible - imaging the human brain with light. J. Cereb. Blood Flow Metab. 23 (1), 1–18. http://doi.org/10.1097/01.WCB.0000043472.45775.29.

Pinti, P., Scholkmann, F., Hamilton, A., Burgess, P., Tachtsidis, I., 2018. Current status and issues regarding pre-processing of fNIRS neuroimaging data: an investigation of diverse signal filtering methods within a General Linear Model framework. Front. Hum. Neurosci. 12. http://doi.org/10.3389/fnhum.2018.00505.

Ravicz, M.M., Perdue, K.L., Westerlund, A., Vanderwert, R.E., Nelson, C.A., 2015. Infants' neural responses to facial emotion in the prefrontal cortex are correlated with temperament: a functional near-infrared spectroscopy study. Front. Psychol. 6 (July), 1–12. http://doi.org/10.3389/fpsyg.2015.00922.

RStudio Team, 2016. RStudio. Integrated Development for R. RStudio, Inc., Boston, MA. http://www.rstudio.com/.

Scholkmann, F., Spichtig, S., Muehlemann, T., Wolf, M., 2010. How to detect and reduce movement artifacts in near-infrared imaging using moving standard deviation and spline interpolation How to detect and reduce movement artifacts in near-infrared

imaging using moving standard deviation and spline interpolation, 31, pp. 649–662. http://doi.org/10.1088/0967-3334/31/5/004.
Timeo, S., Brigadoi, S., Farroni, T., 2017. Perception of Caucasian and African faces in 5- to 9-month-old Caucasian infants: a functional near-infrared spectroscopy study. Neuropsychologia 126, 3–9. http://doi.org/10.1016/j.neuropsychologia.2017.09.011.
Wilcox, T., Biondi, M., 2015. fNIRS in the developmental sciences. Wiley. Interdisciplin. Rev.: Cogn. Sci. 6 (3), 263–283. http://doi.org/10.1002/wcs.1343.
Yücel, M.A., Selb, J., Cooper, R.J., Boas, D.A., 2014. Targeted principle component analysis: a new motion artifact correction approach for near-infrared spectroscopy. J. Innovate. Optical. Health. Sci. 7 (02), 1350066. https://doi.org/10.1142/S1793545813500661.
Zhang, Y., Brooks, D.H., Franceschini, M.A., Boas, D.A., 2005. Eigenvector-based spatial filtering for reduction of physiological interference in diffuse optical imaging. J. Biomed. Opt. 10 (1), 011014. https://doi.org/10.1117/1.1852552.