Neurophotonics

Neurophotonics.SPIEDigitalLibrary.org

Investigation of the sensitivity-specificity of canonical- and deconvolution-based linear models in evoked functional near-infrared spectroscopy

Hendrik Santosa
Frank Fishburn
Xuetong Zhai
Theodore J. Huppert

SPIE.

Hendrik Santosa, Frank Fishburn, Xuetong Zhai, Theodore J. Huppert, "Investigation of the sensitivity-specificity of canonical- and deconvolution-based linear models in evoked functional near-infrared spectroscopy," *Neurophoton.* **6**(2), 025009 (2019), doi: 10.1117/1.NPh.6.2.025009.

Neurophotonics 6(2), 025009 (Apr–Jun 2019)

Investigation of the sensitivity-specificity of canonical- and deconvolution-based linear models in evoked functional near-infrared spectroscopy

Hendrik Santosa,[a] Frank Fishburn,[b] Xuetong Zhai,[c] and Theodore J. Huppert[d,*]
[a]University of Pittsburgh, Department of Radiology, Pittsburgh, Pennsylvania, United States
[b]University of Pittsburgh, Department of Psychiatry, Pittsburgh, Pennsylvania, United States
[c]University of Pittsburgh, Department of Bioengineering, Pittsburgh, Pennsylvania, United States
[d]University of Pittsburgh, Departments of Radiology and Bioengineering, Clinical Science Translational Institute, and Center for the Neural Basis of Cognition, Pittsburgh, Pennsylvania, United States

Abstract. Functional near-infrared spectroscopy (fNIRS) is a noninvasive brain imaging technique to measure evoked changes in cerebral blood oxygenation. In many evoked-task studies, the analysis of fNIRS experiments is based on a temporal linear regression model, which includes block-averaging, deconvolution, and canonical analysis models. The statistical parameters of this model are then spatially mapped across fNIRS measurement channels to infer brain activity. The trade-offs in sensitivity and specificity of using variations of canonical or deconvolution/block-average models are unclear. We quantitatively investigate how the choice of basis set for the regression linear model affects the sensitivity and specificity of fNIRS analysis in the presence of variability or systematic bias in underlying evoked response. For statistical parametric mapping of amplitude-based hypotheses, we conclude that these models are fairly insensitive to the parameters of the regression basis for task durations >10 s and we report the highest sensitivity-specificity results using a low degree-of-freedom canonical model under these conditions. For shorter duration task (<10 s), the signal-to-noise ratio of the data is also important in this decision and we find that deconvolution or block-averaging models outperform the canonical models at high signal-to-noise ratio but not at lower levels. [DOI: 10.1117/1.NPh.6.2.025009]

Keywords: functional near-infrared spectroscopy; linear models; sensitivity-specificity; signal-to-noise ratio.

Paper 18053R received Oct. 8, 2018; accepted for publication Apr. 4, 2019; published online May 30, 2019.

1 Introduction

Functional near-infrared spectroscopy (fNIRS) is a noninvasive brain imaging technique that measures evoked hemodynamic changes in the brain using low levels of red to near-infrared light. For these measurements, an arrangement of light sources and detectors is placed on the scalp. Although, depending on an individual's anatomy, the light traveling through the tissue from a source to a detector position (usually 25 to 35 mm away) can reach the outer layers of the cortex of the brain, which provides sensitivity to measure superficial changes in brain activation in between a light source and detector pair. In most studies, two or more wavelengths of light are recorded, which provide information to spectrally distinguish both oxy-hemoglobin (HbO_2) and deoxy-hemoglobin (Hb) changes via modified Beer–Lambert law.[1] Using a grid of optical light sources and detector positions, fNIRS can record the spatial distribution of changes in hemoglobin during functional tasks, providing a measurement of underlying brain activity.

When compared to other popular brain imaging, such as functional magnetic resonance imaging (fMRI), fNIRS offers more portability and the ability to record the brain during more ecologically valid conditions. For example, fNIRS has been employed in functional imaging of infants (see Refs. 2 and 3, for a review), gait (e.g., Refs. 4 and 5), two-person brain connectivity [(hyperscanning) e.g., Refs. 6 and 7], bedside imaging (see Refs. 3 and 8, for a review), etc. However, fNIRS has a lower spatial resolution compared to fMRI (around 2 to 3 cm). This technique is also prone to systemic physiological noise and motion artifacts (see Ref. 9, for review). Isolating the evoked signals from the unwanted noises is a challenge and an area of open investigation (see Ref. 10, for review). Nonetheless, the use of fNIRS has been steadily growing over the past two decades, particularly in niche applications and populations where more conventional modalities such as fMRI are more difficult or constraining.

Similar to fMRI, fNIRS infers changes in brain activity based on the changes in the hemodynamic signal (blood oxygenation level-dependent signal), which relates to the underlying electrophysiological processes via complex neurovascular coupling (i.e., relationship between the local neural activity and the oxygen levels of hemoglobin in the blood of the cerebral cortex). Previous work has shown a strong correlation between fNIRS and fMRI signals temporally[11,12] and spatially.[13,14] Because of this correlation, the time-series analysis of fNIRS data is often similar to that of fMRI and is generally based on mapping the statistical parameters from a regression-based model of the time course of measurements. In this paper, we extend the previous published works of our group,[15,16] detailing these models and the statistical generalizations needed for fNIRS data. For example, compared to fMRI, fNIRS instruments typically sample at a higher rate (5 to 20 Hz); however, fNIRS still measures

*Address all correspondence to Theodore J. Huppert, E-mail: huppertt@upmc.edu

the relatively slow hemodynamic response and this requires modification to the statistical model to account for this oversampling, as detailed in Huppert.[9]

Because it is more portable and less restrictive, fNIRS is widely used with infants, children, and other special populations. This also allows fNIRS to be used in a wider variety of different experimental conditions and environments. Over the past few years, the use of fNIRS in these populations and tasks has called into question the appropriateness of using regression-based canonical models based on the "normal" hemodynamic response function (HRF) observed in fMRI work. In particular, the canonical models make assumptions about the shape and timing of the brain's response, which may or may not be systematically different in special populations. As fNIRS sample rates are faster, it could be more sensitive to these errors or valuable information could potentially be lost in making these assumptions. The objective of this study is to investigate the sensitivity and specificity of these regression models applied to fNIRS data. Here, we use receiver operating characteristic (ROC) analysis to quantify the effects of variability and/or bias in the shape of the underlying hemodynamic response on the regression methods and to determine under what circumstances different models may be preferred. In this work, we have used numerical simulation methods to look at the effects of several HRFs with varying different parameters, including various signal-to-noise ratio (SNR) levels and various durations of the task period.

2 Theory

2.1 *Functional Near-Infrared Spectroscopy Linear Statistical Models*

In fNIRS studies, measurements are made between a series of discrete light emitter and detector positions on the head. The intensity of the light transmitted from an emitter to a detector position is sensitive to changes in the optical absorption of the underlying tissue along the light's path through the tissue. This results in a time course of optical transmission measurements. Concurrent and colocalized measurements at two or more optical wavelengths are then used for estimating the changes in the concentration of HbO_2 and Hb using the modified Beer–Lambert law[1] and given by

$$\Delta \mathrm{OD}_i(\lambda, k) = [\varepsilon^{\mathrm{HbO_2}}(\lambda)\Delta c_i^{\mathrm{HbO_2}}(k) + \varepsilon^{\mathrm{Hb}}(\lambda)\Delta c_i^{\mathrm{Hb}}(k)] l_i \mathrm{PPF}(\lambda), \quad (1)$$

where i is the channel index, λ is the wavelength of the laser source, $\Delta \mathrm{OD}(\lambda, k)$ is the optical density variation at time k, and $\varepsilon^{\mathrm{HbO_2}}$ and $\varepsilon^{\mathrm{Hb}}$ are the absorption coefficients of HbO_2 and Hb. Here, $\Delta c^{\mathrm{HbO_2}}$ and Δc^{Hb} are the changes of HbO_2 and Hb concentration levels, respectively; l is the distance between the source and the detector; and PPF is the partial pathlength factor, which corrects for the increased effective distance traveled by the light due to scattering and the partial volume factor accounting for the fraction of this path that was actually in the volume of interest (brain).[1]

In many fNIRS studies, these changes in hemoglobin are recorded over time during some variation of a repeated cognitive task(s). The fNIRS signals between each source-to-detector pair are analyzed using a general linear regression model to test for statistical differences between the baseline and the task conditions for each scan. This approach is similar to fMRI, although several differences in the structure of noise in fNIRS compared to other modalities should be noted (see Ref. 9, for review). First-level statistical models for examining evoked signal changes are given by a regression model described by the equation

$$Y = X * \beta + \epsilon, \quad (2)$$

where X is the design matrix encoding the timing of stimulus events, β is the coefficient (weight) of that stimulus condition for that source–detector channel, and Y is the vector of measurements. The design matrix (X) can come from either a canonical model of the expected response or a deconvolution model (described in next section). We note that the traditional block-averaging is also described by this Eq. (2) and can be considered a subset of the deconvolution model,[17] in the case of non-overlapping events. In Eq. (2), ϵ is the measurement noise/error term. The validity of this model depends on the properties of ϵ and the matching statistical assumptions (in most cases, ϵ is assumed to be an uncorrelated, normally distributed, zero mean random variable). Details on these assumptions and their effect on the regression results when the noise is nonideal can be found in Ref. 9.

In general, statistical testing of the regression coefficients (β) to either compare these to zero (baseline) or between two task conditions is used for inferring the location and level of brain activity using a Student's t-test. While there are a few approaches to solving Eq. (2), which have been used in fNIRS analysis including generalized linear models (e.g., Refs. 9, 16, 18, and ordinary least-squares regression[19]), in this paper, we will use the NIRS-specific generalized linear model formulation proposed by Barker et al.[15]

2.2 *Prewhitened and Robust Linear Model*

In this work, we used a NIRS-adopted version of the general linear model, as previously described in Barker et al.[15] This approach uses an autoregressive prewhitening method and an iteratively reweighted least squares (AR-IRLS) to control type-I errors in the fNIRS statistical model. This approach has been previously shown to have excellent sensitivity-specificity properties in comparison to ordinary least squares and other general linear regression models for fNIRS data containing physiological noise (serial-correlated noise structures) and motion-related artifacts (heavy-tailed outliers).[15] In brief, this regression model uses an n'th order autoregressive filter (W_{AR}) determined by an Akaike model-order (AIC) selection to whiten both sides of this expression, e.g.,

$$W_{AR} * Y = W_{AR} * X * \beta + W_{AR} * \varepsilon. \quad (3)$$

As described in Barker et al.,[15] the regression model is first solved using robust regression and the residual noise is then fit to an AR model. This filter (W_{AR}) is applied to both sides of the original model and then resolved and repeated until convergence. This AR filter removes serially correlated errors in the data that result from physiological noise and/or motion artifacts. AR whitening, however, does not address the heavy-tailed noise from motion artifacts. To do this, the AR-whitened model is solved using robust weighted regression, which is a procedure to iteratively downweight outliers, such as motion artifacts

$$S \cdot W_{AR} * Y = S \cdot W_{AR} * X * \beta + S \cdot W_{AR} * \varepsilon, \quad (4a)$$

where S is

$$S\left(\frac{r_W}{\sigma}\right) = \begin{cases} 1 - \left(\frac{r_W}{\sigma \cdot \kappa}\right)^2 & \left|\frac{r_W}{\sigma}\right| < \kappa \\ 0 & \left|\frac{r_W}{\sigma}\right| \geq \kappa \end{cases}, \tag{4b}$$

which is the square root of Tukey's bisquare function[20] and is the same model as used in Eq. (4) from Barker et al.,[15] where r_w is the regression residual ($r_w = S \cdot W_{AR} * \varepsilon$). The tuning constant κ is typically set to 4.685, which from theory provides 95% efficiency of the model in the presence of normally distributed errors and σ is the standard deviation of the residual noise in the model.

Using this model, the regression coefficients (β) and their error covariance is estimated, which is used in defining the statistical tests between task conditions and baseline. The regression model is solved sequentially for each data file for each subject. All source–detector pairs within a file are solved concurrently yielding a full covariance model of the noise, which is used in group-level and region-of-interest analyses. The estimate of β and its covariance matrix is given by the expressions

$$\beta = (X^T \cdot W_{AR}^T \cdot S^T \cdot S \cdot W_{AR} \cdot X)^{-1} \cdot X^T \cdot W_{AR}^T \cdot S^T \cdot S \cdot W_{AR} \cdot Y, \tag{5a}$$

$$\mathrm{Cov}_\beta = [(W_{AR} \cdot X)^T * W_{AR} \cdot X]^{-1} \cdot \sigma^2, \tag{5b}$$

$$\sigma^2 = (W_{AR} \cdot Y - W_{AR} \cdot X * \beta)^T \cdot (W_{AR} \cdot Y - W_{AR} \cdot X \cdot \beta). \tag{5c}$$

2.3 *Linear Regression Design Model (X)*

In the fNIRS field, there is a controversy on what type of design model should be used for linear regression analysis. Namely, it is not clear what effect the assumption of a fixed temporal shape of the response or the use of specific time window for computing the contrast from a block average or deconvolution model has on the statistical power for brain signals from special populations (e.g., infants) that may not have the "typical" adult-like response. To examine this question, we note that in Eq. (2), the design model (X) describes the timing of the experimental task and conditions in the experiment and sets up the hypothesis that can be tested from the regression coefficients. This design can take the form of either an unrestricted (full deconvolution) model or an impulse response model. In the impulse response model, the response is assumed to be linearly additive such that (for example) the brain response to a 30-s task is the same as 30 repeated 1-s duration tasks building up on top of each other. In this model, only the hemodynamic impulse response function is estimated, which generally is thought to peak around 6 to 8 s and recover to baseline around 12 to 15 s. This impulse response is assumed to be the same for tasks of any duration. Thus, an advantage of this model is that it can be used even if the task blocks have differing durations, as is often the case in self-paced paradigms. In comparison, for the full deconvolution model, the entire time course of the block is estimated. For example, in a 30-s duration task, the response over a window of about 45 s (task plus 15-s recovery to baseline) would be used. Because the full deconvolution model does not assume the response to be linearly additive over the task block, it can be used for examining the saturation or habituation effects of a prolonged task duration. However, in this 30-s duration example, the full deconvolution model would require three times more regression terms compared to the impulse response version (modeling 45 s of data instead of just the 15-s impulse window). Note that this discussion of additive linear assumptions applies to blocked design or slow event-related designed experiments. In rapid event-related designs, both the full deconvolution and impulse response models assume linearity. Examples of somewhat simplified full and impulse response models are given in Eqs. (6a) and (6b).

Full deconvolution model

$$\begin{bmatrix} Y_k \\ Y_{k+1} \\ Y_{k+2} \\ Y_{k+3} \\ Y_{k+4} \\ Y_{k+5} \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} * \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix} + \epsilon. \tag{6a}$$

Impulse response model

$$\begin{bmatrix} Y_k \\ Y_{k+1} \\ Y_{k+2} \\ Y_{k+3} \\ Y_{k+4} \\ Y_{k+5} \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 1 & 1 \\ 1 & 1 \\ 0 & 1 \\ 1 & 0 \\ 1 & 1 \end{bmatrix} * \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix} + \epsilon. \tag{6b}$$

The full model design matrix is constructed where each column is a vector of 1s at the task onset of each block. Each column shifts the onset by one entry [Eq. (6b)]. If the vectors of coefficients (β) were plotted down the columns, this would look like the time course of either the full or the impulse response. Note that the model described in Eq. (6a) reduces to the block-averaging equation in the limit that the events are nonoverlapping (specifically, there are no off-diagonal terms in the matrix $X^T \cdot X$ in the Gauss–Markov equation).[21] Under this condition, the regression model reduces to a weighted backprojection ($\beta = 1/n \cdot X^T \cdot Y$), which is the block-averaging operation in terms of the estimation of the beta weights. In block-averaging, the variance of the coefficients is estimated separately for each point, whereas in regression its assumed to be normally distributed with parameter sigma^2 [Eq. (5b)]. However, this distinction is moot when computing the contrast over a time window, as the parametric t-test, which is often used in these block-averaging results, reintroduces the normally distributed assumption. Thus, as commonly implemented by many literatures, e.g., in the HOMER NIRS software, the Student's t-test contrast from block-averaging and deconvolution are equivalent (see Ref. 19, for more details).

For either the full deconvolution model or the impulse response model, a basis set (denoted as matrix Γ) can be used to reparameterize the model such that

$$Y = X * \Gamma * \beta' + \epsilon. \tag{7}$$

This models the time course (column of the coefficient vector) as a linear combination of the basis. This basis set matrix can be either a complete (lossless) or a restricted (e.g., smoothing or canonical) basis. In this paper, we have explored several of the commonly used basis sets (e.g., canonical HRF; see Secs. 2.3.1–2.3.5, for details).

For either the original Eq. (2) or the basis-set reparametrized Eq. (7) models, then, the estimated model coefficient vector

$\beta_i(k)$ is used to calculate the t-value for a two-tailed t-test to test the null hypothesis $c^T\beta_i(k) = 0$.[22] In this study, the t-statistics of channel i at time step k are obtained using

$$t_i(k) = \frac{c^T\beta_i(k)}{\sqrt{\hat{\sigma}_i^2(k)c^T[X^T(k)X(k)]^{-1}c}}, \tag{8}$$

where c is a vector contrast for selecting the coefficient interest and σ^2 is the mean squared error of the residual. The contrast vector (c) encodes the hypothesis to be tested. For example, in the deconvolution model, where there is a coefficient (β) for each estimated time point, the contrast vector $c = [0\,0\,1\,1\ldots\,0\,0]^T$ would compute the summed contrast over a window of time.

2.3.1 *Canonical hemodynamic response function*

The canonical HRF (also commonly referred to as the "double gamma function" in fMRI literature) models a hemodynamic response with an undershoot period. This is one possible basis set that can be used in Eq. (7). This is defined by the following equation:

$$\text{HRF} = \frac{b_1^{a_1} * t^{(a_1-1)}}{\Gamma(a_1)} * e^{(-b_1*t)} - c * \frac{b_2^{a_2} * t^{(a_2-1)}}{\Gamma(a_2)} * e^{(-b_2*t)}, \tag{9}$$

where b_1 (default 1 s^{-1}) and b_2 (default 1 s^{-1}) are the dispersion time constants for the peak and undershoot periods, and a_1 (default 4 s) and a_2 (default 16 s) are the peak time and undershoot time. Here, c (default 1/6) is the ratio of the height of the main peak to the undershoot, and $\Gamma(.)$ is the scalar value of the gamma function and is a normalizing factor.

An extension of the canonical model is to add the derivatives in the dispersion (b_1) and onset (a_1) parameters. In this context, the expanded basis set can be viewed as a first-order approximation on a Taylor-series expansion allowing small variations in the value of these parameters. In these models, a total of three (or six in the case of second derivatives) regression vectors are used for each task condition. Using Eq. (8), the statistical test of the resulting model is then based on the Student's t-test for nonzero values of these coefficients. In this work, we have explored the case in which all three (main term and first derivatives) coefficients are used to define the statistical test (e.g., $c = [1\,1\,1]$), which we denote as the "fixed effects" (FE) derivative model and the case in which two derivative terms are ignored after solving and only the main term is considered in the contrast (e.g., $c = [1\,0\,0]$), which we denote as the "random effects" (RE) derivative model to indicate that these additional derivative terms are used as nuisance regressors.

2.3.2 *Gamma function*

The gamma HRF basis set models a hemodynamic response without an undershoot period. This is defined by the equation:

$$\text{HRF} = \frac{b_1^{a_1} * t^{(a_1-1)}}{\Gamma(a_1)} * e^{(-b_1*t)}, \tag{10}$$

where b_1 (default 1 s^{-1}) is the dispersion time constants and a_1 (default 6 s) is the peak time.

2.3.3 *Boxcar function*

The boxcar model uses a constant amplitude block for the duration of the task event or stimuli pattern. Boxcar function is created by a vector of zeros except during the task duration where it is equal to a constant value (e.g., 1). A limitation of the boxcar model is that it does not model transitions from baseline to task and is equivalent to comparing the average of the response magnitude during the entire task period to the entire non-task period. The transition periods where the brain response is changing from baseline to task magnitudes are either lumped into the "task" (thereby diluting the estimate of the average brain activity) or lumped into the "baseline" estimate (increasing the estimate of the variance of the baseline). This will also create discontinuities in the time course of the residual error of the model, which can affect the performance of the autoregressive prewhitening and the control of type-I errors. We note that this regression model is the least realistic of the canonical basis sets examined in this paper because it assumes the brain activity to change instantaneously. In this work, we used a boxcar shifted by 4 s to better match the expected peak of the hemodynamic response.

2.3.4 *Finite impulse response*

The canonical basis models (canonical HRF, gamma, and boxcar) provide a limited support to the estimate of the hemodynamic shape and impose smoothness and timing information in the estimate. In contrast, the finite impulse response (FIR) model or deconvolution model uses an identity operator for the basis set (or equivalently uses no basis set) and models the entire temporal shape of the hemodynamic response providing complete support for any timing or shape of response. In the full deconvolution model (denoted as FIR in this work), the model is estimated over the full window of the task duration as demonstrated in Eq. (6a). We again note that the full deconvolution model reduces to a block-average model for non-overlapping events. In the impulse response version (denoted as FIR-IRF in this work), the FIR basis models the impulse response window (16-s duration in this work) and is convolved with duration of the stimulus task to yield the model of the data [e.g., Eq. (6b)]. In both cases, Eq. (8) is used to compute the Student's t-test over a time window of the estimated response. This window must be chosen *a priori*. In this work, we used the window of 4 s after the onset of the task to 8 s after the cessation of the task.

2.3.5 *Nonlinear impulse response estimation procedure*

Finally, in this work, we also introduced a nonlinear impulse response estimation procedure. In this model, a canonical model and first derivatives in dispersion and onset was fit to the data using our standard AR-IRLS regression procedure. The residual error (unweighted/unwhitened) was computed from the model fit. The parameters of the canonical model were then updated by adjusting the shape based on the coefficients of the two derivative terms and by noting that these coefficients were related to a first-order Taylor series in dispersion and onset. A regression model using the new updated canonical model and updated derivatives was then fit to the residual of the original model. This was repeated until convergence as a steepest decent minimization.

3 Methods

In this work, simulations were performed to compare the performance of the various models. To model realistic ranges of hemodynamic impulse responses, we used the experimentally recorded data from a finger-tapping experiment given in Huppert et al.[23] These data consisted of responses from 11 healthy right-handed subjects during a 2-s finger-tapping task. These data were selected because they demonstrated a range of onset, time-to-peak of the response, and the presence/absence of post-stimulus undershoots, which had been shown to correlate with the intersubject variability observed in concurrent fMRI. Using these normalized response from HbO_2 and Hb (Fig. 5 in Ref. 23), we performed a principal component decomposition, which was then used to create synthetic data as a linear combination of the eigenvectors. A total of 2200 new HRF shapes had been generated from these original data and are shown in Fig. 1. This set of hemodynamic responses had a considerably larger overall support then typically observed in fNIRS studies with a time to peak that varied between about 3 and 23 s (the typical response is usually thought to be between 6 and 8) and a recovery that varied from 6 to 35 s (typical 12 to 15 s). These simulated responses had between 20% and 100% (perfect match) overlap with the original canonical response model. Although not exhaustive, we feel that this set of simulations covers a large variety of expected hemodynamic responses extending beyond just finger-tapping.

For each simulation, one of the 2200 different impulse response shapes was taken as the "biological truth" and convolved with a randomly generated stimulus-timing paradigm of a specific duration task between 1 and 30 s. The duty cycle of the task was kept constant at 20% for all simulations such that the shorter duration blocks had more trials than the longer duration blocks. The events were non-overlapping such that block-averaging and deconvolution were equivalent in these data. The convolution of the selected impulse response and the stimulus block timing was added to the experimental resting-state (no task) fNIRS data taken from Perlman et al.[24]

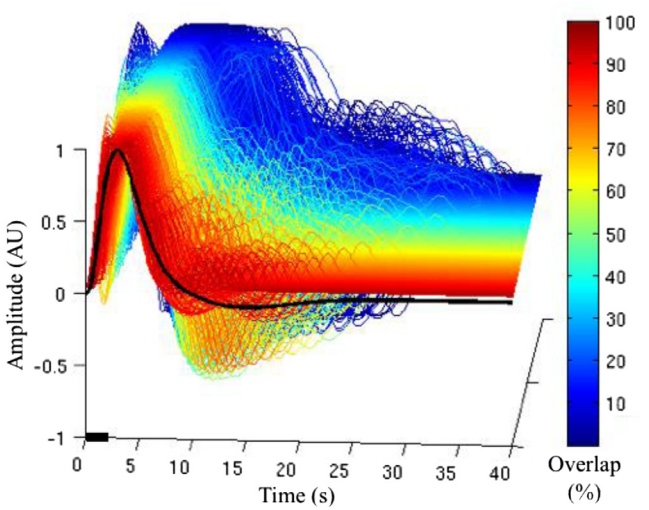

Fig. 1 HbO_2 (thick solid line) of 2-s finger-tapping task with the range of interstimulus interval between 4 and 20 s (average = 12 s), which is averaged from 2200 responses (see Ref. 23, for experiment details). The thin dotted line and thin dashed line represent 25 to 75 and 5 to 95 percentiles, respectively. The thick line shows the average (normalized) evoked response for 2-s task period.

These data were measured in 12 channels in the forehead from 8 sources and 4 detectors in children. A specific contrast-to-noise ratio (CNR) between 5:1 [high contrast] and 1:100 [extremely low contrast] was generated. In our experience, CNR of 1:2 was reasonable for a cognitive task in children data (e.g., Ref. 24).

The ROC curve analysis was run by generating and analyzing the simulated time courses where activation had been added to half of the fNIRS source–detector pairs. This was repeated 900 times using a randomly selected resting-state data file and randomly generated stimulus timing to generate a ROC curve for each of the eight different regression models (described in next section). A separate ROC curve was generated for each of the seven different SNR levels (SNR = 0.01, 0.1, 0.3, 0.5, 1.0, 2.0, 5.0), and eight different task durations (1-, 2-, 5-, 10-, 15-, 20-, 25-, 30-s durations), and all 2200 impulse response models for a total of 985,600 ROC curves (7×10^9 total simulations). In this study, we compared eight different regression models: (i) canonical, (ii) canonical with derivatives of random effects [canonical+RE(deriv)] (iii) canonical with derivatives of fixed effects [canonical+FE(deriv)], (iv) gamma function, (v) boxcar function, (vi) finite impulse response with impulse response version (FIR-IRF), (vii) full deconvolution FIR model, and (viii) nonlinear. All analyses, including ROC curve generation by using various basis sets, were done using our NIRS Brain AnalyzIR toolbox.[16]

ROC curves were generated from the various schemes (i.e., regression models, durations, and SNRs) by varying the estimated p-value threshold for activation from 0 to 1, and then calculating the true-positive rate (TPR) or sensitivity and false-positive rate (FPR) or (1-specificity). Ideally, the ROC curve will climb quickly toward to the top-left, meaning the model correctly predicted the class. The ideal condition or perfect test will have a highest area under the curve (AUC) of 1. Additionally, the AUC value of 0.5 represented the random chance or worthless test. Furthermore, we also estimated the control of the type-I error by showing the relationship between the actual FPR and the expected theoretical error reported by MATLAB (denoted as p-hat). The ideal condition ("truth") showed similar values between p-hat and FPR, where the slope of that condition was equal to 1. A large positive slope meant that the model has a high FPR, whereas, a small positive slope meant that the model has high false-negative rate.

4 Results

4.1 *Sensitivity of Regression to Systematic Bias in the Hemodynamic Response Function Shape*

Using the range of derived hemodynamic shapes (shown in Fig. 1), we have examined the sensitivity and specificity of the various regression bases to systematic bias in the shape of the underlying hemodynamic response compared to the basis used in recovery. For consistency across simulations, the bias is defined relative to the canonical HRF model as the R-squared fit between the simulated and the canonical models, which we denote as the "percent-overlap". In Figs. 2(a)–2(d), we show the AUC from ROC analyses for the (i) canonical (blue solid line), (ii) canonical+RE(deriv) (red dashed line), (iii) FIR-IRF (yellow dashed line), (iv) FIR (dark-green and dotted line), and (v) nonlinear (light-blue and dash-dot line) deconvolution models as a function of the match between the generative response shape and the idea canonical HRF

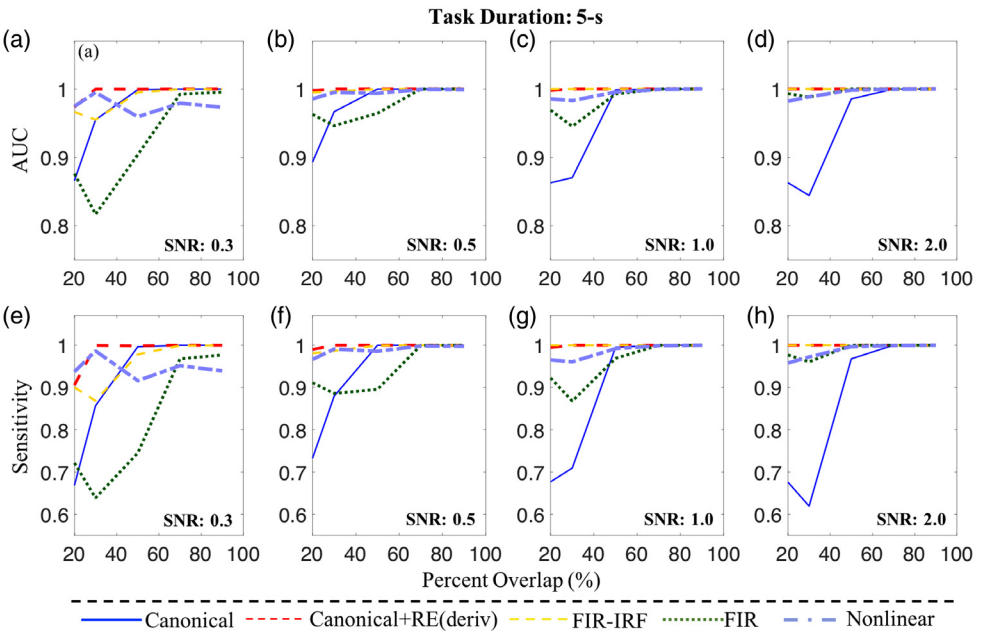

Fig. 2 Comparison of AUC and sensitivity at $p = 0.05$ from ROC analysis at various mismatch levels (relative bias with the canonical HRF model) between 20% and 100% overlap for 5-s task duration. An overlap of 100% means a perfect match to the canonical HRF. Overlap is mathematically defined here as the R^2 fit. In each panel, it shows the mismatch or (overlap) for both AUC (first row) and sensitivity (second row) using five selected regression models: canonical, canonical+RE(deriv), FIR-IRF, FIR, and nonlinear. First to fourth columns show the performance at SNR levels of 0.3, 0.5, 1.0, and 2.0, respectively.

model. In Figs. 2(e)–2(f), we show the sensitivity of the ROC models (at $p = 0.05$). The results are presented at SNR = 0.3, 0.5, 1.0, and 2.0 levels as shown in panels (a) and (e), (b) and (f), (c) and (g), (d) and (h), respectively. In all panels, the x-axis is the overlap between these models (100% overlap equals an exact match to the canonical HRF). A task duration of 5 s is used in these regression models. The rationale for showing this cross section at 5-s duration of the total results will be more apparent from the further results sections.

As shown in Fig. 2, as expected, the canonical HRF model's sensitivity falls off as mismatch between the generative and recovery responses is increased, especially for simulations with <50% overlap with the canonical model [see panel (e) and (f) (SNR: 0.3, 0.5, 1.0, and 2.0), for detail]. Below about a 50% overlap, the canonical model fails, but adding derivatives to this model (as random effects) recovers some of this loss in sensitivity and an increase of the AUC. The nonlinear model does not ever perform better than the canonical model with derivatives and, thus, is not recommended. At high SNR levels (1:0 and 2:0), the FIR (full deconvolution) and FIR-IRF (impulse response model) are able to handle large mismatches. However, this is not substantially better than the canonical model with derivatives at this 5-s task duration. As the SNR is lowered, the FIR and FIR-IRF models fail faster than the canonical models. This is due to the higher degrees of freedom in the FIR models. Thus, even though there is a mismatch between the simulated and the recovered HRF shapes, the performance of canonical models is better than the FIR models at lower SNR levels. The FIR-IRF response model is more stable than the FIR (full deconvolution/block-averaging) model at moderate SNR levels.

We note that a ROC analysis quantifies the sensitivity and specificity of the model to test null hypothesis. In the regression model of fNIRS, the null hypothesis is that the magnitude of the evoked signal change during the task is not statistically different from the baseline period. As demonstrated by these results, the rejection of this null hypothesis does not require a "perfect" alternative hypothesis. That is, it is not required to get the time window of the contrast or the shape of the response exactly right to still have the power to reject the null hypothesis.

4.2 *Sensitivity of Linear Models to Hemodynamic Response Function Mismatch per Task Duration*

Although adding derivatives (as random effects) to the canonical HRF model recovers some of the loss of sensitivity due to systemic mismatch between the underlying and the recovered hemodynamic shapes at the short task duration (5 s), as shown in Fig. 2, this effect is negated as the duration of the task increases. Figure 3 shows the comparison of the area under the ROC curve (AUC) for all (448) possible configurations from eight different regression models, seven different SNR levels, and eight different task durations. It is noted the ROC analysis is repeated 900 times for every configuration. This figure has eight panels whereas each of these panels represents eight different task durations, that is, 1-, 2-, 5-, 10-, 15-, 20-, 25-, and 30-s durations, as shown in Figs. 3(a)–3(h), respectively. For every panel, it describes the AUC values (y-axis) for seven different SNR levels of 0.01, 0.1, 0.3, 0.5, 1.0, 2.0, and 5.0 (x-axis). As mentioned earlier, the range of AUC is 0.5 (worst) to 1 (best). Eight different regression models have been compared in every panel: (i) canonical (blue solid line), (ii) canonical+RE(deriv) (red dashed line), (iii) canonical + FE(deriv) (green dotted line), (iv) gamma function (black and dash-dot line), (v) boxcar function (magenta solid line), (vi) FIR-IRF (yellow dashed line),

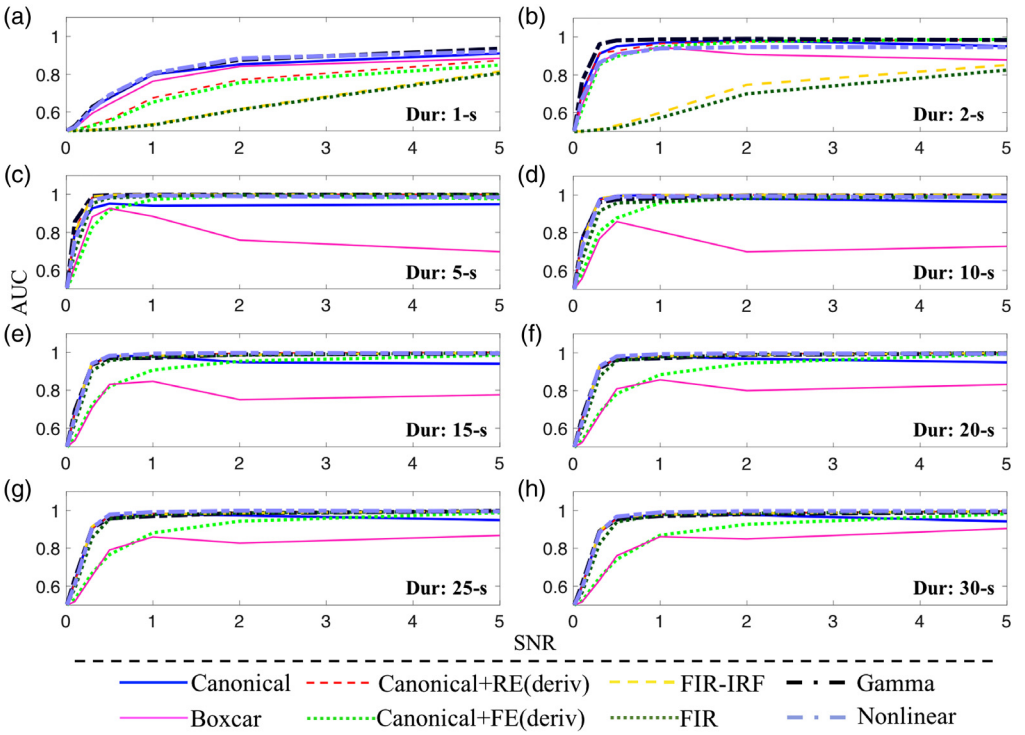

Fig. 3 Comparison of AUC from ROC analysis for all (eight) different regression models at various SNR levels (*x*-axis) from 900 repeated simulations. Eight different activation or task periods of 1, 2, 5, 10, 15, 20, 25, and 30 s are investigated in panels (a)–(h), respectively. In every panel, it shows the comparison of AUC values at seven SNR levels of 0.01, 0.1, 0.3, 0.5, 1.0, 2.0, and 5.0 for all (eight) different HRF models (i.e., canonical, canonical+RE(deriv), canonical+FE(deriv), gamma, boxcar, FIR-IRF, IRF, and nonlinear).

(vii) FIR (dark-green and dotted line), and (viii) nonlinear (light-blue and dash-dot line).

As the task duration increases, the exact timing of the transients of the response at the onset and recover are washed out by the steady-state behavior of the response during the duration. We found that above ~10-s task durations, the models all had about the same performance even when there was a mismatch between the true underlying HRF shape and the regression basis. For task durations of 2 to 5 s, the performance of the models differed mainly at low SNR levels (<SNR 1.0). Similar to the findings in Fig. 2, the FIR and FIR-IRF models lost their sensitivity quickest as SNR decreased, which was most pronounced at the shortest duration tasks (1 to 2 s). We noted that the boxcar model had a bit of odd behavior with a peak in its sensitivity at about SNR = 0.5–2 but then a loss in sensitivity at high SNRs. This was actually because this boxcar basis set had sharp edges to the response (e.g., a binary regressor). This resulted in discontinuities in the residual of the model after fitting and much of the response signal at the edges of the response are considered noise, particularly at higher SNR levels. As described previously, the boxcar model, which can only model an instantaneous change in the brain response between the baseline and the task period, was an unrealistic model. These discontinuities in the model residual at the transition points were more pronounced at higher SNRs.

4.3 *Receiver-Operatingor Characteristic Analysis for Short Task Duration*

Figure 4 shows the comparison of the ROC analysis for 5-s task duration using six different SNR levels, that is, 0.1, 0.3, 0.5, 1.0, 2.0, and 5.0, as shown in Figs. 4(a)–4(f), respectively. Similar to the previous figure, eight different regression models have been compared in every panel. In detail [SNR: 0.1 in Fig. 4(a)], the ROC performance shows the AUC in descending order: gamma (AUC: 0.86), canonical+RE(deriv) (AUC: 0.82), FIR-IRF (AUC: 0.81), canonical (AUC: 0.80), nonlinear (AUC: 0.76), FIR (AUC: 0.70), boxcar (AUC: 0.63), and canonical+FE (deriv) (AUC: 0.60). As SNR levels increase (>0.5), the choice of basis used in the regression models has little effect on the ROC analysis [except boxcar function for reasons previously noted; see Figs. 4(c)–4(f)].

5 Discussion

Activation analysis in fNIRS adopts the same approach from the fMRI field, that is, to attempt to accurately model the hemodynamic response elicited by the task design. Most fNIRS or fMRI task-based studies have been primarily focused on estimating the amplitude of evoked responses across different task conditions or in reference to a baseline. The accuracy and robustness of the hemodynamic response model have an important role in determining the sensitivity and specificity of the resulting estimates of activation. In addition, the shape of the HRF can be characterized by several parameters: amplitude or height, rise time, time to peak followed by an undershoot, and full width at half maximum, for both fMRI[25–28] and fNIRS studies.[29] Many different HRF estimation methods have been proposed with various different parameters using several functions, such as canonical response, FIR, and nonlinear. However, a rigorous comparison of the performance of these various approaches has not been done prior to the present study.

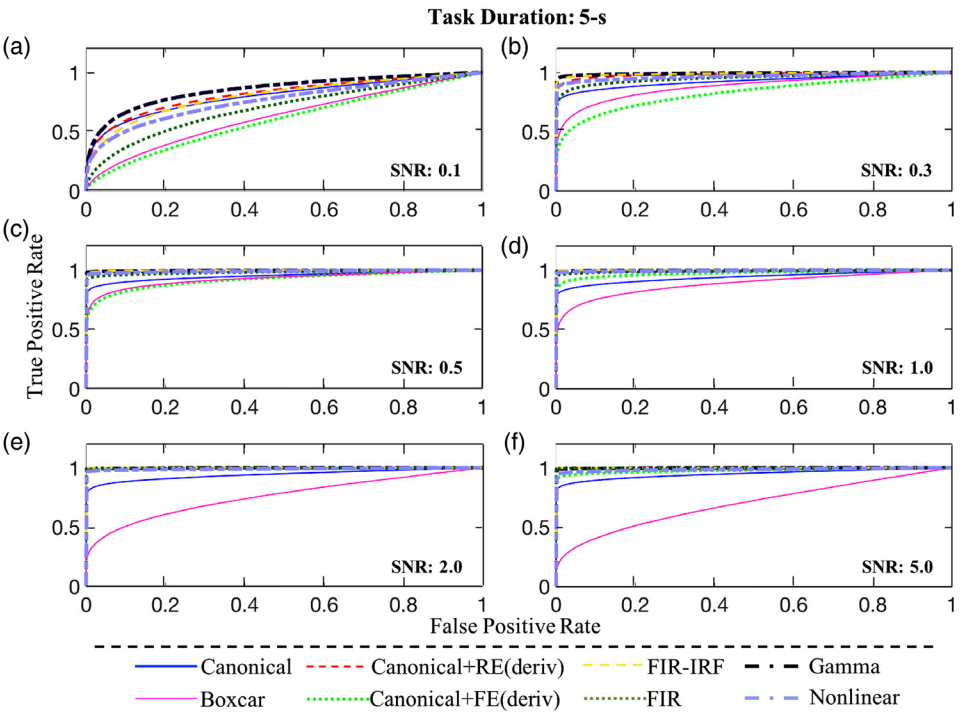

Fig. 4 ROC curves [TPR or sensitivity against FPR or (1-specificity)] for all (eight) different regression models at various SNR levels for 5-s task duration. Six different SNR levels of 0.1, 0.3, 0.5, 1.0, 2.0, and 5.0 are investigated in panels (a)–(f), respectively. In every panel, it shows the comparison of all (eight) different regression models: canonical, canonical+RE(deriv), canonical+FE(deriv), gamma, boxcar, FIR-IRF, IRF, and nonlinear.

A central theme of statistics is that you do not prove a hypothesis; you can only disprove a null one. In linear regression as it pertains to fNIRS brain imaging and this work, the null hypothesis is that the magnitude of the fNIRS signals during the task period does not differ from the baseline period. To test this null hypothesis, we offer an alternative such as the mean over a specific window of time during the response differs from the baseline period. If the mean over this time window differs from the baseline, then we can reject the null hypothesis. However, this does not mean that this is the "best" time window and there could have been an even better choice of alternative hypothesis that could have led us to reject the null hypothesis with even more confidence. In this context, the choice of basis set used in the regression (canonical, boxcar, FIR, etc.) simply frames the alternative hypothesis. We do not need to get this alternative hypothesis perfectly correct to still have the statistical power to reject the null hypothesis. When we use a canonical model, we are not ruling out that there could be a better-shaped model or different time window that may have allowed us to reject the null hypothesis even more soundly. The better our alternative hypothesis matches the data, the more sensitive (higher TPR) of the statistical test. However, the simpler (lower degrees of freedom) the alternative hypothesis, the more specificity (lower FPR) the test has. Thus, the performance of a linear regression model is a trade-off between reducing the false-negative rate of the model by using an accurate alternative hypothesis and reducing the FPR by using a model with fewer degrees of freedom such as the canonical model.

In this study, we found that the short task durations (<5 s) are most affected by the choice of basis set used within the regression model, especially for low SNR level (SNR <0.5). This is mainly due to the fact that the shorter tasks place stronger emphasis on the transients of the response. As the task duration increases, the transients are diluted by the steady-state behavior of the response. At short durations (see Fig. 2), a simple canonical model is more affected by the mismatch between the basis set used in recovery and the underlying true evoked signal. Although arbitrary shapes could be modeled with the FIR (deconvolution) model, this model only works well at higher SNR levels. Thus, at low SNR (<0.5), the canonical model has better performance in ROC analysis despite mismatches in the response shape. These mismatches are further negated by the inclusion of derivatives in the model when used as random effects (e.g., used in the regression but not included as part of the contrast). When derivatives are included directly as part of the contrast ("fixed effects" version), the results are much worse and the false discovery rate increases.

5.1 *Response Magnitude versus Timing Hypotheses*

We note that these findings only apply if the primary hypothesis of the study is testing the magnitude of the hemodynamic response, that is, if the hypothesis is comparing the evoked brain signal during a task to baseline or is comparing the amplitude of two tasks to each other. In this work, we did not look at how statistical tests can be used to examine if two tasks or groups have different evoked response timings from each other.

If two responses had different underlying timings, the use of a single basis set (including the FIR model using the same temporal window to define the contrast in both responses) would provide differing sensitivities to the two responses. Thus, rejection of the null hypothesis that the two responses had the same amplitude cannot distinguish between two responses with the

same timing but different amplitudes or two responses with the same amplitude but differing timings. When viewed as a first-order Taylor-series expansion on the timing of the canonical basis set, the inclusion of derivatives to the canonical model can be used to test amplitude (testing the main coefficients) and first-order differences to the timing (using tests on the coefficients assigned to the derivatives). This however, would only test a null hypothesis that the two responses were not statistically different to the first order in terms of the onset or dispersion of the canonical shape. In this work, we had proposed a nonlinear regression model but found that these models required a minimum level of SNR to be useful as currently implemented. When testing magnitudes of the evoked response, we could not recommend the use of the nonlinear model in this current work. However, future work applying more robust nonlinear estimation methods, such as regularized methods, may improve the utility of these nonlinear models and allow for better testing of timing differences.

5.2 *Limitations of this Study*

In this work, our results and conclusions are based on the testing of the null statistical hypothesis that the magnitude of the brain response during the task does not differ from the magnitude of the baseline period. In both our statistical tests and our simulations to generate ROC curves, we have assumed a linearly additive hemodynamic model. It is known that the hemodynamic response can be nonlinear, such as observed during long task durations demonstrating habituation effects or tasks with very short interstimulus intervals. In these cases where nonlinearities may exist, using a linear model to test and reject the null hypothesis is still statistically valid as a first-order test but will lose sensitivity (increase false-negatives) as the nonlinearity becomes more pronounced. Alternative methods such as the Volterra series[30] can be used to model such nonlinearities as better alternative hypothesis in these scenarios. An interesting future extension of this work would be to examine how these linear regressions begin to break down as the assumption of a linear additive hemodynamic response is violated. However, in this current work, our simulations are restricted only to the linear additive case.

A further conclusion of this work is that simpler (lower degrees of freedom) regression models are favorable particularly at lower SNR. We found that the increased sensitivity of these simple models often outweighs the loss in specificity due to the usage of a canonical model that is not a perfect match to the data. The results show that the canonical model performed well down to about a 50% overlap, which covers a fairly large variation in responses. However, this conclusion is not intended to say that the canonical model is necessarily the "best" for all data. For example, if there is a systematic bias in the shape of the brain response (e.g., infants may have different shape than adults), then using a slightly modified canonical model will recover some of the loss in specificity due to this mismatch. However, the conclusion of the paper is that, based on the AUC of the ROC analysis, it is better to have a slightly mismatched canonical shape than it is to use a model with high degrees of freedom.

6 Conclusions and Recommendations

Based on the simulations and analysis presented in this work, we conclude as follows:

- For task durations longer than about 10 s, the choice of basis set for the regression model is less important. For these longer tasks, regression models that used basis sets with lower degrees of freedom (even if there is up to about 50% mismatch in the shape of the response) have a better performance based on the ROC analysis for testing the hypothesis of differing response amplitudes.
- For tasks shorter than about 5 s, the sensitivity (TPR) of the model depends more on the choice of basis set. In this case, adding derivatives to the basis set to allow small variations in the modeled response recovers the sensitivity at the slight expense of reduced specificity. In our simulations, the best choice of basis set has been a canonical model with first derivatives included as random effects in the model. However, it is important not to use these derivative terms to define the contrast as these lead to high false-positives.
- Although the FIR and FIR-IRF models can adapt to any arbitrary underlying response timing, these models are more sensitive to noise and have worse performance at low SNR levels (<0.5) due to their higher degrees of freedom. At low SNR, the canonical model with derivatives is a better model compared to FIR, based on the ROC analysis even in the presence of a mismatch in the shape of the response.
- We do not recommend using a boxcar function as the basis set in the regression model for any task duration and any SNR level.
- The nonlinear HRF model, which is introduced in this work, did not outperform the canonical model with derivatives and was more affected by low SNR. In practice, nonlinear models are not useful for testing amplitude-based hypotheses. We found that the errors/uncertainties due to the nonlinear fitting procedures outweigh the slight improvements in the sensitivity gained by having a more accurately shaped response model.
- Our overall recommendation is that low degree-of-freedom canonical HRF model with derivatives (as random effects) provides the best choice for basis set in the linear model used for fNIRS analysis and can tolerate a moderate degree of mismatch between the underlying shape and the model assumptions. This represents the best trade-off between sensitivity and specificity of the methods tested in this work.

Disclosures

None of the authors has any financial conflicts of interest to disclose related to this work.

Acknowledgments

The authors acknowledge the funding from the National Institutes of Health (R01EB013210 [TJH]).

References

1. M. Cope et al., "Methods of quantitating cerebral near infrared spectroscopy data," *Adv. Exp. Med. Biol.* **222**, 183–189 (1988).

2. S. Lloyd-Fox, A. Blasi, and C. E. Elwell, "Illuminating the developing brain: the past, present and future of functional near infrared spectroscopy," *Neurosci. Biobehav. Rev.* **34**(3), 269–284 (2010).
3. G. Greisen, T. Leung, and M. Wolf, "Has the time come to use near-infrared spectroscopy as a routine clinical tool in preterm infants undergoing intensive care?" *Philos. Trans. A Math. Phys. Eng. Sci.* **369**(1955), 4440–4451 (2011).
4. M. Chen et al., "Neural correlates of obstacle negotiation in older adults: an fNIRS study," *Gait Posture* **58**, 130–135 (2017).
5. A. L. Rosso et al., "Neuroimaging of an attention demanding dual-task during dynamic postural control," *Gait Posture* **57**, 193–198 (2017).
6. X. Cui, D. M. Bryant, and A. L. Reiss, "NIRS-based hyperscanning reveals increased interpersonal coherence in superior frontal cortex during cooperation," *Neuroimage* **59**(3), 2430–2437 (2012).
7. T. Liu et al., "Inter-brain network underlying turn-based cooperation and competition: a hyperscanning study using near-infrared spectroscopy," *Sci. Rep.* **7**(1), 8684 (2017).
8. H. Obrig, "NIRS in clinical neurology—a 'promising' tool?" *Neuroimage* **85**(Pt1), 535–546 (2014).
9. T. J. Huppert, "Commentary on the statistical properties of noise and its implication on general linear models in functional near-infrared spectroscopy," *Neurophotonics* **3**(1), 010401 (2016).
10. F. Scholkmann et al., "A review on continuous wave functional near-infrared spectroscopy and imaging instrumentation and methodology," *Neuroimage* **85**, 6–27 (2014).
11. G. Strangman et al., "A quantitative comparison of simultaneous BOLD fMRI and NIRS recordings during functional brain activation," *Neuroimage* **17**(2), 719–731 (2002).
12. X. Cui et al., "A quantitative comparison of NIRS and fMRI across multiple cognitive tasks," *Neuroimage* **54**(4), 2808–2821 (2011).
13. A. Sassaroli et al., "Spatially weighted BOLD signal for comparison of functional magnetic resonance imaging and near-infrared imaging of the brain," *Neuroimage* **33**(2), 505–514 (2006).
14. A. T. Eggebrecht et al., "A quantitative spatial comparison of high-density diffuse optical tomography and fMRI cortical mapping," *Neuroimage* **61**(4), 1120–1128 (2012).
15. J. W. Barker, A. Aarabi, and T. J. Huppert, "Autoregressive model based algorithm for correcting motion and serially correlated errors in fNIRS," *Biomed. Opt. Express* **4**(8), 1366–1379 (2013).
16. H. Santosa et al., "The NIRS brain AnalyzIR toolbox," *Algorithms* **11**(5), 73 (2018).
17. A. M. Dale and R. L. Buckner, "Selective averaging of rapidly presented individual trials using fMRI," *Hum. Brain Mapp.* **5**(5), 329–340 (1997).
18. J. C. Ye et al., "NIRS-SPM: statistical parametric mapping for near-infrared spectroscopy," *Neuroimage* **44**(2), 428–447 (2009).
19. T. J. Huppert et al., "HomER: a review of time-series analysis methods for near-infrared spectroscopy of the brain," *Appl. Opt.* **48**(10), D280–D298 (2009).
20. A. E. Beaton and J. W. Tukey, "The fitting of power series, meaning polynomials, illustrated on band-spectroscopic data," *Technometrics* **16**, 147–185 (1974).
21. A. M. Dale, "Optimal experimental design for event-related fMRI," *Hum. Brain Mapp.* **8**(2–3), 109–114 (1999).
22. K. J. Friston, *Statistical Parametric Mapping: The Analysis of Functional Brain Images*, Academic Press, Amsterdam (2007).
23. T. J. Huppert et al., "A temporal comparison of BOLD, ASL, and NIRS hemodynamic responses to motor stimuli in adult humans," *Neuroimage* **29**(2), 368–382 (2006).
24. S. B. Perlman, T. J. Huppert, and B. Luna, "Functional near-infrared spectroscopy evidence for development of prefrontal engagement in working memory in early through middle childhood," *Cereb. Cortex* **26**(6), 2790–2799 (2016).
25. T. Arichi et al., "Development of BOLD signal hemodynamic responses in the human brain," *Neuroimage* **63**(2), 663–673 (2012).
26. P. S. Bellgowan, Z. S. Saad, and P. A. Bandettini, "Understanding neural system dynamics through task modulation and measurement of functional MRI amplitude, latency, and width," *Proc. Natl. Acad. Sci. U. S. A.* **100**(3), 1415–1419 (2003).
27. R. B. Buxton et al., "Modeling the hemodynamic response to brain activation," *Neuroimage* **23**(Suppl1), S220–S233 (2004).
28. E. Formisano and R. Goebel, "Tracking cognitive processes with functional MRI mental chronometry," *Curr. Opin. Neurobiol.* **13**(2), 174–181 (2003).
29. K. S. Hong and H. D. Nguyen, "State-space models of impulse hemodynamic responses over motor, somatosensory, and visual cortices," *Biomed. Opt. Express* **5**(6), 1778–1798 (2014).
30. K. J. Friston et al., "Nonlinear responses in fMRI: the Balloon model, Volterra kernels, and other hemodynamics," *Neuroimage* **12**(4), 466–477 (2000).

Biographies of the authors are not available.