Functional magnetic resonance imaging: imaging techniques and contrast mechanisms

Alistair M. Howseman† and Richard W. Bowtell

1Wellcome Department of Cognitive Neurology, Institute of Neurology, Queen Square, London WC1N 3BG, UK
2Magnetic Resonance Centre, School of Physics and Astronomy, University of Nottingham, Nottingham NG7 2RD, UK

Functional magnetic resonance imaging (fMRI) is a widely used technique for generating images or maps of human brain activity. The applications of the technique are widespread in cognitive neuroscience and it is hoped they will eventually extend into clinical practice. The activation signal measured with fMRI is predicated on indirectly measuring changes in the concentration of deoxyhaemoglobin which arise from an increase in blood oxygenation in the vicinity of neuronal firing. The exact mechanisms of this blood oxygenation level dependent (BOLD) contrast are highly complex. The signal measured is dependent on both the underlying physiological events and the imaging physics. BOLD contrast, although sensitive, is not a quantifiable measure of neuronal activity. A number of different imaging techniques and parameters can be used for fMRI, the choice of which depends on the particular requirements of each functional imaging experiment. The high-speed MRI technique, echo-planar imaging provides the basis for most fMRI experiments. The problems inherent to this method and the ways in which these may be overcome are particularly important in the move towards performing functional studies on higher field MRI systems. Future developments in techniques and hardware are also likely to enhance the measurement of brain activity using MRI.

Keywords: fMRI; echo-planar imaging; BOLD contrast; brain mapping

1. INTRODUCTION

Functional neuroimaging using nuclear magnetic resonance (NMR) is approaching the end of its first decade as a scientific discipline. During that time the technique has evolved with great speed from being a potentially useful tool into arguably the most powerful method available for generating maps of neuronal activity in the human brain. This development has seen the technique move from the domain of the magnetic resonance imaging (MRI) research laboratory into the arena of dedicated centres for functional neuroimaging populated by neuroscientists, psychologists and psychiatrists with, in some cases, few MRI specialists. A large body of research has already been generated using functional MRI (fMRI) to study brain function in healthy volunteers. Among others, the following fields have been investigated: vision (Tootell et al. 1998); motor (Ashe & Ugurbil 1994), language (Binder 1997), memory (Fletcher et al. 1997); emotion (Phillips et al. 1997; Schneider et al. 1997) and pain (Davis et al. 1997). Many studies have also been performed in patient populations, although these entail additional and varied difficulties. The technique has been used in the fields of stroke (Leifer et al. 1998; Thulborn et al. 1998), amputation and phantom limbs (Lotze et al. 1998), presurgical planning (Bittar et al. 1998; Lehericy et al. 1998; McCarthy 1998), epilepsy (Jackson et al. 1994; Graveline et al. 1998) and various psychiatric disorders (see Fu & McGuire, this issue). The extent to which the theoretical, statistical (see Petersson et al., both papers in this issue) and practical difficulties associated with these studies can be overcome will define the clinical value of fMRI and this in turn will ultimately determine the number of MRI systems installed with functional imaging capability.

The remarkable speed with which the fMRI technique has become a valuable tool in imaging neuroscience is largely due to the fact that several of the core ingredients which make up fMRI methodology were already at an advanced level of development when the potential to image brain function using MRI first became apparent. In particular, by the early 1990s the established field of functional imaging using positron emission tomography (PET) (Lassen et al. 1991) had seen the development of ideas about (i) the design of experiments to generate images of cognitive processing, and (ii) methods of processing data to yield statistical images reflecting brain activity associated with those cognitive functions (for overview, see Frackowiak et al. 1997). The acquisition of functional imaging data using a non-invasive method (MRI) with spatial and temporal resolution superior to that of PET was rapidly embraced by this research community. The rapid growth of fMRI has also been facilitated by recent developments in NMR imaging methods and instrumentation. MRI has been in widespread clinical use since the early 1980s and MRI manufacturers now provide highly optimized MRI systems. Two technical factors have been especially important in the development of fMRI. First, the speed of
the imaging method needs to be such that data from the entire brain can be acquired in a time less than the duration of the brain states induced by the presentation of stimuli. This requirement is fulfilled by echo-planar imaging (EPI), which was initially proposed in the earliest days of MRI (Mansfield 1977), but which was not available on clinical MRI systems until recently because of the technical difficulties associated with EPI (for overview of EPI, see Schmitt et al. (1998)). The second aspect of scanner technology which has been especially important in allowing sophisticated fMRI studies to be performed is the ever-increasing computing power available, which facilitates the acquisition, processing and storage of very large data sets.

Although the first fMRI experiments (Belliveau et al. 1991) used an exogenous gadolinium-based contrast agent this technique was rapidly superseded by the discovery that deoxyhaemoglobin could be used as an endogenous contrast agent (Ogawa et al. 1990). This so-called blood oxygenation level dependent (BOLD) contrast has proved to be a very sensitive MRI marker of neuronal activity at magnetic field strengths of 1.5 T and above. The sensitivity of BOLD contrast is such that experiments can be performed in individual subjects, with temporal resolution on the scale of hundreds of milliseconds to seconds and spatial resolution of 3 mm or less, in almost all cortical and sub cortical structures in the brain. This article will briefly describe the mechanisms which generate contrast in fMRI and will discuss the different imaging techniques which are employed. There remain outstanding issues regarding the level of accuracy with which BOLD contrast reflects neuronal activity. Furthermore BOLD contrast remains essentially non-quantifiable. These and other problems associated with the fMRI technique will be addressed, including the image artefacts which pose an ever greater problem at higher magnetic field strength.

2. BOLD CONTRAST IN MRI

(a) Principles of MRI and BOLD contrast

NMR image intensity reflects the concentration of water within a sample and is also dependent on the chemical and physical environment in which the water molecules reside (for review, see Gadman 1995). Image contrast in MRI can be varied through the choice of imaging parameters, and in fMRI contrast is tailored to optimize the signal dependence on deoxyhaemoglobin concentration ([HHb]). The MRI signal arises from the stimulation of transitions between the nuclear spin energy levels of the hydrogen atoms in water, placed in a large magnetic field. These transitions are induced by the application of radio-frequency (RF) energy at a specific (resonance) frequency which is proportional to the strength of the magnetic field. An NMR signal is detected, at the resonance frequency, as a voltage across a tuned RF coil. Images of the distribution of water are created using magnetic field gradients which impose a linear relationship between the NMR frequency and spatial location. NMR relaxation is the process by which the perturbed nuclear spin system returns to thermal equilibrium after the application of a pulse of RF energy. A relaxation property which is dependent on the range of NMR frequencies within an image voxel, and is described by a time constant $T_2^*$, is the source of BOLD contrast. $T_2^*$ is affected by random processes involving interactions between nuclear spins as well as by magnetic field inhomogeneity, which can result from either macroscopic magnetic susceptibility differences in the head or the microscopic effects of paramagnetic substances. While biological samples are primarily diamagnetic, deoxyhaemoglobin is paramagnetic and its concentration affects the value of $T_2^*$ in a voxel due to the microscopic magnetic field gradients established in its vicinity. This mechanism can give rise to BOLD contrast on MR images which are sensitive to magnetic field inhomogeneity, that is, $T_2^*$-weighted images (Ogawa et al. 1990).

By using EPI it is possible to image changes in blood oxygenation levels very rapidly (Turner et al. 1991).

A robust way to generate BOLD contrast is to continually acquire MR images while subjects are presented with blocks of sensory stimulation, of duration approximately 30s, interspersed with intervals of rest (Ogawa et al. 1992; Kwong et al. 1992; Bandettini et al. 1992). Simply, the average baseline (resting) condition image is subtracted from the image corresponding to the average of the stimulation condition. The difference images show BOLD contrast in regions where the stimulated neuronal activity leads to a reduction in [HHb]. This occurs because a localized increase in regional cerebral blood flow (rCBF) associated with neuronal firing is accompanied by a relatively smaller increase in the use of oxygen, an effect first described by PET studies using $^{15}$O labelling (Fox & Raichle 1986; Fox et al. 1988). The reduction in [HHb] leads to smaller local field gradients, which results in less signal loss due to spin dephasing i.e. an increase in $T_2^*$ and a consequent increase in the level of the MRI signal (Ogawa et al. 1990, 1992).

The magnitude of the signal change which can be generated by BOLD contrast has been the subject of vigorous debate since the earliest reports of fMRI. In several early fMRI studies which made use of fast gradient echo imaging techniques a component of $T_2^*$ contrast was inadvertently included in the functional signal change and this could lead to large (tens of per cent) changes (Frahm et al. 1993). However, it quickly became apparent (Kim et al. 1994; Franh et al. 1994) that these signals, which are due to blood in large vessels flowing into the imaging slice, are less well localized to neuronal activity than the BOLD signal. Most fMRI experiments performed today use sequences which ensure that the $T_2^*$ component is minimal (Howsman et al. 1999).

Despite the clarification that $T_2^*$ contrast should be avoided, the debate regarding signal magnitude continues. There are so many factors affecting the measured signal changes (how a region of interest is selected, field strength, several imaging parameters etc.) that comparisons between studies are difficult. However, this issue is important because it is related to which part of the vasculature yields the dominant signal change and therefore how accurately the spatial distribution of fMRI signals reflects neuronal activity. In general, BOLD contrast represents a qualitative rather than a quantitative measure of brain activity because the relationship between the signal change in $T_2^*$-weighted images and
the underlying vascular geometry and distribution of deoxyhaemoglobin is not easily modelled. Essentially the problem is that many image voxels contain very heterogeneous distributions of vessels. The voxel size in a typical fMRI data set is of the order of 3 mm x 3 mm x 3 mm (27 μ). At this scale any voxel may contain blood vessels ranging from capillaries (5 μm diameter) and venules (10–30 μm) up to pial veins (100–400 μm) and larger draining veins. Only vessels on the venous side of the circulation are subject to oxygenation changes and only those close enough to the site of neuronal activity will contribute to BOLD contrast. Further downstream the effects are gradually diluted. The signal change measured in a voxel will therefore reflect how the underlying physiological events alter the distribution of [rHb] and how the geometry of the vessels impact on the physics of the BOLD contrast mechanism.

If contributions to BOLD contrast were limited to intravascular signals then the small fractional blood volume in the brain would severely limit the MRI signal changes due to [rHb] changes. However, because the mechanism by which BOLD occurs involves the magnetic field gradients which surround deoxyhaemoglobin, these gradients can extend into the surrounding tissue and so a higher proportion of water can be affected (Ogawa et al. 1990). This effect has been analysed with a Monte Carlo simulation of a biophysical model which assumes a homogeneous distribution of vessels (Ogawa et al. 1993). The model indicates that, with 75% of the blood volume in a voxel found in capillaries and 25% in venules, the BOLD signal change in a $T_2^*$ image acquired at an optimal echo time (TE) at 4 T would be approximately 8%. Additionally the model predicts that at higher field strength the (extravascular) capillary component of BOLD contrast becomes relatively more significant than that from venules and pial veins. At a lower field strength of 1.5 T it has been suggested that the intravascular water is the dominant source of the BOLD signal (Boxerman et al. 1995), a theory which is supported by experimental evidence that when large spoiler (diffusion) gradients are applied most of the BOLD signal is removed as the intravascular component is dephased. However, in a series of experiments in rats at 2.0 T and humans at 1.5 T using intravascular contrast agents, Zhong et al. (1998) argue that there is also a significant extravascular component of BOLD contrast at these field strengths. Although the theoretical predictions are that BOLD contrast should be less than a few per cent at field strengths up to 2 T, there is much experimental evidence of larger changes in $T_2^*$-weighted images (Lai et al. 1993; Frahm et al. 1994; Howseman et al. 1999). It has been proposed that these larger percentage changes may occur, if a significant fraction of a voxel is filled by a large vessel, as a result of simple bulk phase changes occurring in the vessel during activation (Haacke et al. 1994).

There are techniques, including the use of diffusion gradients, which allow partial discrimination between BOLD signals which arise purely from the capillary bed and those which occur in larger vessels (Boxerman et al. 1995; Lee et al. 1995; Song et al. 1996). Although it is desirable that fMRI contrast should reflect activity in grey matter it is probably unrealistic to expect that good sensitivity can be maintained at a field strength of 1.5 T if all the signal from the larger vessels is removed. Nearly all fMRI experiments are performed using $T_2^*$-based imaging. However, it is also possible to generate contrast using $T_2$-weighted imaging with a spin echo sequence. In $T_2$ imaging the dephasing which leads to signal decay is due only to random processes and not to dephasing due to static inhomogeneities. [rHb] can affect $T_2$ because of the diffusion of water molecules through the microscopic field gradients. The contrast due to functional activation on $T_2$ images is much smaller than on $T_2^*$ images, hence its limited use, although simulations suggest (Ogawa et al. 1993; Weisskoff et al. 1994) that this is partly because large vessels contribute much less to the signal. Bandettini et al. (1994b) estimate that the ratio between the contrast generated by gradient echo and spin echo fMRI in the same region is about 2, when both techniques are optimized.

(b) Relationship between BOLD and physiology

A comprehensive review of how the relationship between neuronal activity, rCBF and metabolic activity affects PET data is given by Jueptner & Weiller (1995). With PET the key issue is how well rCBF changes correlate with neuronal activity. In fMRI the relationship between image contrast and neuronal activity is dependent on a complex set of factors which determine the spatio-temporal changes in [rHb]. An increase in blood flow and an increase in blood volume accompany the onset of neuronal firing. The balance between these processes and the rate at which oxygen is metabolized ([rCMRO2]) determine the local [rHb] although the correlation between [rHb] and neuronal firing may be nonlinear and temporarily varying. Evidence about the relationship between blood flow increases and the use of oxygen has been provided by direct optical measurements in the exposed visual cortex of cat brain (Malonek & Grinvald 1996). These experiments show that the response to sensory stimulation is an initial phase (a, 3 s) of hypo-oxygenation which is highly localized to individual cortical columns followed by an extended phase (several seconds) of hyperoxygenation which is more widely dispersed (up to 5 mm). Broadly, the temporal characteristic of this response is in accord with that measured by fMRI, although the optical data suggest that a rather larger initial dip in the BOLD response would be expected than that which has been measured (Menon et al. 1995). Crucially, the optical data raise the possibility that the main (positive) part of the BOLD response is not so well localized and therefore the true functional resolution of fMRI may be limited not by the imaging resolution itself but by the spatial extent of the physiological changes arising within the vicinity of neuronal firing. More restricted functional localization by using the initial dip has been verified with fMRI at high field in human visual cortex (Hu et al. 1997) and simulations support the hypothesis that there can be a localized $T_2^*$ effect due to increased oxygen extraction prior to changes in rCBF and blood volume (Dymond & Norris 1997). However, before designating the delayed positive part of the BOLD response as being too widespread to provide high-resolution localization, it is salutary to consider the study of Yang et al. (1997b) in which they show that positive BOLD contrast has the same spatial
pattern and extent as the measured electrical activity in individual rat whisker barrels.

The temporal characteristics of the BOLD response are important not only because they relate to the physiological events underpinning BOLD contrast but because they also impact on experimental design and data processing. The form of the haemodynamic response function (HRF) measured in fMRI can vary in timing from brain region to region, but is broadly similar. An initial very small signal decrease of about 1 s duration, which is more easily measured at high field (Menon et al. 1993; Hu et al. 1997), is followed by the BOLD signal increase which peaks between 4 and 8 s after the onset of stimulation (Friston et al. 1994). After the end of stimulation the signal exhibits a small undershoot which returns to baseline over tens of seconds (Janz et al. 1997; Fransson et al. 1998). A report that during extended stimulation the BOLD signal decays to baseline due to a recoupling of flow and metabolism (Frahm et al. 1996; Kruger et al. 1996) has not been confirmed to be a general phenomenon in other studies (Bandettini et al. 1997; Howseman et al. 1998). Consequently it is currently widely accepted that the HRF can be convolved with the experimental design as a model for expected activity during fMRI analysis; that is, the haemodynamic response which leads to BOLD is time invariant during the duration of an experiment. However, it has also been shown that there may be nonlinear components in the magnitude of the response if the rate at which distinct individual stimuli are presented is high (Friston et al. 1998a). This analysis is based on experimental data from a direct comparison between PET and fMRI (Rees et al. 1997), which showed that as the rate at which words were presented to a subject increased, a measured linear increase in rCBF was not mirrored by a linear increase in the amplitude of BOLD, the BOLD response starting to saturate with word rates above 30 per minute. These data strongly suggest that the extraction of oxygen increases nonlinearly as the neuronal activity increases because the linear increase in rCBF is expected to correlate well with the neuronal firing rate (Jueptner & Weiller 1995). This finding has clear implications for theoretical models of the coupling between rCBF and oxygen metabolism and especially needs to be considered in event-related fMRI studies (Rosen et al. 1998; Josephs & Henson, this issue).

In figure 1 the excellent correspondence between the activation images produced using PET and fMRI in the Rees et al. (1997) study is shown.

A recent model incorporating the effects of dynamic changes in blood oxygenation and blood volume provides an explanation for the form of the BOLD HRF (Buxton & Frank 1997; Buxton et al. 1998). This model is based on the premise that there is a strong coupling between blood flow and oxygen metabolism but that the biophysics of oxygen extraction determines that a large increase in flow is required to support a small increase in oxygen metabolism. This so-called balloon model, which assumes that the capillary volume remains fixed, but allows for an expandable volume in the venous part of the vascular bed, can predict the transient aspects of the BOLD response. Perhaps one of the most important aspects of such work in future will be an integrated model of the spatial and temporal behaviour of BOLD because recent studies have further highlighted the spatio-temporal complexity of the BOLD response. Using sustained visual stimulation, Chen et al. (1998) found that there was evidence for transient responses as seen by Frahm et al. (1996) in large vessels, but a sustained response in cortical grey matter. Also, measurement of the delay between the onset of haemodynamic change and when maximum signal change occurs has shown that the capillary-based signal reaches a maximum several seconds before that from larger vessels (Lee et al. 1995).

The broad BOLD HRF imposes a limit to the temporal resolution of fMRI in detecting individual brain events, although the ultimate resolution which may be obtained with good experimental design may be rather better than may be expected from an HRF of several seconds lag and width. In the same brain region the detectability of discrete responses is no better than 3–5 s (Bandettini 1993; Kim et al. 1997). However, differences in the onset of activity in different regions are detectable to approximately an order of magnitude better than this if some signal averaging is used (DeYoe et al. 1992; Hykin et al. 1995; Savoy et al. 1995). The measurement of differences between regions should be treated with some caution, however, because they are likely to reflect different latencies in the HRF as well as differences in the timing of neuronal activity. It has been shown that BOLD contrast can be generated by stimuli as short as a few tens of milliseconds (Savoy et al. 1995). The differences between responses to different sorts of single events can be measured to an accuracy much greater than the HRF width. By comparing the responses to words which were either novel or had previously been seen, Friston et al. (1998b) have shown that fMRI can discriminate between dynamics in a single region on a 100 ms time-scale.

In experiments which aim to image functional segregation of relatively large cortical areas, the current limited understanding of the degree to which BOLD contrast accurately reflects the spatial extent of neuronal activity is perhaps unimportant. The BOLD technique yields signal changes located at least within 10 mm of the site of neuronal firing, which is quite adequate in many cases. However, in experiments in which higher functional resolution is required then difficulties may be encountered in generating contrast with sufficient resolution. An example of this is a study which attempted to image the functional differences between the strip and interstrip activity in area V2 in human visual cortex but found that $T_2^*$ contrast was not sufficiently specific to yield functional differences (Thulborn et al. 1997). In this same study there is evidence that, by using $T_2^*$ contrast to be preferentially sensitive to the microvasculature, higher functional resolution in the fMRI data can be obtained. This interesting prospect will warrant extended study as more investigations are performed on high-field systems.

3. fMRI Techniques

Several high-speed imaging methods have been used for fMRI, most of which are based on the principles of EPI. Although to some extent the diversity of techniques currently being used at different sites reflects the sequences

Phil. Trans. R. Soc. Lond. B (1999)
provided by the different system manufacturers, a range of imaging options is desirable and there is no single optimal imaging method for all fMRI experiments. Although relatively high speed is a requirement for fMRI, this leads to several trade-offs which have to be made and ultimately some temporal resolution may be sacrificed to improve other aspects of image quality. The major technical difficulty in the generation of EPI images is the switching of very large currents through the system's gradient coils.

The imaging technique that is most commonly used is multislice EPI (for a comprehensive review of EPI, see Schmitt et al. (1998)). Figure 2 schematically shows how image volume data are acquired in an epoch-based fMRI experiment using multislice EPI. Each image is acquired by first selectively exciting an NMR signal from a thin slice of the head, using a shaped RF pulse and a slice selection field gradient, and subsequently sampling the evolution of this signal in the presence of a rapidly switched magnetic field gradient. The switched gradient waveform generates a series of echoes of the NMR signal, which are progressively phase encoded by the action of a second orthogonal field gradient, applied as a series of 'blips' occurring at the zero crossings of the switched gradient. The set of echoes formed in this manner contains all the information necessary to construct a two-dimensional image of the slice. The image matrix size is set by the product of the number of echoes acquired and the number of data points sampled in each echo, typical values being $64 \times 64$ for fMRI experiments. Using the $k$-space formalism, in which the imaging sequence is viewed in terms of the measurement of a range of spatial frequencies ($k$) describing the object structure, the sequential acquisition of these echoes corresponds to a raster-like scan through $k$-space. Data at the centre of $k$-space, whose magnitude largely defines the image contrast, are acquired at a TE whose minimum value is set by the time needed to generate half the total number of echoes. The sequence of the form shown in figure 2 is known as gradient echo EPI, as there is no refocusing of the dephasing effects of magnetic field homogeneities. Gradient echo EPI images thus naturally show the $T_2$ weighting which is most sensitive to the level of blood oxygenation. Examples of images acquired using this method are shown in figure 3. The images have a $256 \times 128$ matrix size and are of higher resolution than usually achieved with EPI (data courtesy of Dr Steve Roberts, SMIS Ltd, Guildford, UK). These very high quality EPI images have been acquired using a head gradient and RF coil insert in a 1.5T magnet. Using a smaller diameter gradient coil means that higher gradient strengths and faster switching times can be achieved. The total acquisition time for each image in figure 3 is $130$ ms.

Volume data are built up by repeating the process of image acquisition at different slice positions. The total volume acquisition time is given by the product of the number of slices and the time between slice acquisitions. Whole-brain coverage can be achieved in approximately $5$ s. Repetition of this process allows the acquisition of a number of data volumes during the different states of a functional paradigm as shown in figure 2. The time between repeated acquisition of the same slice (TR) controls the effect of $T_1$ relaxation on the image intensity, with short TR values giving reduced signal, and greater likelihood of the manifestation of $T_1$-related in-flow effects.

Many early fMRI studies were performed using fast gradient echo, Fourier imaging methods such as FLASH (Haase et al. 1986) because EPI was not so widely available as it is today. These techniques can still find application because higher resolution images can be generated than with EPI, but this is at a cost of an increased time for the acquisition of each image. With these techniques, rather than scanning all $k$-space after a
single RF excitation, multiple excitations are used, with the data required for image reconstruction being acquired line by line. The elevated acquisition times of these techniques lead to a reduction in the temporal resolution with which the time-course of the haemodynamically mediated signal change can be monitored and therefore they are not compatible with the 'event-related' fMRI experiments (Josephs et al. 1998a) that are currently coming into wider use.

Very high resolution structural MRI scans with good grey–white matter contrast are usually acquired as part of an fMRI study so that the EPI data can be overlaid onto images showing cortical structure. These high-resolution images are also usually normalized into a standard brain space (Friston et al. 1995) so that Talairach coordinates can be reported for functional activations. For studies involving cortical areas which are close to large draining vessels it is sometimes also useful to acquire MR venograms, which are images consisting only of slow flow in blood vessels (Haacke et al. 1994). Where very high functional resolution is required, such as in somatosensory cortex, an accurate delineation of the local vasculature can prove useful when determining the precise location of functional signal change.
(a) **Artefacts in fMRI images**

A particular disadvantage of using any $T_2$ imaging method, sensitive to microscopic field inhomogeneities to generate image contrast, is that the technique will also be sensitive to macroscopic field inhomogeneities such as those found at tissue, air and bone interfaces. The difference in the magnetic susceptibility of these substances leads to localized magnetic field gradients being established which are inherent to the anatomical structure of the head. In the human brain these gradients are largest in the frontal pole above the air-filled sinuses, in the inferior temporal lobes where there are complex

---

*Phil. Trans. R. Soc. Lond. B* [1999]
Figure 4. Two slices from an inversion recovery (1200 ms) data set acquired in the coronal orientation at 3 T.
(a) Two-shot interleaved EPI acquisition. (b) Single-shot EPI acquisition. In both cases voxel dimensions are
3 mm × 3 mm × 3 mm, matrix size 64 × 64 and gradient switching frequency 1.9 kHz. The activation shown resulted
from finger tapping which was visually cued using a checkerboard pattern reversing at 4 Hz. It is clear from these two
slices showing motor and visual cortex, that the distortion in the two-shot images (a), which have an effective frequency
per point of 60 Hz, is less than that in the single-shot images (b) which have a 30 Hz point separation in the ( medio-lateral)
phase encode direction.

boundaries between tissue, the petrous bone and the air passages of the ear and at the level of the cerebellum and
brain stem. The consequences of these localized field gradients, which operate over several voxels in the
images, are a loss of MR signal due to spin dephasing across the width of a voxel and a relocation of signal due to
the additional magnetic field gradient, i.e. image distortion. An example of geometric distortion can be seen in the frontal lobe in figure 3c. The loss of signal on $T_2^*$-weighted images is directly related to the TE. This effect can be seen in the anterior part of the temporal lobe in figure 3d where the TE is 60 ms. By shortening the TE this signal loss can be minimized, but at the cost of a reduction in the sensitivity to BOLD contrast. By using spin echo images this signal loss can be eliminated, although, as discussed earlier, the resultant $T_2$ contrast is less sensitive to brain activity. A compromise between these conflicting requirements is to use a technique known as asymmetric spin echo EPI in which $T_2^*$ contrast with a shorter effective TE can be achieved (Stables et al. 1998).

The geometric distortion in EPI images is very severe because of the long read-out time after each RF pulse.
The problem is insignificant with techniques such as FLASH because each read-out time is short. The form of the
distortion is strongly dependent on the orientation in which the images are acquired. In figure 4 severe distortion
in the brain stem is seen in coronal EPI images acquired at 3 T. Geometric distortion affects spin echo EPI in a similar way as in gradient echo EPI. Shorter EPI acquisition times reduce the distortion, but these are limited by the gradient specifications of the system. Post-processing methods to reduce geometric distortion using pre-acquired field inhomogeneity maps have been proposed (Bowtell et al. 1994a; Jezzard & Balaban 1993; Reber et al. 1998) although these have yet to find widespread use, perhaps because the severity of this problem in fMRI is still not widely recognized. However, as the demands for ever more precise localization of activity increase, then the removal of geometric distortion from the EPI data when co-registering onto the structural images will become very important.

Another artefact which is characteristic of EPI data is a low intensity ghost image which appears halfway across
the image field of view. The source of this so-called Nyquist ghost is that during EPI processing the odd and
even echoes need to be time reversed with respect to each other because of the polarity reversal of the imaging
gradient. This time reversal introduces artefacts due to imperfections in the timing of the experiment, the
generation of eddy currents, nonlinearities in the receiver
filters and field inhomogeneity (Schmitt et al. 1998). Removal of the Nyquist ghost during post-processing
reduces its level to typically around 5%. An additional complication is that the ghost level may change during
the acquisition of a time-series of images and it has been shown that with the confound of stimulus correlated
motion (see § 4) false activation due to correlated fluctuations in the Nyquist ghost can occur (Grooten et al.
1998).

(b) Choice of imaging parameters in experimental design

Among the important considerations when selecting an imaging method for fMRI are speed, spatial resolution,
signal-to-noise ratio (SNR), level of BOLD contrast and minimization of false activation artefacts. Additionally
the design of the experiment in terms of its temporal characteristics (both timing of stimuli and overall
duration) and the extent of brain coverage required, may affect the choice of imaging method. If whole-brain
coverage is required then some form of EPI method is essential, the only appropriate use of (low flip angle)
FLASH methods being where a high-resolution study of a known activation focus is required. For fMRI using $T_2^*$ contrast the TE should be set to approximately the value of the $T_2^*$ of grey matter to maximize BOLD contrast, although the statistical significance of activation is a broad function of TE (Yang et al. 1998b). The slice thickness in multislice studies is another important parameter. The signal loss due to susceptibility gradients described above is reduced if thinner slices are acquired. However, very thin slices (<1 mm) have low SNR and the number of slices needed to cover the entire brain becomes excessive. In a recent study (Howseman et al. 1999), it has been shown that in an area of the brain where there are no susceptibility artefacts, using broader slices (5–8 mm) provides optimal sensitivity per unit time. In the same study there is also evidence that using small gaps between slices makes the technique less sensitive to the effects of $T_1$ saturation when there is subject motion. Again in the

Phil. Trans. R. Soc. Lond. B (1999)
same study, the order in which the slices are acquired has been shown not to affect the data and so the most important aspect of slice order is how the timing of stimuli should be presented so as to avoid differential sensitivity in different parts of the brain based on the timing of the experiment alone (Price et al. 1999).

(c) **Variants on the EPI technique**

As described above the most widely used fMRI method at present is multislice EPI with either contiguous or interleaved slice acquisition. There are, however, several variants on EPI which have been used for fMRI and can potentially offer advantages of some form.

(i) **Segmented EPI**

One of the biggest limitations in EPI is (by MRI standards) the relatively low spatial resolution. For fMRI performed on systems using a body gradient coil, typical EPI matrix sizes are 64 × 64 with a resolution of 3 mm. Image matrices of 128 × 128 can be generated, with voxel sizes of 1.5 mm, but these require a long sampling time and lead to a higher degree of distortion. An alternative approach to improving the spatial resolution is to use two or more RF pulses and acquire a subset of the total number of required k-space lines after each. Typically the k-space lines would be interleaved because this minimizes artefacts when the sets of lines are conjoined and Fourier transformed to generate the final image. In figure 4 a reduction in geometric distortion when using segmented EPI compared with single-shot EPI can be seen. Two-shot segmented EPI has also been used to generate very high resolution 256 × 256 matrix images at 3 T where the read-out time needs to be kept short (Clare et al. 1996). Images from this study are shown in figure 5, in which very localized functional activation is seen with a visual stimulus presented in a small fraction of the right visual field. It should also be noted that the principle of k-space segmentation can be applied to any of the other techniques discussed below.

(ii) **Three-dimensional (3D) phase-encoded EPI**

An alternative approach to imaging the entire brain is to make use of the principle of phase encoding in the third dimension rather than acquiring data one slice at a time. This means that data are sampled from the entire 3D volume simultaneously but that many RF excitations of the system are necessary to build up a complete 3D set of k-space data, which is then processed using a 3D Fourier transform. An image matrix with N points in the third dimension can be acquired in the same time needed to image N slices in multislice EPI. In principle the technique has advantages because higher SNR can be obtained. The $T_2$ contrast between grey and white matter can also easily be manipulated, which may be advantageous in image realignment. A comparison between this method and 2D multislice imaging (Porter et al. 1997) showed similar activation images. However, through the use of navigator echoes it should be possible to achieve an improvement in the significance of the fMRI data with the 3D method and reduce their dependence on motion (Ramsey et al. 1998). Additionally, if high enough resolution is used, it should be possible to reduce signal drop-out using 3D EPI (Wang et al. 1999).

(iii) **Spiral scanning EPI**

In the spiral version of EPI, k-space is sampled in a spiral rather than on a rectilinear trajectory. This method makes less demands on the gradient hardware and is rather less sensitive to some artefacts, in particular there is no Nyquist ghost. Field inhomogeneity causes image blurring rather than geometric distortion. The technique can be applied in either two dimensions, in which case multislicing as in conventional EPI is used, or as a 3D spiral in k-space. A comparison of both of these methods with conventional multislice EPI for functional imaging has been performed (Yang et al. 1998b) and this suggests that the temporal stability of the fMRI time-series can be better with spiral EPI. One disadvantage of the spiral techniques is that the data points in k-space have to be

*Phil. Trans. R. Soc. Lond. B (1999)*
regridded, which slows down the image processing. Another version of EPI which has been evaluated is a 2D spiral with phase encoding in the third dimension, which improves the signal drop-out problem and gives higher contrast to noise in the functional images (Yang et al. 1996; Lai & Glover 1998). Segmented spiral techniques with navigator echoes have also been successfully used for fMRI (Glover & Lai 1998). 

(iv) Echo-planar imaging

A final method worthy of mention is echo-planar imaging (EPI) (Mansfield et al. 1988; Harvey & Mansfield 1996), which acquires all the data required to generate a volume image of the brain in a time of approximately 100 ms. EPI is technically very demanding and the use of a dedicated head gradient coil is very advantageous. It is unlikely to ever result in image matrices much larger than $32 \times 32 \times 32$; however, it does have the capacity to image BOLD contrast in the entire brain simultaneously. An example of the potential of this technique performed at 3 T (Hykin et al. 1995) is shown in figure 6. The data illustrate how EPI can be used to measure brain activity and has the potential to simultaneously monitor the dynamics of functional change over large cortical volumes.

4. LIMITATIONS OF fMRI METHODS

As a technique for imaging human brain function, fMRI has already demonstrated several advantages over PET and in the coming years $H_2^{15}O$ activation studies with PET may become redundant. The improvements in spatial and temporal resolution and sensitivity with fMRI are important, but probably the most significant advantage of fMRI is the absence of ionizing radiation. There is no limit with fMRI on the number of scans that a subject can undergo. This allows repeated measurements to be made, facilitating longitudinal studies over time and allowing the assessment of the reproducibility of activation images (Noll et al. 1997; McGonigle et al. 1999). Although it is possible that fMRI may largely replace PET for activation studies there are a number of limitations in using fMRI and currently there are certain experiments which can be better performed with PET. In §3(a) one of the most important limitations was discussed, the effect of field inhomogeneity rendering small volumes of brain tissue invisible. For example in the inferior temporal lobe, geometric distortion and signal drop-out are so severe on $T_2^*$-weighted EPI images that some language studies which have been successful with PET are difficult to reproduce with fMRI. Other than this difficulty, which is inherent to the imaging itself, other potential limitations are (i) the somewhat confined environment in which subjects lie in an MRI scanner; (ii) the need to operate other equipment in a large static magnetic field; (iii) acoustic noise generated by the imaging experiments; (iv) subject movement and physiological artefacts; and (v) possible peripheral nerve stimulation by the imaging gradients.

The internal diameter of a typical MRI scanner is 60 cm and the bore is between 1.5 and 2 m long. The subject’s head is placed in an RF coil with a diameter and length of 30 cm. Many fMRI studies involve the presentation of a visual stimulus and in some cases video presentation is required. This is commonly achieved using a projection system placed outside the magnet room, directed onto a screen which the subject views via a mirror. If the RF coil has an aperture above the eyes this system is very successful. However, particularly at higher fields, it is difficult to produce such open designs and it is necessary to make use of some form of goggles, either using prisms or optical fibre technology. Other types of stimulation apparatus need to be built to operate in a magnetic field. The restricted access of the MRI scanner also makes the performance of motor tasks more difficult, especially so because it is imperative that the subject does not move their head in tandem with the task.

EPI generates a loud ringing noise of about 100 dB SPL at frequencies of 1–2 kHz due to the switching of large gradient coil currents (up to 500 A) in the static field. For the subjects’ protection it is imperative that ear defenders or plugs are worn during an fMRI experiment. The sound generated by the scanner means that auditory stimulation experiments are more difficult than with PET. By piping sound through ear defenders which provide about 20 dB attenuation, subjects can hear stimuli over the scanner noise. The sound created by the scanner is constant throughout the duration of the experiment and so the resultant activation of auditory cortex is largely cancelled out when a comparison between experimental conditions is made, although there is some evidence of acoustic masking of activation by scanner noise (Shah et al. 1997; Bandettini et al. 1998). This problem can be overcome at a cost of reduced data acquisition rates, by acquiring images in each volume very rapidly and then leaving gaps in time between clustered volume acquisitions (Hall et al. 1999; Eden et al. 1999). The use of a head-only gradient coil can reduce the noise level considerably.

The problem of subject movement during scanning has bedevilled fMRI since its earliest days, leading to the suggestion that many apparent activation results were the result of motion artefact (Hajnal et al. 1995). Today it is accepted that this is generally not the case, although the problem of motion artefacts is an important one. Even after precautions are taken to minimize head movement through the use of head restraints, head movement of 1–2 mm during scanning is common. In uncooperative patients it may be very much worse. Post-processing realignment methods (Friston et al. 1995) correct for rigid body translations and rotations of the head. However, these methods cannot account for all of the effects of movement. Errors in the realignment occur due to the limited width of the interpolation kernels which are constrained by computing time (Ashburner et al. 1997). An additional problem is the effect of $T_1$ relaxation, which can cause intensity fluctuations if there is movement during the scanning of the brain volume (Friston et al. 1996). This effect manifests itself differently in different imaging techniques and 3D EPI methods may not have the sensitivity to very small movements which multislice methods have. Even after correction, low-frequency drifts in the voxel time-course in fMRI data sets are quite common, although these may be largely due to very subtle fluctuations in magnetic field homogeneity (Smith et al. 1998). The most severe problem of motion arises in experiments which involve a motor task. This is because although only very small movement of the head may
result from hand movements for example, this movement may be highly correlated with the experimental design. This can lead to entirely false activations because changes in image intensity are then highly correlated with the comparison specified in the analysis. These apparent activations are typically found at sharp image boundaries and in areas of field inhomogeneity. Speaking or swallowing during the experiment introduce particularly severe movement-related problems. Even if the experimental timing is such that the verbal response falls during an interval between the acquisition of images it is unlikely that the subject’s jaw will return to the same place and because of the complexity of tissue, bone and air-filled cavities in this part of the head, artefacts due to magnetic field changes in inferior brain structures can be very severe (Birn et al. 1998). Whether the problem of speaking during an fMRI acquisition can be overcome remains to be seen.

Improvements in the speed of image acquisition using EPI will inevitably involve the use of larger switched magnetic field gradients, but these are ultimately limited in strength by the need to avoid stimulation of nerves in the subject’s body (Schmitt et al. 1998). Such stimulation results from the voltage, and hence current, induced in tissue as a result of the rapidly changing magnetic fields. The allowable rate of change of field (dB/dt) is limited by statutory guidelines to levels that largely prevent stimulation of peripheral nerves and which are consequently much less than the higher threshold at which stimulation of cardiac muscle might occur. Current fMRI experiments using EPI operate below these statutory levels, although modern gradient system technology will allow switching rates above the threshold. For a given gradient strength the peak dB/dt values in the body are much lower when using a head gradient insert, rather than a whole-body gradient coil, and such coils therefore offer the best route to reduced image acquisition times.

5. FUTURE DIRECTIONS OF fMRI METHODS

Most fMRI experiments are currently carried out at field strengths of 1.5 T, although there are a significant number of systems operating at higher magnetic fields of between 3.0 and 4.0 T (approximately 20 to 30 worldwide). Future developments of fMRI techniques and their applications are likely to be concentrated on these high-field systems which offer the potential of increased BOLD contrast as well as greater sensitivity in all MR imaging and spectroscopy experiments. New techniques are also being developed to provide more direct measurements of brain activity than BOLD contrast. These include spin labelling methods (see § 5(b)) and the use of hyperpolarized xenon to measure perfusion (Albert et al. 1994). Using optical pumping it is possible to massively increase the polarization of xenon (and helium), which can subsequently be introduced as a blood tracer with potentially high sensitivity. In vivo NMR spectroscopy allows the measurement of the concentration of brain metabolites and this technique has been applied to the measurement of metabolic flux during functional activity (Shulman et al. 1993). It has further been suggested (Shulman & Rothman 1998; Rothman et al., this issue) that the measurement of glutamate to glutamine neurotransmitter}

\( ^{13} \text{C} \) NMR spectroscopy allows an interpretation of fMRI studies more closely related to neurobiological processes rather than purely in psychological terms. Measurement of diffusion anisotropy via MRI-based diffusion tensor imaging allows the measurement of nerve fibre tract orientation in the brain (Toe et al. 1997). This potentially could lead to the non-invasive mapping of connectivity between cortical areas and illuminate links between brain structure and function. Preliminary studies using diffusion tensor imaging in conjunction with fMRI have been performed in the visual system (Werring et al. 1998b). Furthermore, in a study of motor recovery, fMRI activation has been correlated with preserved diffusion anisotropy (Werring et al. 1998a) suggesting that cortical pathways have been spared within the vicinity of a traumatic focal lesion.

The future should also see further integration of fMRI with other modalities measuring brain activity. The acquisition of EEG data while a subject lies in an MRI scanner has already been demonstrated (Ives et al. 1993; Lemieux et al. 1997; Allen et al. 1998) and offers the prospect of combining the excellent spatial resolution of the fMRI technique with the millisecond temporal resolution of EEG. The fMRI activation associated with interictal epileptic spikes has already been measured. The sensitivity of the electrophysiological signal appears sufficient to combine fMRI with normal event-related brain potentials. Transcranial magnetic stimulation has also been combined with fMRI (Bolting et al. 1998; Josephs et al. 1998b). This will allow the direct determination of the haemodynamic response resulting from transcranial magnetic stimulation-evoked neuronal activity and for transient dysfunction to be introduced during the course of an fMRI experiment.

In current fMRI studies the processing of the data is performed after completion of the scan on a computer remote from the MR imaging system. Typically it may take several hours to perform full data processing including image realignment. However, with the most up to date computer technology it is possible to perform online fMRI data processing if some of the stages of processing, such as image realignment, are removed (Cox & Jesmanowicz 1998). This opens up the opportunity to perform fMRI experiments in real time, which may have applications in presurgical planning for example, where the neurologist could monitor a patient’s responses and adapt the stimulus presentation while the subject remains in the scanner.

(a) The use of high field strength (> 3 T) for fMRI

Higher field strength in MRI offers greater intrinsic SNR, which increases approximately linearly with field strength, and in many circumstances this increase can be translated into better spatial or temporal resolution in imaging and greater sensitivity in spectroscopy (see Gadian 1995). The best field strength for implementing fMRI experiments has been the subject of some debate. The ratio of BOLD signal change to noise that can be generated in an fMRI experiment depends upon the fractional change in the \( T_2^* \) relaxation rate (\( \Delta R_2^* \)) occurring during activation, as well as the base image SNR, and is maximized when TE is approximately \( T_2^* \). Simulations predict that \( \Delta R_2^* \) should vary approximately
quadratically with field strength for small vessels and linearly for large vessels (Fisel et al. 1991; Ogawa et al. 1993), with experimental measurements in grey matter displaying a field dependence falling between these two extremes (Turner et al. 1993; Bandettini et al. 1994a). In measurements on visual cortex the ratio of $\Delta R_2^*$ to $R_2^*$ has, however, been found experimentally to vary more slowly with field strength (Gati et al. 1997), but in conjunction with the linear increase in SNR this variation does lead to a hyperlinear increase in BOLD contrast-to-noise ratio with field strength, which makes high-field fMRI advantageous. The greater sensitivity to the capillary-based component compared with the effects of larger vessels (Gati et al. 1997) is an important benefit of working at high field. Ugurbil et al. (this issue) describe applications which are enabled by operation at high field.

It is not always possible, however, to achieve the predicted increase in contrast or to acquire data satisfactorily from the entire brain, because the increased large-scale susceptibility effects which occur at high magnetic field can prevent translation of the increased signal strength into improved image SNR and can limit the usable echo time to sub-optimal values. The use of asymmetric or standard spin echo techniques, which at high field can give sufficient BOLD contrast with less or no signal drop-out offers a way around this problem. The specific problems relating to EPI at high field have seen considerable efforts already (Bowtell et al. 1994b; Kim et al. 1996; Hu & Le 1996; Yang et al. 1997a; Jesmanowicz et al. 1998) and particularly on the very high-field systems of 4.0 T and above, the optimization of EPI represents an important and very challenging hurdle if fMRI is to reap the potential benefits of the higher field. At high field a technique known as BURST imaging (Hennig & Hodapp 1993) may be useful in fMRI (Jakob et al. 1998) because BURST has the attractive feature of being a quiet imaging method and also is much less sensitive to image distortion than EPI. Its primary disadvantage is a relatively low SNR but at high field it may be advantageous for auditory studies.

In the light of the above, the recent successful production of human head images at the exceptionally high field of 8 T (Robitaille et al. 1998) is a very important development. While there are clearly a number of issues relating
to the penetration depth of higher frequency RF into the head and RF power deposition, as well as the increased susceptibility effects and the issue of the safety of exposure to very high static magnetic fields, it seems almost certain that the drive towards ever higher field strength for biological NMR will continue.

(b) Measurement of tissue perfusion with MRI

The direct measurement of cerebral perfusion has been addressed by a number of subtly different but related MRI techniques in recent years. These are referred to as arterial spin labelling (ASL) methods and involve the tagging of spins in a slab of tissue and measuring, in an imaging slice, their transfer from the circulation into brain tissue. The tagging of spins can be achieved by altering their magnetization state non-invasively, an advantage over other methods which use contrast agents.

These ASL methods (e.g. EPISAR (Edelman et al. 1994), FAIR (Kwong et al. 1994), QUIPS (Wong et al. 1996)) are all subtraction methods where the signal differences in the imaging slice are between tagged and untagged measurements. Measuring small differences on a large baseline signal is inherently subject to many sources of error. The signal changes produced by all of these methods depend upon the transit times of the tagged blood into the imaging slice and the $T_1$ relaxation rates of blood and brain water to generate a value for CBF. Therefore the accuracy of the CBF measurements and the specificity of the functional maps of CBF change may not be very high. However, these techniques show potential for making (MRI) quantitative and it has been shown that similar functional maps can be achieved with perfusion methods as with BOLD (Yang et al. 1998a; Zhu et al. 1998), albeit with longer acquisition times and lower SNR. Considerable effort is currently being expended to refine further these methods and overcome the experimental and theoretical difficulties associated with them. It seems unlikely that ASL methods will become generally used for functional imaging studies because of their lower sensitivity but these techniques are already proving valuable in cross-validating the measurement of brain activity with PET and fMRI. In a recent study by Davis et al. (1998), perfusion-sensitive MRI and the breathing of carbon dioxide were used to calibrate BOLD contrast. This allows the measurement of changes in oxidative metabolism, and using a checkerboard stimulus it was found that oxygen consumption increased by 16% whereas blood flow increased by 45%, confirming the differential increase in rCBF and rCMRO$_2$ but demonstrating that oxidative metabolism is a significant component of functionally induced metabolic activity.

In concluding this review of the MRI methods which are being used in the measurement of human brain activity we hope that we have left the reader with a sense of the direction in which these techniques are heading. During its lifetime, MRI has never failed to amaze and delight its practitioners with the new applications and techniques which it holds in store. The use of MRI in functional neuroimaging has been one of its most spectacular successes and there are surely further surprises ahead about the ways in which we are able to harness this quantum physics phenomenon to explore the workings of the human mind.

REFERENCES


Bowtell, R., Mansfield, P., Coxon, R. J., Harvey, P. R. & Glover, P. M. 1994b High-resolution echo-planar imaging at 3.0 T. MAGMA 2, 241–245.


Phil. Trans. R. Soc. Lond. B (1999)


Hykin, J., Coxon, R., Bowtell, R. W., Glover, P. & Mansfield, P. 1995 Temporal differences in functional activation between
separate regions of the brain investigated with single shot EPI. In *Proceedings of the 3rd meeting of ISMRM, Nice*, p. 451.


