Provisional chapter

The Neuro-Metabolic Underpinnings of fMRI BOLD Dynamics

Christopher W. Tyler and Lora T. Likova

Additional information is available at the end of the chapter

1. Introduction

The primary indicator of brain activity in the fMRI technique is known as the blood oxygenation level dependent (BOLD) responses, which derives from the hemodynamic response of the local blood vessels recorded throughout the cortex. The goal of this review is to describe new approaches to the estimation of the neural signals underlying the BOLD signal. A proper understanding of the metabolic pathway underlying the fMRI BOLD signal is a necessary precursor to an analytic capability for neural signal estimation from the BOLD waveforms. Any such estimation must be based on a model of the known neural population dynamics underlying the BOLD metabolic signal generation, which may be progressively refined as more information becomes available about both neural response characteristics and the metabolic cascade. Given adequate signal/noise ratio, it is possible to develop approaches that overcome the temporal limitations of BOLD signal and are able to reveal relevant properties of the underlying neural signals. This analysis can provide a direct linkage between the live assessment of the functioning brain and the direct neurophysiological recordings in other species.

FMRI analysis techniques for estimating the BOLD signal typically employ the Generalized Linear Model (Boynton et al, 1996), which incorporates the convolution approach to the estimation of the underlying neural signal. Convolution is based on the assumption of a unitary BOLD waveform kernel that generates the straightforward prediction of the BOLD response waveform for any stimulus type or duration in any brain area. In fact, however, major deviations from a standard BOLD waveform may be found, even in the same cortical regions, for variations in stimulus conditions. D'Avossa, Shulman & Corbetta (2003), for example, reported strong differences in waveform when the response to the motion or color of a cue/stimulus pairing was modulated by attention. Such local waveform differences most likely

derive from differences in the *neural* signals driving the BOLD activation, since the metabolic and hemodynamic processes that mediate the paramagnetic signals should be invariant on the scale of typical voxel dimensions within a given cortical region.

The BOLD signal measured by fMRI has low temporal resolution (0.5 – 5 *s*) relative to other methods for mapping human brain function (such as high-density EEG analysis). Estimation of single parameters of the BOLD waveform, such as response delay alone, can improve the temporal resolution for the neural signal delays to 100 *ms* or better for narrowly targeted brain regions (Menon, Luknowsky & Gati, 1998; Henson et al., 2002) but this requires the assumption that the BOLD signal has a unitary waveform, which is often not the case (Aguirre et al., 1998; D'Avossa, Shulman & Corbetta, 2003; Handwerker, Ollinger & D'Esposito, 2004; Likova & Tyler, 2007, 2008; Tyler & Likova, 2009). Even minor deviations from a stable waveform violate the assumptions of such single-parameter analysis and invalidate the delay measure. A more advanced approach to neural signal estimation is therefore needed.

2. The chain of metabolic processing

Although much effort has gone into the analysis of the temporal dynamics of the BOLD signal as a proxy for the underlying neural activity that elicits it, the link between neural activity and the hemodynamic response is nevertheless indirect, involving a chain of metabolic processes mediated by the astrocytic glial cells in the cerebral cortex. Here, we focus on the role of the metabolic pathways mediated by the glial cell in coupling the neural activity to the BOLD responses in the blood vessels, and the consequent implications for the processes governing the fMRI BOLD dynamics.

In general terms, the stimulus impinging on the subject generates a sequence of neural responses starting with the transduction into a neural signal within the sensory receptors. The first element in this chain is the astrocytes surrounding the neuron, which provide glucose to the neuron and replenish its supply by ATP metabolism fueled by oxygen from neighboring blood vessels. This signal then propagates to the brain and activates various populations of neurons within the voxels being analyzed by the fMRI technique (Logothetis, 2003). For instance, the neural signals arriving from the visual pathway generate synaptic activation of the populations of neurons in the primary visual cortex. This synaptic activation generates an energetic demand for the restoration of the neurotransmitter molecules carrying the activation signals across the synapses. The chain of cortical processing progresses from the neural events through the metabolic demand mediated by the glucose/lactate cycle in neighboring astrocyte glial cells to the processes of oxygen delivery by the adjacent arterioles and capillaries that is detected by the imaging technology. The first element in this metabolic chain is the local astrocyte processes, which provide glucose to the neuron and replenish its supply by an ATP metabolism, fueled by diffusion through the astrocyte cytoplasm from their endfeet connections to neighboring blood vessels (Magistretti & Pellerin, 1999; Magistretti, 2009). The integrated metabolic demands are met primarily by the astrocytes, which integrate the required energy consumption over time and space and make a complementary metabolic demand on the adjacent vasculature.

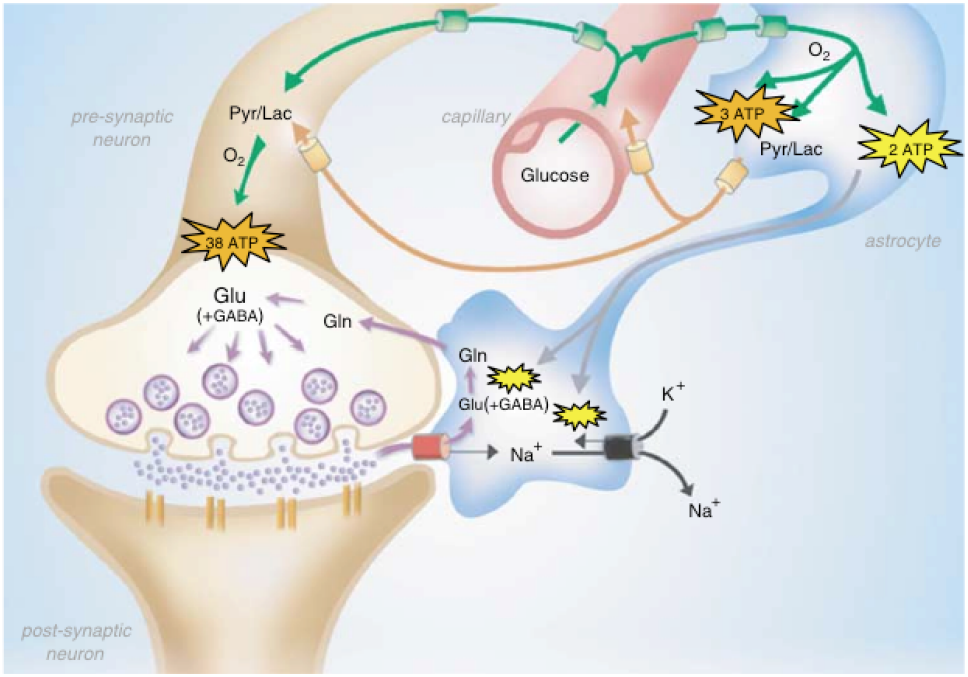

Figure 1. The astrocytes as the substrate for the neurovascular coupling of the neural metabolism. (From Hyder et al., 2006, with permission).

In detail, the pathways are complex, involving lactate, glucose and glycogen metabolic mechanisms, mediated by both intracellular astrocyte and supplementary extracellular transport (Dienel & Cruz, 2008; Gandhi et al., 2009) but the connection between the local glucose metabolism close to the synapse and the oxygen-based hemodynamics in the blood vessel remains unclear. Three hypotheses for this neuro-hemodynamic coupling may be advanced (see Fig. 1), although all three remain to be tested:

- *aerobic glucose metabolism.* This is the concept of a direct coupling between the neural glucose metabolism and the vascular oxygen supply, in which the neural metabolism is supported by oxygen transported from the blood vessels either within the encapsulating astrocytes or through the extracellular cytoplasm (having passed through the astrocytic sheath) to reach the site of the neural synapses and provide the oxidative metabolism of the glucose to reconstitute the ATP used in the neural response. Studies in rat cortex have demonstrated a linear relationship between neural activity, glutamatergic neurotransmitter flux and the cerebral rate of oxygen metabolism (Hyder et al., 2002; Smith et al., 2002). Since cells are predominantly linear summators of the excitatory and inhibitory transmitter release across their synaptic population, the energetic demand driving the BOLD signal is most closely coupled to the net transmitter signal imping-

ing on the cells, and hence to intracellular potential in the cells. The problem with this hypothesis is that aerobic metabolism is a process with a time constant of the order of minutes (Margaria et al., 1933), which is too slow to account for the 5 s time constant of the BOLD response, despite its high efficiency.

- *anaerobic glucose metabolism.* This is the concept that the entire neural metabolic process is based on anaerobic glucose delivery from the blood vessel, and that the variation in the oxygenation fraction of the hemoglobin is an epiphenomenon. On this hypothesis, the metabolic demand the site of the neural synapse generates a signal in the astrocytes to release nitric oxide in the filopodia (endfeet) wrapping the arterioles, stimulating an increase in the arteriolar volume with a consequent increase in the proportion of oxyhemoglobin in that region of tissue. The ~5 *s* time constant of decay of the nitric oxide in the presence of hemoglobin (Hakim et al., 1996) meshes well with that of the BOLD response dynamics. The nitric oxide release signal is presumably mediated by a calcium wave traveling 'antidromically' through the astrocyte (Bezzi et al., 1998; Koehler et al., 2006), with its diffusion time accounting for the 1-2 *s* onset delay in the onset of the BOLD signal.

- *anaerobic stimulation of the combined metabolic pathway.* This is a mixed concept in which the anaerobic metabolic demand from the neural glucose metabolism stimulates the nitric oxide (NO)-mediated arteriolar dilation (Burke & Bührle, 2006; Kitaura et al., 2007) with the consequent increase in both glucose and oxygen transport into the astrocytes (and extracellular cytoplasm). On this concept, the slow diffusion of the oxygen and glucose molecules to the synaptic sites is irrelevant to the BOLD response time course. The critical factor is that the metabolic demand generated by the neural glycolysis is fast enough to elicit an NO signal to the arterioles that, together with the NO decay time constant, generates an arteriolar volume time course compatible with the measured BOLD dynamics. On this interpretation, the question of how long the transported products take to reach their metabolic targets, to provide the needed aerobic recovery from the anaerobic depletion, is inaccessible by the BOLD signal probe. It can only discriminate the slower events resulting from the arteriolar volume changes.

Thus, the most likely basis of the metabolic demand driving the cortical BOLD signal is the energetic load deriving from the total conductance changes in the postsynaptic membrane generated by a range of processes subsequent on transmitter release at the synaptic inputs to each neuron. The summed metabolic demand in the nexus of active cortical neurons adjacent to a capillary forms the drive for the metabolic response in that region of cortex. Hence, the transmitter release is tightly coupled to the activation of the post-synaptic receptors on the recipient cell membrane and consequently to the energetic demands of the membrane receptor activation (and to a lesser extent to the subsequent recycling of the transmitter molecules). The majority of these energetic demands are met by the conversion of glutamate to glutamine in the neighboring astrocytes (Sibson et al., 1998, 2001; Rothman & Schulman, 1998). The glutamine is then taken up by the neurons for reconversion to glutamate for use as a transmitter, releasing energy within the neuron in the feedback loop.

3. The time course of astrocytic responses

A key factor in understanding the role of astrocytes in the metabolic pathway supporting neural activity is their time course. It may be emphasized that the astrocyte metabolic processes are slow relative to the intracellular signal dynamics, as are the processes of hemodynamic oxygen supply. The time constant of a biological process may be defined as the unit area of the response to a sufficiently brief input event (or, equivalently, the unit area of the temporal derivative of the response to a step input.) The time constant of the astrocyte responses at the cell body is known to be of the order of several seconds (Kelly & van Essen, 1974; Filosa, Bonev & Nelson, 2004; Metea & Newman, 2006). The hemodynamic response of the blood vessels to expand in response to the neural metabolic demand is mediated by control of the arteriole diameter with the enveloping astrocyte endfeet (Magistretti & Pellerin, 1999; Hyder et al., 2006; see Fig. 1) with a very similar time course to that of the slow astrocyte responses that must underlie the observed hemodynamics. The post-neural processing stages are often modeled as a linear hemodynamic response kernel convolved with the presumed neural signal. However, this approach overlooks the key role of the pre-hemodynamic processes of the glial and other intermediaries as just described. To reflect the contributions of these intermediary processes, therefore, we will refer to these as the 'metabolic response kernel' (MRK) incorporating both the glial and hemodynamic components of the metabolic recovery processes.

4. The form of the metabolic response kernel

The time domain approach of the present analysis allows the extraction of the maximum possible information about the temporal evolution and any processing nonlinearities of the neural signals underlying particular BOLD activation profiles. The analysis in the following sections reveals that much information about the neural response properties is reflected in the BOLD signals, even if the time-resolution is insufficient to reproduce the exact neural signal. This capability is particularly clear in the case that the full metabolic response kernel (MRK) is monophasic (see Section 5). (Other MRK forms are a biphasic form with an initial positive lobe of the response rebounding to a subsequent negative lobe before returning to the baseline level (Buxton, 2001), or the triphasic form in which the positive lobe is both preceded and followed by a negative lobe (Thompson et al., 2003). In either case, the neural response properties are difficult to distinguish from the metabolic ones, proportionately to the area of the extra lobes in the MRK.)

However, it is an established property of the BOLD response that it is largely sustained for an appropriately sustained neural response (Boynton et al., 1996; Birn, Saad & Bandettini, 2001; Glover, 1999; Logothetis, 2003; Shmuel et al., 2006). The implication of this result is that the MRK as a whole is monophasic, since convolution of a sustained input with a biphasic impulse response for the subsequent processing will inevitably result in a transient rather than sustained BOLD response.

The monophasic assumption was tested for the rat cortex by both Martindale et al. (2003) and de Zwart et al. (2005), who showed that the empirical dispersion of the BOLD response generated a delay increasing with distance from the activation site, but always well-fit by a monophasic model of the BOLD impulse response. Similarly, direct measurements of cerebral blood flow and the concentrations of oxygenated and deoxygenated hemoglobin in the human brain (Hoge et al., 2005) reveal only a monophasic temporal waveform for each of these contributors to the BOLD response. These results are all compatible with the inference of a dominant monophasic positive BOLD response in cat LGN and cortex, as reported in Thompson et al. (2003, 2004, 2005). This monophasic form also seems to be a fair approximation in the case of human fMRI because the canonical response kernel (commonly termed the HRF) provided in the SPM software package, although biphasic, has a negative lobe of less than 10% of the amplitude of the positive lobe. It is therefore only a minor modification of this standardized kernel to assume that it has no significant negative lobe, which is the assumption made for the following analysis. (As will be explained below, the residual biphasic component in most published reports can be equally attributed to *neural* rather than hemodynamic rebound signals.)

5. Implications for the variety of BOLD response waveforms

Armed with this monophasic assumption for the MRK, we show how several properties of the neural signal are reflected in the recorded BOLD waveform (Tyler & Likova, 2011). This demonstration assumes a linear relationship between the neural response and the BOLD waveform, in order to make its properties clear before introducing the nonlinear aspects of the analysis. The analysis will demonstrate how the principles of the polarity, latency, transience and number of phases may be reflected in the BOLD response when they are present in the underlying neural population response:

5.1. Polarity

For any monophasic neural response, the polarity of the BOLD response will be an accurate reflection of the polarity of the neural response, regardless of the difference in their time courses to any order of magnitude.

5.2. Latency

Any delay in the neural response will also be reflected in the consequent BOLD response. Of course, the metabolic processing sequence may introduce additional delays, but neural delays such as response reaction times or perceptual ambiguity delays should be accurately reflected in the BOLD waveform once the inherent delays of the MRK are taken into account.

5.3. Transience

As described in Section 4, transience of the BOLD response for a sustained stimulus implies a transience of the underlying neural response. For example, in most visually responsive cortical

neurons the onset of a sustained light is known to generate neural responses consisting of an initial transient followed by a smaller sustained response that is often of much lower amplitude. Such stimulation will generate a transient BOLD response even though the stimulation and photoreceptor response are sustained. (It is for this reason that typical stimulation in fMRI experiments is repetitive, since rapid repeated stimulation will generate a series of transients that combine to form an effectively sustained neural response.)

5.4. Number of phases

For a monophasic MRK (HRF), the BOLD waveform will have the same number of phases as the input stimulus, if the neural input is balanced for positive and negative lobes. Thus, the fact that the measured BOLD waveform is typically biphasic does not imply that the MRK is necessarily biphasic. The negative lobe may derive from a biphasic *neural* response to a stimulus rather than to blood dynamics.

Some of these properties of the BOLD response are illustrated in the simulation of Fig. 2, for which the MRK is assumed to be the gamma function shown in Fig. 2A, right, as the basis function for the formal analysis (see Section 9.1). Note that gamma bases are statistical descriptors of the occurrence of discrete particles (Stacy, 1962; Farwell & Prentice, 1977) that have a simple analog implementation that is a cascade of identical lowpass filters (De Vries & Principe, 1992; Celebi & Principe, 1995; Chen, 2006). In terms of molecular diffusion processes within neurons, therefore, the gamma function represents an optimal description of the cooperative process of the arrival of effector molecules at the channels controlling current flow through the cellular membrane (Shao, 1997).

6. Direct forward modeling approaches

A few previous studies have made estimates of the effects of different neural models on the form of the BOLD response dynamics. For example, Mechelli et al. (2001) report a simulation study of the estimated regional cerebral blood flow (rCBF) and blood-oxygenation-dependent (BOLD) signals as a function of the duration, onset asynchrony and relative amplitudes of two brief stimuli. They included a basic model of neuronal dynamics and varied one parameter of this model – the amplitude of a slow late transient – to show its effect on the simulated BOLD responses. This exercise constitutes an unvalidated forward model of the effect of only one parameter of a simulated neural response.

However, what Mechelli et al. (2001) offer as the analysis of the effects on the BOLD waveform is described as "the BOLD parameter estimates", which is actually a single-valued function with no specification of which parameter(s) is/are being estimated. Since no comparison is made with empirical data or their noise limitations, the fact that this study includes a proposed model for the neural dynamics does not qualify it as a validated procedure for estimating the neural population dynamics underlying the local BOLD signals (which is the goal of the present chapter).

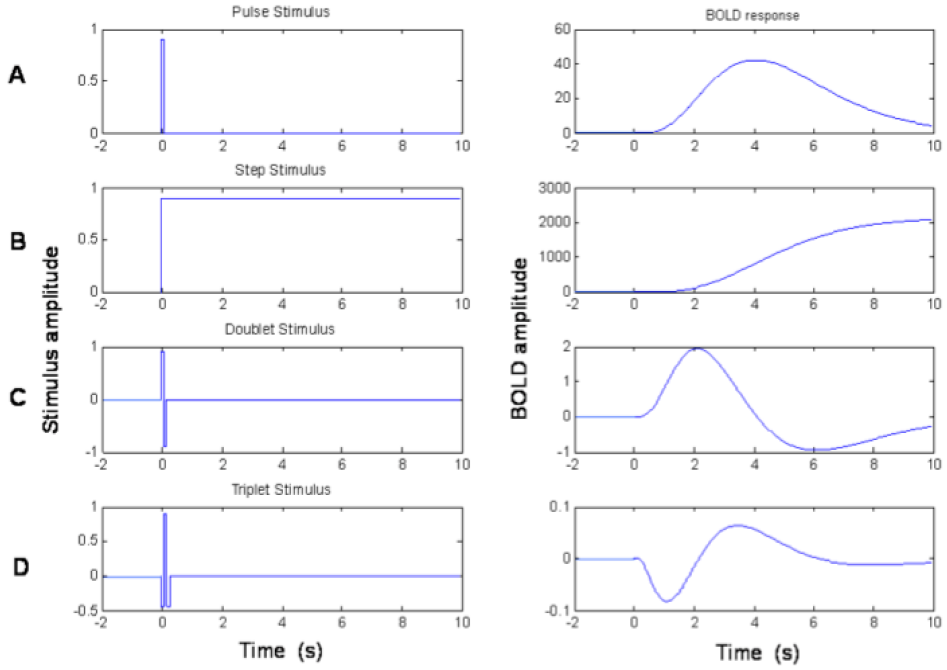

Figure 2. Left panels: (A) neural impulse response, (B) step response, (C) balanced doublet and (D) balanced triplet response. Right panels present convolution of each of these responses with the MRK shown at upper right. Note that differences in neural response characteristics (left panels) at the time scale of 100 *ms* generate profound changes in the simulated BOLD waveforms (right panels) on a much longer timescale, which in turn are diagnostic of the differences in the neural signals.

Similarly, Buxton et al. (2004) extend their balloon model of the hemodynamic response leading to the BOLD signal by proposing a model of the neural response to account for the temporal nonlinearity in BOLD responses as a function of duration. This neural response incorporates a slow subtractive inhibitory component to the net neural signal, which has the effect of producing a neural response consisting of an initial transient followed by a sustained plateau. Model responses for three kinds of stimuli – a single short pulse, two short pulses and one long pulse, are offered as a demonstration of the properties of this model. As with Mechelli et al. (2001), no attempt is made to compare the model outputs with actual BOLD recordings, so the Buxton et al. (2004) study again does not qualify as a validated procedure for estimating the neural population dynamics underlying the local BOLD signals.

Both Mechelli et al. (2001) and Buxton et al. (2004) include parameters intended to account for the temporal nonlinearity of short duration responses (which do not fall linearly as response duration is reduced; Boynton et al., 1996; Birn, Saad & Bandettini, 2001). Both studies demonstrate the required lack of reduction in a single example of a short-duration response, but neither study provides a validation of either the waveform or the amplitude response function relative to empirical BOLD data. In principle, although either model could provide a platform

for such validation or for the further estimation of the dynamics of the underlying neural population response, they neither do so nor suggest procedures by which such estimation could be achieved.

7. Analytic model of the neural-BOLD coupling

To start the analysis, we developed a specific model structure of the processes leading to the BOLD paramagnetic signal of fMRI recordings. This model incorporates aspects of the biophysical processes that are not present in the linear convolution analyses of Friston et al. (1997, 1998, 2000), although somewhat condensed in comparison with the biophysical/metabolic derivations of Mechelli et al. (2001), Buxton et al. (2004) or Sotero & Trujillo-Bareto (2007, 2008).

The metabolic demand through the biophysical chain to the measured BOLD signal has two sources of nonlinearities. The major source of the BOLD waveform measured by fMRI is the summed local field potential in the cortical region contributing to the signal for each voxel. Because each astrocyte meets the metabolic demand for multiple synapses, it sums the metabolic demand generated by the underlying neural activity across the local neural sources, and is thus nonlinear with respect to the formalism of the General Linear Model. The metabolic demand is no longer specifiable as the convolution of the stimulus waveform with a single impulse response function but is the complex sum of many such convolution processes. In addition, the coupling itself also exhibits saturation nonlinearities when pushed beyond the region of small-signal linearity.

The neural responses within each voxel are treated as generated by sets of homogeneous populations with similar signal waveforms $N_i(t)$ within each population (Fig. 3, where N is the neural signal variation over time t for each population i). Each neural response then generates a local metabolic demand $M_i(t)$ that may have a nonlinear relationship to the neural signal waveform (Chatton, Pellerin & Magistretti, 2003). The MRK will be convolved with a nonlinear transform of the presumed neural signal to provide an estimate of the neural metabolic demand that is being met by the combined glial and hemodynamic metabolic response. The fMRI analysis also has a finite dynamic response time, but it will be treated as incorporated in the MRK of the glial/hemodynamic response.

These nonlinearities provide the opportunity to evaluate an analytic model of the neural signal leading to the BOLD activation. The model assumes the presence of neuronal subpopulations having response dynamics with various decay time constants in response to the stimulus. Pooling among the subpopulation responses can then explain the multiple decay characteristics of the recorded local field potentials (LFPs). Subsequent convolution with a characteristic metabolic response kernel then generates the predicted response $Y(t)$ for the BOLD waveform for the region of cortex generating the LFP signal, accounting for concurrently recorded BOLD waveforms.

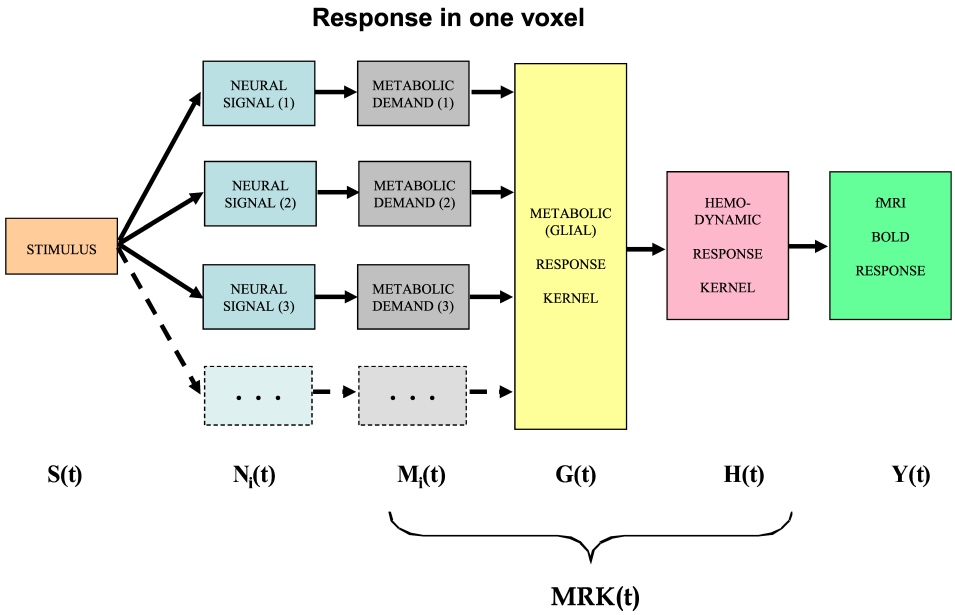

Figure 3. Block diagram of the main processing stages that lead up to the BOLD signal. The boxes represent processes denoted by capitalized functions of time. The *i* subscript indicates that the stage incorporates multiple components in parallel within the voxel, as indicated by the parallel boxes in these columns. Dashed boxes indicate an array of further components. See text for further details.

The neural responses within each voxel are modeled as sets of homogeneous populations $N_i(t)$ with similar signal waveforms within each population. Each neural population response then generates a local metabolic demand $M_i(t)$ that may have a nonlinear relationship to the neural signal waveform. The integrated metabolic demands are met primarily by the astrocytes, which integrate the required energy consumption over time and space and make a complementary metabolic demand $G(t)$ on the adjacent vasculature. The hemodynamic processes $H(t)$ replenish the energy depletion in the astrocytes, leading to the paramagnetic response that generates the BOLD signal of the differential precession rates of the oxygenated/deoxygenated hemoglobin moieties.

8. Implications for the analysis of BOLD fMRI signals

To evaluate the neural contribution to the differential BOLD response waveforms within the same cortical regions, Likova & Tyler (2007) developed an "instantaneous stimulus paradigm" to evoke BOLD signals in response to instantaneous stimulus transitions. It would typically be assumed that such transitions, being instantaneous, would all generate the same BOLD waveform (effectively equivalent to the MRK) for all different stimulus configurations. Thus, any significant deviation from that prediction in the BOLD waveform elicited in the same

cortical area can contribute to revealing the specifics of the underlying neural processing and enhance the understanding of the networks of extended perceptual responses to complex stimulus configurations. Indeed, the instantaneous stimulus paradigm generated striking differences in the BOLD waveforms properties (e.g., latency, sign, amplitude and width) even within the same brain areas as a function of the stimulus type (Likova & Tyler, 2007).

Generally speaking, there is overall homogeneity of the cortical blood supply on the scale of fMRI voxels (~2 *mm* cubes), although at a finer scale the blood supply may be separated into multiple layers supplied by different local cortical sources (Duvernoy, 1999). Although there have been reports of different temporal dynamics of the BOLD response recordable at high field strength from different cortical layers (Silva & Koretsky 2002; Jin & Kim, 2008; Siero et al., 2011) and degrees of capillary branching order (Tian et al., 2010). Nevertheless, all these studies are subject to the caveat that the response at different depths were mediated by ***neural*** stimulation, and there was no validation of how the neural signals may or may not have varied with cortical depth in these studies. Moreover, these reports may be characterized by three summary statements: i) all the reported BOLD responses are of similar form, predominantly monophasic; ii) the times-to-peak are similar over cortical depth, varying by less than 1 *s* from the mean peak time of ~5 *s*; iii) there are minor waveform differences with either early or late negativities of the order of 10-20% of the amplitude of the positive peak. The two-photon study of the relationship to capillary structure (Tian et al., 2010) found that BOLD peak times were ***invariant*** with cortical depth, although there was a minor degree of depth dependency in the onset time of the BOLD response and the strength of the initial dip. Overall, therefore, it seems that BOLD dynamics show only minor variations with cortical depth, and that even these minor variations have an unknown degree of contribution from neural response variations to the differential cortical layer dynamics.

Any reported variations in the vascular dynamics are, however, minor relative to the dramatic waveform differences across the cortex observed by Likova & Tyler (2007), as illustrated in Fig. 4. Widely separated regions across the cortex form sets of regions (indicated by the colored dots) with major differences in BOLD waveforms across the different sets. These striking waveform variations are therefore most likely attributable to differences in the underlying neural dynamics, not to spatial variations in vascular dynamics. No study of BOLD waveform variations has suggested that differences as large as these could be attributable to hemodynamic variations. Consequently, these results imply that fMRI signals contain much more information about the neural processing than is commonly appreciated, and thus have the potential to capture them through an appropriate approach.

9. Nonlinear dynamic forward optimization

However, until recently there has been no method of transcending the BOLD temporal limitations in order to estimate the dynamics of the neural signals underlying the measured fMRI waveforms. Tyler & Likova (2009, 2011) therefore proposed a Nonlinear Dynamic Forward Optimization (NDFO) approach for the time-resolved estimation of the neural signals

underlying the particular characteristics of the temporal BOLD waveforms for a particular stimulus processed by a particular cortical region. The philosophy of this approach is to utilize the information available from neurophysiological studies of the neural population dynamics and biochemical studies of the metabolic pathway coupling to the measurable blood response to provide Bayesian priors as to the likely temporal structure of the component neural signals. Such a forward modeling approach provides a compact account of the measured waveform with the minimal number of neural predictors, based on prior knowledge of the expected temporal properties of neural signals and of their consequent metabolic demand. In the case of the neural signals, the goal is to estimate the amplitude and time course of each of the neural components whose metabolic effects, when summed, account for the measured BOLD waveform for a particular stimulation condition and cortical region.

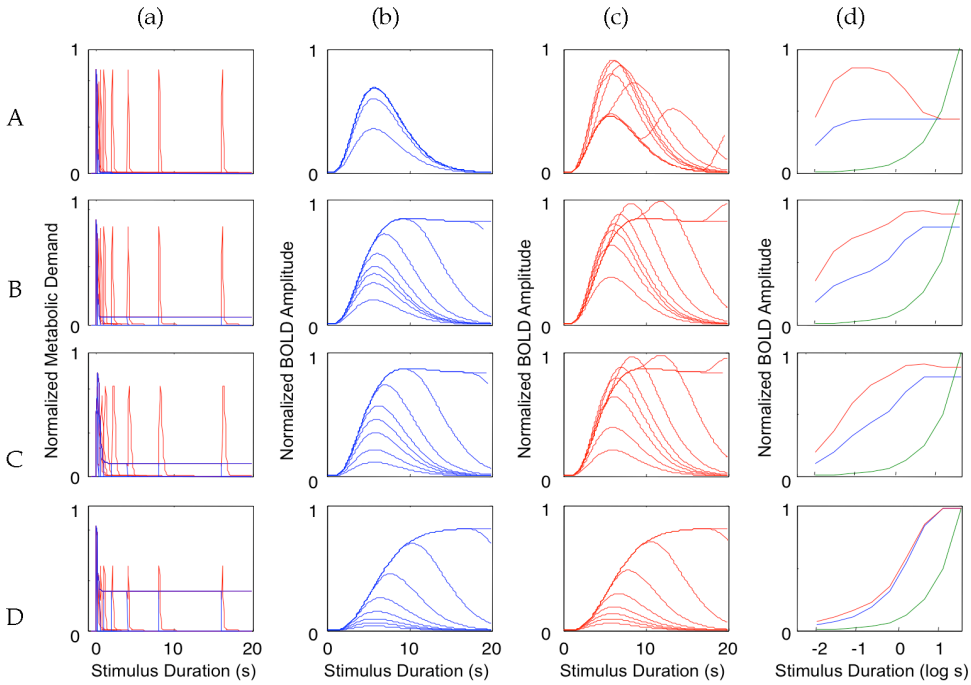

Figure 4. BOLD waveforms in different brain regions to an instantaneous figure/ground stimulation consisting of the asynchronous replacement of center (square) and surround fields of random dots in two sequential screen updates (i.e., separated by only a 60 *ms* delay). Colored spots indicate sets of brain regions with similar BOLD waveforms (ordered in terms of peak delay), which differ radically across regions coded by different colors. (From Likova & Tyler, 2007).

Rather than simply characterizing the behavior of the BOLD waveform (Birn, Saad & Bandettini, 2001; d'Avossa, Shulman & Corbetta, 2003; Fox et al., 2005; Grotz et al., 2009) or attempting to infer the potentially complex properties of the underlying neural mechanisms from the form

of the BOLD response by deconvolution (Glover, 1999; Logothetis, 2003; Logothetis & Wandell, 2004), the concept of forward modeling is to incorporate as much knowledge as possible about the likely neural substrate and optimize the remaining details to best fit the BOLD waveform. Here the predictors are non-linear because there is a nonlinear relationship between the neural responses and the metabolic demand that they generate, although the summative property of the paramagnetic signals throughout a voxel implies that we can assume that the component metabolic demands sum linearly together (Fig. 3). The dynamic forward modeling approach may thus incorporate a variety of possible nonlinearities into the structure of the model. The neural model that we investigate in the present version of the analysis is the sum of a positive and a negative component based on delayed gamma functions convolved with the stimulus waveform

9.1. Analytic framework for the neural temporal response

The starting point for the NDFO modeling is a gamma function model of the neural signal, whose first effect in terms of the cascade of BOLD dynamics is to create a metabolic demand G(t) in the neighboring glial cells (see Fig. 3). Gamma probability density functions have the analytic form $a.\ t^{k-1}.\ e^{-t/\tau}/(k)$, where t is the time dimension, a is a scaling parameter and k and τ are generic waveform parameters. They may be termed simply "gamma functions" to emphasize their analytic rather than statistical properties. In engineering the same function is known as the n-pole filter function (with n=k), and is used to describe the temporal dynamics of a wide range of processes. For the present application, the temporal gamma function is assumed to have integer powers of k and corresponds to the solution of differential equations with real (non-imaginary) roots. The gamma function has the analytic advantage over many other functions, such as the Gaussian, that it is by definition causal because it has the value of zero at t=0, and is defined as zero for $t < 0$ (i.e., the full specification *is* $y=a.\ t^{k-1}.\ e^{-t/\tau}$ for $t >= 0$; $y=0$ for $t < 0$, where the gamma scaling factor is folded into the scaling constant *a)*. Its shape progresses from highly asymmetric around the peak for small k to approximately Gaussian and symmetric for large k (see Fig. 4).

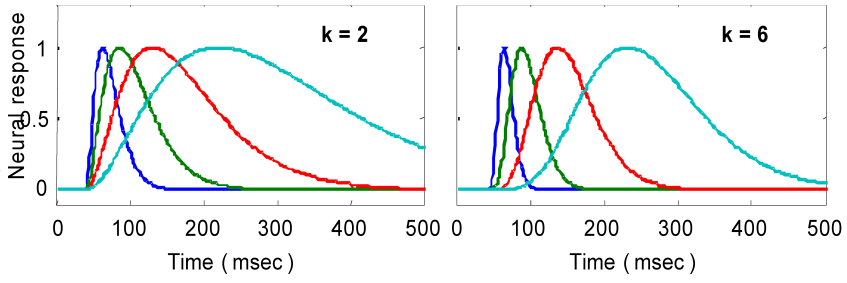

Figure 5. Examples of delayed gamma function step responses with exponents of k=2 (left) and 6 (right). Successive functions (colors) introduce a wide range of peak latencies for the neural signal estimates (with time constants increasing in factors of 2 and a fixed delay Δt of 40 ms).

A key feature of this formalism is that the peak latency is determined by the time constant, **τ**, and is proportional to the width (at half height) of the response peak, which can be estimated to good accuracy by the methods described in the next section.

9.2. Neural model

A comprehensive model of the BOLD requires an accurate model of the intracellular potential dynamics deriving from the sensory stimulation. Based on the gamma function formalism, we propose to use the nonlinear neural response model jointly specified in eqs. 1 and 2:

$$N_{\mathrm{P}}(t) = [|N(t)|_{+} + \lambda \cdot |N(t)|_{-}]e^{-t/\eta} + \varepsilon(t) \tag{1}$$

where (t) is the net source of additive noise and the function is governed by the parameter set $P = (a, k_n, \tau_n, b, \Delta t, \eta, \lambda)$ (see Table 1).

The nonlinear neural signal $N_r(t)$ in eq. 1 is the sum of half-rectified positive and a negative components based on delayed gamma functions, convolved with the stimulus waveform:

$$N(t) = a \cdot S(t) \otimes n(t - \Delta t) \tag{2}$$

where

$$n(t) = (1-t) \cdot t^{k_n - 1} e^{-t/\tau_n} - \frac{b}{\tau_n} t^{k_n} e^{-t/\tau_n}$$

(Note the convention that time series functions are capitalized, impulse response kernels are lower case, and vectors are bold face.) The neural impulse response expression $n(t)$ is set up so that its convolution with a step function is equivalent to the sum of a pure transient and a pure sustained component. In addition, the expression is specified with an additional transmission delay, **Δτ**, that delays the response relative to the stimulus without affecting its waveshape with the parametrization specified in Table I, which defines the parameters in vector **p** of this equation. To illustrate the properties of the model, we analyze the effect of varying the inhibitory ratio implied by the negative component weight b, and the offset/onset gain ratio λ in Fig. 6.

a	scaling constant (a fitting parameter but not a waveshape parameter)
k_n	integer exponent governing the rising phase
τ_n	time constant of falling phase in the neural response
Δt	transmission delay before response onset
b	sustained/transient ratio in the step response
η	time constant of nonlinear gain control in the neural response
λ	ratio of offset to onset gain in the neural response

Table 1. Nonlinear Forward Model Parameters

One issue that arises is how to measure the latency Δt of the delayed gamma functions of Fig. 5. A simple derivation can show that the peak latency t_{peak} of these responses is specified by the expression ($k\tau+\Delta t$). Thus if k and Δt, both of which are well-determined from neurophysiological studies in monkey cortex, are set to the means of their Bayesian priors on this basis, the peak latency t_{peak} can be determined from the value of τ, which can be accurately derived from the model optimization.

9.3. Metabolic demand

Since little is known about the glial dynamics of transmitter recovery, we may pursue two options as to their effects. One option is that the metabolic demand driving the BOLD response derives from the transmitter recovery cycle following the activation by an axonal spike. Since axonal spikes represent only the positive aspect of the intracellular voltage and since 90% of cortical synapses are excitatory (Shank & Aprison, 1979; Wang & Floor, 1994), the signal transmitted from one cortical stage to the next may be treated as a half-wave rectified version of the dynamic neural signal (i.e., only the positive component in eq. 1) with λ set to zero. To illustrate the properties of our model, this prediction is shown as the blue curves in Fig. 6A(a) (indexed in row/column notation), which is an overlay of the model estimates of the neural responses to stimulus pulses that double in duration from 8 *ms* to 16 *s* (eight doublings). For this example, the neural response has balanced excitation and inhibition, so even the prolonged pulses generate only an initial transient response, with the negative lobe at offset being thresholded out by the rectification. (Note that the local metabolic demand, $M_i(t)$, has the same time course in this model as the transmitter recovery from which it derives. The energetic processes required for the recovery to the initial state, however, form a chain of glial metabolic response, $G(t)$, that may have substantially slower time course at one or more stages.) The other option is to consider the instantaneous metabolic demand of both excitatory and inhibitory cells, or both 'on' and 'off' cells, implying that the signal generating the metabolic demand is a full-wave rectified version of the intracellular voltage (i.e., the full expression of eq. 1 with $\lambda > 0$).

9.4. Metabolic coupling

Having provided relevant variables for the linear and nonlinear components of the neural response dynamics, we may now consider the issue of the metabolic coupling with the neural signal to generate the BOLD response. This coupling has been modeled extensively over the past two decades, with the best-known example being the Buxton-Friston *balloon model* (Friston et al., 2000; Buxton et al., 2004) and the most elaborated version being by Sotero & Trujillo-Barreto (2007, 2008). However, these models are not well-validated by empirical human studies (because they contain too many interdependent variables to allow the assessment of each separately), and too little is known of the dynamics of transmitter recovery and/or the nonlinearities in the process at present to securely assign time constants to the astrocytic component relative to the hemodynamic component of the metabolic coupling. We will therefore treat the entire chain from the metabolic demand to the magnetic resonance signal in the traditional fashion, as a unitary linear kernel. (As stated above, this kernel is often termed the hemodynamic response function HRF, but in view of its likely substantial astrocyte contribution, we give it the more general term of the metabolic response kernel, MRK). As

more information becomes available, nonlinear aspects of the metabolic coupling may readily be incorporated into the analysis.

10. Characteristics of the model

We may now evaluate the response to these two options for the nonlinearity of the metabolic demand through the biophysical chain of the metabolic processes to the measured BOLD signal (Table 1). The main goal is to estimate the properties of the neural signal processing, and it will be seen that there is sufficient information to provide a rich analysis of these properties, and to account for the empirical nonlinearities of the BOLD signal, as long as the metabolic supply chain conforms to the linearity assumption.

For this demonstration, we assume an MRK of the form:

$$\mathrm{MRK}(t) = t^{k_m - 1} \cdot \mathrm{e}^{-t/\tau_m}, \tag{3}$$

where k_m and $\boldsymbol{\tau}_m$ are the metabolic waveform parameters.

Thus, the forward model structure was the convolution of the nonlinear neural signal to the boxcar stimulus of variable duration and the MRK to generate the model BOLD response:

$$\mathrm{BOLD}(t) = \mathrm{N}(t) \otimes \mathrm{MRK}(t) \tag{4}$$

The results of the simulation study are shown in Fig. 6, where the capital letters code for different sets parameter values and the lower-case letters code for different stages of the simulation output. For a given row, column (a) shows the assumed metabolic demand, column (b) plots the BOLD responses over duration for the half-wave-rectified model of metabolic demand, column (c) plots the BOLD responses over duration for a fully-rectified model, and column (d) plots the duration summation curves for the peak amplitudes. The other parameter values were, a=0.8, $\boldsymbol{\tau}_n$=40 *ms*, k_n=4, Δt=0, =∞, λ=-1. In the first three panels in each row, the successive curves represent responses to a doubling of the stimulus duration relative to that for the previous curve, while the fourth panel plots the peak amplitude of each of the successive curves in columns (b) and (c), together with the linear prediction for a purely sustained response (green curves; BOLD waveform not shown).

The development of the peak amplitude in the BOLD temporal summation series for a purely transient neural response is shown as the blue summation curve in Fig. 6A(d). The critical point of this plot is that the asymptotic corner of the summation curve occurs at 40 *ms*, which is the value of the time constant assumed for the neural signal in this example. Thus, the form of the BOLD amplitude summation series (Fig. 6, column (d)) provides a direct empirical estimate of the time constant of neural integration down to the millisecond range. There is no limit in principle to the temporal resolution that can be achieved by this technology since it is estimated from the amplitude variation of the BOLD signal as a function of stimulus duration, not from its temporal aspects.

The reduction in peak amplitude is captured in the red temporal summation curve of Fig. 6A(d), showing a reduction by a factor of two for long durations. (The green curve in Fig. 6A(d) represents the values expected for fully proportional linear summation of the energy in the stimulus pulse; it is an accelerating curve due to the logarithmic abscissa.) The case of the full-rectification model of the metabolic demand is shown in Fig. 6A(c). Here the second neural response peak (i.e., that from the stimulus offset) plays a key role in varying the BOLD waveform, which first extends in time and then shows a two-peaked structure with reduced amplitude for long-duration stimuli.

Fig. 6B(a) shows the half-and fully-rectified version of the metabolic demand to the same pulse duration series, where the neural inhibition is now assumed to be reduced in energy by 1.5% relative to the excitation. This small imbalance is magnified by the convolution with the sustained stimulus, and thus it results in a sustained component that is 12% of the amplitude of the initial transient (blue curve in Fig. 6B(a)) and then into an almost fully sustained set of BOLD response functions (Fig. 6B(b), compared to A(b)). Thus, the form of the BOLD response functions can be strongly diagnostic of even slight variations in the properties of brief neural signals. Moreover, the nature of the metabolic demand function (half-or fully-rectified) has a big impact on the form of the BOLD response, determining whether or not an offset peak occurs at the tail of the responses even when they are sustained (Fig. 6B(b vs. c)). Such a peak has been reported in some studies (d'Avossa et al, 2003; Fox et al., 2005) but is not always evident. Thus it remains an empirical question to what extent rectification is representative of BOLD waveforms; intermediate forms of the rectification model are required to capture the empirical properties in detail.

Note that the amplitude series in Fig. 6B(b) and (c) show bands of denser packing of the functions, where the amplitude changes are not spaced in proportion to the doublings of stimulus duration. Viewed in terms of the sequence of BOLD waveforms in Fig. 6B(b) and (c), the regions of dense packing form an intermediate "shelf" or partial asymptote in the peak amplitude summation plots of Fig. 6B(d). It is again evident that the onset of this intermediate shelf in the summation curve corresponds to the 40 *ms* integration time of the underlying neural signal, while the second asymptote at higher amplitude corresponds to the ~5 *s* integration time of the MRK (HRF). Accurate measurement of such summation functions can therefore provide discriminative characteristics that, when interpreted through the nonlinear model structure, can provide estimates of both the neural *and* the metabolic time constants in the neural-to-BOLD signal chain.

This point is emphasized by the response set in row C of Fig. 6, which probes the effect of varying the time constant of the neural transient. The key difference from the parameters used in row B of Fig. 6, is that the neural time constant for row C was doubled from 40 *ms* to 80 *ms* (and the excitation/inhibition imbalance was also increased to 7% to maintain the same form of offset peak). It is evident that (i) the summation curve (Fig. 6C(d)) takes a measurably different form, and that (ii) the accuracy of estimation of the neural time constant is limited not by the BOLD time constant but by the variability of the BOLD amplitude measures. For example, this analysis shows that the neural time constant is estimable to within about 0.1 log units if the BOLD response functions can be measured to an achievable accuracy of about 10%.

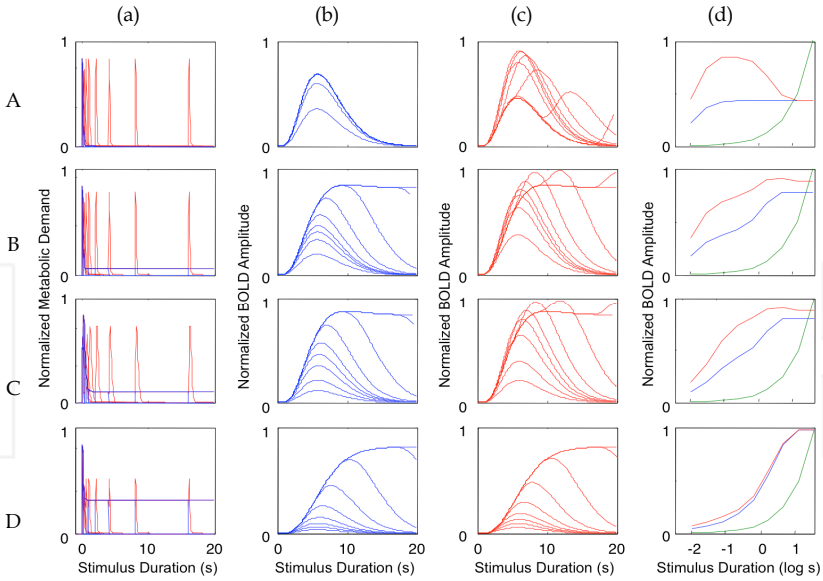

Figure 6. Simulations of four different types of BOLD response for monophasic metabolic demand signals (and a monophasic MRK). The rows represent the results for A: metabolic demands with a purely transient time course, B-D: responses with a mixed transient and sustained time course, with the sustained component at respectively 12%, 18% and 50% of the amplitude of the transient component (based on different ratios of neural excitation/inhibition). For each type, column (a) shows the assumed metabolic demand, column (b) plots the BOLD responses over duration for the half-wave-rectified model of metabolic demand, column (c) plots the BOLD responses over duration for a fully-rectified model, and column (d) plots the duration summation curves for peak BOLD response amplitude (blue curve: half-rectified model, red curve: fully-rectified model, green curve: pure linear summation). Note the use of the logarithmic abscissa in column (d) to focus the analysis on the brief duration regime. The progression of the model BOLD responses with stimulus duration and the form of the summation curves are diagnostic of both the relative weighting between the sustained and transient components of the neural signal and the form of rectification feeding the metabolic demand.

The final case of duration summation analyses (Fig. 6D(d)) shows the NDFO predictions of increasing the excitation/inhibition imbalance of the neural response to 20% (increasing the sustained level to 50% of the transient amplitude), illustrative of a system that is predominantly sustained in nature. Under these conditions, the impact of the initial transient becomes essentially negligible, and the summation curves (Fig. 6D(d)) become indistinguishable from proportional summation (i.e., they run parallel to the green curve). This manipulation illustrates that the power of the NDFO analysis depends on the neural processing being predominantly transient, and that the properties of the underlying neural mechanisms would not be accessible to this form of analysis in predominantly sustained systems. Luckily, however, the well-established deviation from proportionality for short-duration stimuli (Birn, Saad & Bandettini, 2001) implies that neural signals are, in practice, predominantly transient and are therefore typically amenable to this form of NDFO analysis (Tyler & Likova, 2009, 2011).

11. Multicomponent analysis of the neural signal contributions

In a study of the fMRI components, Tyler et al. (2008) recorded the sets of BOLD waveforms generated by fields of dynamic noise-patterned stimuli in different eccentricity bands across the visual field (Fig. 7A). Responses were analyzed in a corresponding series of retinotopic

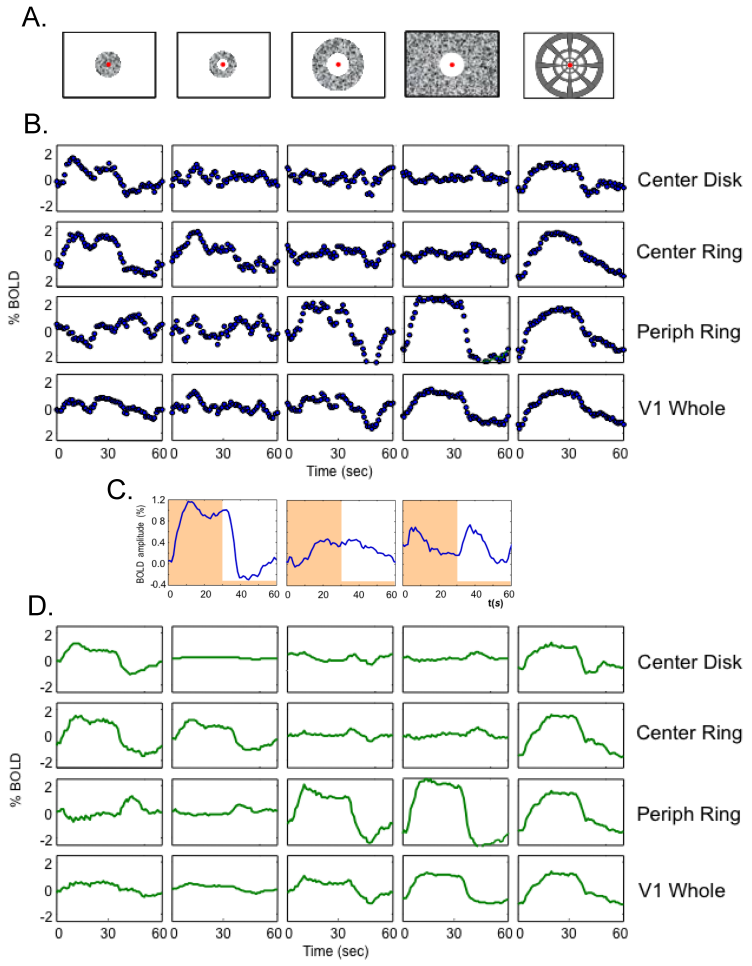

Figure 7. A. The set of spatial noise stimuli used in the ICA study. The eccentricities of the three defining radii were 1, 3.5 and 7 for the inner hole, the central disk/inner edge, and the outer edge of the peripheral annulus, respectively. B. Variety of BOLD response waveforms obtained for the four matching ROIs designated at right, in a block design of 30s on/30s off for fields of dynamic visual noise. Many of these responses differ substantially from the form of the typical GLM and from each other, particularly for the responses to the fine spatial structure of the scaled grid in the last column. C. The first three ICs derived from the individual voxel analysis throughout V1. D. Weighted sums of the three ICs optimized to account for each of the BOLD waveforms in the upper plots.

regions of interest (ROIs) defined bilaterally on V1 (on the basis of separate retinotopic mapping stimuli). An example of the variety of cortical responses obtained in retinotopic area V1 combined for both hemispheres is shown in Fig. 7B. The responses vary not just in amplitude across stimulus types, but markedly in the waveform of the responses. For some stimuli, the same area may show a classic boxcar response, a double-peaked on-response, a rounded on-response, or a negative on-response. Since there is no reason to expect substantial differences in the BOLD dynamics in different regions of V1 (see Section 8), variation in response from single cortical regions is difficult to explain by variations in the hemodynamics of the blood oxygenation and implies a neural origin of the differences in BOLD response profiles for the different stimulus types. This inference is particularly strong for the radial grid stimulus, which spans the areas of all the other stimuli, and should therefore be expected to match their average waveform.

The BOLD waveforms of Fig. 7B were analyzed by an independent components (IC) model of the variations in temporal response over all the voxels in V1 to determine the minimum number of temporal components that could account for the data (see Tyler et al., 2008, for details). There was a relative drop in the component weighting beyond that for the first three ICs, implying that the three ICs in Fig. 7C represented the responses of three stable response populations contributing to the overall response pattern, while further ICs represented noise or spatially inhomogeneous aspects of the responses. Fig. 7D shows that the IC model does a good job of capturing this variety of response patterns with the weighted sums of the first three ICs, which account for 92% of the variance in the empirical responses.

These results support once again the important point that BOLD waveforms within particular cortical regions may vary dramatically as a function of the stimulus type, despite the fact that the metabolic/hemodynamic response kernels are expected to remain invariant across the region (see Section 8). Consequently, differences in the BOLD response waveform in this V1 ROI are best interpreted as being due to differences in the neural response waveforms to the different stimuli.

12. Neural response component analysis

Under the assumption that the neural components signals in local regions of cortex are well-approximated by sets of delayed gamma-function components, the components should be resolvable into the same component vocabulary, i.e., delayed gamma functions, that approximate known components of the neural dynamics. Note that the mean BOLD waveform has an unusual double-peaked form, rising at the stimulus offset as well as the onset, and that the off-period, by contrast, has a single-peaked form (Fig. 8A). This overall waveform could not be fitted by any single gamma-function convolution. As a result, the novel analysis developed by Tyler et al. (2008) consisted of (i) the simultaneous optimization of the set of neural components (ii) through the Buxton-Friston balloon model, together with (iii) the rise and fall parameters of MRK. The results of the analysis for the V1 dataset of Fig. 7 were sufficient to account for >95% of the BOLD variance overall (Fig. 8), which drew from the whole of the V1

ROI. The IC analysis breaks the overall waveform down into independent components that are differentially expressed across the individual voxels of V1 (capturing the temporal waveform variety of the specified ROIs within V1 shown in Fig. 7). The first IC (Fig. 8B) shows the classic form, following the predicted linear form for the balloon model (blue curve) accurately. The optimized neural signal for the fit to the first IC (red curve) is close to a boxcar function. The second IC has a very different form (Fig. 8C), with an early negative BOLD peak at the stimulus onset and no corresponding rise at stimulus offset. The estimated neural signal has the same nonlinear characteristic, although the negative peak has a much shorter latency. The third IC is even more non-linear (Fig. 8D), with approximately equal positive BOLD peaks at stimulus onset and offset, as though responding to the stimulus events through a full-wave rectified nonlinearity. Again, the estimated neural signal is a reflection of the same form of nonlinearity with much shorter latency.

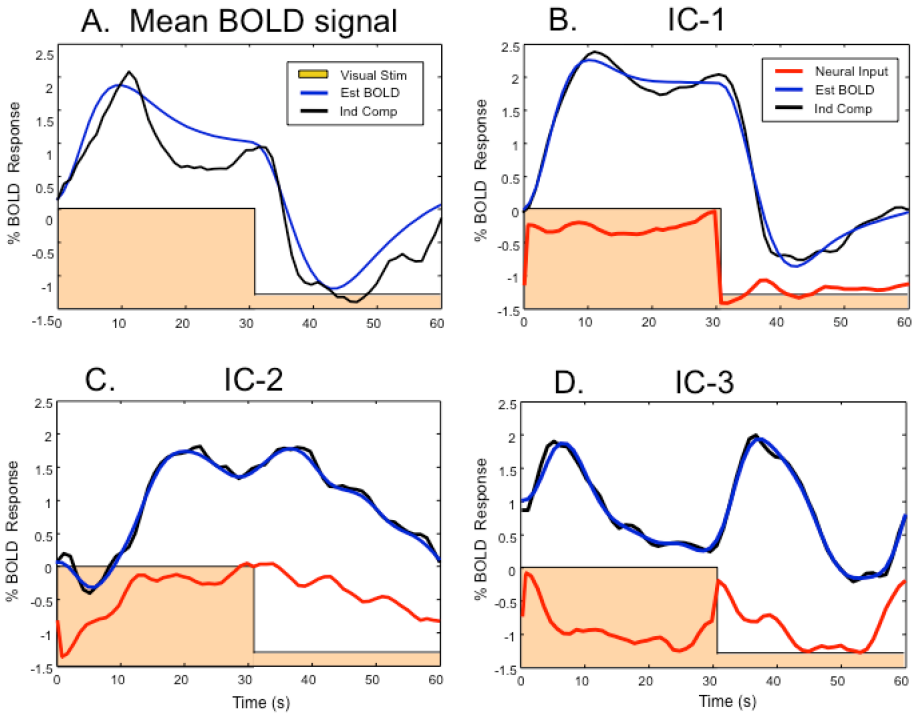

Figure 8. Dynamic forward modeling optimization of multiple neural components to the BOLD responses for cortical area V1. A. The average BOLD waveforms across the whole of V1 (black) with the fit for the Buxton-Friston balloon model. B-D. Plots of the optimized neural response estimates (red) for each of the three BOLD ICs (black curves), with the fit of the balloon model output to each component (blue curves). Note the wide variety of temporal properties of the neural signals (red curves) selected by the optimization to account for the waveform differences of the BOLD components. (from Tyler, Kontsevich & Ferree, 2008, with permission.)

13. Discussion

This chapter illustrates how a range of neural waveform parameters can be estimated from the observed variety of BOLD waveform effects, despite the wide difference in their time courses from seconds to milliseconds. Our novel analysis of the neural signal underlying the BOLD response waveforms constitutes a "temporal microscope" for the neural signals in the cortex generating the recorded BOLD waveforms. This form of estimation depends on the assumption, reviewed in Section 8, that temporal variations in the BOLD waveforms are due to changes in the underlying *neural* activity rather than the parameters of the hemodynamic responses to the metabolic demand. Generally, studies that analyze BOLD temporal variations have operated on the basis that they are likely to reflect variations in the hemodynamics of the blood control mechanisms. We argue that this is an implausible assumption in general because decades of neurophysiological studies have shown that different response networks contain neurons with a wide variety of different temporal responses to the same stimuli (fast transient, slow transient, sustained, inhibitory, etc.), and that there are substantial variations in the response dynamics across stimulus conditions in any given local cortical response. Conversely, there is little convincing evidence for differences in the hemodynamic response parameters among different cortical areas, since even studies finding variations in hemodynamic waveforms across the cortex have typically used stimulus-driven activation of the neurons to mediate the metabolic demand, and therefore are not able to dissociate the neural variation from the hemodynamic variation.

For clarity in this enterprise, we have assumed that the metabolic response is both linear and monophasic, have illustrated (Fig. 2) a variety of BOLD response properties that could arise from such neural nonlinearities. We are not claiming to have proven that these BOLD properties are entirely determined at the neural level but that, conversely, claims that they are purely properties of the vascular hemodynamics *per se* must be considered suspect until they are replicated in paradigms that remove the neural component of the system. One such approach would be stimulation of the hemodynamic response by direct infusion of nitric oxide (NO) in the vicinity of the blood vessels. This experiment has apparently not been attempted, although suppression of the nitric oxide with application of the NO synthase inhibitors (Burke & Bührle, 2006; Kitaura et al., 2007) has been shown to completely abolish the BOLD response while only marginally affecting the local field potentials, establishing a critical role for NO in neurovascular coupling.

Consequently, the present chapter has tried to redress the balance by considering how the BOLD signal variation could arise from the effects of plausible nonlinearities in the variety of neural population responses to standard types of stimulus presentation. In particular, the nonlinearity of the transient responses at the neural response offset could be positive (rectifying) rather than negative (linear), and in either case this could show reduced amplitude relative to the onset response (adaptive gain control). In this example of the NDFO analysis, the offset responses are much smaller than the positive responses at stimulus onset. This analysis demonstrates that estimation of the neural response dynamics for each stimulus type is well

within the capability of one session of normal fMRI methodology, but requires the appropriate combination of experimental design and theoretical analysis.

The final analysis of Figs. 7 and 8 describes a search for independent response components in occipital area V1 in a relatively standard block-design stimulus paradigm incorporating a variety of spatial stimulation patterns (Tyler, Kontsevich & Ferree, 2008). This analysis began with the assumptions of the balloon model of vascular hemodynamics, but the results revealed dramatic response nonlinearities, as expressed through an independent components analysis of the response variation across the cortical space of V1 and the stimulus variety. Also, not shown in these figures is the way in which these nonlinear components were expressed across the cortical space, which implied that they were functional response components dependent on the relation between the stimulated and unstimulated regions, rather than on structural differences in the vasculature. Thus, although the balloon model was an initial assumption of the analysis, the results confront the issue of whether its assumed nonlinearities are hemodynamic or neural in origin, given this result of profound functional nonlinearities across the space of V1, which are *a fortiori* of neural origin.

The neural signal components estimated to underlie the three primary ICs of Fig. 8 provide insight into the nature of the neural nonlinearities involved. Unlike the stable (linear) boxcar of Fig. 8B, the other two estimated neural waveforms have initial transients, one representing a half-wave rectification and the other a full-wave rectification. In neural signals, such nonlinearities are extremely well known, and indeed are characteristic of neural processing in general (as "on" responses and "on-off" responses, respectively), while it is difficult to conceptualize how such nonlinearities could arise from the hemodynamics *per se*. A more global mechanism that has such transient, rectified character is the top-down attentional mechanism, by which activation may be enhanced in regions of recent stimulus change in a transient fashion (Liu, Pestilli & Carrasco, 2005). Such top-down mechanisms are again neural rather than hemodynamic in nature. The present data make it clear that a rich array of such neural nonlinearities should be expected to contribute to the fMRI signals recorded from the brain, although they are not sufficient to distinguish between the top-down and bottom-up hypotheses for the neural signals underlying the nonlinear BOLD responses revealed by this paradigm. Nevertheless, further enhancements to the analysis should allow the technique to answer many questions about the neural mechanisms involved in the BOLD response dynamics.

As knowledge evolves, more complex models of the neural/metabolic coupling and the hemodynamic response could be easily incorporated in our model. These all represent Bayesian information that, if well established, can be used to refine the model structure and enhance the fitting process when available. However, our reading of the literature is that a) the first-order specification of the neural/BOLD coupling is well approximated by a linear response kernel, and b) that estimates of the second-order effects are contaminated by the assumption that the modeling has been purely hemodynamic, and has not taken into account potential and actual nonlinearities in the *neural* signals at the time scale of the BOLD signal. Thus, despite the best efforts of the proponents of elaborated hemodynamic modeling, there is no secure information about the nonlinearities of this process for the human brain in vivo.

Finally, we note that a technique with a philosophy similar to the present approach has been successfully applied to the estimation of net *spatial* receptive field structure of small cortical regions by Dumoulin & Wandell (2008), although they used a linear rather than nonlinear model of the sequence of processes. Their spatial estimates were based on a model of the temporal signal to be expected as a stimulus swept across each defined point on the retina. Like us, they take the temporal stimulus waveform, convolve it with a spatiotemporal model of the response of the underlying neural population and then with a model of the metabolic response function to provide a basic forward model of the temporal BOLD response that is optimized to the measured BOLD response at each cortical location. Our approach takes the temporal analysis several steps further towards biological plausibility, and focuses on the temporal rather than spatial aspect of the neural population response.

14. Conclusion

The conceptualizations and techniques introduced in this chapter provide an analytic capability for resolving the timing and neural signal estimation underlying the BOLD waveforms recorded throughout the cortex. Any such attempt must be based on a model of the known neural dynamics of the neural populations underlying the BOLD metabolic signal generation, which may be progressively refined as more information becomes available, both about the underlying neural response characteristics and about the subsequent metabolic cascade. Given adequate signal/noise ratio, the present analysis shows that it is possible to develop approaches that overcome the temporal limitations of BOLD signal and are able to reveal relevant properties of the underlying neural signals. In combination, these approaches represent a notable advance in the capabilities of the fMRI technology, providing a direct linkage between the live assessment of the functioning brain and the direct neurophysiological recordings in other species, or even in the human brain.

Acknowledgements

Supported by NSF SLC grants 0846229 and 0846230, AFOSR grant #FA9550-09-1-0678 and CDMRP grant XWH-11-2-0066.

Author details

Christopher W. Tyler and Lora T. Likova

Smith-Kettlewell Eye Research Institute, USA

References

[1] Aguirre GK, Zarahn E, D'Esposito M (1998) The variability of human BOLD hemodynamic responses. NeuroImage, 8:360-9.

[2] Barth M, Norris DG (2007) Very high-resolution three-dimensional functional MRI of the human visual cortex with elimination of large venous vessels. NMR Biomed, 20:477-84.

[3] Bezzi P, Carmignoto G, Pasti L, Vesce S, Rossi D, Rizzini BL, Pozzan T, Volterra A (1998) Prostaglandins stimulate calcium-dependent glutamate release in astrocytes. Nature, 391:281–285.

[4] Birn RM, Saad ZS, Bandettini PA (2001) Spatial heterogeneity of the nonlinear dynamics in the FMRI BOLD response. NeuroImage, 14:817-26.

[5] Boynton GM, Engel SA, Glover GH, Heeger DJ (1996) Linear systems analysis of functional magnetic resonance imaging in human V1. J Neurosci, 16:4207-21.

[6] Buracas GT, Boynton GM (2002) Efficient design of event-related fMRI experiments using m-sequences. NeuroImage, 16: 801-13.

[7] Burke M, Bührle C (2006) BOLD response during uncoupling of neuronal activity and CBF. NeuroImage, 32:1-8.

[8] Buxton RB (2001) The elusive initial dip. NeuroImage 16:953-8.

[9] Buxton RB, Uluda K, Dubowitz DJ, Liu TT (2004) Modeling the hemodynamic response to brain activation. Neuroimage, 23 Suppl 1:S220-33.

[10] Celebi S, Principe JC (1995) Parametric least squares approximation using gamma bases. IEEE Trans Signal Proc, 43:781–784

[11] Chatton JY, Pellerin L, Magistretti PJ (2003) GABA uptake into astrocytes is not associated with significant metabolic cost: implications for brain imaging of inhibitory transmission. Proc Natl Acad Sci U S A, 100:12456-61.

[12] Chen S (2006) Signal processing. Euro Trans Telecomm, 17:99-110.

[13] d'Avossa G, Shulman GL, Corbetta M (2003) Identification of cerebral networks by classification of the shape of BOLD responses. J Neurophysiol, 90:360-71.

[14] De Vries B, Principe JC (1992) The Gamma Model: A new neural model for temporal processing. Neural Networks, 5:565-576.

[15] de Zwart JA, Silva AC, van Gelderen P, Kellman P, Fukunaga M, Chu R, Koretsky AP, Frank JA, Duyn JH (2005) Temporal dynamics of the BOLD fMRI impulse response. Neuroimage, 24:667-77.

[16] Dienel GA, Cruz NF (2008) Imaging brain activation: Simple pictures of complex biology. Ann NY Acad Sci, 1147:139-170.

[17] Dumoulin SO, Wandell BA (2008) Population receptive field estimates in human visual cortex. NeuroImage 39, 647–660.

[18] Duvernoy HM, Bourgouin P (1999) The Human Brain: Surface, Three-Dimensional Sectional Anatomy with MRI, and Blood Supply. Springer: New York.

[19] Efron B, Tibshirani RJ (1993) An Introduction to the Bootstrap. New York: Chapman and Hall.

[20] Farwell VT, Prentice RL (1977) A study of distributional shape in life testing, Technometrics 19:69–75.

[21] Filosa JA, Bonev AD, Nelson MT (2004) Calcium dynamics in cortical astrocytes and arterioles during neurovascular coupling. Circ Res, 95:73-81.

[22] Fox MD, Snyder AZ, Barch DM, Gusnard DA, Raichle ME (2005) Transient BOLD responses at block transitions. NeuroImage, 28:956-66.

[23] Friston KJ (1997) Transients, metastability, and neuronal dynamics. NeuroImage, 5:164-71

[24] Friston KJ, Fletcher P, Josephs O, Holmes A, Rugg MD, Turner R (1998) Event-related fMRI: characterizing differential responses. Neuroimage, 7:30-40.

[25] Friston KJ, Mechelli A, Turner R, Price CJ (2000) Nonlinear responses in fMRI: the Balloon model, Volterra kernels, and other hemodynamics. NeuroImage, 12:466-77.

[26] Gandhi GK, Cruz NF, Ball KK, Dienel GA (2009) Astrocytes are poised for lactate trafficking and release from activated brain and for supply of glucose to neurons. J Neurochem, 11:522-36

[27] Glover G (1999) Deconvolution of impulse response in event-related BOLD fMRI. Neuroimage, 9:416-29.

[28] Grotz T, Zahneisen B, Ella A, Zaitsev M, Hennig J (2009) Fast functional brain imaging using constrained reconstruction based on regularization using arbitrary projections. Magn Reson Med, 62:394-405.

[29] Handwerker DA, Ollinger JM, D'Esposito M (2004) Variation of BOLD hemodynamic responses across subjects and brain regions and their effects on statistical analyses. NeuroImage, 21:1639-51.

[30] Hakim TS, Sugimori K, Camporesi EM, Anderson G (1996) Half-life of nitric oxide in aqueous solutions with and without haemoglobin. Physiol Meas, 17: 267

[31] Hansen KA, David SV, Gallant JL (2004) Parametric reverse correlation reveals spatial linearity of retinotopic human V1 BOLD response. NeuroImage, 23: 233-41

[32] Hegdé J, Van Essen DC (2004) Temporal dynamics of shape analysis in macaque visual area V2. J Neurophysiol, 92:3030-42.

[33] Henson RN, Price CJ, Rugg MD, Turner R, Friston KJ (2002) Detecting latency differences in event-related BOLD responses: application to words versus nonwords and initial versus repeated face presentations. NeuroImage, 15:83-97.

[34] Hoge RD, Franceschini MA, Covolan RJ, Huppert T, Mandeville JB, Boas DA (2005) Simultaneous recording of task-induced changes in blood oxygenation, volume, and flow using diffuse optical imaging and arterial spin-labeling MRI. Neuroimage, 25:701-7.

[35] Hyder F, Rothman DL, Shulman RG (2002) Total neuroenergetics support localized brain activity: implications for the interpretation of fMRI. Proc Natl Acad Sci USA, 99:10771–10776.

[36] Hyder F, Patel AB, Gjedde A, Rothman DL, Behar KL, Shulman RG (2006) Neuronal-glial glucose oxidation and glutamatergic-GABAergic function. J Cereb Blood Flow Metab, 865-77.

[37] Kelly JP, Van Essen DC (1974) Cell structure and function in the visual cortex of the cat. J Physiol, 238:515-47.

[38] Kitaura H, Uozumi N, Tohmi M, Yamazaki M, Sakimura K, Kudoh M, Shimizu T, Shibuki K (2007) Roles of nitric oxide as a vasodilator in neurovascular coupling of mouse somatosensory cortex. Neurosci Res, 59:160-71.

[39] Koehler RC, Gebremedhin D, Harder DR (2006) Role of astrocytes in cerebrovascular regulation J Appl Physiol, 100:307–317.

[40] Likova LT, Tyler CW (2007) Instantaneous stimulus paradigm: Cortical network and dynamics of figure-ground organization. Human Vision and Electronic Imaging, SPIE 6492, 64921E.

[41] Likova LT, Tyler CW (2008) Occipital network for figure/ground organization. Exp Brain Res, 189:257-67.

[42] Lingnau A, Ashida H, Wall MB, Smith AT (2009) Speed encoding in human visual cortex revealed by fMRI adaptation. J Vis, 9:3.1-14.

[43] Liu T, Pestilli F, Carrasco M (2005) Transient attention enhances perceptual performance and fMRI response in human visual cortex. Neuron, 45: 469-477.

[44] Logothetis NK (2003) The underpinnings of the BOLD functional magnetic resonance imaging signal. J Neurosci, 23:3963-3971

[45] Logothetis NK, Wandell BA (2004) Interpreting the BOLD signal. Ann Rev Physiol, 66:735-769

[46] Magistretti PJ, Pellerin L (1999) Cellular mechanisms of brain energy metabolism and their relevance to functional brain imaging. Philos Trans R Soc Lond B Biol Sci, 354:1155-63.

[47] Magistretti PJ (2009) Role of glutamate in neuron-glia metabolic coupling. Am J Clin Nutr, 90:875S-880S.

[48] Margaria R, Edwards HT, Dill DB (1933) The possible mechanism of contracting and paying the oxygen debt and the role of lactic acid in muscular contraction. Am J Physiol, 106:689–714.

[49] Martindale J, Mayhew J, Berwick J, Jones M, Martin C, Johnston D, Redgrave P, Zheng Y (2003) The hemodynamic impulse response to a single neural event. J Cereb Blood Flow Metab, 23:546–555;

[50] Mechelli A, Price CJ, Friston KJ (2001) Nonlinear coupling between evoked rCBF and BOLD signals: a simulation study of hemodynamic responses. NeuroImage, 14:862-72.

[51] Menon RS, Luknowsky DC, Gati JS (1998) Mental chronometry using latency-resolved functional MRI. Proc Natl Acad Sci USA, 95:10902-7.

[52] Metea MR, Newman EA (2006) Glial cells dilate and constrict blood vessels: a mechanism of neurovascular coupling. J Neurosci, 26:2862-70.

[53] Meyer EP, Ulmann-Schuler A, Staufenbiel M, Krucker T (2008) Altered morphology and 3D architecture of brain vasculature in a mouse model for Alzheimer's disease. Proc Natl Acad Sci USA, 105:3587-92.

[54] Mulligan SJ, MacVicar BA (2004) Calcium transients in astrocyte endfeet cause cerebrovascular constrictions. Nature, 431:195-9.

[55] Prentice RL (1974) A log-gamma model and its maximum likelihood estimation. Biometrika 61:539–544.

[56] Shank R, Aprison M (1979) Biochemical aspects of the neurotransmitter function of glutamate. In: Glutamic Acid: Advances in Biochemistry and Physiology. Raven: New York, 139–150.

[57] Shao XM (1997) Parametric survival analysis for gating kinetics of single potassium channels. Brain Res, 770: 96-104

[58] Shmuel A, Augath M, Oeltermann A, Logothetis NK (2006) Negative functional MRI response correlates with decreases in neuronal activity in monkey visual area V1. Nature Neurosci, 9: 569-577.

[59] Shulman RG, Rothman DL (1998) Interpreting functional imaging studies in terms of neurotransmitter cycling. Proc Natl Acad Sci USA 95: 11993–11998.

[60] Sibson NR, Dhankhar A Mason GF, Rothman DL, Behar KL, Shulman RG (1998) Stoichiometric coupling of brain glucose metabolism and glutamatergic neuronal activity. Proc Natl Acad Sci, USA 95:316–321.

[61] Sibson NR, Mason GF, Shen J, Cline GW, Herskovits AZ, Wall JE, Rothma DL, Shulman RG (2001) In vivo 13C NMR measurement of neurotransmitter glutamate cycling, anaplerosis, TCA cycle flux in rat brain during [2–13C]glucose infusion. J Neurochem, 76:975–989.

[62] Silva AC, Koretsky AP (2002) Laminar specificity of functional MRI onset times during somatosensory stimulation in rat. Proc Natl Acad Sci USA, 99:15182-7.

[63] Smith AJ, Blumenfeld H, Behar KL, Rothman DL, Shulman RG, Hyder F (2002) Cerebral energetics and spiking frequency: the neurophysiological basis of fMRI. Proc Natl Acad Sci USA, 99:10765–10770.

[64] Sotero RC, Trujillo-Barreto NJ (2007) Modelling the role of excitatory and inhibitory neuronal activity in the generation of the BOLD signal. NeuroImage, 35:149-65.

[65] Sotero RC, Trujillo-Barreto NJ (2008) Biophysical model for integrating neuronal activity, EEG, fMRI and metabolism. NeuroImage, 39:290-309.

[66] Stacy EW (1962) A generalization of the gamma distribution. Ann Math Stat, 33, 1187–1192.

[67] Tian P, Teng IC, May LD, Kurz R, Lu K, Scadeng M, Hillman EM, De Crespigny AJ, D'Arceuil HE, Mandeville JB, Marota JJ, Rosen BR, Liu TT, Boas DA, Buxton RB, Dale AM, Devor A (2010) Cortical depth-specific microvascular dilation underlies laminar differences in blood oxygenation level-dependent functional MRI signal. Proc Natl Acad Sci USA, 107:15246-51.

[68] Thompson JK, Peterson MR, Freeman RD (2003) Single-neuron activity and tissue oxygenation in the cerebral cortex. Science, 299:1070-2.

[69] Thompson JK, Peterson MR, Freeman RD (2004) High-resolution neurometabolic coupling revealed by focal activation of visual neurons. Nat Neurosci, 7:919-20.

[70] Thompson JK, Peterson MR, Freeman RD (2005) Separate spatial scales determine neural activity-dependent changes in tissue oxygen within central visual pathways. J Neurosci, 25:9046-58.

[71] Tyler CW, Kontsevich LL, Ferree TC (2008) Independent components in stimulus-related BOLD signals and estimation of the underlying neural responses. Brain Res, 1229: 72-89.

[72] Tyler CW, Likova LT (2009) Neural signal estimation through time-resolved functional imaging. In: Brain Mapping Research Progress, Girard IC, Andre JS (Ed), Nova Scientific Publishers, 1-31.

[73] Tyler CW, Likova LT (2011) Estimating neural signal dynamics in the human brain. Frontiers Syst Neurosci, 5:33.

[74] Wang Y, Floor E (1994) Dynamic storage of glutamate in rat brain synaptic vesicles. Neurosci Lett, 180:175–178.

INTECH