Q & A for fMRI basics

Summarized by jerryzhu@siu.edu

Jul 2013

References/Acknowledgements:

1. Handbook of Functional MRI Data Analysis
2. Functional Magnetic Resonance Imaging
3. Multimodal Neuroimaging Training Program

Q1: Describe the hardware for MRI. How is MRI signal collected from a scanner? What is the difference between the T1, T2 and T2* decay? Taking the 2-D gradient-echo pulse sequence as an example, how are gradient fields applied to acquire spatial information to reconstruct images?

The three major components of an MRI scanner are the static magnetic field, radiofrequency coils, and gradient coils.

Nearly all MRI scanners today create their static magnetic field through electromagnetism. They also use liquid helium (usually) to cool down the temperature of wires to achieve superconductivity, thus requiring little electricity. Note that the static magnetic fields in MRI scanners are always active even when no images are being collected.

Radiofrequency (RF) coils are turned on and off during scanning to send and receive electromagnetic fields at the resonance frequency of the atomic nuclei within the static magnetic field.

Gradient coils superimpose magnetic gradients onto the strong static magnetic field, causing the MR signal to become spatially dependent in a controlled fashion. Typically they are turned on briefly after the excitation process.

Additionally shimming coils produce high-order compensatory magnetic fields to correct for the inhomogeneity of the magnetic field. The shim fields are adjusted once for each subject and then left on for the duration of a scanning session.

Typically one computer coordinates all the above hardware components according to pulse sequences, and another computer reconstructs and analyzes images.

Finally computers are also used to present stimuli to subjects and collect their responses. Many scanners also have equipment dedicated to monitoring/recording physiological measures like heart rate, skin conductance.

All matter is composed of atoms, each of which has an atomic nucleus and a cloud of electrons. An atomic nucleus in turn has protons and neutrons (with the exception of the hydrogen atom that has only one proton and no neutron). The atomic nucleus that has odd-numbered atomic mass (e.g. 1H, 23Na) exhibits nuclear magnetic resonance. In a normal condition, such nuclei spin because of thermal energy and the spins are in random orientations and tend to cancel out each other. Within the strong static magnetic field, the spins would precess around an axis that is either parallel to the magnetic field (most nuclei stay in this low-energy state) or antiparallel to the magnetic field (high-energy state) at a frequency (i.e. Larmor frequency), which is a function of the static magnetic field B0 (for a specific scanner, it is fixed and cannot be changed, e.g. either 1.5T or 3T) and the nucleus of interest for an MRI study (usually it is 1H but could be different). Again, the precessions are in different phases and tend to cancel out each other.

The RF coils apply energy to the nuclei that can pertubate spins to flip to a higher energy state (transverse plane). The flipping angle is determined by the nature of nucleus, RF magnetic field strength B1 (which is orthogonal to B0 and is far smaller than B0), and the duration of RF. Note that spins only respond to RF at a frequency matched to the Larmor frequency. The frequency of RF is built into RF coils, which could be replaceable. This process is known as excitation.

Then the RF is turned off. Flipped spins cause a change in flux in a transverse receive coil; flux change in turn induces a voltage across the coil, thus leading to MRI signal. The detected signal is also at the Larmor frequency. The MRI signal does not last forever; it decays over time, generally within seconds. This process is known as relaxation. Two primary mechanisms contribute to the loss of MRI signal: on one hand, spins gradually lose the energy absorbed during the excitation and return to the direction of B0, causing the recovery of longitudinal magnetization. The time constant associated with this longitudinal relaxation is called T1; on the other hand, flipped spins initially precess at about the same phase. Over time, the coherence of spins is lost and they gradually de-phase. As a result their collective contribution into the transverse detector diminishes. The cause of this transverse relaxation could be due to spin-spin interaction and field inhomogeneity. The signal loss by the spin-spin interaction only is called T2 decay, and loss by combined effects of spin-spin interaction and field inhomogeneity is called T2* decay. The shapes of T2 and T2* decay curves are similar but T2* is always faster than T2. Note that T1 and T2 are due to independent processes and generally T1 > T2.

Spatial information is acquired by the application of magnetic field gradients. In 2-D gradient-echo pulse sequence, a gradient field in the longitudinal direction, Gz, is applied at the

same time of a RF pulse to select a slice in a brain volume. Then Gy gradient is turned on to change the phases of precessing spins. Finally Gx gradient changes the frequency of spins when we read out signals. Each RF pulse fills a line in the k-space and the above process is repeated in the direction of ky for the number of times equal to the number rows in an in-plane image.

After k-space is filled, a 2-D inverse Fourier transformation is performed to convert raw data from k-space to image space, thus completing a single slice image collection. The process is then repeated to a whole brain volume.

Q2: What are the physiological changes underlying fMRI? What is the BOLD contrast in fMRI? Describe the shape of the canonical hemodynamic response function and its linear property.

A cognitive activity is realized through firing of neurons and activities of glial cells that are supporting cells. This neural activity requires energy in the form of ATP to restore the action potential, regenerate neurotransmitters, maintain resting potential, and etc. Because the brain does not store energy, it must create ATP through the oxygen and glucose, both of which are supplied through increased blood flow.

Increased blood flow is believed to be initiated when active neurons release substances (e.g. NO) that diffuse to the extracellular space and reach nearby blood vessels. These neurovascular-coupling substances cause the vessels to dilate and finally increased blood flow.

Blood vessels that are close to the active neuron are more sensitive and responsive in terms of dilation than distal vessels. In other words the blood flow changes and blood vessel changes are local/specific to the neural activity. This is good news to the fMRI spatial resolution.

So there is a connection between neuronal activities and blood flow changes. How is the blood flow change related to fMRI signal? The key role lies in the hemoglobin in the blood stream. In the artery blood, hemoglobin exists mainly in the form of oxygenated hemoglobin (Hb, i.e. binding oxygen molecules); when blood oxygen is extracted, it becomes deoxygenated hemoglobin (dHb). It has been found dHb distorts local magnetic field, but Hb does not. That means, dHb speeds T2* decay and decreases MRI signal. In fact, Ogawa (1990) showed that the gradient-echo images of the brains of animals breathing pure oxygen were different from those of animals breathing normal air, suggesting that the difference in signal on T2*-weighted images

is a function of the amount of dHb. This is called blood-oxygenation-level dependent (BOLD) contrast.

When oxygen and glucose are extracted at the surface of capillaries, changing Hb in the blood flow to dHb, the change seems to increase the concentration of dHb and one might expect a decrease in the MRI signal. However, because of the oversupply of fresh blood flow, it actually decreases the concentration of dHb and in turn decreases MRI signal loss due to T2* effects (i.e. increases MRI signal).

To summarize, when neurons become active, the activity results in an increase in blood flow. Because the blood is usually oversupplied, the extraction of oxygen from blood actually decreases the concentration of dHb. The decrease of dHb concentration in the cascade finally leads to an increase in MRI signal!

The change in the MRI signal triggered by neural activity is known as the hemodynamic response. Although neural responses occur within tens of milliseconds following a sensory stimulus, the response of hemodynamic change is sluggish. The shape of HRF features an initial dip (1-2s after the stimulus onset), a rise to peak (4-6s), a peak, a fall from peak (12-20s), and an undershoot (up to 20s or more).

The initial dip is due to early extraction of oxygen from blood and the accumulation of dHb in capillaries before a rush of blood flow, thus possibly providing more spatial specificity to neural activities. However it has also been argued that the initial decrease of MRI signal may reflect the initial flushing of dHb from arteries and transient concentration of dHb in the downstream venous system. If this latter model is true, the initial dip is not a better marker local to neural activities.

After a short latency, the metabolic demands of increased neuronal activity over baseline levels results in an oversupply of Hb, a decreased concentration of dHb, an increase in MRI signal, and finally a maximum signal intensity known as the peak. The height of the peak is the most common feature of interest, since it is most directly related to the amount of neuronal activity in the tissue. For BOLD fMRI, the maximum observed amplitude is about 5% for primary sensory stimulation, whereas signals of interest in cognitive studies are often in the 0.1–0.5% range.

Following cessation of neuronal activity, blood flow decreases more rapidly than blood volume, resulting in a relatively great amount of dHb and a MRI signal below the baseline level. This effect is known as post-stimulus undershoot.

Importantly, there is substantial variability in each of these features of the HRF across brain areas and across individuals. After all, different brain areas are irrigated by different blood vessels. Subjects with different ages and healthy conditions have variable hemodynamic responses.

Research has shown that the hemodynamic response is roughly a linear system. Boynton (1996) presented flickering checkerboard patterns with different visual contrast levels and durations. Higher contrast levels elicited larger amplitude of hemodynamic responses, and the response to a longer stimulus could be predicted by the sum of the response to multiple shorter stimuli. Dale and Bucker (1997) presented clusters of one, two, or three stimuli at interstimulus intervals of either 2 or 5 s. They found that the responses to the second and third trials in the set were generally similar to that of the first trial, especially under the 5-s interval condition. All these results suggest that HRF has the properties of scaling and superposition when multiple stimuli are presented in succession. Notice that the two experiments also confirm that if the

stimulus duration in a block design or the interstimulus interval in an event-related design is too short, a refractory period following stimulus presentation can decrease the hemodynamic response to the subsequent stimulus.

Q3: Explain each preprocessing step (i.e. slice time correction, motion correction, normalization and smoothing).

Most fMRI data are acquired using 2D pulse sequences that acquire images one slice at a time. The slicing order could be ascending or descending and interleaved or sequential. First, a descending order is preferred to an ascending one, because if an ascending acquisition is used, the blood flowing up into the brain will be repeatedly excited and saturated.

Second, both sequential and interleaved acquisitions have advantages and disadvantages. In reality, slices are not perfectly rectangular, and so there is some overlap between spatially adjacent slices. In a sequential acquisition, one slice is acquired just after its neighbor, and the overlapping part is excited twice in quick succession. This may lead to some "saturation" in the overlapping region and resulting loss of signal. The saturation issue can be avoided in an interleaved acquisition in that there is a longer interval between the acquisitions of spatially adjacent slices. However in case of a transient large head movement, a sequential acquisition could have one slice affected at the time of movement; the rest slices remain fine; instead an interleaved acquisition could have problems in a series of slices at and after the movement.

Regardless of the slicing order used, each slice is acquired at a different time point within the TR. Further fMRI statistical models assume all slices in a brain volume are collected at the same time. Timing differences are especially problematic for interleaved sequences, in which spatially consecutive slices are not acquired successively. Slice timing correction is also more important for event-related designs than for block designs; the former depend upon accurate modeling of the timing of experimental events, whereas the latter measure changes in BOLD activity over long intervals so that slice-timing correction is less critical.

To perform slice-timing correction, one can choose a reference slice and then interpolate the data in all other slices to match the timing of the reference slice. The reference slice could be the first acquired slice or the slice collected at 1/2TR or an average slice. If you have a structure you are interested in a priori, say, hippocampus, it may be wise to choose a slice close to that structure, to minimize any possible interpolation errors. Linear or polynomial interpolation is less accurate than sinc or Fourier interpolation. As an alternative to the method of interpolation based on a reference slice, one can also add a temporal derivative to the GLM model so that each slice is compared to a time-shifted HRF.

Slice-timing correction is more effective for data acquired at relatively short TRs than for data acquired at long TRs (e.g. >3 s), though the need for accurate interpolation is greater at longer TRs because of the larger intervals between successive acquisitions.

Finally, doing slice-timing correction before motion correction could spread the influence of movement to many slices; instead, doing motion correction first could introduce slight timing uncertainty. In general for interleaved slice acquisition, slice-timing correction should be done first; for sequential acquisition, motion correction first.

Even the best subjects can still have head movement in the scanner due to breathing, for example. Also, most fMRI experiments are portioned into a number of relatively short runs to reduce subject fatigue and to overcome scanner drift. During the breaks between runs, subjects typically relax and talk to the experimenters, often resulting in considerable head motion. And many experimental tasks require subjects to make motor responses, which may in turn induce (task-related) head motion.

Head motion may cause a very large, abrupt change in MRI signal for voxels, especially those at the edge of the brain, or those around ventricles, resulting in artifactual activations outside the brain or in ventricles.

For motion correction, brain volumes in the time series are coregistered to a single reference volume using a rigid body transformation algorithm. The rigid body transformation assumes that the size and shape the imaged brains do not change throughout an experiment. This is a plausible assumption in fMRI studies, although inhomogeneities in the magnetic field may distort images. Therefore there are six parameters in this linear transformation: xyz translation and rotation around xyz axes, i.e. roll, pitch, yaw (note: scaling and shearing are also linear transformation but not used in the rigid body transformation). The goal is to use optimized estimation method to minimize a pre-defined cost function; as a result a set of translation and rotation parameters is determined as the likely amount of head motion. After that, spatial interpolation is applied to all the volumes with estimated motion parameters in a way similar to temporal interpolation described in slice-timing correction.

Regarding the reference volume, there does not seem to be any appreciable benefit of using a mean image rather than a single image. When using a single image as the reference, the image from the middle of the time series is preferred because the middle image should be the closest (on average) to any other images in the time series. Again sinc or Fourier interpolation is more accurate than linear interpolation. Alternatively, the calculated motion parameters can be included as a term in the GLM to remove head motion from the model. In this strategy, motion is estimated but not interpolated during the preprocessing.

Motion correction is generally more effective for small head movement (smaller than one voxel size). For large motion, a researcher may consider tossing out the subject's data.

In some cases fMRI data are collected from an individual with the goal of understanding that single person, e.g. in the case of planning a brain surgery to remove tumor. However, in most cases, we wish to compare different groups of subjects, e.g. patients and controls. For intersubject comparison and signal averaging across subjects to be feasible, each subject's brain must be warped into a common space because individual brains are highly variable in their size and shape. This process is known as normalization.

In order to conduct normalization, we need a common reference frame in which to place the different individual brains. Talairach space is defined by a set of anatomical landmarks: the middle point of anterior commissure is the origin; the plane along the anterior commissure and posterior commissure and orthogonal to the midline sagittal plane is axial plane; coronal plane is the one orthogonal to both sagittal and axial plane. The space also has bounding box delineated by the most extreme portions of the brain in each direction. A major problem is that there is no MRI scan available for this atlas. Instead researchers rely on other templates (e.g. MNI152) that have been aligned to the Talairach space using landmark-based registration.

First, we coregister each subject's functional images to the T1 weighted anatomical image because functional images are lower-resolution, lacking fine anatomical details. This process may involve automatic or manual skull stripping of the T1 anatomical image. Or in general, segmentation of the T1 anatomical image into gray matter, white matter, CSF, and skull can improve the coregistration.

Second, each subject's anatomical image is transformed to a template brain using 12-parameter linear transformation (6 rigid body + 3 scaling + 3 shearing) and nonlinear transformation.

Third, the estimated parameters from the first and second steps can be concatenated to normalize functional images.

It is worthy to note that one can normalize functional images directly to an EPI template, in which case the functional and anatomical coregistration are not necessary (i.e. bypassing Step 1 and 2 in the above method). This approach is largely driven by the high-contrast features at the edges of the brain, meaning that although the overall outline of the brain is accurate, structures within the brain may not be accurately aligned. Normalization can be done before or after statistical analysis; each preprocessing stream is valid. Furthermore, structural alignment does not imply functional alignment because of individual difference in physiology and functional organization. By normalization, small but otherwise meaningful variations among individuals' functional neuroanatomy may be lost. Investigators interested in individual differences may wish to consider alternatives to normalization. Finally data from such population as children, the elderly and patients may require special treatment in normalization due to their unique brain features.

Spatial smoothing involves the application of a Gaussian filter to spread the intensity at each voxel in the image over nearby voxels. It may seem incomprehensible that, having put so much effort into acquiring fMRI data with the best possible resolution, we would then blur the images by smoothing, which amounts to reduce spatial resolution. However there are a number of reasons researchers choose to apply spatial smoothing to fMRI data.

First, smoothing can increase SNR. By smoothing, one can average out high-frequency noises that are presumably independent across each voxel. In contrast, fMRI signals have spatial correlation, due both to functional similarity of adjacent brain regions and to blurring introduced

by the vascular system, and thus are less influenced by smoothing. As a result, the signal-noise-ration in the data is increased after smoothing.

Second, smoothing improves the validity of later statistical techniques. For example, the random field theory, a method to control the multiple comparison problem, requires a specific degree of spatial smoothness in the data. Smoothing also increase the normality of data, which are assumed by many common statistical tests.

Third, smoothing reduces the anatomical variability across subjects for group analyses in standard space.

The amount of smoothing imposed by a Gaussian kernel is determined by the width of the distribution, which is further described by the full width at half-maximum (FWHM) in millimeters. The larger FWHM, the greater smoothing. If the kernel size is too small, it has little effect on SNR; if the kernel size is too large, meaningful activation in smaller structures could be attenuated and lost, such as nuclei within the midbrain, where only a single voxel may be significantly active. It is recommended to use twice the voxel dimensions as a starting point for the smoothing kernel size.