Neuroinformatics (2022) 20:665–675
https://doi.org/10.1007/s12021-021-09551-6

ORIGINAL ARTICLE

Reproducible Inter-Personal Brain Coupling Measurements in Hyperscanning Settings With functional Near Infra-Red Spectroscopy

Bizzego Andrea[1] · Azhari Atiqah[2] · Esposito Gianluca[1,2,3]

Accepted: 21 October 2021 / Published online: 29 October 2021

Abstract

Despite a huge advancement in neuroimaging techniques and growing importance of inter-personal brain research, few studies assess the most appropriate computational methods to measure brain-brain coupling. Here, we focus on the signal processing methods to detect brain-coupling in dyads. From a public dataset of functional Near Infra-Red Spectroscopy signals (N=24 dyads), we derived a synthetic control condition by randomization, we investigated the effectiveness of four most used signal similarity metrics: Cross Correlation, Mutual Information, Wavelet Coherence and Dynamic Time Warping. We also accounted for temporal variations between signals by allowing for misalignments up to a maximum lag. Starting from the observed effect sizes, computed in terms of Cohen's d, the power analysis indicated that a high sample size ($N > 150$) would be required to detect significant brain-coupling. We therefore discuss the need for specialized statistical approaches and propose bootstrap as an alternative method to avoid over-penalizing the results. In our settings, and based on bootstrap analyses, Cross Correlation and Dynamic Time Warping outperform Mutual Information and Wavelet Coherence for all considered maximum lags, with reproducible results. These results highlight the need to set specific guidelines as the high degree of customization of the signal processing procedures prevents the comparability between studies, their reproducibility and, ultimately, undermines the possibility of extracting new knowledge.

Keywords Brain-brain coupling · Similarity measures · Reproducibility · Functional Near Infra-Red Spectroscopy · Inter-personal coupling · Hyperscanning

Introduction

Once enmeshed in a single-brain framework in which only intrapersonal brain activity can be examined, the emergence of inter-brain coupling has since steered the field of social neuroscience towards a multi-brain paradigm (Hasson et al., 2012). Inter-brain coupling, a measure of temporal alignment of brain signals across two or more individuals, has garnered abundant acclaim in recent years due to its revolutionary approach in computing the neuronal correlates underpinning between-persons dynamics paradigm (Nam et al., 2020). This concept was first conceived in a seminal study by Hasson et al. (2004) where the authors employed a functional Magnetic Resonance Imaging (fMRI) paradigm to measure participants' brain activity during a movie screening before computing inter-brain coupling across participants. Hasson et al. (2004) showed that inter-brain coupling occurred in higher-order social cortices that spanned beyond sensory processing modules which borne the proposition that inter-brain coupling was fundamental for individuals to experience a "shared social world". Whereas single-brain approaches have overlooked the bidirectional flow of relevant social information that transpires between two persons, hyperscanning, the simultaneous recording from two or more individuals, offers a means by which these processes can be studied (Hasson & Frith, 2016). In just two decades, hyperscanning has made significant strides in illuminating inter-brain coupling across abundant topics including shared

✉ Esposito Gianluca
gianluca.esposito@unitn.it; gianluca.esposito@ntu.edu.sg

[1] Department of Psychology and Cognitive Science, University of Trento, Trento, Italy

[2] Psychology Program, School of Social Sciences, Nanyang Technological University, Singapore, Singapore

[3] Lee Kong Chian School of Medicine, Nanyang Technological University, Singapore, Singapore

social attention (Bilek et al., 2015), joint decision-making (Fallani et al., 2010; Tang et al., 2016) and cooperative tasks (Toppi et al., 2016; Liu et al., 2016). Hypercanning studies have demonstrated that inter-brain coupling is a unique interpersonal phenomenon that could not be solely accounted for by the activity of any one brain (Chatel-Goldman et al., 2013). Inter-brain coupling has rapidly become the theoretical foundation upon which the neural correlates of interpersonal interactions are investigated. Although mainly focusing on brain-brain coupling, hyperscanning studies have also been conducted on peripheral signals to quantify coupling of the Autonomic Nervous System (Bizzego et al., 2020a; Golland et al., 2019; Mønster et al., 2016).

Functional Near-Infrared Spectroscopy (fNIRS) is particularly used in hyperscanning (Zhang et al., 2018; Chen et al., 2020; Azhari et al., 2020a) as it bears resemblance to fMRI in measuring the changes in oxygenation levels of blood in the brain to determine brain activity indirectly (Huppert et al., 2006). However, fNIRS is similar to the EEG in its ability to assess cortical brain areas only (Scholkmann et al., 2013). In spite of its comparatively lower temporal resolution to the EEG, fNIRS possesses a significant suite of advantages including (i) resistance to motion-induced noise, (ii) portability and (iii) lower costs, which enables ecological acquisition of brain responses in social settings and special application on infant populations (Czeszumski et al., 2020).

While there is currently no standardized experimental protocol to investigate brain-brain coupling in hyperscanning paradigms, the use of emotional auditory stimuli alongside an investigation of prefrontal cortex (PFC) responses is common in fNIRS studies. For instance, Chong et al. (2019) demonstrated that haemodynamic changes in the left PFC was evident among nursing students when they were exposed to emotional auditory stimuli. In another study, Azhari et al. (2020c) similarly showed that different audio-visual pairings of salient human vocalisations with environmental scenes elicited significant changes in the left PFC. Along the same vein, Bigliassi et al. (2015) illustrated that different genres of music, which elicit different emotional states in listeners, can be captured as changes in haemodynamic signals within the PFC. These studies motivates the investigation of brain activity in the PFC in empirical paradigms that employ emotional auditory stimuli.

Computational Approaches

The key computational paradigm of hyperscanning studies considers the similarity between two signals as a proxy to quantify the coupling between the processes that generated the two signals. Applied to physiological or brain signals from two or more individuals, similarity measures are used to investigate the interpersonal coupling: how coordinated the individuals are in responding to stimuli or other environmental cues (Azhari et al., 2019, 2020b).

Despite tremendous advances in this field, analytical approaches to fNIRS still need standardization (Pinti et al., 2019; Bizzego et al., 2020b). In particular, the lack of consensus in computational approaches to investigate signal similarity elicits profound problems of reproducibility that exist across both brain and physiological hyperscanning studies. Three key computational aspects emerge to be of paramount concern. First, there is a lack of cross-examination of the different metrics that are used to quantify the similarity between signals. The thriving variety of metrics and approaches that have been used attests to the nascent and exploratory phase in which this field currently resides (Czeszumski et al., 2020). Second, there is ambiguity in how to make the computation of the similarity robust to localized temporal variations between two signals (Mauss et al., 2005; Kettunen et al., 1998). Third, there remains a lack of standardised assessment on the statistical aspects required to efficiently detect similarity: studies have been conducted on large variations of sample sizes and using different statistical frameworks. In this study we aim at assessing the validity and reliability of the main computational approaches, which is the priority towards reliable investigations of interpersonal coupling.

Similarity Metrics

The present study examines four commonly used similarity metrics: (1) Cross Correlation, (2) Mutual Information, (3) Wavelet Coherence, and (4) Dynamic Time Warping. Cross Correlations (CC) employs a simplified approach by computing the extent to which two time-series data are correlated to each other and it has largely been used to investigate similarity between signals. CC of BOLD signals have been used to study coupling in various social scenarios (Hasson & Frith, 2016; Pan et al., 2020; Funane et al., 2011). Cross-correlations have also been adopted to compute similarity of EEG (Kawasaki et al., 2013), fMRI (Bilek et al., 2015) and peripheral physiological signals (Mauss et al., 2005; Bizzego et al., 2020a). Mutual Information (MI) quantifies the amount of shared information between two interdependent signals, in terms of how knowing one signal minimises the uncertainty of the other signal (Shannon, 1948). It has been applied to physiological (Faes et al., 2014) and brain signals (Pravitha & Bruce, 2010), to compute brain network connectivity (Sun et al., 2018; Salvador et al., 2007; Hinrichs et al., 2006) and coordinated physical activity (Trendafilov et al., 2020). Wavelet Coherence (WC) is a computational method that has garnered considerable attention in hyperscanning studies. WC is essentially the correlation between the Wavelet Transforms (Torrence & Compo, 1998) of two signals, and it is used to analyse similarities in time frequency space (Grinsted et al., 2004). WC has been especially applied to fNIRS hyperscanning studies that involve active social engagement (Zhang et al., 2020; Jiang et al., 2012; Ikeda et al., 2017). Dynamic time warping (DTW) was developed to obtain a reliable distance between signals able to account for slight temporal shifts that would have otherwise led to convoluted distance computations (Berndt & Clifford, 1994; Sakoe & Chiba, 1978). The DTW algorithm produces localized non-linear modifications (warping) of one signal to optimize the pairing between the samples of the two signals and correcting misalignments. It has been broadly applied to investigate physiological (Gashi et al., 2018) and brain signals in EEG (Karamzadeh et al., 2013), fMRI (Meszlenyi et al., 2017) and fNIRS (Azhari et al., 2019) studies.

Lag

The hemodynamic response observed in fNIRS is affected by inter-subject variability and is altered by other physiological components, not related to the brain processing of the stimuli (Huppert, 2016; Seghouane & Ferrari, 2019; Taylor et al., 2018; Abibullaev et al., 2013). To cope with the effects of these alterations, in hyperscanning studies it is customary to include some flexibility in the temporal alignment of signals, which is implemented by allowing some lag between the two signals.

However, there is a lack of information on the role of such lag, and its value, on the extent of similarity observed (Meszlenyi et al., 2017). For instance, Hasson and Frith (2016) examined the relationship between synchronised finger tapping and inter-brain coupling as partners were engaged in reciprocal adaptation to each other. Static analyses (i.e., with no lag allowed) generated a negative correlation; on the opposite, analyses that allowed a 1 s lag between the partners generated a positive correlation. Mounting evidence from neuroimaging studies likewise points to the differences in inter-brain coupling activity that exists on time scales ranging from a few seconds to minutes (Kiviniemi et al., 2011; Handwerker et al., 2012; de Pasquale et al., 2010).

Statistical Methods

Finally, there is a need to assess the statistical validity of similarity computations to ensure that analyses possess sufficient power to reject false positives. Three issues to the statistical validation of hyperscanning studies can be identified. The first issue is the impossibility of obtaining a ground truth for inter-brain coupling and the difficulty in designing case-control studies, which are inherently highly demanding in terms of resources and subjects (Liu et al., 2018). To overcome the lack of a control condition, most studies adopt randomization approaches (Golland et al., 2015; Golland et al., 2014) to statistically evaluate the significance of the similarity measure. Second, the observed effect sizes can be very small (Lombardo et al., 2016), due to noise artefacts or low signal quality, but also to other physiological components in the brain signals, not necessarily associated with the processing of stimuli (Huppert, 2016). Besides, designing hyperscanning experiments with high sample size is highly demanding as the datapoint is the dyad (or group), which involves the number of participants is multiplied. Combined together, small effect sizes and small sample sizes compromise the possibility of observing statistically relevant results. Third, hyperscanning studies customarily involve some functional interpretation of the results by investigating different cerebral regions, via multiple signals localized on the scalp. The use of multiple channels poses an obstacle when correcting for multiple hypotheses to determine the regions implicated in coupling (Ayrolles et al., 2020).

This Study

The aim of this study is to inform researchers on the good practices and pitfalls towards reproducible hyperscanning studies and to provide benchmarks for a further development of the discipline.

To address the above discussed research gaps, the present study compares the four commonly used similarity metrics (i.e. CC, MI, WC, DTW) across several lags and examine the statistical power of a published fNIRS hyperscanning dataset (Azhari et al., 2020b). Azhari et al. (2020b) measured prefrontal cortical inter-brain coupling of adult-adult spousal pairs while simultaneously listening to auditory stimuli.

From this dataset, we derived two synthetic experimental conditions: (a) with signals from both spouses while simultaneously listening to the same auditory stimulus; (b) with signals from a member while listening to an auditory stimulus and signals from the other member while listening to static noise. We then compare the two synthetic conditions, assuming that reproducible metrics and procedure should be effective in detecting a significant greater similarity in the first condition.

This assumption is validated by the robust inter-brain coupling shown to exist in romantic dyads in previous studies (Kinreich et al., 2017; Pan et al., 2017). In an fNIRS hyperscanning study, Pan et al. (2017) revealed that couples showed significantly greater inter-brain coupling than pairs of strangers and even friends. Similarly, Kinreich et al. (2017) demonstrated in their EEG hyperscanning study that inter-brain coupling during naturalistic social interaction was evident in couples but not in strangers.

First, we compare all similarity metrics at multiple lags, also considering the appropriate normalization method for the brain signals. Starting from the effect size observed, we perform a power analysis to inform about the sample size required to obtain statistically relevant results. Finally, we present an alternative statistical approach based on bootstrapping and validate the results on the selected metrics and lag.

Methods and Materials

Dataset

This study was conducted based on signals retrieved from a published dataset (Azhari et al., 2020b) which investigated the coupling of prefrontal brain activity between spouses when listening to audio stimuli.

The dataset included fNIRS signals from 24 couples (N = 48, Mean age = 34.0, SD = 5.61) consisting of 24 mothers (N = 24, Mean age = 33.0, SD = 5.56) and 24 fathers (N = 24, Mean age = 35.0, SD = 5.59).

Each couple participated in two experimental conditions and the order of those conditions was counterbalanced across couples: together-condition, in which partners were presented with auditory stimuli at the same time in the same room, and separate-condition, in which partners listened to auditory stimuli in separate rooms and at different times.

Six audio stimuli were presented: 3 to reflect negative valence sounds: “female cry”, “infant cry, low pitched”, and “infant cry, high pitched”; 2 positive valence sounds: “female laugh”, and “infant laugh”; and “static noise”. Each audio stimulus was presented three times for each experimental condition, in a randomized order. In this study, only signals from the together-condition were used.

Brain activity of the prefrontal-cortical regions was acquired from two light wavelengths (760nm and 850nm) using a non-invasive fNIRS neuroimaging system, with a sampling rate of 7.81Hz. The system collected brain activity from 20 channels (i.e. one channel is comprised of one source-detector pairing) which covered four main prefrontal-cortical areas (Fig. 1): medial left, medial right, frontal left and frontal right.

The fNIRS channels provided were pre-processed to automatically remove channels with high signal noise (e.g. due to artefacts produced by head movements) and eliminate baseline shift variations. Finally, haemoglobin concentrations were determined using the modified Beer-Lambert law and the signals were visually inspected by two independent experts for validation. All oxygenated and deoxygenated hemoglobin signals were checked by the independent coders to ensure that signals were of typical waveforms, where the concentration of oxygenated haemoglobin rises after cortical activation, whereas that of deoxygenated haemoglobin decreases.

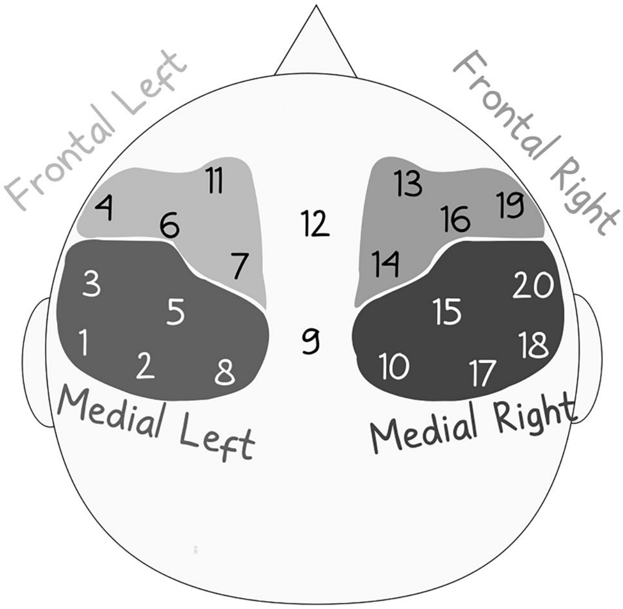

Fig. 1 Location of the 20 fNIRS channels and four clusters

Signal Processing

The 20 channels are grouped into four brain regions (Fig. 1): Medial Left, Medial Right, Frontal Left and Frontal Right. The brain activity of each cluster was obtained by computing the average of the normalized channels that composed each cluster. Assessing the optimal procedure to compute the cluster signals was out of the scope of this study; however, to ensure good quality cluster signals, a cluster signal was computed only if at least 3 channels were available. This step was performed to simplify the downstream analysis, as it reduced the number of brain activity signals from 20 to four.

When considering physiological signals from different individuals, it is common practice to normalize the collected signals to reduce the effect of individual characteristics. In this study, we wanted to consider the impact of signal normalization before computing the similarity measures. While several methods have been proposed for normalization, this study focused on a simple standardization via a transformation to z-score: $z = (X - \mu)/\sigma$, where X is the original signal, μ is the mean and σ is the standard deviation. We therefore computed the similarity measures on (i) the original cluster signals and on (ii) the standardized cluster signals.

Similarity Metrics

In this study, we considered four similarity metrics, each one representing a different theoretical framework proposed in literature to measure similarity between signals.

The first similarity metric is Cross Correlation (CC) which is one of the key similarity metrics that is used in signal processing. The computation of CC is based on the `scipy` Python package (v1.5.2). The second similarity metric is Mutual Information (MI), which pertains to the domain of entropy measures and is used to quantify the exchange of information between complex systems. The computation of the MI is based on the `JIDT` Java library (v1.5) (Lizier, 2014). The third similarity metric is Wavelet Coherence (WC), which leverages on the representation of a signal in time-frequency domain. The computation of WC is based on the `pycwt` Python module (v0.3). The fourth similarity metric is the Dynamic Time Warping distance (DTW), a well known algorithm which finds an optimal match between the samples of the two signals before computing their distance. The computation of DTW distance is based on the `dtw` R library (v1.22) (Giorgino, 2009). To impose some constraints on the extent of warping allowed, we imposed the Sakoe-Chiba windows type (Sakoe & Chiba, 1978), with asymmetric paths, while allowing the initial and final samples of the signals to not necessarily be paired. Finally, the DTW distance was multiplied by -1 to obtain a measure consistent with the other metrics, which reported measures of similarity.

To account for anticipations or delays of the physiological activity between the dyadic members, some studies proposed to compute the similarity between shifted version of the signals, up to a maximum lag that is empirically defined (Mauss et al., 2005). In this study we investigated the impact of the value of the lag, testing 4 maximum lags: 0, 1, 2, and 5 seconds. Given a maximum lag λ, a vector of similarity measures is computed with one signal which is shifted between $-\lambda$ s to $+\lambda$ s with increments of 0.128 s. The maximum value is then selected as the similarity measure for the given lag λ. This process was applied to all measures except DTW, as the DTW algorithm is specifically designed to account for misalignment in the signals. Therefore, when computing the DTW distance, the time shift was used to constrain the maximum distance allowed between paired samples to optimize the alignment between the two signals.

Since each auditory stimulus is presented three times, similarity is computed for each repetition and then averaged across the three trials, to obtain the similarity measure of each stimulus. The similarity measures of the five stimuli were then averaged to compute the similarity measure for each dyad in both conditions. This step was performed to reduce the complexity of the downstream analysis.

Analytic Plan

The scope of this study was to identify the appropriate similarity measures and procedures to compute interpersonal brain coupling. We considered a number of combinations of similarity measures and parameters, namely: 4 similarity measures (CC, MI, WC, DTW); 4 maximum lags (0, 1, 2, 5 seconds); 2 standardization versions (normalized and non-normalized signals); for the 4 brain regions (Frontal Left, Frontal Right, Medial Left, and Medial Right).

To compare the efficacy of the different similarity measures and parameters, we computed the similarity in two conditions:

A. Between the signals of the two members of the dyad, while both were listening to the same auditory stimulus (experimental condition);
B. Between the signal of one member who was listening to the auditory stimulus and the signal of the other member who was listening to the “static noise” stimulus, randomly assigning the members to the stimulus or to the “static noise” (synthetic control condition);

The approach of deriving a synthetic control condition was derived from Golland et al. (2014, 2015), with the difference that, instead of randomizing the dyads, we consider the same dyad while listening to two stimuli, one of which (i.e. static noise) does not convey any auditory cue.

Our assumption was that, for the brain regions involved in the processing of the auditory stimulus, higher similarity should be measured in condition A, compared to condition B. In fact, in condition A, the signals from the two members of the dyad should show similar patterns, as a similar brain activity should be elicited by the same auditory stimulus. Instead, in condition B, the signals were collected while the two members were listening two different stimuli - and during two distinct time intervals. Under this assumption, we identified which similarity measures and parameters were more appropriate by analysing which combinations were more efficient in detecting the difference between the two conditions.

First, we computed the difference between the two conditions in terms of Cohen’s d. We expect that the metrics and parameters which are able to detect coupling would report a $d > 0$, with higher values associated with a more efficient capability of detecting coupling. The differences were statistically assessed using Mann-Whitney tests ($\alpha = 0.05$). The resulting effect sizes, computed in terms of Cohen’s d, were used to perform a power analysis to estimate the number of dyads that would require obtaining significant results. The range of the d values computed from standardized cluster signals was higher than those computed from non-standardized signals. We therefore decided to continue the analysis focusing only on similarity measures computed from standardized signals.

In the second part of the analysis, we explored the use of a bootstrap procedure to improve the statistical assessment of the differences between the two conditions. Bootstrap (Fisher & Hall, 1991) is asymptotically more accurate than the standard intervals (e.g., sample variance) to estimate

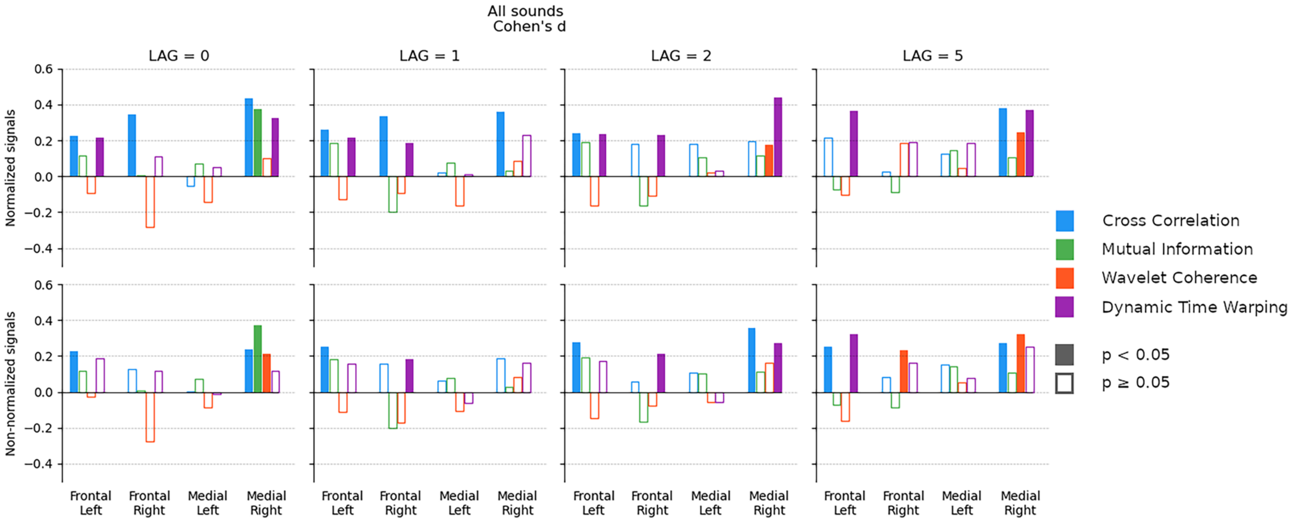

Fig. 2 Cohen's *d* for the four similarity measures, maximum lags and brain regions, when normalizing the signals before computing the measure (top) and when using the original signal (bottom)

confidence intervals and is used when the sample size is insufficient (Dwivedi et al., 2017).

The bootstrap consist of repeating the following steps N times: (i) selection of a subset of k samples, and (ii) inference of a property (e.g. mean) on the subset, thus obtaining a new distribution with N values of the selected property. In this study, $N = 100$, $k = 6$ (corresponding to 25% of the sample size) and the inferred measure was the mean. Cohen's d was computed to quantify the difference between the distributions of the means in condition A and B; Mann-Whitney tests were used to assess significance ($\alpha < 0.05$). We then investigated the role of maximum lag λ on the computed ds and selected the similarity measure and λ that were more efficient in detecting the differences between the two conditions (CC, $\lambda = 2$s).

Finally, using the selected measure and λ, we computed similarity in the original channels in the two conditions, computed the Cohen's d, and applied the bootstrap method to assess the significance of the differences.

Since the purpose of employing the Mann-Whitney test was to guide the identification of the most appropriate

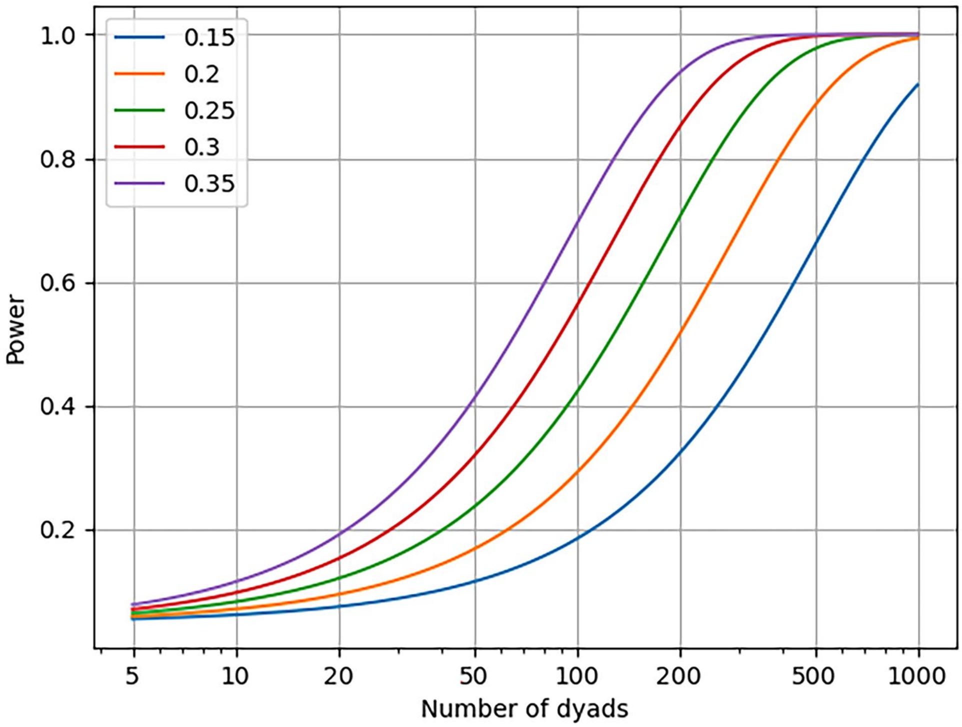

Fig. 3 Results of the power analysis for different values of the effect size *d* within the range of the obtained Cohen's *d*. A logarithmic scale is used for the horizontal axis

measures, and in light of the many parameters that were considered, we decided to avoid correcting for multiple hypotheses.

All analyses were computed based on custom code, written in Python and based on the `physynch` module (Bizzego et al., 2021).

Results

Effect Sizes and Normalization

In the first part of the analysis (Fig. 2), we computed the effect size (Cohen's d) and assessed the statistical significance of the difference between the two conditions, for all similarity metrics, maximum lags and brain regions.

From the analysis, it emerged that standardizing the signal before computing the similarity results in slightly higher ds. The most efficient similarity metrics resulted the CC and DTW; however, in many cases the computed d had negative values, or the difference resulted not significant, as opposed to what was assumed.

Power Analysis

In fact, the subsequent power analysis ($alpha < 0.05$) with effect sizes ranging from $d \simeq 0.15$ to $d \simeq 0.35$ (Fig. 3), highlighted that a higher sample size would be required ($N \simeq 150, d \simeq 0.35$; $N = 600, d \simeq 0.15$) to obtain adequate statistical power (0.8).

Performing hyperscanning experiments is highly demanding and achieving such sample sizes is difficult. In the subsequent part of the analysis we explored the use of bootstrap methods to overcome the issues associated with a small sample size.

Bootstrap

With the bootstrap procedure, more differences resulted significant (Fig. 4). Negative ds were still observed, in particular for WC and MI. For a maximum lag of 5 seconds, all measures (except MI and WC for the Frontal Left and MI and CC Frontal Right brain regions) report positive and significant ds. Qualitatively, CC and DTW showed more consistent patterns across different lags and brain regions; WC failed to report positive d except when allowing a maximum lag of 5 seconds.

Further investigating the relation between maximum lag and ds (Fig. 4), we observe that with CC and DTW obtained greater effect sizes, and ds computed on DTW increased with increasing values of lag.

WC produced only 6 positive and significant ds and appeared to be the least appropriate similarity measure. MI performed better than WC but led to 6 non significant ds. CC and DTW showed similar performances and were both indicated as appropriate to quantify coupling.

Similarity of Channels

For the computation of similarity between channels, we selected CC, with a maximum lag of 2 seconds. CC was preferred to DTW due to its simpler mathematical formulation.

Similarity of individual channels (Fig. 5) resulted in significance for 12 channels, out of 18 (66.7%); the Medial Left brain region had the higher number of non significant channels (3), followed by the Frontal Right (2) and Medial Right (1).

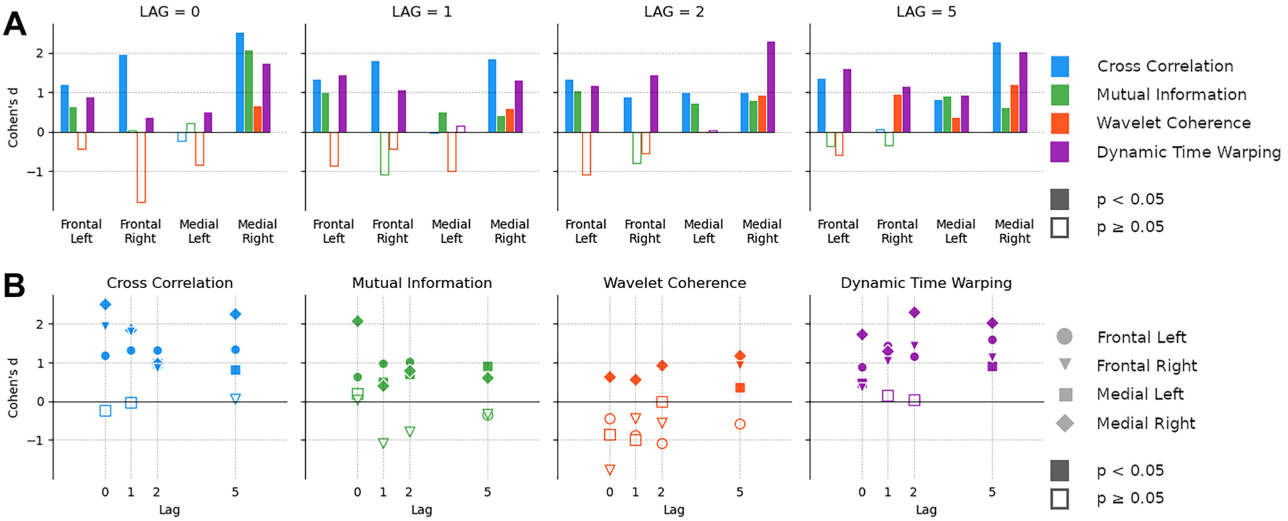

Fig. 4 Bootstrap Cohen's d: **A**. by maximum lag and brain regions; **B**. by similarity measure and maximum lag

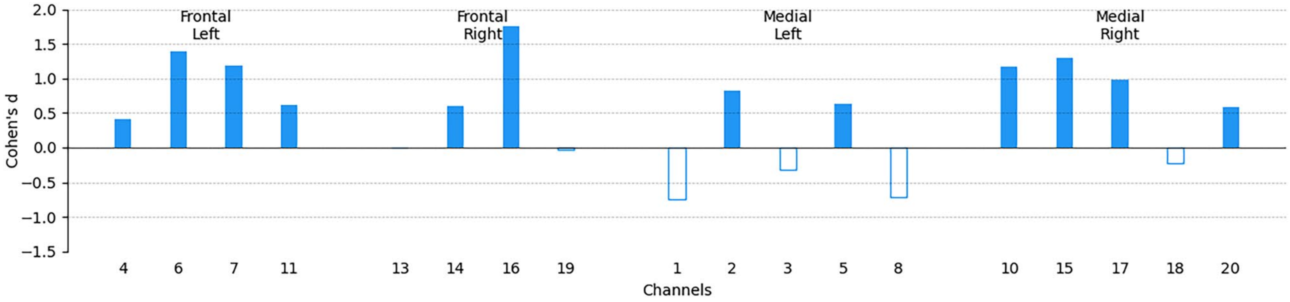

Fig. 5 Bootstrap Cohen's *d* for each channel, using Cross Correlation with no lag

Discussion

In the present study, signals from the published dataset were arranged to obtain two experimental conditions to allow for the adoption of case-control statistical approaches. Four similarity metrics and four lag values (i.e. 0, 1, 2, 5 seconds) were used to assess differences between the two conditions, for each of the auditory stimuli and prefrontal clusters.

Statistical analyses led to the emergence of CC and DTW as the most efficient similarity metrics. These observations are further supported by findings from bootstrapping analyses which identified that CC and DTW consistently depicted greater Cohen's *d* with increasing lag parameters. Conversely, WC and MI metrics still displayed negative Cohen's *d* even at a maximum lag duration of 5 seconds. We note that WC is the predominant metric used to assess interactive hyperscanning studies, due to its ability to extricate fleeting phase-locked processes between interacting dyadic partners (Ikeda et al., 2017). Unlike these interactive studies, the passive presentation of auditory stimuli of standardised lengths in the present study might have favoured CC and DTW metrics. Having homogeneous time-series lengths due to a discrete stimulus duration ensures that normalised DTW indexes are easily interpreted and highly corroborative of CC values.

Finally, CC generated 12 significant channels, with the Medial Left cluster being the region with the least number of significant channels. These findings demonstrate that similarity results vary according to the specific metric, lag parameter and brain region being investigated.

Some similarity metrics and some brain regions showed similar Cohen's *d* values, independently of the lag. These apparently counter-intuitive findings have a methodological interpretation. In fact, the lag used to compute the similarity between signals is the maximum lag that is allowed, not imposed. As a consequence, it might happen that while allowing 5-seconds lag, the actual selected lag is lower. That would cause the measures with different maximum lags to indicate similar Cohen's *d*.

It is worth noting that, with the exception of manipulating the lag parameters, all other parameters in the present study were set to default. Optimising specific parameters in each of these analyses could have led to improved results although the fine-tuning of these parameters could inadvertently heighten the risk of false positives, bias and otherwise result-driven approaches.

Power analyses, based on the observed effect size, revealed that a sample size ranging from N=150 to N=600 is required to achieve robust statistical power. However, since a majority of hyperscanning studies in the prevailing literature do not fullfill the recommended sample size, alternative analytical techniques that account for smaller sample sizes need to be considered. In this study, we proposed the use of an asymptotic bootstrapping method to estimate confidence intervals in a small sample size (Dwivedi et al., 2017; Fisher & Hall, 1991) without any assumption regarding its functional form (Efron & Tibshirani, 1994). Despite its frequent use in studies with small sample sizes, the validity of the proposed bootstrapping technique needs to be investigated and compared against other alternative approaches. One such alternative is bayesian boostrapping, which has been found to produce uncertainty intervals that are more reliable compared to its standard bootstrapping counterpart (Mostofian & Zuckerman, 2019), or the bias-corrected bootstrap method (Sadoun et al., 2020; Carpenter & Bithell, 2000). Future studies should conduct systematic comparisons between standard bootstrapping and alternative methods to arrive at the most optimal technique to analyse small datasets in hyperscanning studies. Given the small sample of the present study, it is worth noting that p-values were not corrected for multiple hypotheses across channels as that would have rendered the results entirely insignificant. The difficulty in obtaining a large sample size in hyperscanning studies raises a salient point of contention in regard to the statistical corrections that need to be made at various levels from bootstrapping to p-value adjustment.

Beyond analytical approaches for small sample sizes, the experimental design of hyperscanning studies deserves significant attention. This present study allowed for the testing of a null hypothesis as it incorporated a control stimulus (i.e., static noise), which was later contrasted against experimental stimuli of human vocalisations (i.e, cries, laughter). In interactive hyperscanning studies, some examples of introducing a control condition are by having a cooperative and competitive condition or a joint compared to independent condition (Wass et al., 2018). We recommend that a similar inclusion of a control condition be adopted in all hyperscanning studies so that null and alternative hypotheses may be tested.

As with all experiments, this study has several limitations that should be noted. First, the common implementation of a passive listening task using in single-scanning fNIRS studies and its clear physiological link to the PFC provides ground for computational research to be conducted using this paradigm. However, despite the ubiquitous use of emotional auditory stimuli in single-scanning fNIRS studies, the physiological underpinning of inter-brain coupling in response to emotional auditory stimuli is less well-known compared to that of more established tasks like synchronised finger-tapping, which has been extensively researched for the past two decades (Jäncke et al., 2000; Konvalinka et al., 2010; Repp, 2005; Vesper & Richardson, 2014). Therefore, the findings from this study must be interpreted with caution and future research may apply our proposed computational methods to well-established paradigms in inter-brain synchrony research such as dyadic joint finger-tapping. Despite this limitation, the present study has contributed significantly to our understanding of computational approaches to synchrony in response to real-world emotional auditory stimuli that go beyond artificial finger-tapping tasks. As there is still a lack of what is considered a "de facto protocol" for studies in the relatively nascent field of fNIRS hyperscanning, future work needs to be conducted to arrive at a 'gold-standard' empirical paradigm from which conclusive computational analyses can be obtained.

Second, the study solely focused on coupling and did not investigate the effect of physical copresence of dyadic partners on each other. Such an endeavour would have required an additional computational analysis to quantify the effect of presence of the dyadic partner (Golland et al., 2014; Golland et al., 2015) and would have reduced effect sizes even further. Third, this study did not assess causality, i.e., how one partner influences and adapts to the other partner, although such analyses would be highly relevant in interactive hyperscanning studies. Fourth, a small dataset numbering only 24 pairs of couples was used which reduced the statistical power of the analyses. Future comparisons across similarity metrics should be conducted on hyperscanning studies with larger datasets.

Information Sharing Statement

The fNIRS data that support the findings of this study are available in Dataverse with the identifier https://doi.org/10.21979/N9/KF1JOG. The analyses were based on custom code written in Python. The code to reproduce the findings is publicly available at https://gitlab.com/abp-san-public/fnirs-synch-methods.

Funding A.B. was supported by a Post-doctoral Fellowship within MIUR programme framework "Dipartimenti di Eccellenza" (DiPSCO, University of Trento). G.E. was supported by NAP SUG 2015, Singapore Ministry of Education ACR Tier 1 (RG149/16 and RT10/19).

Declarations

Ethics Approval Not applicable.

Conflict of Interest The authors declare no conflict of interest.

References

Abibullaev, B., An, J., Jin, S.-H., Lee, S. H., & Moon, J. I. (2013). Minimizing inter-subject variability in fnirs-based brain-computer interfaces via multiple-kernel support vector learning. *Medical Engineering & Physics, 35*(12), 1811–1818.

Ayrolles, A., Brun, F., Chen, P., Djalovski, A., Beauxis, Y., Delorme, R., Bourgeron, T., Dikker, S., & Dumas, G. (2020). HyPyP: a hyperscanning python pipeline for inter-brain connectivity analysis. *Social Cognitive and Affective Neuroscience, 16*(1–2).

Azhari, A., Gabrieli, G., Bizzego, A., Bornstein, M. H., & Esposito, G. (2020a). Probing the association between maternal anxious attachment style and mother-child brain-to-brain coupling during passive co-viewing of visual stimuli. *Attachment & Human Development*, pages 1–16.

Azhari, A., Leck, W., Gabrieli, G., Bizzego, A., Rigo, P., Setoh, P., et al. (2019). Parenting stress undermines mother-child brain-to-brain synchrony: A hyperscanning study. *Scientific Reports, 9*(1), 1–9.

Azhari, A., Lim, M., Bizzego, A., Gabrieli, G., Bornstein, M. H., & Esposito, G. (2020). Physical presence of spouse enhances brain-to-brain synchrony in co-parenting couples. *Scientific Reports, 10*(1), 1–11.

Azhari, A., Rigo, P., Bornstein, M. H., & Esposito, G. (2020c). Where sounds occur matters: Context effects influence processing of salient vocalisations. *Brain Sciences, 10*(7).

Berndt, D. J. & Clifford, J. (1994). Using dynamic time warping to find patterns in time series. In *Proceedings of the 3rd International Conference on Knowledge Discovery and Data Mining*, page 359-370. AAAI Press.

Bigliassi, M., León-Domínguez, U., & Altimari, L. R. (2015). How does the prefrontal cortex listen to classical and techno music? a functional near-infrared spectroscopy (fNIRS) study. *Psychology & Neuroscience, 8*(2), 246–256.

Bilek, E., Ruf, M., Schäfer, A., Akdeniz, C., Calhoun, V. D., Schmahl, C., et al. (2015). Information flow between interacting human brains: Identification, validation, and relationship to social expertise. *Proceedings of the National Academy of Sciences, 112*(16), 5207–5212.

Bizzego, A., Azhari, A., Campostrini, N., Truzzi, A., Ng, L. Y., Gabrieli, G., et al. (2020). Strangers, friends, and lovers show different physiological synchrony in different emotional states. *Behavioral Sciences, 10*(1), 11.

Bizzego, A., Balagtas, J. P. M., & Esposito, G. (2020). Commentary: Current status and issues regarding pre-processing of fnirs neuroimaging data: An investigation of diverse signal filtering methods within a general linear model framework. *Frontiers in Human Neuroscience, 14,* 247.

Bizzego, A., Gabrieli, G., Azhari, A., Setoh, P., & Esposito, G. (2021). Computational methods for the assessment of empathic synchrony. In Esposito, A., Faundez-Zanuy, M., Morabito, F. C., and Pasero, E., editors, Progresses in Artificial Intelligence and Neural Systems. Springer Singapore.

Carpenter, J., & Bithell, J. (2000). Bootstrap confidence intervals: when, which, what? a practical guide for medical statisticians. *Statistics in Medicine, 19*(9), 1141–1164.

Chatel-Goldman, J., Schwartz, J.-L., Jutten, C., & Congedo, M. (2013). Non-local mind from the perspective of social cognition. *Frontiers in Human Neuroscience, 7,* 107.

Chen, Y., Zhang, Q., Yuan, S., Zhao, B., Zhang, P., & Bai, X. (2020). The influence of prior intention on joint action: an fnirs-based hyperscanning study. *Social Cognitive and Affective Neuroscience, 15*(12), 1351–1360.

Chong, J. S., Lu, C. K., & Tang, T. B. (2019). Study of emotional state effect on decision making by using fNIRS. *2019 IEEE International Circuits and Systems Symposium (ICSyS).*

Czeszumski, A., Eustergerling, S., Lang, A., Menrath, D., Gerstenberger, M., Schuberth, S., et al. (2020). Hyperscanning: a valid method to study neural inter-brain underpinnings of social interaction. *Frontiers in Human Neuroscience, 14,* 39.

de Pasquale, F., Della Penna, S., Snyder, A. Z., Lewis, C., Mantini, D., Marzetti, L., et al. (2010). Temporal dynamics of spontaneous meg activity in brain networks. *Proceedings of the National Academy of Sciences, 107*(13), 6040–6045.

Dwivedi, A. K., Mallawaarachchi, I., & Alvarado, L. A. (2017). Analysis of small sample size studies using nonparametric bootstrap test with pooled resampling method. *Statistics in Medicine, 36*(14), 2187–2205.

Efron, B. & Tibshirani, R. J. (1994). *An introduction to the bootstrap*. CRC Press.

Faes, L., Nollo, G., Jurysta, F., & Marinazzo, D. (2014). Information dynamics of brain-heart physiological networks during sleep. *New Journal of Physics, 16*(10).

Fallani, F. D. V., Nicosia, V., Sinatra, R., Astolfi, L., Cincotti, F., Mattia, D., et al. (2010). Defecting or not defecting: how to human behavior during cooperative games by eeg measurements. *PloS one, 5*(12), 1–9.

Fisher, N. I., & Hall, P. (1991). Bootstrap algorithms for small samples. *Journal of Statistical Planning and Inference, 27*(2), 157–169.

Funane, T., Kiguchi, M., Atsumori, H., Sato, H., Kubota, K., & Koizumi, H. (2011). Synchronous activity of two people's prefrontal cortices during a cooperative task measured by simultaneous near-infrared spectroscopy. *Journal of Biomedical Optics, 16*(7).

Gashi, S., Di Lascio, E., & Santini, S. (2018). Using students' physiological synchrony to quantify the classroom emotional climate. In *Proceedings of the 2018 ACM International Joint Conference and 2018 International Symposium on Pervasive and Ubiquitous Computing and Wearable Computers*, page 698-701.

Giorgino, T. (2009). Computing and visualizing dynamic time warping alignments in R: The dtw package. *Journal of Statistical Software, 31*(7), 1–24.

Golland, Y., Arzouan, Y., & Levit-Binnun, N. (2015). The mere co-presence: Synchronization of autonomic signals and emotional responses across co-present individuals not engaged in direct interaction. *PloS one, 10*(5).

Golland, Y., Keissar, K., & Levit-Binnun, N. (2014). Studying the dynamics of autonomic activity during emotional experience. *Psychophysiology, 51*(11), 1101–1111.

Golland, Y., Mevorach, D., & Levit-Binnun, N. (2019). Affiliative zygomatic synchrony in co-present strangers. *Scientific Reports, 9*(1), 1–10.

Grinsted, A., Moore, J. C., & Jevrejeva, S. (2004). Application of the cross wavelet transform and wavelet coherence to geophysical time series. *Nonlinear Processes in Geophysics, 11*(5–6), 561–566.

Handwerker, D. A., Roopchansingh, V., Gonzalez-Castillo, J., & Bandettini, P. A. (2012). Periodic changes in fMRI connectivity. *NeuroImage, 63*(3), 1712–1719.

Hasson, U., & Frith, C. D. (2016). Mirroring and beyond: coupled dynamics as a generalized framework for modelling social interactions. *Philosophical Transactions of the Royal Society B: Biological Sciences, 371*(1693), 20150366.

Hasson, U., Ghazanfar, A. A., Galantucci, B., Garrod, S., & Keysers, C. (2012). Brain-to-brain coupling: a mechanism for creating and sharing a social world. *Trends in Cognitive Sciences, 16*(2), 114–121.

Hasson, U., Nir, Y., Levy, I., Fuhrmann, G., & Malach, R. (2004). Intersubject synchronization of cortical activity during natural vision. *Science, 303*(5664), 1634–1640.

Hinrichs, H., Heinze, H.-J., & Schoenfeld, M. A. (2006). Causal visual interactions as revealed by an information theoretic measure and fMRI. *NeuroImage, 31*(3), 1051–1060.

Huppert, T. J. (2016). Commentary on the statistical properties of noise and its implication on general linear models in functional near-infrared spectroscopy. *Neurophotonics, 3*(1).

Huppert, T. J., Hoge, R. D., Diamond, S. G., Franceschini, M. A., & Boas, D. A. (2006). A temporal comparison of bold, asl, and nirs hemodynamic responses to motor stimuli in adult humans. *Neuroimage, 29*(2), 368–382.

Ikeda, S., Nozawa, T., Yokoyama, R., Miyazaki, A., Sasaki, Y., Sakaki, K., & Kawashima, R. (2017). Steady beat sound facilitates both coordinated group walking and Inter-Subject neural synchrony. *Frontiers in Human Neuroscience*, *11*.

Jäncke, L., Loose, R., Lutz, K., Specht, K., & Shah, N. J. (2000). Cortical activations during paced finger-tapping applying visual and auditory pacing stimuli. *Brain Research. Cognitive Brain Research, 10*(1–2), 51–66.

Jiang, J., Dai, B., Peng, D., Zhu, C., Liu, L., & Lu, C. (2012). Neural synchronization during face-to-face communication. *Journal of Neuroscience, 32*(45), 16064–16069.

Karamzadeh, N., Medvedev, A., Azari, A., Gandjbakhche, A., & Najafizadeh, L. (2013). Capturing dynamic patterns of task-based functional connectivity with EEG. *NeuroImage, 66,* 311–317.

Kawasaki, M., Yamada, Y., Ushiku, Y., Miyauchi, E., & Yamaguchi, Y. (2013). Inter-brain synchronization during coordination of speech rhythm in human-to-human social interaction. *Scientific Reports, 3*(1), 1–8.

Kettunen, J., Ravaja, N., Näätänen, P., Keskivaara, P., & Keltikangas-Järvinen, L. (1998). The synchronization of electrodermal activity and heart rate and its relationship to energetic arousal: A time series approach. *Biological Psychology, 48*(3), 209–225.

Kinreich, S., Djalovski, A., Kraus, L., Louzoun, Y., & Feldman, R. (2017). Brain-to-Brain synchrony during naturalistic social interactions. *Scientific Reports, 7*(1).

Kiviniemi, V., Vire, T., Remes, J., Elseoud, A. A., Starck, T., Tervonen, O., & Nikkinen, J. (2011). A sliding Time-Window ICA reveals spatial variability of the default mode network in time. *Brain Connectivity, 1*(4), 339–347.

Konvalinka, I., Vuust, P., Roepstorff, A., & Frith, C. D. (2010). Follow you, follow me: continuous mutual prediction and adaptation in joint tapping. *The Quarterly Journal of Experimental Psychology, 63*(11), 2220–2230.

Liu, D., Liu, S., Liu, X., Zhang, C., Li, A., Jin, C., et al. (2018). Interactive brain activity: Review and progress on EEG-Based hyperscanning in social interactions. *Frontiers in Psychology, 9,* 1862.

Liu, N., Mok, C., Witt, E. E., Pradhan, A. H., Chen, J. E., & Reiss, A. L. (2016). NIRS-based hyperscanning reveals inter-brain neural synchronization during cooperative jenga game with face-to-face communication. *Frontiers in Human Neuroscience, 10,* 82.

Lizier, J. T. (2014). Jidt: An information-theoretic toolkit for studying the dynamics of complex systems. *Frontiers in Robotics and AI, 1,* 11.

Lombardo, M. V., Auyeung, B., Holt, R. J., Waldman, J., Ruigrok, A. N. V., Mooney, N., et al. (2016). Improving effect size estimation and statistical power with multi-echo fMRI and its impact on understanding the neural systems supporting mentalizing. *NeuroImage, 142,* 55–66.

Mauss, I. B., Levenson, R. W., McCarter, L., Wilhelm, F. H., & Gross, J. J. (2005). The tie that binds? coherence among emotion experience, behavior, and physiology. *Emotion, 5*(2), 175.

Meszlenyi, R. J., Hermann, P., Buza, K., Gál, V., & Vidnyánszky, Z. (2017). Resting state fMRI functional connectivity analysis using dynamic time warping. *Frontiers in Neuroscience, 11,* 75.

Mønster, D., Håkonsson, D. D., Eskildsen, J. K., & Wallot, S. (2016). Physiological evidence of interpersonal dynamics in a cooperative production task. *Physiology & Behavior, 156,* 24–34.

Mostofian, B., & Zuckerman, D. M. (2019). Statistical uncertainty analysis for small-sample, high log-variance data: Cautions for bootstrapping and bayesian bootstrapping. *Journal of chemical theory and computation, 15*(6), 3499–3509.

Nam, C. S., Choo, S., Huang, J., & Park, J. (2020). Brain-to-brain neural synchrony during social interactions: A systematic review on hyperscanning studies. *Applied Sciences, 10*(19), 6669.

Pan, Y., Cheng, X., Zhang, Z., Li, X., & Hu, Y. (2017). Cooperation in lovers: An fNIRS-based hyperscanning study. *Human Brain Mapping, 38*(2), 831–841.

Pan, Y., Novembre, G., Song, B., Zhu, Y., & Hu, Y. (2020). Dual brain stimulation enhances interpersonal learning through spontaneous movement synchrony. *Social Cognitive and Affective Neuroscience*, *16*.

Pinti, P., Scholkmann, F., Hamilton, A., Burgess, P., & Tachtsidis, I. (2019). Current status and issues regarding pre-processing of fnirs neuroimaging data: an investigation of diverse signal filtering methods within a general linear model framework. *Frontiers in Human Neuroscience, 12,* 505.

Pravitha Ramanand, M. C. B. & Bruce, E. N. (2010). Mutual information analysis of eeg signals indicates age-related changes in cortical interdependence during sleep in middle-aged vs. elderly women. *Journal of Clinical Neurophysiology*, 27(4):274.

Repp, B. H. (2005). Sensorimotor synchronization: a review of the tapping literature. *Psychonomic Bulletin & Review, 12*(6), 969–992.

Sadoun, A., Chauhan, T., Mameri, S., Zhang, Y. F., Barone, P., Deguine, O., & Strelnikov, K. (2020). Stimulus-specific information is represented as local activity patterns across the brain. *NeuroImage, 223,*.

Sakoe, H., & Chiba, S. (1978). Dynamic programming algorithm optimization for spoken word recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing, 26*(1), 43–49.

Salvador, R., Martinez, A., Pomarol-Clotet, E., Sarró, S., Suckling, J., & Bullmore, E. (2007). Frequency based mutual information measures between clusters of brain regions in functional magnetic resonance imaging. *NeuroImage, 35*(1), 83–88.

Scholkmann, F., Holper, L., Wolf, U., & Wolf, M. (2013). A new methodical approach in neuroscience: assessing inter-personal brain coupling using functional Near-Infrared Imaging (fNIRI) hyperscanning. *Frontiers in Human Neuroscience, 7,* 813.

Seghouane, A.-K., & Ferrari, D. (2019). Robust hemodynamic response function estimation from fnirs signals. *IEEE Transactions on Signal Processing, 67*(7), 1838–1848.

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell System Technical Journal, 27*(3), 379–423.

Sun, C., Yang, F., Wang, C., Wang, Z., Zhang, Y., Ming, D., & Du, J. (2018). Mutual information-based brain network analysis in post-stroke patients with different levels of depression. *Frontiers in Human Neuroscience, 12,* 285.

Tang, H., Mai, X., Wang, S., Zhu, C., Krueger, F., & Liu, C. (2016). Interpersonal brain synchronization in the right temporo-parietal junction during face-to-face economic exchange. *Social Cognitive and Affective Neuroscience, 11*(1), 23–32.

Taylor, A. J., Kim, J. H., & Ress, D. (2018). Characterization of the hemodynamic response function across the majority of human cerebral cortex. *NeuroImage, 173,* 322–331.

Toppi, J., Borghini, G., Petti, M., He, E. J., De Giusti, V., He, B., et al. (2016). Investigating cooperative behavior in ecological settings: an EEG hyperscanning study. *PloS one, 11*(4).

Torrence, C., & Compo, G. P. (1998). A practical guide to wavelet analysis. *Bulletin of the American Meteorological Society, 79*(1), 61–78.

Trendafilov, D., Schmitz, G., Hwang, T.-H., Effenberg, A. O., & Polani, D. (2020). Tilting together: An Information-Theoretic characterization of behavioral roles in rhythmic dyadic interaction. *Frontiers in Human Neuroscience*, 14.

Vesper, C., & Richardson, M. J. (2014). Strategic communication and behavioral coupling in asymmetric joint action. *Experimental Brain Research, 232*(9), 2945–2956.

Wass, S. V., Noreika, V., Georgieva, S., Clackson, K., Brightman, L., Nutbrown, R., et al. (2018). Parental neural responsivity to infants' visual attention: How mature brains influence immature brains during social interaction. *PLOS Biology, 16*(12).

Zhang, M., Ding, K., Jia, H., & Yu, D. (2018). Brain-to-brain synchronization of the expectation of cooperation behavior: A fNIRS hyperscanning study. In *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 546–549. IEEE.

Zhang, X., Noah, J. A., Dravida, S., & Hirsch, J. (2020). Optimization of wavelet coherence analysis as a measure of neural synchrony during hyperscanning using functional near-infrared spectroscopy. *Neurophotonics, 7*(1).

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.