THE SPATIAL, TEMPORAL, AND INTERPRETIVE LIMITS OF FUNCTIONAL MRI

PETER A. BANDETTINI

Since the inception of functional magnetic resonance imaging (fMRI) in 1991, an explosive growth in the number of users has been accompanied by steady widening of its range of applications. A recent search of the National Library of Medicine database for articles with fMRI or BOLD (blood oxygenation-dependent) in the title revealed more than 1,000 citations. Improvements continue in pulse sequence design, data processing, data interpretation, and the tailoring of cognitive paradigms to the unique advantages and limits of the technique. This chapter describes the receding limits of fMRI. Specifically, the limits of spatial resolution, temporal resolution, interpretability, and implementation are discussed. The goal is to give the reader a perspective of the evolution of fMRI in the past 9 years and a sense of excitement regarding its ultimate potential.

A user of fMRI primarily is interested in extracting at least one of three types of neuronal information: where neuronal activity is happening, when it is happening, and the degree to which it is happening. To extract this information optimally, an understanding of some of the more esoteric details is necessary, which are presented in this chapter. First, the basics of fMRI contrast are discussed. Second, the key of fMRI interpretation, the neuronal–hemo-
dynamic transfer function, is described. Third, an overview of methods by which neuronal activation is played out in fMRI subjects and subsequently measured is provided. In this section, the popular technique of event-related fMRI (ER-fMRI) is described in detail, along with emerging methods of neuronal information extraction. Fourth, the issues of temporal and spatial resolution are discussed. Fifth, the limits of interpretation and the potential for further neuronal–hemodynamic information extraction are discussed. Lastly, some implementation limits are finally given as a practical guideline.

CONTRAST IN fMRI

Several types of physiologic information can be mapped with fMRI. This information includes baseline cerebral blood volume (1–3), changes in blood volume (4), baseline and changes in cerebral perfusion (5–10), and changes in blood oxygenation (11–17). Recent advances in fMRI pulse sequence and experimental manipulation have allowed quantitative measures of cerebral metabolic rate of oxygen (CMRO₂) changes and dynamic, noninvasive measures of blood volume with activation to be extracted from fMRI data (18–20).

Blood Volume

In the late 1980s, the use of rapid MRI allowed tracking of transient signal intensity changes over time. One application of this utility was to follow the T2*- or T2-weighted signal intensity as a bolus of an intravascular paramagnetic contrast agent passed through the tissue of interest (2). As it passed through, susceptibility-related dephasing increased then decreased as the bolus washed out. The area under these signal attenuation curves is proportional to the relative blood volume. In 1990, Belliveau and colleagues (4) took this technique one step further and mapped blood volume during rest and during activation. The first maps of brain activation obtained with fMRI were demonstrated with this technique. As soon as the technique was demonstrated, it was rendered obsolete (for brain activation imaging) by the use of an endogenous and oxygen-sensitive contrast agent—hemoglobin.

Blood Oxygenation

As early as the 1930s, it was known that hemoglobin is paramagnetic and deoxyhemoglobin is diamagnetic (21). In 1982, it was discovered that changes in blood oxygenation change the T2 of blood, but it was not until 1989 that
CH6 Contrast in the Detection of Neuronal Activity

Cerebral Tissue Activation

Local Vasodilation

Increase in Cerebral Blood Flow and Volume

Oxygen Delivery Exceeds Metabolic Need

Increase in Capillary and Venous Blood Oxygenation

Decrease in Deoxy-hemoglobin

Deoxy-hemoglobin: paramagnetic

Ox-hemoglobin: diamagnetic

Decrease in susceptibility-related intravoxel dephasing

Increase in T2 and T2*

Local Signal Increase in T2 and T2* – weighted sequences

FIGURE 26.1. The cascade of hemodynamic and magnetic resonance imaging events that occur following brain activation.

this knowledge was used to image in vivo changes in blood oxygenation (22). Blood oxygenation-dependent contrast, coined BOLD contrast by Ogawa et al. (23), was used to image the activated brain for the first time in 1991. Interestingly, Ogawa et al. predicted its utility for functional brain imaging; however, they predicted a signal decrease rather than a signal increase, as implied by some earlier positron emission tomography (PET) results by Fox and Raichle (24) suggesting that the oxygen extraction fraction decreased during activation. The first results of the use of BOLD contrast were published in 1992 (13,15,23). Because of its sensitivity and ease of implementation, the contrast used to observe susceptibility changes with changes in blood oxygenation is the most commonly used functional brain imaging contrast, and this is the technique primarily discussed in this chapter. The cascade of events that follow brain activation and lead to BOLD signal changes is shown in Fig. 26.1.

Blood Perfusion

An array of new techniques exist for mapping cerebral blood perfusion in humans. Arterial spin labeling-based perfusion mapping MRI techniques are similar to those applied in other modalities, such as PET and single-photon emission computed tomography (SPECT); in-flowing blood is tagged and then allowed to flow into the imaging plane. The radiofrequency (RF) tagging pulse is usually a 180-degree pulse that “inverts” the magnetization.

Generally, these techniques can be subdivided into those that use continuous arterial spin labeling, which involves continuously inverting blood flowing into the slice, and those that use pulsed arterial spin labeling, which periodically inverts a block of arterial blood and measures the arrival of that blood into the imaging slice. Examples of these techniques are echo-planar imaging with signal targeting and alternating RF (EPISTAR) (25) and flow-sensitive alternating inversion recovery (FAIR) (10,26). Recently, a pulsed arterial spin-labeling technique known as quantitative imaging of perfusion using a single subtraction (QUIPSS) has been introduced (27).

Hemodynamic Specificity

With each of the above-mentioned techniques for imaging volume, oxygenation, and perfusion changes, the precise type of observable cerebrovascular information can be more finely delineated. Although this information is typically more than the cognitive neuroscientist requires, it is useful to give an abbreviated summary of how specific MRI can actually be. Regarding susceptibility contrast imaging, spin-echo sequences are more sensitive to small susceptibility compartments (capillaries and red blood cells), and gradient-echo sequences are sensitive to susceptibility compartments of all sizes (28–31). Outer-volume RF saturation removes in-flowing spins (32), thereby reducing non-susceptibility-related inflow changes when short-repetition time (with high flip angle) sequences are used. Diffusion weighting or “velocity nulling,” involving the use of $b > 50 s/mm^2$, reduces the intravascular signal (33), thereby reducing, but not eliminating, large-vessel effects (intravascular effects are removed but extravascular effects remain) in gradient-echo fMRI and all large-vessel effects in spin-echo fMRI. Performing BOLD contrast fMRI at high field strengths has the same effect as diffusion weighting in the context of susceptibility-based contrast because the $T2^*$ and $T2$ of venous blood becomes increasingly shorter than the $T2^*$ and $T2$ of gray matter as field strength increases; therefore less signal arises from venous blood at higher field strengths. (34). Figure 26.2 is a schematic diagram summarizing the pulse sequence selectivity of the specific aspects of the vasculature.

Cerebral Metabolic Rate of Oxygen

Recently, advances in mapping activation-induced changes in the CMRO2 with fMRI have been developed (18,20, 35–37). The basis for such measurement is that BOLD and perfusion contrast can be explained by the combination of a handful of parameters. The key, then, is either to constrain the contrast or manipulate the physiologic state such that the number of parameters reduces to one or two. Normalization by means of a hypercapnia has evolved as a method for mapping changes in CMRO2 (18). The basic idea is that when the brain is activated, increases in flow, volume, and oxygenation are accompanied by an increase in CMRO2. When a subject at rest is undergoing a hypercapnic stress (5% CO2), the cerebral flow, volume, and oxygenation increase without an accompanying increase in CMRO2; there-
Hemodynamic Specificity

**FIGURE 26.2.** The vascular tree, including arteries (left) and arterioles, capillaries, and veins (right). If the inside of the vessel drawing is filling in, the signal has an intravascular contribution. Arterial spin labeling (ASL) is differentially sensitive to the arterial–capillary region of the vasculature, depending on the inversion time (TI) used and whether or not velocity nulling (otherwise called diffusion weighting) gradients are used. A small amount of velocity nulling and a TI of about 1 s make ASL techniques selectively sensitive to capillaries. Susceptibility-based techniques, including gradient-echo and spin-echo, are also differentially sensitive to specific aspects of the vasculature. Gradient-echo techniques are sensitive to susceptibility perturbers of all sizes; therefore, they are sensitive to all intravascular and extravascular effects. Spin-echo techniques are sensitive to susceptibility perturbers about the size of a red blood cell or capillary, so that they are sensitive to intravascular effects in vessels of all sizes and to extravascular capillary effects. Velocity nulling makes gradient-echo sequences sensitive to extravascular capillary-to-vein effects, and makes spin-echo sequences selectively sensitive only to capillary effects. See color version of figure.

fore, less oxygen is extracted from the blood stream, so that the blood oxygenation change, relative to the perfusion change, is greater than with brain activation. By comparing the ratio of the (simultaneously measured) perfusion and BOLD signal changes during hypercapnia and during brain activation, CMRO₂ information can be derived.

**HEMODYNAMIC TRANSFER FUNCTION**

The hemodynamic transfer function is referred to here as the combined effect on the fMRI signal change by the spatial and temporal variation in neuronal–vascular coupling, blood volume, blood flow, blood oxygenation, hematocrit, and vascular geometry, among other things. A goal of fMRI method development is to characterize this transfer function completely (i.e., its spatial, temporal, pulse sequence, subject, physiologic, and pharmacologic dependencies), so that more precise inferences can be made about underlying neuronal activation location, magnitude, and timing. The ultimate limits of fMRI depend on this characterization. This goal is particularly relevant in the context of understanding pharmacologic effects on brain function.

After the onset of activation, or rather after the neuronal firing rate has passed an integrated temporal–spatial threshold, either direct neuronal, metabolic, or neurotransmitter-mediated signals reach arteriole sphincters and cause dilation. The time for this initial process to occur is likely to be less than 100 ms. After vessel dilation, the blood flow rate increases by 10% to 200%. The time for blood to travel from arterial sphincters through the capillary bed to pial veins is about 2 to 3 s. This transit time determines how rapidly the blood oxygenation saturation increases in each part of the vascular tree. As shown in Fig. 26.2, depending on the pulse sequence used, different aspects of the hemodynamics are manifested.

**Location**

In resting state, hemoglobin oxygen saturation is about 95% in arteries and 60% in veins. The increase in hemoglobin saturation with activation is largest in veins, changing from about 60% to 90%. Likewise, capillary blood saturation changes from about 80% to 90%. Arterial blood, already saturated, shows no change. This large change in saturation is one reason why the strongest BOLD effect is usually seen in draining veins.

The second reason why the strongest BOLD effect is seen in draining veins is that activation-induced BOLD contrast is highly weighted by blood volume in each voxel. Because capillaries are much smaller than a typical imaging voxel, most voxels, regardless of size, likely contain about 2% to 4% capillary blood volume. In contrast, because the size and spacing of draining veins are on the same scale as most imaging voxels, it is likely that veins dominate the relative blood volume in any voxel that they pass through. Voxels that pial veins pass through can have 100% blood volume, whereas voxels that contain no pial veins may have only 2% blood volume. This stratification in blood volume distribution strongly determines the magnitude of the BOLD signal.

As suggested in Fig. 26.2, different pulse sequence weightings can give different locations of activation. For example, Fig. 26.3 shows the activation in the motor cortex with two different functional MRI contrast weightings collected in the same plane—perfusion and BOLD. Although much overlap is seen, the hot spots vary by as much as 10 mm. The perfusion change map is sensitive primarily to capillary perfusion changes, whereas the BOLD contrast activation map is weighted mostly by veins. A potential worry regarding fMRI location is that venous blood, flowing away from the activated area, may maintain its elevated oxygen
Latency

One of the first observations made regarding fMRI signal changes is that after activation, the BOLD signal takes about 2 to 3 s to begin to deviate from baseline (16,38). Because the BOLD signal is highly weighted toward venous oxygenation changes, with a flow increase, the time for venous oxygenation to begin to increase is about the time that it takes blood to travel from arteries to capillaries and draining veins, which is 2 to 3 s. The hemodynamic “impulse response” function has been effectively used to characterize much of the BOLD signal change dynamics (39–41). This function has been derived empirically by performing very brief and well-controlled stimuli. In addition, it can be derived by deconvolving the neuronal input from the measured hemodynamic response (42,43). This type of analysis assumes that the BOLD response behaves in a manner that can be completely described by linear systems analysis, which is still an open issue. Regardless, observed hemodynamic response to any neuronal activation can be predicted with a reasonable degree of accuracy by convolving expected neuronal activity timing with the BOLD “impulse response” function. This function has typically been mathematically described by a \( y \) variate function (39).

If a task onset or duration is modulated, the accuracy to which one can correlate the modulated input parameters to the measured output signal depends on the variability of the signal within a voxel or region of interest. In a study by Savoy et al. (44) addressing this issue, variability of several temporal sections of an activation-induced response was determined. Six subjects were studied, and for each subject, 10 activation-induced response curves were analyzed. The relative onsets were determined by finding the latency with which the correlation coefficient was maximized with each of three reference functions representing three parts of the response curve: the entire curve, the rising section, and the falling section. The standard deviations of the whole curve, rising phase, and falling phase were found to be 650, 1,250, and 450 ms, respectively.

Across-region differences in the onset and return to baseline of the BOLD signal during cognitive tasks have been observed. For example, during a visually presented event-related word stem completion task, Buckner et al. (45) reported that the signal in visual cortex increased about 1 second before the signal in the left anterior prefrontal cortex. One might argue that this observation makes sense from a cognitive perspective because the subject first observes the word stem and then, after about a second, generates a word to complete this task. Others would argue that the neuronal onset latencies should not be more than about 200 ms. Can inferences of the cascade of brain activation be made on this time scale from fMRI data? Without a method to constrain or work around the intrinsic variability of the onset of BOLD signal over space, such inferences should not be made in temporal latency differences below about 4 s.

Lee et al. (46) were the first to observe that the fMRI signal change onset within the visual cortex during simple visual stimulation varies from 6 to 12 s. These latencies were also shown to correlate with the underlying vascular structure. The earliest onset of the signal change appeared to be in gray matter, and the latest onset appeared to occur in the largest draining veins. Similar latency dispersions in motor cortex have been observed. In one study, latency differences, detected in visual cortex with the Hilbert transform, did not show a clear correlation of latency with evidence for draining veins (47).

Figure 26.4 is a summary of the sources of temporal variability. Figure 26.4A shows a plot of the average time course from the motor cortex as a result of 2-second finger tapping. As mentioned, the first source of variability is the intrinsic noise in the time series signal. The standard deviation of the signal is on the order of 1%. The second source of variability is that of the hemodynamic response. As mentioned, this ranges from 450 to 1,250 ms, depending on whether one is observing the rising phase of the signal or saturation as far as a centimeter away. When brain activation is observed on a scale of centimeters, this has not been a major concern. Nevertheless, this issue is discussed in detail later in the chapter.
FIGURE 26.4. Demonstration of several of the limits of functional magnetic resonance imaging temporal resolution. Echo-planar imaging was performed at 3 T by using a Bruker Biospec 3T/60 equipped with a local head gradient coil. A time course series of axial images (matrix size = 96 × 96, field of view = 20 cm, echo time = 40 ms, repetition time = 500 ms, flip angle = 80 degrees) through the motor cortex was obtained. Bilateral finger tapping was performed for 2 s, followed by 18 s of rest. These figures demonstrate that the upper temporal resolution is determined by the variability of the signal change in time and space.

A: Time course of the signal elicited by tapping fingers for 2 s. The standard deviation at each point is in the range of 1% to 2%. The standard deviation of the hemodynamic change, in time, is in the range of 450 to 650 ms.

B: Map of the dot product (a measure of the activation-induced signal change magnitude) and the relative latencies or delays of the reference function (the plot in A was used as the reference function) at which the correlation coefficient was maximized. The spatial distribution of hemodynamic delay has a standard deviation of about 900 ms. The longest delays approximately match the regions that show the highest dot product and the area where veins are shown as dark lines in the T2*-weighted anatomic image.

C: Histogram of relative hemodynamic latencies. This was created from the latency map in (B). The falling phase. The third source of variability is the latency spread over space.

The plot in Fig. 26.4A was used a reference function for correlation analysis and allowed to shift ± 2 s. Figure 26.4B is a histogram of a number of voxels in an activated region that demonstrated a maximum correlation with the reference function at each latency (relative to the average latency) to which the reference function was shifted. The spread in latencies is more than 4 s. Figure 26.4C includes a map of dot product (measure of signal change magnitude) and latency; the regions showing the longest latency roughly correspond to the regions that show the largest signal changes. Although these largest signal changes are likely downstream draining veins, it is important to note that this approximate correlation between latency and magnitude is extremely weak. Many very small signal changes show very long latencies. It is also interesting to note that the inverse, that many large signal changes show short latencies, is typically not true. This implies that many downstream vessels may be almost fully diluted back to resting state oxygenation, therefore showing only a small signal change but still a large latency. Again, work is ongoing to characterize this effect better.

Magnitude

As previously discussed, the magnitude of the fMRI signal change is influenced by many variables across subjects, neuronal systems, neuronal systems, and voxels. Making a complete and direct correlation between neuronal activity and fMRI signal change magnitude in a single experiment will remain impossible until all the variables can be characterized on a voxel-related basis. Because of these primarily physiologic variables, the magnitude of BOLD signal changes on brain activation maps typically ranges from 1% to 5% [at, say, 1.5 tesla (T), gradient-echo sequence, echo time of 40 ms]. The picture is not that bleak, though. In the past several years, considerable progress has been made in characterizing
the magnitude of the fMRI signal changes with underlying neuronal activity.

The progressive series of studies was as follows: First, as mentioned previously, it was clear that areas that showed significant BOLD signal change were in the appropriate neuronal area corresponding to specific, well-characterized tasks. Second, inferred neuronal modulation was carried out by systematically varying some aspect of the task. Clear correlations between BOLD signal change magnitude and visual flicker rate, contrast, word presentation rate, and finger tapping rate were observed (13,48–50). This parametric experimental design represented a significant advance in the manner in which fMRI experiments were performed, enabling more precise inferences, not about the BOLD signal change with task modulation. Nevertheless, of course, the degree of neuronal activation (i.e., integrated neuronal firing over a specified area) was still inferred.

Recently, several more intriguing studies have emerged correlating measured neuronal firing rate with well-known stimuli in animals (51) and humans (52,53) and demonstrating a remarkably high correlation between BOLD signal change and electrophysiologic measures.

**Linearity**

Related to the topic of signal change magnitude is that of BOLD signal change linearity. It has been found that, with very brief stimulus durations, the BOLD response shows a larger signal change magnitude than expected if one assumes that it behaves as a linear system (54,55). This greater than expected BOLD signal change is generally specific to stimulus durations below 3 s. Reasons for nonlinearities in the event-related response can be neuronal, hemodynamic, or metabolic in nature. The neuronal input may not be a simple boxcar function. Instead, an increased neuronal firing rate at the onset of stimulation (neuronal “bursting”) may cause a slightly larger amount of vasodilation that later plateaus at a lower steady-state level. The amount of neuronal bursting necessary to change the hemodynamic response significantly, if a linear neuronal–hemodynamic coupling is assumed, is quite large. For example, to account for the almost double functional contrast for the experimental related to the linear convolution-derived single-event responses, the integrated neuronal response greater than 2 s must double. If it is assumed that neuronal firing is at a higher rate only for about the first 50 ms of brain activation, the neuronal firing rate must be 40 times greater than steady state for this duration.

BOLD contrast is highly sensitive to the interplay of blood flow, blood volume, and oxidative metabolic rate. If, with activation, any one of these variables changes with a different time constant, the fMRI signal may show fluctuations until a steady state is reached (56,57). For instance, an activation-induced increase in blood volume would slightly reduce the fMRI signal because more deoxyhemoglobin would be present in the voxel. If the time constant for blood volume changes were slightly longer than that of flow changes, then the activation-induced fMRI signal would first increase, then be reduced as blood volume later increased. The same could apply if the time constant of the oxidative metabolic rate were slightly slower than that of flow and volume changes. Evidence for an increased oxidative metabolic rate after 2 min of activation is given by Frahm et al. (57), but no evidence suggests that the time constant of the increase in oxidative metabolic rate is only seconds longer than the flow increase time constant—as would be required to be applicable only to relatively high-amplitude single-event responses. These hemodynamics, which may also differ on a voxel-related basis, remain to be characterized fully.

**SCANNER-RELATED ISSUES**

A complete discussion of all scanner-related issues and potential solutions is beyond the scope of this chapter. A rudimentary yet necessary description of the most basic problems and solutions is presented in the following sections. Most practitioners of functional MRI typically undergo a painful, frustrating, and prolonged period of learning about all scanner-related limitations and issues. Some give up hope completely. Those who are determined usually emerge hopeful again, and with a much better “feel” for what can and cannot be done in regard to brain imaging. This learning process also applies to understanding the physiology of the signal, but typically the greatest anguish arises in the context of MRI pulse sequences and hardware.

In the first place, all the categories listed below are more or less linked. In this section, an attempt is made to walk the reader informally through the trade-offs and issues involved in performing an fMRI experiment.

**Acquisition Rate**

Image acquisition rate is ultimately limited by how fast the signal can be digitized and by how rapidly the imaging gradients can be switched. MRI can be logically divided into single-shot and multishot techniques. In single-shot imaging, spins are excited with a single excitation pulse and all the data necessary to create an image are collected at once. Echo-planar imaging (EPI) is the most common single-shot technique; with one “echo,” a single “plane” is acquired. Multi-shot techniques are most commonly used for high-resolution anatomic imaging. In clinical scanning (with multishot imaging), a single “line” of raw data is acquired with each RF excitation pulse. Because of the relatively long time it takes for the longitudinal magnetization to return to equilibrium (characterized by the T1 of the tissue), a certain amount of time, between 50 and 500 ms, is spent waiting between shots; otherwise, soon no signal would be
left. It would be “saturated.” Because of this necessary waiting time, multishot techniques are typically slower than single-shot techniques. For a 150-ms “waiting time,” or repetition time (TR), an image with 128 lines of raw data would take 150 ms multiplied by 128, or 19.2 s.

In the case of EPI, the entire data set for a plane is typically acquired in about 20 to 40 ms. In the context of performing a BOLD experiment, the echo time (TE) is about 20 to 40 ms. Along with some additional time for applying other necessary gradients, the total time for an image to be acquired is about 60 to 100 ms, so that 10 to 16 images can be acquired in a second. Improvements in digital sampling rates and gradient slew rates will allow small improvements, but essentially, this is about the upper limit for imaging humans.

In the context of an fMRI experiment with EPI, the typical image acquisition rate is determined by how many slices can temporally fit into a TR. For whole-brain imaging, approximately 20 slices (5-mm thickness) are required to cover the entire brain. This allows a TR of about 1.25 to 2 s at minimum. This image collection rate is more than adequate to capture most of the details of the slow and dispersed hemodynamic response.

**Spatial Resolution**

The spatial resolution is also primarily determined by gradient strength, digitizing rate, and time available. For multishot imaging, as high a resolution as desired can be achieved if one is willing to wait. One can keep on collecting lines of data with more RF pulses. For EPI, the signal decay rate (described by T2* with gradient-echo EPI and by T2 with spin-echo EPI) plays a significant role in determining the resolution. One can sample for only so long before the signal has completely decayed away. For this reason, the resolution of EPIs is generally much lower than that of multishot images. To get around this problem, two further strategies are commonly used. The first is multishot EPI, in which the full EPI acquisition is used multiple times (but many fewer times than in typical clinical multishot imaging) and interleaved to increase the resolution. The second is to perform an operation called **conjugate synthesis**, which basically makes use of the fact that, in raw data space, half of the data is redundant. This allows at most twice the resolution, with some cost in signal to noise and image quality. An example of multishot EPI is shown in Fig. 26.5.

**Signal to Noise**

The signal to noise and the functional contrast to noise are influenced by many variables. These include, among others, voxel volume, TE, TR, flip angle, receiver bandwidth, field strength, and RF coil used. Not considering fMRI for a moment, the image signal to noise is increased with larger voxel volume, shorter ET, longer RT, larger flip angle (to 90 degrees), narrow receiver bandwidth, higher field strength, and smaller RF coil. That said, in the context of fMRI, the functional contrast to noise is optimized with a voxel volume equaling the size of the activated area, TE = gray matter T2*, short TR (optimizing samples per unit time), flip angle = Ernst angle = Cos (–TR/T1), narrow receiver bandwidth, high field strength, and smaller RF coil. Of course, all of these variables come at some expense to others.

**Stability**

Theoretically, the noise, if purely thermal in nature, should propagate similarly over space and across time. In fMRI, this is not at all the case because each image is essentially captured in 40 ms and the time series is collected in minutes. Stability is much more of an issue on the longer time scale.

---

**FIGURE 26.5.** An example of multishot echo-planar imaging. Number of excitations ranged from 1 to 8. The image resolution increases but the signal to noise and functional contrast to noise decrease. In addition, instabilities are introduced into the time course by the use of multishot imaging.
Flow and motion are correlated in many areas with cardiac and respiratory cycles. Subject movement and scanner instabilities also contribute. Single-shot techniques have generally better temporal stability than multishot techniques because, with multishot techniques, the image is collected over a larger time scale; instabilities on a longer time scale enter into the image creation itself. This leads to nonrepeatable ghosting patterns that generally decrease temporal signal to noise ratio. Work is ongoing to characterize and reduce temporal instabilities for both single-shot and multishot imaging techniques (58,59). Correction techniques include cardiac gating (60), navigator pulses (61,62), and raw data reordering (63,64).

**Image Quality**

Image quality issues that are the most prevalent are image warping and signal dropout. Although books can be written on this subject, the description here is limited to the bare essentials.

Image warping is fundamentally caused by either or both of two factors, $B_0$-field inhomogeneities and gradient nonlinearities. A nonlinear gradient causes nonlinearities in spatial encoding, so that the image is distorted. This is primarily a problem when local, small-gradient coils are used that have a small region of linearity that drops off rapidly at the base and top of the brain. With the growing prevalence of whole-body gradient coils for performing EPI, this problem is no longer a major issue. If the $B_0$ field is inhomogeneous, as is typically the case with imperfect shimming procedures, particularly at higher field strengths, the protons are processing at frequencies different from those expected in their particular location. This causes image deformation in the areas of poor shim, particularly with the long readout window or acquisition time of EPI. A solution is either to shim better (65,66) or map the $B_0$ field and perform a correction based on this map (67).

Signal dropout is related to the problem described above in that it is also caused by localized $B_0$-field inhomogeneities, typically at interfaces of tissues having different susceptibilities. If within a voxel, because of the $B_0$ inhomogeneities, spins are of different frequencies, their signals cancel each other out. Several strategies exist for reducing this problem. One is, again, to shim as well as possible at the desired area. Because of still imperfect shimming procedures, this is usually not satisfactory. The other is to reduce the voxel size (increase the resolution), so that stratification of different frequencies is reduced within a voxel. The third is to choose the slice orientation such that the smallest voxel dimension (in many studies, the slice thickness is greater than the in-plane voxel dimension) is perpendicular to the largest $B_0$ gradient. For this reason, many studies are performed with the use of sagittal or coronal slice orientations.

As with many of the topics discussed, much more can be said, but the goal here is simply to provide an introduction and references to additional reading material.

**BEST RESULTS SO FAR**

The primary discussion up to this point has focused on the limits imposed by the scanner and the hemodynamics. In this section, some of the most successful, thought-provoking, and innovative fMRI studies, from a methodologic perspective, performed as of September 2000 are discussed. The best results in temporal and spatial resolution are presented.

**Temporal Resolution**

As explained in previous sections, MR images can be acquired at an extremely rapid rate; therefore, scanner-related limits are not the prime determinant of the upper limits of temporal resolution in fMRI. The key to increasing the temporal resolution in fMRI is either to characterize the hemodynamic response better or to work around its limits. The results described here are examples of this work during the past few years.

To obtain information about relative onsets of cascaded neuronal activity from hemodynamic latency maps, it is possible to determine relative latency changes on modulation of the task timing. In a study of Savoy et al. (68), activation onset latencies of 500 ms were discernible when they used a visual stimulation paradigm in which the left and right hemifields were stimulated at relative delays of 500 ms. First, the subject viewed a fixation point for 10 s. Then, the subject’s left visual hemifield was activated 500 ms before the right. Both hemifields were activated for 10 s, then the left hemifield stimulus was turned off 500 ms before the right.

Although with careful choice of region of interest, from which the time course plot is made, these onset differences can be shown, maps of latency cannot reveal the onset differences because, as mentioned, the variability over space, which is about 4 s, dominates the inserted 500-ms variability from left to right hemifield.

To map the relative latency differences between hemifields, it is necessary to modulate the relative stimulation timing. As an extension of their results, the left—right onset order was switched so that, in the second run, the right hemifield was activated and turned off 500 ms before the left. Latency maps were made for each onset order and subtracted from each other to reveal a clear delineation between the right and left hemifields that was not apparent in each of the individual maps. This operation and the resulting relative latency map is shown in Fig. 26.6. These maps are of the change in onset of one area relative to another, not of absolute latency. It is also useful to note that the standard deviation of these maps is reduced simply to the standard
FIGURE 26.6. The use of latency maps and task modulation to extract relative latencies. Activation within a region of visual cortex is shown. In one condition (left), the right visual hemifield stimulation precedes the left by 500 ms. In the other condition (middle), the left visual hemifield precedes the right hemifield stimulation by 500 ms. Latency maps from both these conditions show an across-voxel spread of $\pm 2.5\ s$, which is too large to identify clearly the relative latencies across hemifields. However, once the data are normalized for this intrinsic variance by directly comparing the hemodynamic response from the two different lags within individual voxels, the offset between the left and right hemifields can be observed (right). This demonstration suggests that normalization of the hemodynamic lag can allow small relative temporal offsets to be identified. These normalized offsets can then be compared across regions to make inferences about neuronal delay. For this experiment, the repetition time was 400 ms.

deviation of the latencies in each voxel, not the standard deviation of the latencies over space. Maps such as these can be extremely useful in determining which regions of activation are modulated relative to other areas with a specific task timing modulation.

Published work by Menon et al. (69), Kim et al. (70), Richter et al. (71), and Bandettini (72) has explored the temporal resolution limits of fMRI. The results of Menon et al. (69), similar to those mentioned above, indicate that relative brain activation timings on the order of 50 ms can be discerned.

In the study of Richter et al. (71), a parametrically varied event-related mental rotation task was used. Each mental rotation task was presented individually. A high correlation between task duration and event-related width was demonstrated. The longer the task took to accomplish (larger rotation angle), the wider the event-related response was shown to be in specific parietal locations.

Spatial Resolution
The hemodynamic point spread function was first considered and characterized by Engel et al. (73–75). Localization to 1.1 mm was determined.

The first successful mapping of ocular dominance columns in humans was published by Menon et al. (76). Their intriguing results show that the optimal way to pull out differences in activation across closely spaced units is to perform very brief stimuli so as not to reduce the dynamic range of the oxygenated blood that is flowing away beyond the unit of activation.

In regard to MRI pulse sequence, it is important to note that mapping cortical columns multishot imaging is required (76). Performance of multishot imaging requires either navigator echoes or shot-to-shot phase-correction schemes. If these are not performed, temporal stability is seriously compromised.

Many have argued that some aspects of the BOLD signal change dynamics are more spatially localized to neuronal activity. Specifically, the evasive “pre-undershoot” has been indicated as such (77,78). The rationale is that this transient “dip” in the signal that occurs 0.5 to 2 s after the onset of activation and quickly gives way to the much larger signal increase is secondary to direct extraction of oxygen from the blood by adjacent activated tissue. Recently published work has demonstrated the utility of such an approach for mapping cortical columns in animals (79–81).

The highest-resolution fMRI performed with single-shot EPI was obtained by Jesmanowicz et al. (82). Here, a partial k-space strategy was used to obtain a presumed 256 × 256 resolution. The actual resolution achieved is debatable because T2* effects reduced the resolution below that implied by the matrix size.

NEURONAL ACTIVATION INPUT STRATEGIES

Much of this chapter has been devoted to the basics and some esoteric issues regarding fMRI. This section provides an overview of the types of neuronal input strategies with which fMRI has been used to extract information about what the brain is doing. Given a question of brain function, what are the available strategies that one can use to design their paradigm? A schematic summary of these strategies appears in Fig. 26.7.

Block Design
A block design paradigm was the first used in fMRI and is still the most prevalent neuronal input strategy. Borrowed
Neuronal Activation Input Strategies

1. Block Design
2. Frequency Encoding
3. Phase Encoding
4. Single Event
5. Orthogonal Block Design
6. Free behavior Design

FIGURE 26.7. Overview and schematic depiction of types of neuronal input strategies available in functional magnetic resonance imaging. In addition, parametric designs, which involve systematically varying some aspect of the intensity of the neuronal input, can be applied to any of the design strategies.

from previous PET studies, it involves having a subject alternately perform a task for at least 10 s, then a control task for a similar time, so that the hemodynamic response reaches a steady state in each condition. This is a useful technique in that it is easy to implement, and standard statistical tests can be used to compare each condition.

Phase and Frequency Encoding

Phase encoding of a stimulus involves varying some aspect of the stimulus in a continuous and cyclic manner. This strategy has been most successfully used in retinotopic mapping (75,83,84). In this type of study, the eccentricity of a visual stimulus ring is continuously varied; then, after the most extreme eccentricity is reached, the cycle is repeated. The data are typically subjected to Fourier analysis, and the areas that show a signal change temporal phase that correlates with the stimulus phase are mapped. This is a powerful technique because it makes use of the entire time series, in that there are no “off” states, and lends itself to Fourier analysis. This method has also been used for somatotopic mapping (85) and tonotopic mapping (86).

Frequency encoding is much less common but can be achieved for certain types of stimuli. The method is to designate a specific on–off frequency for each type of stimulus used. Again, Fourier analysis reveals the most power under a spectral peak corresponding to the brain area specific to the particular on/off frequency. The utility of this method has been demonstrated in mapping left and right motor cortex by cueing the subject to perform a finger-tapping task at different on–off rates for each hand (87).

Orthogonal Designs

Orthogonal task design is a powerful extension of block design. The basic concept is that if one designs two different task timings to create BOLD responses that are orthogonal to each other, then these tasks can be performed simultaneously during a single time series collection with no cross-task interference, so that comparison is much more precise. This technique was pioneered by Courtney et al. (88). In their study, six orthogonal tasks were designed into a single time series. This type of design also lends itself to event-related studies.

Parametric Designs

As mentioned in the section on magnitude, parametric designs are powerful in that more precise statements can be made about relative neuronal activity. A parametric task design simply involves systematically varying some aspect of the task during the time series. This may be a finger-tapping rate, stimulus contrast or flicker rate, cognitive load, or attention demand, and instead of mapping the magnitude of the change with a single task, the slope of the change corresponding to a task is mapped. In this manner, relative brain activation magnitude may be teased out of the time series.

Event-Related Designs

Before 1995, a critical question in event-related fMRI was whether a transient cognitive activation could elicit a significant and usable fMRI signal change. In 1996, Buckner et al. (45) demonstrated that event-related fMRI in fact lends itself quite well to cognitive neuroscience questions. In their study, a word stem completion task was performed; a block design and an event-related strategy were used. Robust activation in the regions involved with word generation were observed in both cases.

Given the substantial number of recent reports of event-related fMRI (40–42,65,89–112), it can probably be said that this is one of the more exciting developments in fMRI since its discovery.

The advantages of event-related activation strategies are many (113). These include the ability to randomize task types in a time series more completely (114–116), the ability to analyze fMRI response data selectively based on measured behavioral responses to individual trials (111), and the option to incorporate overt responses into a time series. Separation of motion artifact from BOLD changes is possible by use of the temporal response differences between motion effects and the BOLD contrast-based changes (91).

When a constant ISI is used, the optimal interstimulus interval (ISI) is about 10 to 12 s. Dale and Buckner (43) have shown that responses to visual stimuli, presented as rapidly as once every 1 s, can be adequately separated by
using overlap correction methods. Overlap correction methods are only possible if the ISI is varied during the time series. These results appear to demonstrate that the hemodynamic response is sufficiently linear that deconvolution methods can be applied to extract overlapping responses. Burock et al. (95) have demonstrated that remarkably clean activation maps can be created with an average ISI of 500 ms. If one assumes that the hemodynamic response is essentially a linear system, there appears to be no obvious minimum ISI in attempts to estimate the hemodynamic response. Dale has suggested that an exponential distribution of ISIs with a mean as short as psychophysically possible is optimal for estimation (100). Of course, the speed at which one can present stimuli depends on the study being performed. Many cognitive tasks may require a slightly longer average presentation rate.

Future work in event-related experimental optimization rests on what further information can be derived from these responses. Between-region, between-voxel, between-subject, and stimulus-dependent variations in amplitude, latency, shape, and responsivity of the event-related fMRI responses are still relatively uncharacterized. Reasons for these differences are also still unclear.

Free Behavior Designs

With many types of cognitive neuroscience questions, it is not possible to constrain the timing or performance of a task precisely. It is necessary to allow the subject to perform the task “freely” and obtain a continuous measurement of the performance, then use the measurement as a reference function for subsequent time series analysis. Examples of this type of design are emerging. As an example, skin conductance changes are difficult to predict or control. In a study by Patterson et al. (117), skin conductance was simultaneously recorded during an array of tasks and during “rest.” The skin conductance signal change was then used as a reference function in the fMRI time series analysis. In several cortical and subcortical regions, signal changes were observed that were highly correlated with the skin conductance changes. Without the use of this measurement, such signal changes would have appeared as noise. It is thought that this type of design will become more prevalent as methods to monitor subject performance or state precisely become more sophisticated.

CONCLUSION

This chapter has attempted to combine a review of the fundamentals of fMRI with a glimpse of the state of the art. Starting with the basics of fMRI contrast, the discussion moved on to hemodynamic transfer function—the basis of understanding fMRI signal change. Characteristics related to the hemodynamic transfer function include location, latency, magnitude, and linearity. Then, perhaps less provocative but still important issues of working with an MRI scanner and understanding some practical limitations were discussed. A sampling of best results, those successfully bringing into play many of the features of experimental design and analysis already mentioned, was presented. The chapter ended with a brief overview of neuronal input strategies, or rather, ways in which one can activate the brain in the context of an fMRI experiment.

Functional MRI is about 9 years old and apparently still at the beginning of its growth curve in terms of users and applications. Clinical applications are just beginning, whereas cognitive neuroscience applications are in full swing. The field of fMRI continues to develop along intertwining paths of understanding signals, creating tools, and refining the questions being asked.

REFERENCES


54. Boynton GM, Engel SA, Glover GH, et al. Linear systems analy-


