MINIREVIEW

FUNCTIONAL NEUROIMAGING

Gregory S. Berns*

Department of Psychiatry and Behavioral Sciences, Emory University School of Medicine and Tech/Emory Biomedical Engineering Dept. & School of Psychology, Georgia Institute of Technology

(Received in final form July 12, 1999)

Summary

Functional neuroimaging represents an area of brain imaging that has undergone tremendous advancements in the last decade. It is now possible to design experiments that elucidate the functional interplay between brain regions that give rise to specific human cognitive processes. Positron emission tomography (PET) and functional magnetic resonance imaging (fMRI) form the core technologies that have allowed such studies. This article reviews the basis of these techniques, their strengths and limitations, the underlying neurophysiology, and the future of functional neuroimaging.

Key Words: PET, fMRI, cognitive neuroscience, brain, metabolism, hemodynamics

Functional neuroimaging refers to a class of techniques that noninvasively measure correlates of neural activity. Positron emission tomography (PET) and functional magnetic resonance imaging (fMRI) are the two technologies most commonly used today to study the human brain “in action.” The explosion of information about brain function in the last decade has resulted in large part from these two techniques.

The study of brain function has now led to the general acceptance of a new field: cognitive neuroscience. Cognitive neuroscience is simply the study of the neural basis of cognition, and PET and fMRI are two of the predominate tools of the field. Partly because of its roots in cognitive psychology, the field has been concerned mainly with the neural basis of human cognition, as opposed to nonhuman cognition.

The bias is understandable because these techniques represent the first window into actual brain function. Contrast this with preexisting tools. Conventional MRI, while revealing images in exquisite detail of the brain, can only provide static pictures of anatomy. The search for anatomical differences

---

*Address correspondence to: Gregory S. Berns, M.D., Ph.D., Department of Psychiatry and Behavioral Sciences, Emory University School of Medicine, 1639 Pierce Dr., Suite 4000, Atlanta, GA, USA. Tel: 404-727-2556, Fax: 404-727-3233, email: gberns@emory.edu.
that correlate with behavior, normal or abnormal, has had limited success and only in diseases with obvious brain pathology, e.g. Alzheimer’s disease (1). Measurements of neuronal activity, which require the placement of electrodes in, or around neurons, have been used to a great extent in animals and have provided a wealth of information about neuronal function. For obvious reasons, it is not possible to do this in humans except in patients undergoing brain surgery. Even then, one must know where to place the electrodes.

Cognition is a distributed process. When one speaks of a particular cognitive function, e.g. memory, it is not localizable to a single brain region. Although this may be a result of not knowing the organization of the brain sufficiently well, most evidence points towards networks of regions functioning with a particular choreography that gives rise to function. Therefore, to study the neural basis of cognition, one needs a technique that can measure neural function simultaneously in the entire brain. Quantitative electroencephalography (EEG) and magnetoencephalography (MEG) are recent tools that have this capability, but with limited spatial resolution. Although it may never be possible to measure the state of every neuron in the brain, PET and fMRI provide enough spatial and temporal resolution to make meaningful conclusions about the role specific brain regions play during cognition.

Correlates of Neural Activity

It has been known for over 100 years that blood flow to the brain increases in a regionally specific manner according to mental activity. The father of modern psychology, William James, was aware of observations relating regional brain pulsation to mental activity (2). Paul Broca, known primarily for his observations of the effects of left frontal lesions on language and which now bears the eponym, Broca’s area, performed several experiments relating regional brain temperature to cognitive function (3). But it was not until the 1950’s when Seymour Kety and Louis Sokoloff developed the autoradiographic technique for quantitatively measuring regional blood flow that specific cognitive functions could be directly mapped in the living brain (4).

Both PET and fMRI rely on the empirical relationship between neuronal activity and regional blood flow. Stated simply, blood flow increases in areas where neuronal activity increases. Most cognitive neuroscience studies implicitly assume the verity of this relationship. The relationship between neural activity and the hemodynamic response is far more complicated than a simple direct relationship. We lack fundamental data regarding the nature of this relationship, and so there is no agreed upon physiological model.

Neural activity can describe several aspects of function. The brain contains four broad classes of cells: endothelial cells that form the blood-brain barrier, epithelial cells that secrete cerebrospinal fluid, glial cells, and neurons. The glial, or supporting cells, comprise the bulk of the brain matter and do not actively transmit electrical impulses. However, they have crucial roles in the transmission of action potentials by regulating the extracellular environment. Neurons transmit action potentials and are generally believed to be the cell type functionally responsible for cognition. Action potentials, generated near the soma, are transmitted through the axon to the dendritic tree. Thus there are three potential neuronal sites that contribute to metabolism and ultimately blood flow: soma, axon, synapse. Most of the action occurs in the synapses, and these are most directly correlated with the metabolic requirements of the neuron (5,6).

Controversy surrounds the putative blood flow-metabolism relationship. Blood flow increases in both a spatially and temporally correlated fashion with synaptic activity. It is not clear, however, if blood flow is coupled to oxidative metabolism. This is surprising since the brain consumes 20% of the oxygen inhaled. Local oxygen consumption does not increase to the same extent as local glucose utilization, suggesting that neuronal activity may be an anaerobic process (7,8). One theory states that
neuronal activation results in a transient increase in the extracellular pH, which stimulates glycolysis in the surrounding glial cells (5). This leads to the production of both pyruvate and lactate, with the lactate becoming available to diffuse back to the neuron, where it can be oxidatively metabolised. This model postulates that the glial cells serve as an energy reserve for the synapses, and oxidative metabolism occurs well after the electrical activity, replenishing the neuronal energy supplies. Optical imaging suggests that blood flow is, in fact, coupled to oxidative metabolism (9). Results suggest that a transient local decrease in oxygen concentration occurs in response to synaptic metabolic requirements (oxidative metabolism). The increase in blood flow occurs after a delay of 1-3 seconds, but overreacts, leading to the observation that the blood flow increase outpaces oxygen consumption. Unfortunately, there is little direct evidence for this reactive hyperemia, and it does not explain why glucose metabolism appears more tightly coupled to blood flow than oxygen consumption.

Despite the lack of understanding of the relationship between neuronal activity, metabolism, and blood flow, the empirical relationship between these parameters appears both reliable and reproducible in a variety of contexts. Increases in neuronal activity, specifically synaptic activity, result in a local increase in blood flow. Because oxygen consumption does not increase to the same extent, the increase in blood flow results in a lower concentration of deoxyhemoglobin. This relationship forms the basis of fMRI.

PET

Positron emission tomography, or “PET,” was developed from in vivo autoradiographic techniques. An autoradiographic procedure would typically inject an animal with a biologically interesting compound synthesized with a radioisotope (e.g. $^3$H). When the animal was sacrificed, the local tissue radioactivity was easily quantified. Although autoradiography yields exquisitely detailed pictures of brain activity, it can only be applied in animals, and the animals must be sacrificed to obtain the brain tissue. Although these techniques were in use in the 1950’s and 1960’s, the development of an in vivo method applicable to humans awaited technological advancements from the “silicon revolution,” namely the availability of high-quality inexpensive crystal detectors, and the huge advancements in computing power realized in the late 1970’s.

PET requires three basic technologies: the production of positron emitting compounds, the ability to detect simultaneously-emitted gamma rays, and the computational power to reconstruct the sources of emission. Positrons, or positively charged electrons (antimatter), have a particular advantage over other radioactive compounds. When a positron encounters an electron, they annihilate each other, and their collective energy is transformed into two high-energy photons, which are emitted in exactly opposite directions. Because the photons travel 180° apart, it is easy to arrange a ring of detectors to determine where the annihilation occurred. When two detectors are activated simultaneously, then one knows that the emission occurred somewhere along the line connecting the two detectors (Fig. 1). By collecting the counts over a period of time, say 60 s, then it becomes possible to reconstruct the geometry of the source.

Positrons are produced indirectly, by the radioactive decay of particular isotopes. The most commonly used isotopes ($^{14}$C, $^{15}$O, $^{18}$F, $^{13}$N) are produced in a cyclotron by the bombardment of targets with high-energy protons. This results in a gas, e.g. $^{15}$O$_2$, which can then be used in any chemical reaction, e.g. oxidation-reduction reaction with product H$_2$O$_2$. After appropriate purification procedures, these compounds can then be injected intravenously into a human subject, and they flow to the brain in about 20 s. The isotope undergoes radioactive decay by positron emission, and the half-life depends on the particular isotope (2 mins for $^{15}$O).

Because the photons emitted during positron decay are fairly high-energy (511 keV gamma rays), they tend to pass through matter with relative ease. A specialized detector, called a scintillation detector,
Schematic diagram of a ring of PET detectors. When a positron meets an electron, they annihilate each other and produce two gamma-rays travelling in opposite directions. A coincidence circuit isolates pairs of detectors so that the emission can be localized to the line connecting them.

is required to accurately count the decays in a directional fashion. PET scanners consist of rings of these detectors arranged in parallel planes. An individual detector would be constructed from a scintillating crystal, either bismuth germanate (BGO) or lutetium oxyorthosilicate (LSO), and amplification electronics. When a gamma-ray enters the crystal, it loses its energy through either the photoelectric or Compton effect, which results in the production of electrons. These electrons further interact with the crystal, resulting in the production of visible wavelength photons. These photons are then detected and amplified by a photomultiplier tube and converted into an electrical pulse. A “coincidence circuit” allows for the identification of the detector that picks up the 180°-emitted gamma-ray.

Depending on the injected molecule, a particular regional distribution will occur. In the case of H$_2^{15}$O, it will follow the regional blood flow. Other compounds will cross the blood-brain barrier and bind to specific receptors, in which case the distribution of radioactivity will reflect receptor concentration. $^{18}$Fluoro-deoxyglucose ($^{18}$FDG), a commonly used tracer, is metabolized by hexokinase during glycolysis like glucose. Unlike glucose-6-phosphate, FDG is not metabolized further and thus accumulates intracellularly, yielding a measurement of local metabolic activity (6,10).

Most PET neuroimaging studies can be grouped into one of three categories: metabolic, blood-flow, or receptor. Metabolic studies use $^{18}$FDG to measure regional glucose metabolism. $^{18}$FDG, like all $^{18}$F-compounds, has the advantage of a relatively long half-life (110 mins.). This allows for the synthesis to be performed in one location, the subject injection in another, and the scanning in yet another. In fact, one can have a subject doing a particular task in a location remote from the PET scanner, and inject $^{18}$FDG, which will be trapped in brain regions according to the local metabolic rate. This has an obvious advantage in situations in which placing the subject in the scanner would alter the conditions of the task. For example, $^{18}$FDG is used commonly in sleep-studies. The main disadvantage is that the long half-life results in effectively no temporal resolution. This offers a time-averaged snapshot of a particular brain state, and the state is averaged over 20-60 mins.

Blood-flow studies use H$_2^{15}$O to measure changes in local brain blood flow (11-13). As noted before, blood flow is an indirect measure of local synaptic activity. Because $^{15}$O has a short half-life (2 mins.), several administrations can be performed in one session. A typical H$_2^{15}$O study would have 8-16
injections and scans for each subject. The experimental design would manipulate what the subject does during each scan. Each scan lasts about 1 min., with 8-10 mins. between scans (5 half-lives). $H_2^{15}O$ studies not only allow for multiple conditions to be studied, but they also allow for the repetition of conditions, increasing statistical power. The main disadvantage is that because of the short half-life, the $H_2^{15}O$ must be produced reliably and in close proximity to the scanner.

Receptor studies use radioligands—chemicals incorporating a positron emitting isotope into a molecule whose pharmacokinetics are already known. Ideally, these ligands bind specifically to one receptor type. Most of these studies are of the mapping type, which shows the distribution of a particular receptor in the brain (e.g. D2 dopamine receptor). Here, the measured radioactivity reflects both the local concentration of receptors ($B_{max}$) and the affinity of the ligand for the receptor (measured by $K_D$, the equilibrium dissociation constant). If the ligand acts as a competitive antagonist, then the apparent affinity is also affected by the concentration of the endogenous neurotransmitter. The analysis can be simplified by considering the ratio $B_{max}/K_D$, termed the binding potential (BP). Ligands undergo both specific and nonspecific binding. Typically, one is interested only in the specific binding, that is, to the receptor of interest. By using a reference tissue, which is known to have a low receptor concentration, then the nonspecific binding can be subtracted out, e.g. cerebellum has no D2 receptors. In this case, the difference in distribution volume for the two tissues is directly proportional to the binding potential. These molecules require a more involved synthesis than either water or $^{18}FDG$, and this is compounded by the problems of radiation exposure to chemists and a race against the clock as the isotope decays. The end product must meet several requirements: high specific activity (the amount of radioactivity/mole), high radiochemical purity, high chemical purity, and sterility. $^{18}F$-ligands are easier to synthesize because of the long half-life, but $^{11}C$-ligands (20 min. half-life) have a higher potential for biological relevance.

**Functional MRI**

Functional magnetic resonance imaging, or fMRI, refers to a variant of MRI that is sensitive to local changes in deoxyhemoglobin concentration. As noted earlier, the regional blood flow increase apparently surpasses the oxygen consumption. This results in an apparent decrease in deoxyhemoglobin. In the 1930's, Linus Pauling had observed that the amount of oxygen carried by hemoglobin is inversely proportional to the degree to which it perturbed a magnetic field. This property of differential paramagnetism was finally demonstrated in vivo in the late 1980's, and fMRI was born (14,15).

Fundamentally, MRI measures properties of hydrogen atoms. The behavior of a hydrogen atom depends exquisitely on its immediate environment, namely its host molecule and its location in the molecule. Consider a single proton (H$^+$). Quantum theory states that atomic nuclei possess a property called "spin," and because the proton is charged, the spinning charge produces a small magnetic field. Normally the spin axes are oriented randomly, but when a substance is placed in a strong external magnetic field, the axes align themselves in the direction of that field. Some of the spin axes will point up and some will point down. If an equal number point both directions, then there will be no net magnetization; however, if there is an unequal distribution of up and down spins, then the substance as a whole will acquire a magnetization. This is the magnetic susceptibility of the substance and is highly dependent on the molecular makeup. The magnetization does not happen instantaneously, and the time a substance takes to reach its net magnetization is termed T1.

MRI works by using a strong external magnetic field to align the H$^+$ spins and then perturbs these spins with a radiofrequency (RF) pulse. Just like a spinning top, the protons do not spin exactly along a single axis, even in an external field. Instead there is a slow wobble, or precession, and the rate of precession is proportional to the strength of the external field. Because the precession frequency is
proportional to the magnetic field, the application of a slight gradient in the field results in a linear variation in precession frequency. This means that the precessional frequency encodes spatial position. The precession can be accentuated by applying an RF pulse at the same frequency of precession, the resonant frequency. When this happens, the precession axis gets knocked out of its original orientation. If enough protons get bumped, then the tissue will acquire a slight magnetization perpendicular to the external field, termed the transverse magnetization. When the RF pulse is turned off, the transverse magnetization decays back to zero, and the time this takes is referred to as the relaxation time, T2. By varying the field gradients and RF pulses, a three dimensional volume can be encoded and subsequently reassembled. The specifics, however, are not directly relevant to the neurophysiology, and the interested reader is referred to other references (16).

The RF pulse drives all the protons to precess in a synchronous fashion, meaning they are in phase. As soon as the pulse is turned off, the spins start to dephase. The dephasing results from two sources: one intrinsic to the substance and one from external field inhomogeneities. Intrinsically, the spinning protons themselves create small inhomogeneities in the local magnetic field. Regions in which protons align with the magnetic field create a local increase in field strength, thus increasing the precession frequency. Oppositely aligned protons decrease the magnetic field and decrease precession frequency. This slight variation in precession frequency means that in the absence of an external RF sine wave, protons will dephase because of the slightly different precession frequencies. The rate of dephasing determines the relaxation time, T2, which in turn depends on the molecular environment of the proton. The T2 of gray matter is about 60 ms (17,18). The second source of dephasing arises from slight inhomogeneities in the externally applied magnetic field. Any substance that possesses either paramagnetic (deoxyhemoglobin) or ferromagnetic (iron) properties will distort the field. These extrinsic sources result in an apparently faster dephasing, termed T2*.

Functional MRI exploits the fact that deoxyhemoglobin has paramagnetic properties and oxyhemoglobin does not. Deoxyhemoglobin disturbs the local magnetic environment, causing the surrounding protons to dephase even faster than they would otherwise (Fig. 2). Recall that neuronal activity leads to an overactive increase in blood flow, which actually decreases the amount of deoxyhemoglobin relative to oxyhemoglobin. Because less deoxyhemoglobin means less rapid spin dephasing, this increase in blood flow appears as an increase in MR signal.

The temporal resolution of fMRI is determined both by the hemodynamic response and the physical constraints of the scanner magnetic fields. The hemodynamic response generally lags the neural activity by 3-5 s and may extend upwards to 10 s. The rate at which the scanner can acquire images

![FIG. 2.](image)

Schematic diagram of the effect of hemoglobin (Hb) on the local magnetic field of brain tissue. Only deoxyHb has paramagnetic properties and locally distorts the field, leading to faster spin dephasing.
is influenced by the desired resolution. Generally, the more slices and the finer the resolution within each slice, the longer a whole-brain acquisition takes. While an individual slice can be acquired in as little as 100 ms, whole-brain imaging requires about 3 s.

**The Measurement of Cognition: Issues of Experimental Design**

Neither PET nor fMRI typically can measure absolute neural activity. Although PET can yield measures of absolute metabolism or blood flow, this requires arterial blood sampling, which is substantially more invasive and is not commonly performed anymore. Consequently, neuroimaging experiments must be designed to quantify relative changes in activity. But relative to what? The most commonly used approach, subtractive mapping, extends an idea proposed more than 130 years ago. Franciscus Donders, a Dutch physiologist, proposed a method to study specific cognitive processes (19). Armed only with subjects’ response times, he suggested that longer response times represented added mental processing. He subtracted the time to respond to any light from the time needed to respond to a particular colored light. He concluded that this difference, about 50 ms, represented the time to process color. One can easily imagine this exact experiment being repeated in either a PET or fMRI scanner, and instead of subtracting response times, subtracting brain activity maps.

The subtractive approach to neuroimaging has been used to great success over the past 10 years. One must be clear about the assumptions. Two principles operate here. First, one assumes that brain activity scales in a linear fashion. If a particular brain region requires $x_A$ amount of activity to perform function $A$ and $x_B$ amount of activity for function $B$, then doing both $A$ and $B$ would require $x_A + x_B$ activity. This is a powerful approach because we can ignore any baseline activity requirement. Assuming the baseline activity remains unchanged, we can determine the regional activity requirements associated with $B$ simply by doing the subtraction: $(x_A + x_B) - x_A$. In actual practice, a paired t-test or ANOVA is used instead of a direct subtraction, yielding statistical probabilities. The assumption here is that levels of “brain activity” are additive in a linear fashion. Undoubtedly, this is not the case, but it is not yet clear under what circumstances it is a reasonable approximation. Secondly, and potentially more problematic, is the assumption that cognitive processes are additive. Also referred to as the “pure insertion” hypothesis, this assumes that the addition of another cognitive process does not alter the original one. What if identifying colored lights does not simply turn on the putative color processing module in the brain, but alters the entire network of activity, including the structures that identify any light? For the measurement of a single parameter like response time, this is less relevant, but in brain activity maps, which may have thousands of measurements, some regions may go up and others may go down. What does decreased regional activity imply about the function of that region?

Given the uncertainty in these assumptions, it is surprising that such impressive imaging results have been obtained at all. An optimist might argue that because linear designs do work, then the assumptions have not been disproved. However, there is ample neurophysiological evidence for the nonlinearity of neural activity, and there are too many unknowns in the equation coupling neural activity to blood flow. On the assumption of pure insertion, there is ample evidence that people alter strategies when performing additional tasks (20-22). What then do neuroimaging experiments measure? One possibility is that linear assumptions are good enough, and the aforementioned nonlinearities contribute relatively little. This does not appear to be the case either, since varying something as simple as task repetition rate can dramatically change the activity pattern (23). If repetition rate affects activity maps, then one can quickly imagine a host of parameters that might affect the observed image: motivation, monetary payment, physical discomfort, length of time in the scanner, angle subtended by the presentation screen, form of response (button press vs. finger tap), etc. If true, then each experiment must be considered as a case study under a particular set of conditions. This would be consistent with the wide variation in findings under apparently similar, but not identical, tasks.
The Measurement of Cognition: Issues of Data Analysis

As a general principle, good experimental design precedes good statistical analysis. Given the constraints previously outlined, all functional brain imaging studies yield large amounts of data, from which conclusions must be drawn. Most studies rely on subtractive analyses and therefore use various types of t-tests and ANOVA's, i.e. linear parametric statistics. However, the large amounts of data points being analyzed (1000's) means that the problem of multiple comparisons must be addressed. Consider a typical fMRI study. One might acquire 16 planes of data through the brain, with each plane consisting of a $64 \times 64$ matrix. This yields 65,536 data points, or voxels (volume elements), with about 1/3 of these being voxels in the brain. The application of a t-test with a conservative threshold of $p = 0.001$, still means that approximately 20 voxels will appear significant by chance alone. Increasing the significance threshold will decrease the number of Type I errors, but at the expense of missing truly significant activations (Type II errors).

The basic analysis of imaging data applies standard statistical tests to every voxel in the image in a univariate fashion (24). In other words, each voxel is considered independent from the others. Then, one of several methods is used to correct the significance level for multiple comparisons. The main problem is that the data points are not independent from one another. Points near each other are more likely to display similar activations. Some techniques consider the spatial extent of a particular area of activation. By estimating an overall “smoothness” to the image, one can predict what magnitude and spatial extent constitutes a significant activation. This is the method of gaussian random fields. Fundamentally two factors influence the calculation of significance: the magnitude of activation and the variance of the activation, both within and between subjects. Thus a typical analysis would use a general linear model formulation and would include the experimental effects as fixed variables, subject effects, and a number of covariates. If the experiment was a simple two-condition design, then the significance of the F-value would be equivalent to a paired t-test. More complex experimental designs would examine specific contrasts between conditions. The end result is a statistical parametric map (SPM) which displays the significance of activations, overlaid on a map of the brain.

Brains come in all different shapes and sizes. Unless one is doing a single case study, most experiments attempt to infer some general property of the brain that applies to a particular population of individuals. This requires a pooled analysis, but before this can be done, each brain must be fit into a standardized space. If it were a matter of simple scaling then it would be straightforward. One approach relies on the specification of landmarks, e.g. corpus callosum, ventricles, etc. However, this quickly becomes tedious and is prone to subjective judgements. Low resolution PET and fMRI images make this almost impossible. A set of automated image registration techniques has been developed (25,26). These algorithms iteratively minimize the mismatch error between two images. One can find the optimal way to stretch, shift, and rotate one brain image to match another in order to minimize the difference between them. Further modifications apply nonlinear warping to achieve even better fits. These techniques have been well-validated and allow the comparison of activation across subjects; however, significant problems arise in the study of different populations, like schizophrenia or Alzheimer’s disease. Many brain illnesses are associated with structural change, and so it becomes problematic to compare brain activation patterns if the underlying structure has been altered.

The Future of Neuroimaging

The rapid acceptance of fMRI suggests that it may soon supplant PET for many functional neuroimaging studies. In part, economics is driving this trend: a typical fMRI session might cost about $500 whereas a PET session might cost $3000. Almost any MRI scanner can be appropriately programmed to yield functional images, but relatively few PET centers exist. However, there are important limitations to fMRI suggesting that both techniques will remain important tools for the study
of human cognition. Both fMRI and H\textsubscript{2}\textsuperscript{15}O-PET measure blood-flow changes associated with synaptic activity, but neither of these techniques can distinguish between excitatory and inhibitory activity, or any other neurotransmitters. Current fMRI technology will remain an "activation" methodology, but this is only part of the story. Future PET studies will increasingly focus on neurotransmitter-receptor interactions, which in conjunction with fMRI, will clarify the contribution of excitatory, inhibitory, and modulatory neurotransmitters to brain "activation." We are just beginning to explore the interactions of various neurotransmitter systems and their relationship to behavior (27,28). Functional MRI is also evolving rapidly. Improvements in the speed of image acquisition are pushing the temporal resolution down to 100 ms. Although the hemodynamic response does not occur that quickly, cognitive tasks are being designed that allow one part of the brain to act as a "timekeeper" for another (29). With fMRI to probe activity patterns and PET to map neurotransmitter-receptor interactions, the future of neuroimaging is exciting indeed.

Acknowledgements

This work was supported by a grant from the Stanley Foundation.

References