Multi-time-point analysis: A time course analysis with functional near-infrared spectroscopy

Chi-Lin Yu¹ · Hsin-Chin Chen² · Zih-Yun Yang² · Tai-Li Chou^{1,3}

Published online: 5 February 2020
© The Psychonomic Society, Inc. 2020

Abstract

In the data analysis of functional near-infrared spectroscopy (fNIRS), linear model frameworks, in particular mass univariate analysis, are often used when researchers consider examining the difference between conditions at each sampled time point. However, some statistical issues, such as assumptions of linearity, autocorrelation and multiple comparison problems, influence statistical inferences when mass univariate analysis is used on fNIRS time course data. In order to address these issues, the present study proposes a novel perspective, multi-time-point analysis (MTPA), to discriminate signal differences between conditions by combining temporal information from multiple time points in fNIRS. In addition, MTPA adopts the random forest algorithm from the statistical learning domain, followed by a series of cross-validation procedures, providing reasonable power for detecting significant time points and ensuring generalizability. Using a real fNIRS data set, the proposed MTPA outperformed mass univariate analysis in detecting more time points, showing significant differences between experimental conditions. Finally, MTPA was also able to make comparisons between different areas, leading to a novel viewpoint of fNIRS time course analysis and providing additional theoretical implications for future fNIRS studies. The data set and all source code are available for researchers to replicate the analyses and to adapt the program for their own needs in future fNIRS studies.

Keywords fNIRS · time series · linear model · mass univariate analysis · random forest

Introduction

Functional near-infrared spectroscopy (fNIRS) is a noninvasive tool for recording hemodynamic activity along the scalp time-locked to response events. By measuring the absorption of the near-infrared light (650–950 nm) through the scalp (Villringer & Chance, 1997; Villringer & Dirnagl, 1994), fNIRS can detect hemodynamic changes in the concentration of oxyhemoglobin and deoxy-hemoglobin in brain regions (Ferrari, Mottola, & Quaresima, 2004; Jobsis, 1977; Kleinschmidt et al., 1996; Strangman, Boas, & Sutton,

2002a; Strangman, Culver, Thompson, & Boas, 2002b; Villringer, Planck, Hock, Schleinkofer, & Dirnagl, 1993). Previous studies suggest that fNIRS data are highly consistent with data from the most widely used neuroimaging modality, functional magnetic resonance imaging (fMRI) (Strangman, Boas, et al., 2002a; Strangman, Culver, et al., 2002b), but fNIRS further provides many advantages over other neuroimaging tools. For example, acceptable time resolution of brain activity for determining chronological components of mental processes is an important feature of fNIRS (Rossi, Telkemeyer, Wartenburger, & Obrig, 2012; Wallois, Mahmoudzadeh, Patil, & Grebe, 2012). Based on its temporal resolution, fNIRS can thus provide specific time course information for physical or mental events in the human brain. Other advantages of fNIRS include its portability and robustness to motion, making it suitable for field examination or under dynamic movement (Arenth, Ricker, & Schultheis, 2007; Ferrari et al., 2004; Hoshi, 2003; Irani, Platek, Bunce, Ruocco, & Chute, 2007; Rossi et al., 2012; Wallois et al., 2012).

In fNIRS statistical analysis, the most common approach is to *average* signals across all time points of the event of interest, but this technique is limited by loss of time course information. Statistical methods such as the *t* test or analysis of variance (ANOVA) would then be applied on these averaged

Electronic supplementary material The online version of this article (<https://doi.org/10.3758/s13428-019-01344-9>) contains supplementary material, which is available to authorized users.

✉ Tai-Li Chou
tlchou25@ntu.edu.tw

¹ Department of Psychology, National Taiwan University, No. 1, Sec. 4, Roosevelt Road, Taipei 106, Taiwan

² Department of Psychology, National Chung Cheng University, Chiayi, Taiwan

³ Graduate Institute of Brain and Mind Sciences, National Taiwan University, Taipei, Taiwan

values in order to draw a conclusion (Germon et al., 1999; Hoshi, 2003; Hoshi, Kobayashi, & Tamura, 2001; Isobe et al., 2001; Kleinschmidt et al., 1996; Mehagnoul-Schipper et al., 2002; Okamoto et al., 2004). However, this averaging approach is not sensitive for detecting temporal information of brain activation related to mental activity (Tak & Ye, 2014). In other words, the relationship between cognitive states and the time course of brain signals may not be revealed when using this approach. Other methods such as the general linear model (Penny, Friston, Ashburner, Kiebel, & Nichols, 2011) are therefore adopted to understand the fNIRS time course (Abdelnour & Huppert, 2009; Custo et al., 2010; Koh et al., 2007; Minagawa-Kawai et al., 2010; Plichta et al., 2006; Plichta et al., 2007; Shimada & Hiraki, 2006; Singh & Dan, 2006; Ye, Tak, Jang, Jung, & Jang, 2009). However, the general linear model is still limited in that it converts hemodynamic response function (HRF) curves into a single beta value, which only represents an activation level rather than real-time course information of individual trials.

In order to obtain the time course information for fNIRS data, researchers usually implement *mass univariate analysis*. Under the linear model framework, mass univariate analysis can examine the differences between different experimental conditions at each individual time point (H.-C. Chen, Vaid, Boas, & Bortfeld, 2011; H.-C. Chen, Vaid, Bortfeld, & Boas, 2008; Mihara, Miyai, Hatakenaka, Kubota, & Sakoda, 2008; Yu, Wang, & Hu, 2016). For example, H.-C. Chen et al. (2011) used the one-sample *t* test to directly examine whether increases in oxyhemoglobin concentration significantly differed from zero under each sampled time point, and used the paired *t* test to examine whether increases in oxyhemoglobin concentration differed significantly from one condition to another. Problems with multiple comparisons as a result of hundreds of time points in HRF curves have been somewhat considered in the literature by using significance-adjusted methods, such as the family-wise error rate (FWER) and the Benjamini-Hochberg false-discovery rate (FDR) control procedures (Benjamini & Hochberg, 1995), to reduce the type I errors in inference. For instance, in the aforementioned study by Chen et al. (2011), the testing at each sampled time point is corrected by using one of the FWER controls, namely the Bonferroni test (Dunn, 1961). Based on these procedures of mass univariate analysis, it seems that neuroscientists can make further inferences regarding the time course of fNIRS data.

Although findings obtained by mass univariate analysis have increased our understanding of fNIRS time course, several methodological issues still exist that actually limit inference and generalizability. First, it is apparent that adjusting the level of statistical significance, which is necessary for multiple testing in an fNIRS time course, will decrease the power of the testing procedure under a given effect size. Moreover, even if some studies overcome the power problem in multiple-

hypothesis-testing procedures, the results might not be reliable, due to the presence of structural dependencies and correlated time points of fNIRS data (Efron, 2007; Leek & Storey, 2008). This autocorrelation issue is not considered in mass univariate analysis, which assumes that the covariance across neighboring time points is not informative. Even with the use of the improved extension of the Benjamin-Hochberg procedure (Benjamini & Yekutieli, 2001), which modified the original method and tried to take into account the dependencies between tests, the problem remained unaddressed with this type of highly auto-correlated data (Causeur, Chu, Hsieh, & Sheu, 2012). Furthermore, mass univariate analysis assumes that the relationships between brain signals and psychological states are linear. However, the assumptions of a linear relationship between the brain and behavior are inevitably limited by the problem of individual differences. Consider a situation in which 10 of 20 people have higher hemodynamic changes in the concentration of oxyhemoglobin during one condition and lower during another condition, but the reverse is shown for the other 10 people. These differences will not be detected using mass univariate analysis. Thus these shortcomings must be taken into consideration when using mass univariate analysis, which may not be able to correctly detect the differences between experimental conditions.

The objective of the current study was to provide a novel perspective that we call *multi-time-point analysis* (MTPA). This method aims to extract fNIRS time course information to discriminate signal differences between different underlying cognitive states (i.e., experimental conditions). In general, MTPA treats fNIRS data analysis as a supervised classification problem, where the experimental condition is the variable of primary interest and the signals at each time point are predictors. A tree-based statistical learning method, the random forest (Breiman, 2001), was implemented as the major algorithm. The name MTPA is analogous to the popular technique multi-voxel pattern analysis (MVPA), which also adopts statistical learning algorithms for discriminating cognitive states, in the fMRI domain (Etzel, Gazzola, & Keysers, 2009; Haxby, Connolly, & Guntupalli, 2014; Haynes, 2015; Mitchell et al., 2004; Mur, Bandettini, & Kriegeskorte, 2009; Norman, Polyn, Detre, & Haxby, 2006; Pereira, Mitchell, & Botvinick, 2009). In contrast to MVPA that can “spatially” characterize the information from multi-voxels in fMRI (Norman et al., 2006), the proposed MTPA can “temporally” decode the neural activity from multiple time points in fNIRS.

The following sections are organized as a series of demonstrations on an fNIRS data set. We first apply mass univariate analysis to the data set. The proposed MTPA is then used to analyze the same data set. Finally, we compare the MTPA with the mass univariate analysis, and describe the pros and cons of the two methods.

Materials and methods

Participants

Fifteen native Chinese speakers (10 women), aged 20–29 years (mean = 21.9, SD = 2.5), from National Chung Cheng University were recruited for the experiment. All were compensated NT\$250 for their participation. The data for one subject was excluded due to severe motion. The remaining 14 subjects were right-handed, with normal or corrected-to-normal vision, and no specific disease or cognitive disorder. The experiment was conducted following local institutional review board regulations, and all participants provided their written consent before the experiment.

Materials and functional activation task

Forty-eight Chinese character pairs were semantically related according to their free association values (mean = 0.14, SD = 0.13, range 0.73–0.01), and 24 Chinese character pairs were semantically unrelated, with zero association values. The characteristics of these 72 stimuli were the same as in a prior study (Chou, Chen, Wu, & Booth, 2009b). The subjects were asked to complete a meaning judgment task. In the task, two visual Chinese characters were presented sequentially, and the subjects had to quickly and accurately indicate whether or not the character pair was related in meaning (related or unrelated) by pressing the yes or no buttons with their right hand. Furthermore, 24 pairs of non-characters, which were made by replacing radicals of real characters with other radicals that did not form real Chinese characters, were included as the perceptual control condition. Participants were asked to indicate whether or not the two stimuli were identical by pressing a yes or no button as quickly and as accurately as possible. There were also 24 baseline trials with the first stimulus as a solid square and the second stimulus as a hollow square. Participants were to press a button as soon as the solid square turned into the hollow square. For each trial, subjects first saw a fixation signal (a solid square) (500 ms) presented at the center of the screen, followed by the first character or stimulus (800 ms), a 20-ms blank interval, and the second character or stimulus (3000 s). Each participant received a different randomized sequence of these 120 stimulus pairs. Before the formal experiment, subjects received 15 practice trials to become familiar with the procedure. Note that only the comparison between “Related” and “Unrelated” conditions was used for the present fNIRS analyses.

Apparatus

The experiment was conducted on a personal computer using E-Prime software (Schneider, Eschman, & Zuccolotto, 2002) for presenting stimuli. The multichannel fNIRS optical system

NIRScout (NIRx Medical Technologies, USA) was adopted to monitor cortical hemodynamic changes. During the task, the optical signals were simultaneously collected by the fNIRS electronic control box serving as both the source and the receiver of the near-infrared light. By emitting two wavelengths of near-infrared light (760 nm and 850 nm) through the scalp and analyzing the characteristics of their subsequent absorption and scattering by the modified Beer–Lambert law, fNIRS can measure changes in the concentration of oxyhemoglobin at the brain regions of interest (Strangman, Boas, et al., 2002a; Strangman, Culver, et al., 2002b).

The present study focused on left inferior frontal gyrus (IFG) and left middle temporal gyrus (MTG), which were the most two consistently activated regions in previous language studies (Gow, 2012; Hagoort, 2005; Jefferies, 2013), as the primary regions of interest. Therefore, we utilized four emitters to direct the two wavelengths of near-infrared light through the scalp and four detectors to receive the returning near-infrared light. Eight channels thus formed by pairs of emitters and detectors covered these regions of interest. Specifically, detectors and emitters were placed on subjects’ heads following the international 10–20 system (Homan, Herman, & Purdy, 1987; Koessler et al., 2009; Okamoto et al., 2004). Note that the distance between a detector and an emitter was about 3 cm, which was specified to ensure detection of the near-infrared light following penetration of the neocortex in adults (Hebden & Delpy, 1997). In brief, channels 1 to 4 covered the left IFG, and channels 5 to 8 covered the left MTG (Fig. 1).

Results

Imaging preprocessing

The fNIRS data from eight channels were digitally recorded at a 10 Hz sampling rate. Using HOMER software (Huppert & Boas, 2005), the data were converted into optical density units that were digitized and band-pass-filtered (1 Hz to 0.02 Hz) to reduce noise, and then finally converted to reflect the concentration of oxyhemoglobin for further analysis. The converted data were averaged from each individual trial and analyzed in 16-s epochs including the onset (0 s) to the end (16 s) of the stimuli, indicating that the total number of sampling is 167 time points in each curve. Furthermore, the present study aggregated the signal from channel 1 to channel 4, which covered the IFG in the left hemisphere, as the primary estimated region of interest. In other words, there was only one curve per condition for each subject in the left IFG after aggregating. It is note that the experimental condition was a within-subject factor with two conditions, namely “Related” and “Unrelated”. The major purpose of this fNIRS experiment was to investigate the differences between these two

Fig. 1 Illustration of emitter and detector positions. Filled circles with letters indicate laser detectors, open circles with letters indicate laser emitters, and the gray squares with numbers indicate channels. The emitters and detectors were placed on subjects' heads following the international 10-20 system (Homan et al., 1987; Koessler et al., 2009; Okamoto et al., 2004). Channels 1 to 4 covered the left IFG, and channels 5 to 8 covered the left MTG. Specifically, detector A (FP1) received the near-infrared light from emitter A (AF7) and emitter B (3 cm dorsal sides

of FP1), and respectively formed channel 2 and channel 1. Detector B (3 cm dorsal sides of AF7) formed channel 4 and channel 3. Similarly, detector C (TP7) received the near-infrared light from emitter C (P7) and emitter D (3 cm dorsal sides of TP7), and formed channel 6 and channel 5. Detector D (P7) also formed channel 8 and channel 7. The distance (3 cm) between a detector and an emitter was suggested by Hebden and Delpy (1997)

conditions. Note that the analysis and the results of the left MTG (channel 5 to channel 8) is not shown in the following sections in order to simplify the demonstration steps.

First, we used plots of signals for all subjects in the “Related” and “Unrelated” conditions to identify outliers and further check the overall fNIRS pattern (Fig. 2a). In addition, the average fNIRS curves with pointwise confidence intervals (Fig. 2b) were used to investigate the relationships between these two conditions. Although these results did not correct for multiple comparisons, the “Related” condition generally had higher activation than the “Unrelated” condition in the left IFG.

Mass univariate analysis

Conducting mass univariate analysis, such as a paired t test, at each sampled time point is a widely used approach to obtain the time course information of fNIRS. This method computed signals from individual time points as a dependent variable, and examined the difference in signals between conditions, such as the “Related” and “Unrelated” conditions in the present study. Because there were 167 time points in a curve for a

condition, the mass univariate analysis required hypothesis testing 167 times using a paired t test.

The p value corrections must be performed because of multiple comparisons. With a FWER controlled by the Bonferroni test, the present study did not reveal any significant differences between these two conditions. Using a popular alternative, the Benjamini and Hochberg FDR correction procedure (Benjamini & Hochberg, 1995), there was still no significant difference between these two conditions (Fig. 3). Another extension of the FDR (Benjamini & Yekutieli, 2001), which claimed to account for the dependencies across multiple tests (called *positive dependence*), also revealed no significant differences (Supplementary Figure 1).

Since the aforementioned p -value corrections were somewhat conservative, we also applied nonparametric permutation frameworks, which might be a preferred technique for increasing statistical power (Nichols & Holmes, 2002; Nichols & Hayasaka, 2003). The central idea in the use of nonparametric permutation frameworks was to use the collected data itself to generate an empirical null distribution of the maximum statistics. Through permutation of the labels of the data, a null data set could be generated (Good, 2013), the

Fig. 2 Exploration of fNIRS data for the left IFG. Panel **a** shows the individual fNIRS curves by condition. Panel **b** shows the confidence intervals for mean fNIRS curves in two conditions

Fig. 3 Significance testing results between related and unrelated conditions for mass univariate analysis. The time series is shown on the x-axis. The y-axis depicts the $-2\log_{10}(p \text{ value})$ rather than the p value for visualization. It is thus crucial to note that the larger value on the y-axis indicates a smaller p value. The curve indicates the p value after the Bonferroni correction (green curve) and the BH method (Benjamini & Hochberg, 1995) of FDR correction (dark red curve) at each corresponding time point. The red line indicates the $p = 0.05$ corrected threshold. The p value at each time point is specified as significant only when it reaches the $p = 0.05$ significance level (red line)

model could be fit, and the maximum statistic could be recorded (Nichols & Holmes, 2002; Nichols & Hayasaka, 2003). An empirical null distribution of the maximum statistic could be accordingly provided by repeating this process many times. By computing the 95th percentile of this null distribution, we could obtain a significance threshold with the FWER controlled at a level of $p = 0.05$. In the present study, two non-parametric permutation frameworks, the maximum t -statistic and the maximum suprathreshold temporal cluster size, were adopted to account for multiple comparison problems in the mass univariate analysis (Nichols & Holmes, 2002).

In the nonparametric permutation framework with the maximum t -statistic (Nichols & Hayasaka, 2003), we permuted across the labels, which was the combination of subject and condition in this case. That is, each curve was randomly relabeled with a new label using a combination of subject and condition. There were $28!$ factorial ways to permute labels. Since a good approximation of the permutation distribution could be made with enough relabeling (Nichols & Holmes, 2002), we randomly relabeled the data 10,000 times to compute the statistics. For each relabeling, 167 paired t tests were computed, and the maximum t -statistic was recorded. Thus, a permutation-based null distribution of the maximum t -statistic was generated to assess the significance of the experiment. The 95th percentile of this maximum t -statistic distribution provided a significance threshold appropriate for each time point ($p = 0.05$). The significance threshold was 3.32 in the present study (Fig. 4, left panel). Any time point with a t value greater than 3.32 could be claimed to be

Fig. 4. The results of nonparametric permutation frameworks. The null distribution of the maximum t -statistic and the distribution of the maximum suprathreshold temporal cluster size are shown. Dotted lines

show the 95th percentile for each distribution ($p = 0.05$). No significant difference was revealed in these two nonparametric permutation frameworks

significant at a level of $p = 0.05$. With a FWER controlled by this maximum t -statistic framework, nonparametric permutation results showed no significant difference between two conditions.

In another nonparametric permutation framework, the maximum suprathreshold temporal cluster size (max STCS) was used as the maximum statistic (Nichols & Holmes, 2002). In previous fMRI and PET studies, the suprathreshold spatial cluster size was used as the maximum statistic to control the FWER. Spatial clusters were defined by connected suprathreshold regions (e.g., a group of significant voxels). Large spatial clusters suggested significant differences between the two conditions. Likewise, in the present study, temporal clusters were defined by connected suprathreshold time points. Hence, large temporal clusters could suggest a significant difference between two conditions. We used a permutation procedure similar to the aforementioned one. The only difference was that the recorded maximum statistic was the max STCS. Similarly, a permutation-based null distribution of max STCS was generated. The 95th percentile of this max STCS distribution was used as the significance threshold, which was 37 in the present study (Fig. 4, right panel). Any temporal cluster with size greater than 37 time points could be declared significant at the $p = 0.05$ level. In this framework, no significant temporal cluster was revealed.

In addition to the issues of power, mass univariate analysis assumes that the covariance across neighboring time points is not informative, but this is not the case in highly auto-correlated fNIRS data. Also, the assumption of linear relationships between the brain (fNIRS data) and behaviors (conditions) would also influence the inferences by mass univariate analysis.

To briefly sum up, it was demonstrated that, regardless of the correction method used (i.e., Bonferroni and two versions

of FDR) and the nonparametric permutation framework applied (i.e., maximum t -statistic and maximum STCS-statistic), mass univariate analysis may not provide reliable results with the analysis of fNIRS time course data.

Multi-time-point analysis

Overall objectives: using the signals of time points to discriminate conditions

Due to the aforementioned issues including low power, auto-correlation problems, and the assumptions of linearity, mass univariate analysis was not able to detect the significant differences between conditions, and might further bias the results, conclusions and implications. In order to address these issues, the present MTPA method provided a different perspective that uses time course information to classify two conditions, instead of examining the difference between two conditions at each time point, and adopts a popular ensemble-learning model, the random forest (Breiman, 2001). This perspective transforms the original questions to a supervised classification problem (Friedman, Hastie, & Tibshirani, 2001; James, Witten, Hastie, & Tibshirani, 2013). Thus, we treated each time point as a predictor but not a dependent variable, and experimental conditions were considered as the dependent variable but not an independent variable. In general, this concept of classification was somehow similar to some much more well-known methods, such as logistic regression and discriminant analysis. That is, we used the information of predictors (called x) to make the “fitted” classification (called $\hat{y}$) as close as possible to the actual category (called y). Some indices, such as accurate classification rates, were also used to evaluate the performance of the model. Importantly, MTPA outperformed mass univariate analysis by addressing three

main issues and providing more correct detection. The concepts and procedures of the proposed method are specifically and concretely described as follows.

Data description

First, we let x_{ij} represent the changes in the concentration in oxyhemoglobin of the j th time point for the i th observation, where $i = 1, 2, \dots, n$ and $j = 1, 2, \dots, p$. The index n denotes the total number of observations by the combination of subjects and conditions, and p represents the total time points. X denotes an $n \times p$ matrix whose (i, j) th element is x_{ij} . For example, x_{61} represents the signal of the first time point for the third subject under the second condition (sixth observation). In the present study, the X matrix, which represents “independent variables” in the data, has 28 rows (14 subjects by 2 conditions) and 167 columns, with y_i denoting the i th value of the condition. Similarly, $i = 1, 2, \dots, n$, and n denotes the total number of observations. Y denotes an $n \times 1$ vector that contains all observations (whose i th element is y_i), and it only contains our target “dependent variable”. For example, y_6 represents the second condition for the third subject. The Y of the present data has 28 observations (14 subjects by 2 conditions), which is matched one-by-one to X , labeling “Related” and “Unrelated”. That is,

$$Y = \begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ \vdots \\ y_n \end{pmatrix}; X = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ x_{31} & x_{32} & \cdots & x_{3p} \\ x_{41} & x_{42} & \cdots & x_{4p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix},$$

The data can be specifically presented as the following matrices:

$$Y = \begin{pmatrix} R_1 \\ U_2 \\ R_3 \\ U_4 \\ \vdots \\ U_n \end{pmatrix}; X = \begin{pmatrix} S_{11} & S_{12} & \cdots & S_{1p} \\ S_{21} & S_{22} & \cdots & S_{2p} \\ S_{31} & S_{32} & \cdots & S_{3p} \\ S_{41} & S_{42} & \cdots & S_{4p} \\ \vdots & \vdots & \ddots & \vdots \\ S_{n1} & S_{n2} & \cdots & S_{np} \end{pmatrix},$$

where R indicates the “Related” condition, U indicates the “Unrelated” condition, and S represents the changes in the concentration in oxyhemoglobin. The n in the matrix equals 28, and the total time point p equals 167.

Data partitioning: choosing bandwidth to determine the subset data

Second, we selected some columns of X to perform the first model fitting. The reason that we did not use whole X to predict Y was due to a problem called “ill-posed” or

“underdetermined” (O’Sullivan, 1986). This occurs when the number of time points (p) is greater than the number observations (n), which is typical in fNIRS studies. Instead of using all time points at one time, we selected a bandwidth, which indicated the number of time points in X , to sequentially select neighboring time points for model fitting until all time points were used as predictors. The bandwidths of 2, 3, 4, 5, and more could be implemented in the present study. However, we chose to start from a bandwidth of 2, because it included the two immediate neighboring time points as the bandwidth and considered the strongest autocorrelations between the signals at those neighboring time points. Also, this choice echoed the rationale of a widely used *down-sampling* approach (Chen et al., 2011; Chen et al., 2008; Khan, Hong, & Hong, 2014; Scholkmann, Wolf, & Wolf, 2013; B. Xu et al., 2014; Yu et al., 2016; Zimmermann et al., 2013), which typically averages several neighboring time points to a 5 Hz rate ($\frac{\text{Our sampling rate}}{\text{Bandwidth}} = \frac{10 \text{ Hz}}{2} = 5 \text{ Hz}$). Furthermore, implementing larger bandwidths did not provide additional benefit to constitute a significant improvement over a bandwidth of 2 in the present NIRS data set (the results of a bandwidth of 3 are also shown in Supplementary Figure 2 for demonstration purposes). For these reasons, we suggest using the more parsimonious value, the bandwidth of 2, which was also used for all the demonstrations in the present study. For example, in the following matrices, we selected the first to second time points inclusively as the predictors. The second to third time points inclusively would then be selected as the next predictors, followed by the third to fourth time points, and finally the 166th to 167th time points inclusively would be selected.

$$Y = \begin{pmatrix} R_1 \\ U_2 \\ R_3 \\ U_4 \\ \vdots \\ U_n \end{pmatrix}; X = \begin{pmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \\ S_{31} & S_{32} \\ S_{41} & S_{42} \\ \vdots & \vdots \\ S_{n1} & S_{n2} \end{pmatrix}$$

Model fittings and evaluations within a subset data

When the subset data, such as the above matrices, were determined (based on the bandwidth), we moved on to build the model. The subset data were then randomly divided into a training set to obtain a fitted random forest model, and a testing set to evaluate model performance. Cross-validation procedures such as K -fold and Hangout cross-validations (Bengio & Grandvalet, 2004; Efron, 1983; Refaailzadeh, Tang, & Liu, 2009) were adopted to investigate whether the model was generalizable (rather than only closely fitting with the present data) (Refaailzadeh et al., 2009). Approximately two-thirds of the observations (20) of the entire subset data with an equal number of “Related” and “Unrelated”

conditions were chosen for the training set, and the remaining one-third (8) were automatically assigned to the testing set. The following matrices demonstrate the components of the training and testing sets.

$$Y_{\text{training } 1} = \begin{pmatrix} R_1 \\ U_2 \\ R_3 \\ U_4 \\ \vdots \\ U_{20} \end{pmatrix}; X_{\text{training } 1} = \begin{pmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \\ S_{31} & S_{32} \\ S_{41} & S_{42} \\ \vdots & \vdots \\ S_{20 \ 1} & S_{20 \ 2} \end{pmatrix}$$

$$Y_{\text{testing } 1} = \begin{pmatrix} R_{21} \\ U_{22} \\ R_{23} \\ U_{24} \\ \vdots \\ U_{28} \end{pmatrix}; X_{\text{testing } 1} = \begin{pmatrix} S_{21 \ 1} & S_{21 \ 2} \\ S_{22 \ 1} & S_{22 \ 2} \\ S_{23 \ 1} & S_{23 \ 2} \\ S_{24 \ 1} & S_{24 \ 2} \\ \vdots & \vdots \\ S_{28 \ 1} & S_{28 \ 2} \end{pmatrix}$$

Next, we fitted a random forest model on the training set, and we obtained a trained model. We then moved on to apply the model to the testing set. Classification errors can be reflected by the receiver operating characteristic (ROC) curve or area under the ROC curve (AUC) to show model performance under cross-validations (Friedman et al., 2001; Hanley & McNeil, 1982; James et al., 2013). The ROC curve is particularly popular for simultaneously considering both the false-positive and true-positive rates for classification patterns, and the value of AUC further suggests a value for the overall performance of the model with the subset data in terms of the ROC (Hanley & McNeil, 1982). The index AUC ranging between 0 and 1 is very useful when conducting two categories of classification, such as the two conditions in the present study. A model that performs better than an AUC value of 0.5 indicates performance above the level of chance, and the larger the AUC value, the better the performance in classifying two different conditions (Friedman et al., 2001; Hanley & McNeil, 1982; James et al., 2013). Note that though the AUC is a good evaluation method and is strongly suggested for two-condition classification problems, using classification errors (or classification accuracies) as evaluations will be easy and straightforward for multi-condition problems. In the present study, we applied the model produced by the training set ($X_{\text{training } 1}$ and $Y_{\text{training } 1}$) to the testing set ($X_{\text{testing } 1}$ and $Y_{\text{testing } 1}$), and further obtained the AUC value of the testing set for model performance. Nevertheless, because of a relatively small sample size for typical fNIRS studies (i.e., 14 subjects and 2 conditions, with 28 observations), performing cross-validations many times has been suggested as a way to better estimate the variability in the results (Q.-S. Xu & Liang, 2001). Repeated estimations create a large number of AUC values. For example, we used 100 cross-validations on the subset data (X_{training} , Y_{training} , X_{testing} and Y_{testing}), so we had 100 AUC values in the present study. These AUC values

represented the ability to predict the dependent variable (two conditions).

The above procedure was conducted to build the model of the first to second time points in X. In other words, we had a set of 100 AUC values in this subset data with the first to second time points. We could then move on to the next two time points (the second to third time points), as shown in the following matrices.

$$Y_{\text{training } 2} = \begin{pmatrix} R_1 \\ U_2 \\ R_3 \\ U_4 \\ \vdots \\ U_{20} \end{pmatrix}; X_{\text{training } 2} = \begin{pmatrix} S_{12} & S_{13} \\ S_{22} & S_{23} \\ S_{32} & S_{33} \\ S_{42} & S_{43} \\ \vdots & \vdots \\ S_{20 \ 2} & S_{20 \ 3} \end{pmatrix};$$

$$Y_{\text{testing } 2} = \begin{pmatrix} R_{21} \\ U_{22} \\ R_{23} \\ U_{24} \\ \vdots \\ U_{28} \end{pmatrix}; X_{\text{testing } 2} = \begin{pmatrix} S_{21 \ 2} & S_{21 \ 3} \\ S_{22 \ 2} & S_{22 \ 3} \\ S_{23 \ 2} & S_{23 \ 3} \\ S_{24 \ 2} & S_{24 \ 3} \\ \vdots & \vdots \\ S_{28 \ 2} & S_{28 \ 3} \end{pmatrix}.$$

Model evaluations across subset data

We applied the same procedures, including the model fittings and evaluations, to the subset data with the second to third time points ($X_{\text{training } 2}$, $Y_{\text{training } 2}$, $X_{\text{testing } 2}$ and $Y_{\text{testing } 2}$). Now we had two sets of 100 AUC values. The first set of 100 AUC values was produced by the model with time points 1 to 2 ($X_{\text{training } 1}$, $Y_{\text{training } 1}$, $X_{\text{testing } 1}$ & $Y_{\text{testing } 1}$), and the second set was produced by the model with time points 2 to 3 ($X_{\text{training } 2}$, $Y_{\text{training } 2}$, $X_{\text{testing } 2}$ and $Y_{\text{testing } 2}$). Because the signals at either time point 1 or time point 3 strongly influenced the signal at time point 2, using these two sets of 100 AUC values to estimate the performance of time point 2 gave us a relatively unbiased result. In other words, the first set of 100 AUC values reflected the model performance of time point 2 under consideration of time point 1, and the second set of 100 AUC values reflected the model performance of time point 2 under consideration of time point 3. By combining the AUC values from two sets, we obtained a distribution of 200 AUC values on time point 2 after considering time points 1 and 3. If the 90% confidence interval of the AUC distribution *did not* cover AUC = 0.5 (Friedman et al., 2001; Hanley & McNeil, 1982; James et al., 2013), the signals at time point 2 were indeed able to discriminate the “Related” and “Unrelated” conditions in the present study. Specifically, there was a significant difference between two conditions at time point 2. In contrast, if the 90% confidence interval of the AUC distribution *did* cover AUC = 0.5, there was no significant difference between two conditions at time point 2.

Next, we continued to move our bandwidth window of 2 to sequentially fit the remaining time points, and calculated the confidence interval of the AUC values at each time point using the aforementioned methods until the last (167th) time point. For example, following time points 2 to 3, we fit the model with time points 3 to 4, and obtained the confidence interval of time point 3...and so on. Note that there would be 200 AUC values to calculate the confidence interval for each time point, except for 100 AUC values for the first and last time points. One might argue that the proposed method still encountered the multiple comparison issues because of performing many confidence interval calculations; however, it is worth pointing out that the proposed method already addressed this concern by adopting the cross-validation method to limit type I errors and ensure generalizability (Browne, 2000; Warner, 2012). Thus, the additional corrections were not required with the proposed method.

Summary of the results of model evaluations

Using the proposed MTPA procedure, we obtained the confidence interval of the AUC values at each time point, and we further showed whether the signals at a specific time point were able to detect the differences between “Related” and “Unrelated” conditions (Fig. 5, right panel). The proposed method was sensitive for detecting much more significant time points (e.g., 10 to 12.5 s) relative to mass univariate analysis

(Fig. 5, left panel). In general, the proposed method revealed a total of 82 time points showing significant differences.

Comparisons between the left IFG and the left MTG

In addition to the left IFG (channels 1 to 4), the same MTPA analysis was able to be applied to the left MTG (channels 5 to 8). As shown in Fig. 6, a total of 82 time points showing significant differences were revealed in the left IFG, whereas only 15 significant time points were detected in the left MTG. As to the average AUC curves in these two areas, in the left IFG the AUC values were higher, indicating better performance in discriminating different experimental conditions, than the left MTG at most of the time points. Moreover, the first significant time point did not appear until 7.86 s (83rd time point) in the left MTG, whereas in the left IFG there were 28 significant time points by 7.86 s.

Recap of multi-time-point analysis

Here we summarize in seven concrete steps our proposed MTPA procedure for significance testing of fNIRS as discussed above.

1. Choose a bandwidth. The suggested value is 2, because it includes the two immediate neighboring time points to consider the strongest influences between the signals at

Fig. 5. Results of the differences between “Related” and “Unrelated” conditions for two approaches. The results of mass univariate analysis using AUC evaluations for comparisons are shown in the left panel, and the results of MTPA are shown in the right panel (a total of 82 significant time points were detected using MTPA.). The time series are shown on

the x-axis, and AUC values are shown on the y-axis. The ribbon areas in both panels indicate the 90% confidence interval at each time point, and the curves indicate the averaged AUC values at each time point. The red line indicates the threshold of AUC = 0.5

Fig. 6. The results of the left IFG (LIFG) and the left MTG (LMTG) using MTPA. Panel **a** shows the confidence interval of AUC values at each sampled time point. The ribbon areas in panel **a** indicate the confidence interval of 90% at each time point. The curves indicate the averaged AUC values. The red line demonstrates the threshold of AUC = 0.5.

A total of 82 significant time points are revealed in the left IFG, and a total of 15 significant time points are shown in the left MTG. Panel **b** shows the comparison between the LIFG and the LMTG by averaged AUC values. In both panels, the time series are shown on the *x*-axis, and AUC values are shown on the *y*-axis.

- each two immediate neighboring time points, and it also echoes the widely used down-sampling 5-Hz rate in the literature.
2. Select the data from the first time point to “1 + bandwidth – 1” time point. For example, if the bandwidth is 2, the subset data will be the first and second time points.
3. Fit a random forest model and use cross-validations. The suggestion for cross-validations is randomly sampling 60~70% (two-thirds) of the data to train the model, and the other 30~40% (one-third) to perform the testing (Friedman et al., 2001; James et al., 2013).
4. Randomly perform multiple cross-validations (e.g., 100 times). The more times we implement cross-validation, the more stable results we obtain. In addition, if the sample size is very large, the results will be stable even if only one cross-validation is performed. As discussed, because fNIRS studies often do not have a large sample size, more cross-validations are suggested.

5. Move on to the next bandwidth of time points. Keep moving the bandwidth window until the last time point has been completed (i.e., 167th time point in the present data set).
6. Calculate the 90% confidence interval of AUC values for each time point. Use all AUC values produced by the same time point for the estimation.
7. Where a confidence interval of AUC values does not cover 0.5 (Friedman et al., 2001; Hanley & McNeil, 1982; James et al., 2013), there is a significant difference between experimental conditions.

The steps summarized above show the capability and flexibility of the proposed method. All the analysis can be performed in R (R Development Core Team, 2016) with R packages (Canty, 2002; Causeur et al., 2012). All statistical plots can be produced using ggplot2 (Wickham, 2009). The data set used in the present study and all source code used to perform

the analyses/generate the figures are provided on the MTPA website (<https://github.com/PsyChiLin/MTPAinR>).

Discussion

The present study proposed a novel MTPA perspective, which extracted the time course information from fNIRS to discriminate experimental conditions in order to address three major concerns with mass univariate analysis. First, mass univariate analysis struggles in multiple comparisons because of its low power. Second, autocorrelations of fNIRS time course data are not considered by mass univariate analysis. In addition, the linearity assumptions in mass univariate analysis are not always true in practical situations. The following paragraphs discuss these issues and additional implications for the proposed MTPA.

First, regarding the power of statistical analysis, the MTPA effectively detected the differences between conditions, whereas the mass univariate analysis revealed no significant effects (i.e., three methods of p -value corrections and two nonparametric permutation frameworks). For example, neuroimaging studies with the semantic judgment paradigm consistently found the left IFG activation (P.-J. Chen et al., 2013; P. J. Chen, Gau, Lee, & Chou, 2016; Chou, Chen, Fan, Chen, & Booth, 2009a; Chou, Chen, Wu, et al., 2009b; Fan, Lee, & Chou, 2010; Lee, Booth, Chen, & Chou, 2011). As shown in the demonstrations, we used exactly the same semantic judgment paradigm in the fNIRS experiment. However, mass univariate analysis could not effectively reveal the significant differences in left IFG activation between experimental conditions because of its low power. This issue would become more severe when researchers increased the sampling rates of fNIRS. Thus, mass univariate analysis was not a perfect choice when we aimed to obtain time course information for fNIRS with a higher temporal resolution (Rossi et al., 2012; Strangman, Boas, et al., 2002a; Wallois et al., 2012). Although there was a method for mass univariate analysis to obtain enough power by averaging the original number of time points to a much smaller number of time points (H.-C. Chen et al., 2011; H.-C. Chen et al., 2008; Khan, Hong, & Hong, 2014; Scholkmann, Wolf, & Wolf, 2013; B. Xu et al., 2014; Yu et al., 2016; Zimmermann et al., 2013), we argue that this method may decrease the temporal resolution due to loss of information on the neglected time points. By contrast, without losing any temporal information, MTPA was able to detect a significant difference between conditions with many time points and further led to a conclusion consistent with previous research.

Second, MTPA was able to estimate the results of a specific time point under consideration of multiple neighboring time points instead of focusing only on each individual time point. This perspective allows researchers to address the

autocorrelation issues, which is one of the most important characteristics of fNIRS (Tak & Ye, 2014). Previously, mass univariate analysis examined the difference between two conditions once per time point and treated the statistical test at each time point as an independent test. In contrast, by implementing a bandwidth, MTPA obtained the results for each time point under consideration of neighboring time points to reflect the autocorrelations of fNIRS time course data. Although specifically how to choose an appropriate bandwidth for estimations requires further investigation, the present perspective can provide a relatively unbiased estimation on the auto-correlated data.

Third, MTPA did not assume a linear relationship between brain signals and experimental conditions, allowing it to provide information that could not be obtained from mass univariate analysis. This MTPA perspective is similar to that of MVPA, which also does not assume linearity in the fMRI domain (Etzel et al., 2009; Haxby et al., 2014; Haynes, 2015; Mitchell et al., 2004; Mur et al., 2009; Norman et al., 2006; Pereira et al., 2009). However, some might argue that not assuming linearity is a limitation. That is, MTPA can detect time points showing significant differences between conditions, but cannot show directional differences (e.g., brain activation in condition A greater than or less than that in condition B). Under such circumstances, MTPA would still be helpful for identifying possible time points showing directional differences. For example, researchers could first use MTPA to examine which time points have a significant difference, and then use mass univariate analysis such as paired t tests to further determine the direction. Therefore, MTPA without a strong assumption of linearity could be broadly applied to help answer a variety of research questions.

In addition, by using cross-validation procedures (Refaeilzadeh et al., 2009), MTPA provides more possibilities to generalize across a population. In general, researchers are interested in investigating neural mechanisms in a larger population. However, fNIRS researchers seldom adopt cross-validation approaches to verify their results, and they usually only use the fNIRS data with a small sample size to make conclusions. The previous conclusions that were made without further verification might require careful reconsideration. Even though the best method for cross-validation procedures is still under debate, MTPA provided a reliable procedure with many cross-validations (e.g., 100 cross-validations) to reduce false-positive results in the case of a small sample size, and provided greater generalizability.

Moreover, some theoretical implications can also be provided by MTPA. As shown in Fig. 5, MTPA was able to detect time points showing significant differences between “Related” and “Unrelated” conditions in both the left IFG and the left MTG. The comparisons between these two areas revealed that there were significant differences between conditions at earlier time points in the left IFG than in the left

MTG, and there were more significant time points in the left IFG than in the left MTG. Together, the results of the MTPA suggest that the processing of semantic judgments is more prominent in the left IFG. Our findings further imply a top-down control mechanism in the left IFG during semantic judgments (Badre, Poldrack, Paré-Blagoev, Insler, & Wagner, 2005; Bitan et al., 2006; Fan & Chou, 2012; Fan et al., 2010). Thus, MTPA can make a comparison between different areas, explaining the underlying neural mechanisms of semantic processing.

In summary, the present study proposed a different perspective, which implemented the random forest algorithm for discriminating cognitive states, to obtain time course information from fNIRS. Similar to MVPA, which “spatially” characterizes the neural information measured by high-spatial-resolution fMRI (Norman et al., 2006), the proposed time course analysis, which we call multi-time-point analysis (MTPA), is a method that “temporally” decodes the neural activity measured by high-temporal-resolution fNIRS.

Acknowledgements We would like to thank Chu-Hsuan Kuo for helpful comments and discussion. This research was supported by grants from the Ministry of Science and Technology of Taiwan (MOST 105-2410-H-002-053) to Tai-Li Chou and (MOST 104-2410-H-194-031-MY3) to Hsin-Chin Chen.

References

- Abdelnour, A. F., & Huppert, T. (2009). Real-time imaging of human brain function by near-infrared spectroscopy using an adaptive general linear model. *NeuroImage*, 46(1), 133–143.
- Arenth, P. M., Ricker, J. H., & Schultheis, M. T. (2007). Applications of functional near-infrared spectroscopy (fNIRS) to neurorehabilitation of cognitive disabilities. *The Clinical Neuropsychologist*, 21(1), 38–57.
- Badre, D., Poldrack, R. A., Paré-Blagoev, E. J., Insler, R. Z., & Wagner, A. D. (2005). Dissociable controlled retrieval and generalized selection mechanisms in ventrolateral prefrontal cortex. *Neuron*, 47(6), 907–918.
- Bengio, Y., & Grandvalet, Y. (2004). No unbiased estimator of the variance of k-fold cross-validation. *Journal of Machine Learning Research*, 5, 1089–1105.
- Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society B*, 57, 289–300.
- Benjamini, Y., & Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *Annals of statistics*, 29, 1165–1188.
- Bitan, T., Burman, D. D., Lu, D., Cone, N. E., Gitelman, D. R., Mesulam, M.-M., & Booth, J. R. (2006). Weaker top-down modulation from the left inferior frontal gyrus in children. *NeuroImage*, 33(3), 991–998.
- Breiman, L. (2001). Random forests. *Machine learning*, 45(1), 5–32.
- Browne, M. W. (2000). Cross-validation methods. *Journal of mathematical psychology*, 44(1), 108–132.
- Canty, A. J. (2002). Resampling methods in R: the boot package. *R News*, 2(3), 2–7.
- Causeur, D., Chu, M.-C., Hsieh, S., & Sheu, C.-F. (2012). A factor-adjusted multiple testing procedure for ERP data analysis. *Behavior research methods*, 44(3), 635–643.
- Chen, H.-C., Vaid, J., Boas, D. A., & Bortfeld, H. (2011). Examining the phonological neighborhood density effect using near infrared spectroscopy. *Human Brain Mapping*, 32(9), 1363–1370.
- Chen, H.-C., Vaid, J., Bortfeld, H., & Boas, D. A. (2008). Optical imaging of phonological processing in two distinct orthographies. *Experimental brain research*, 184(3), 427–433.
- Chen, P.-J., Fan, L.-Y., Hwang, T.-J., Hwu, H.-G., Liu, C.-M., & Chou, T.-L. (2013). The deficits on a cortical-subcortical loop of meaning processing in schizophrenia. *Neuroreport*, 24(3), 147–151.
- Chen, P. J., Gau, S. S. F., Lee, S. H., & Chou, T. L. (2016). Differences in age-dependent neural correlates of semantic processing between youths with autism spectrum disorder and typically developing youths. *Autism Research*, 9(12), 1263–1273.
- Chou, T. L., Chen, C. W., Fan, L. Y., Chen, S. Y., & Booth, J. R. (2009a). Testing for a cultural influence on reading for meaning in the developing brain: The neural basis of semantic processing in Chinese children. *Frontiers in human neuroscience*, 3, 27.
- Chou, T.-L., Chen, C.-W., Wu, M.-Y., & Booth, J. R. (2009b). The role of inferior frontal gyrus and inferior parietal lobule in semantic processing of Chinese characters. *Experimental brain research*, 198(4), 465–475.
- Custo, A., Boas, D. A., Tsuzuki, D., Dan, I., Mesquita, R., Fischl, B., ... Wells, W. (2010). Anatomical atlas-guided diffuse optical tomography of brain activation. *NeuroImage*, 49(1), 561–567.
- Dunn, O. J. (1961). Multiple comparisons among means. *Journal of the American Statistical Association*, 56(293), 52–64.
- Efron, B. (1983). Estimating the error rate of a prediction rule: improvement on cross-validation. *Journal of the American Statistical Association*, 78(382), 316–331.
- Efron, B. (2007). Correlation and large-scale simultaneous significance testing. *Journal of the American Statistical Association*, 102(477), 93–103.
- Etzel, J. A., Gazzola, V., & Keysers, C. (2009). An introduction to anatomical ROI-based fMRI classification analysis. *Brain Research*, 1282, 114–125.
- Fan, L.-Y., & Chou, T.-L. (2012). Hierarchical model comparisons on effective connectivity in semantic judgments of Chinese characters. *Chinese Journal of Psychology*, 54(1), 31–46.
- Fan, L.-Y., Lee, S.-H., & Chou, T.-L. (2010). Interaction between brain regions during semantic processing in Chinese adults. *Language and linguistics*, 11(1), 159–182.
- Ferrari, M., Mottola, L., & Quaresima, V. (2004). Principles, techniques, and limitations of near infrared spectroscopy. *Canadian journal of applied physiology*, 29(4), 463–487.
- Friedman, J., Hastie, T., & Tibshirani, R. (2001). *The elements of statistical learning*. New York, NY: Springer.
- Germon, T., Evans, P., Barnett, N., Wall, P., Manara, A., & Nelson, R. (1999). Cerebral near infrared spectroscopy: emitter-detector separation must be increased. *British journal of anaesthesia*, 82(6), 831–837.
- Good, P. (2013). *Permutation tests: a practical guide to resampling methods for testing hypotheses*. New York, NY: Springer.
- Gow, D. W. (2012). The cortical organization of lexical knowledge: a dual lexicon model of spoken language processing. *Brain and Language*, 121(3), 273–288.
- Hagoort, P. (2005). On Broca, brain, and binding: a new framework. *Trends in cognitive sciences*, 9(9), 416–423.
- Hanley, J. A., & McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, 143(1), 29–36.

- Haxby, J. V., Connolly, A. C., & Guntupalli, J. S. (2014). Decoding neural representational spaces using multivariate pattern analysis. *Annual review of neuroscience*, 37(1), 435–456.
- Haynes, J.-D. (2015). A primer on pattern-based approaches to fMRI: principles, pitfalls, and perspectives. *Neuron*, 87(2), 257–270.
- Hebden, J., & Delpy, D. (1997). Diagnostic imaging with light. *British journal of radiology*, 70(1), 206–214.
- Homan, R. W., Herman, J., & Purdy, P. (1987). Cerebral location of international 10–20 system electrode placement. *Electroencephalography and clinical neurophysiology*, 66(4), 376–382.
- Hoshi, Y. (2003). Functional near-infrared optical imaging: Utility and limitations in human brain mapping. *Psychophysiology*, 40(4), 511–520.
- Hoshi, Y., Kobayashi, N., & Tamura, M. (2001). Interpretation of near-infrared spectroscopy signals: a study with a newly developed perfused rat brain model. *Journal of applied physiology*, 90(5), 1657–1662.
- Huppert, T., Boas, D. A. (2005) HomER: Hemodynamic Evoked Response NIRS data analysis GUI. Available from the Photon Migration Imaging Lab, Martinos Center for Biomedical Imaging, <http://www.nmr.mgh.harvard.edu/PMI/>.
- Irani, F., Platek, S. M., Bunce, S., Ruocco, A. C., & Chute, D. (2007). Functional near infrared spectroscopy (fNIRS): an emerging neuroimaging technology with important applications for the study of brain disorders. *The Clinical Neuropsychologist*, 21(1), 9–37.
- Isobe, K., Kusaka, T., Nagano, K., Okubo, K., Yasuda, S., Kondo, M., ... Onishi, S. (2001). Functional imaging of the brain in sedated newborn infants using near infrared topography during passive knee movement. *Neuroscience letters*, 299(3), 221–224.
- James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning*. New York, NY: Springer.
- Jefferies, E. (2013). The neural basis of semantic cognition: converging evidence from neuropsychology, neuroimaging and TMS. *Cortex*, 49(3), 611–625.
- Jobsis, F. F. (1977). Noninvasive, infrared monitoring of cerebral and myocardial oxygen sufficiency and circulatory parameters. *Science*, 198(4323), 1264–1267.
- Khan, M. J., Hong, M. J., & Hong, K.-S. (2014). Decoding of four movement directions using hybrid NIRS-EEG brain-computer interface. *Frontiers in human neuroscience*, 8, 244.
- Kleinschmidt, A., Obrig, H., Requardt, M., Merboldt, K.-D., Dimagl, U., Villringer, A., & Frahm, J. (1996). Simultaneous recording of cerebral blood oxygenation changes during human brain activation by magnetic resonance imaging and near-infrared spectroscopy. *Journal of cerebral blood flow & metabolism*, 16(5), 817–826.
- Koessler, L., Maillard, L., Benhadid, A., Vignal, J. P., Felblinger, J., Vespignani, H., & Braun, M. (2009). Automated cortical projection of EEG sensors: anatomical correlation via the international 10–10 system. *NeuroImage*, 46(1), 64–72.
- Koh, P. H., Glaser, D. E., Flandin, G., Kiebel, S., Butterworth, B., Maki, A., ... Elwell, C. E. (2007). Functional optical signal analysis: a software tool for near-infrared spectroscopy data processing incorporating statistical parametric mapping. *Journal of biomedical optics*, 12(6), 064010.
- Lee, S.-H., Booth, J. R., Chen, S.-Y., & Chou, T.-L. (2011). Developmental changes in the inferior frontal cortex for selecting semantic representations. *Developmental Cognitive Neuroscience*, 1(3), 338–350.
- Leek, J. T., & Storey, J. D. (2008). A general framework for multiple testing dependence. *Proceedings of the National Academy of Sciences*, 105(48), 18718–18723.
- Mehagnoul-Schipper, D. J., van der Kallen, B. F., Colier, W. N., van der Sluijs, M. C., van Erning, L. J. T. O., Thijssen, H. O., ... Jansen, R. W. (2002). Simultaneous measurements of cerebral oxygenation changes during brain activation by near-infrared spectroscopy and functional magnetic resonance imaging in healthy young and elderly subjects. *Human Brain Mapping*, 16(1), 14–23.
- Mihara, M., Miyai, I., Hatakenaka, M., Kubota, K., & Sakoda, S. (2008). Role of the prefrontal cortex in human balance control. *NeuroImage*, 43(2), 329–336.
- Minagawa-Kawai, Y., Van Der Lely, H., Ramus, F., Sato, Y., Mazuka, R., & Dupoux, E. (2010). Optical brain imaging reveals general auditory and language-specific processing in early infant development. *Cereb Cortex*, 21(2), 254–261.
- Mitchell, T. M., Hutchinson, R., Niculescu, R. S., Pereira, F., Wang, X., Just, M., & Newman, S. (2004). Learning to decode cognitive states from brain images. *Machine learning*, 57(1–2), 145–175.
- Mur, M., Bandettini, P. A., & Kriegeskorte, N. (2009). Revealing representational content with pattern-information fMRI—an introductory guide. *Social cognitive and affective neuroscience*, 4(1), 101–109.
- Nichols, T., & Hayasaka, S. (2003). Controlling the familywise error rate in functional neuroimaging: a comparative review. *Statistical methods in medical research*, 12(5), 419–446.
- Nichols, T. E., & Holmes, A. P. (2002). Nonparametric permutation tests for functional neuroimaging: a primer with examples. *Human brain mapping*, 15(1), 1–25.
- Norman, K. A., Polyn, S. M., Detre, G. J., & Haxby, J. V. (2006). Beyond mind-reading: multi-voxel pattern analysis of fMRI data. *Trends in cognitive sciences*, 10(9), 424–430.
- O'Sullivan, F. (1986). A statistical perspective on ill-posed inverse problems. *Statistical science*, 1(4), 502–518.
- Okamoto, M., Dan, H., Shimizu, K., Takeo, K., Amita, T., Oda, I., ... Suzuki, T. (2004). Multimodal assessment of cortical activation during apple peeling by NIRS and fMRI. *NeuroImage*, 21(4), 1275–1288.
- Penny, W. D., Friston, K. J., Ashburner, J. T., Kiebel, S. J., & Nichols, T. E. (2011). *Statistical parametric mapping: the analysis of functional brain images*. London: Academic Press.
- Pereira, F., Mitchell, T., & Botvinick, M. (2009). Machine learning classifiers and fMRI: a tutorial overview. *NeuroImage*, 45(1), 199–209.
- Plichta, M., Herrmann, M., Baehne, C., Ehlis, A.-C., Richter, M., Pauli, P., & Fallgatter, A. (2006). Event-related functional near-infrared spectroscopy (fNIRS): are the measurements reliable? *NeuroImage*, 31(1), 116–124.
- Plichta, M., Herrmann, M., Baehne, C., Ehlis, A. C., Richter, M., Pauli, P., & Fallgatter, A. (2007). Event-related functional near-infrared spectroscopy (fNIRS) based on craniocerebral correlations: Reproducibility of activation? *Human Brain Mapping*, 28(8), 733–741.
- R Development Core Team (2016). R: A language and environment for statistical computing. R Foundation for Statistical Computing. Retrieved from <http://cran.r-project.org/>. Accessed 2 July 2018
- Refaeilzadeh, P., Tang, L., & Liu, H. (2009). Cross-validation. In *Encyclopedia of database systems*, 532–538. New York, NY: Springer.
- Rossi, S., Telkemeyer, S., Wartenburger, I., & Obrig, H. (2012). Shedding light on words and sentences: near-infrared spectroscopy in language research. *Brain and Language*, 121(2), 152–163.
- Schneider W, Eschman A, Zuccolotto A (2002). *E-Prime User's Guide*. Pittsburgh: Psychology Software Tools.
- Scholkman, F., Wolf, M., & Wolf, U. (2013). The effect of inner speech on arterial CO₂ and cerebral hemodynamics and oxygenation: a functional NIRS study. In *Oxygen Transport to Tissue XXXV*, 81–87. New York, NY: Springer.
- Shimada, S., & Hiraki, K. (2006). Infant's brain responses to live and televised action. *NeuroImage*, 32(2), 930–939.
- Singh, A. K., & Dan, I. (2006). Exploring the false discovery rate in multichannel NIRS. *NeuroImage*, 33(2), 542–549.
- Strangman, G., Boas, D. A., & Sutton, J. P. (2002a). Non-invasive neuroimaging using near-infrared light. *Biological Psychiatry*, 52(7), 679–693.

- Strangman, G., Culver, J. P., Thompson, J. H., & Boas, D. A. (2002b). A quantitative comparison of simultaneous BOLD fMRI and NIRS recordings during functional brain activation. *NeuroImage*, 17(2), 719–731.
- Tak, S., & Ye, J. C. (2014). Statistical analysis of fNIRS data: a comprehensive review. *NeuroImage*, 85(1), 72–91.
- Villringer, A., & Chance, B. (1997). Non-invasive optical spectroscopy and imaging of human brain function. *Trends in neurosciences*, 20(10), 435–442.
- Villringer, A., & Dirnagl, U. (1994). Coupling of brain activity and cerebral blood flow: basis of functional neuroimaging. *Cerebrovascular and brain metabolism reviews*, 7(3), 240–276.
- Villringer, A., Planck, J., Hock, C., Schleinkofer, L., & Dirnagl, U. (1993). Near infrared spectroscopy (NIRS): a new tool to study hemodynamic changes during activation of brain function in human adults. *Neuroscience letters*, 154(1), 101–104.
- Wallois, F., Mahmoudzadeh, M., Patil, A., & Grebe, R. (2012). Usefulness of simultaneous EEG–NIRS recording in language studies. *Brain and Language*, 121(2), 110–123.
- Warner, R. M. (2012). *Applied statistics: from bivariate through multivariate techniques: from bivariate through multivariate techniques*. London: Sage.
- Wickham, H. (2009). *ggplot2: elegant graphics for data analysis*. New York, NY: Springer.
- Xu, B., Fu, Y., Shi, G., Yin, X., Wang, Z., & Li, H. (2014). Improving classification by feature discretization and optimization for fNIRS-based BCI. *Journal of Biomimetics Biomaterials and Tissue Engineering*, 19(1), 1–5.
- Xu, Q.-S., & Liang, Y.-Z. (2001). Monte Carlo cross validation. *Chemometrics and Intelligent Laboratory Systems*, 56(1), 1–11.
- Ye, J. C., Tak, S., Jang, K. E., Jung, J., & Jang, J. (2009). NIRS-SPM: statistical parametric mapping for near-infrared spectroscopy. *NeuroImage*, 44(2), 428–447.
- Yu, C.-L., Wang, M.-Y., & Hu, J.-F. (2016). Valence processing of first impressions in the dorsomedial prefrontal cortex: a near-infrared spectroscopy study. *Neuroreport*, 27(8), 574.
- Zimmermann, R., Marchal-Crespo, L., Edelmann, J., Lambercy, O., Fluet, M.-C., Riener, R., Gassert, R. (2013). Detection of motor execution using a hybrid fNIRS-biosignal BCI: a feasibility study. *Journal of neuroengineering and rehabilitation*, 10(1), 4.

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