Experimental design in brain activation MRI: Cautionary tales

Robert L. Savoy*

The Athinoula A. Martinos Center for Biomedical Imaging of the Massachusetts General Hospital, Harvard Medical School, HyperVision, Inc., U.S. P.O. Box #158, Lexington, MA 02420, USA

Available online 5 July 2005

Abstract

The use of functional magnetic resonance imaging (fMRI) in cognitive neuroscience has expanded at an amazing rate in the past 10 years. Current research includes increasingly subtle and specific attempts to dissect the cognitive and emotional mechanisms called into play when humans make decisions. The present essay will briefly review some of the general considerations and domains of information needed when one designs fMRI-based experiments. However, the main theme will be the difficulties associated with designing, conducting, analyzing and interpreting such research. Functional MRI is an unusually complicated technique, and there are numerous ways for experiments to go wrong. As well as demanding exceptional care in maintaining the quality of one’s own research, this makes the universal problem of evaluating other peoples’ research particularly challenging.

Keywords: Functional MRI (fMRI); Experimental design; Error correction in science

1. Introduction

At the 2004 ConNEcs meeting (“2nd Conference on Neuroeconomics”, Münster, Germany, May 27–28, 2004) I was asked to speak about experimental design in functional magnetic resonance imaging (fMRI). In the normal course of my teaching duties, I spend many hours on this topic with each class, emphasizing hands-on attempts at experimental design as a way to illustrate the challenges. So the invitation to speak for about 35 min on experimental design raised problems for me.

After several failed attempts to design a lecture that covered all the things I thought necessary, I abandoned that approach. Instead, I briefly presented a few of the basic constraints associated with fMRI by describing a very simple demonstration study and its analysis [13] and some generic issues in the design of fMRI-based experiments [6,8] and then asked, “What is the most important message to be conveyed in this short time (or these few pages)?” I quickly found an answer to this question. The most important message in my view: Functional MRI is difficult.

This message has a long and distinguished history. In the first direct and detailed physiological study of the hemodynamics of local cerebral blood flow in the exposed brain of an animal, Roy and Sherrington [12] began their report with the same caution. In modern terms, what they said was: We think different investigators get contradictory results because the technology for measuring cerebral hemodynamics is complicated.

As a classic example of how not to write scientific prose, what they actually wrote, in the overwrought scientific writing of their day, was: “One of the marked characteristics of the literature dealing with the cerebral circulation is, we think, the contradictory nature of the results which have been obtained by different investigators. There is no reason, we imagine, for doubting that the cause of these discrepancies is to be found in the great difficulty of avoiding the sources of error which plentifully surround the subject, and in overcoming certain technical difficulties which we shall presently have to refer to. The ease with which one can obtain results upon certain points, on taking up the subject, is itself, we believe, apt to make the inquirer careless in controlling sources of error, which, it may be noted, are some of them not at first sight obvious. We must on this account say more about the technology of our subject than would be necessary were the

* Tel.: +1 617 497 4647; fax: +1 617 497 4627.
E-mail address: savoy@rowland.org.

0361-9230/$ – see front matter © 2005 Published by Elsevier Inc.
doi:10.1016/j.brainresbull.2005.06.008
subject a simpler one” [12]. But no matter how written, the message is the same.

From a scientific standpoint, the Roy and Sherrington paper is remarkable because the investigators reported both the measurable change in cerebral blood volume elicited by the stimulation of a dog’s paw while using mechanical sensors resting on the exposed cortex of the animal, and the visually observed, macroscopic change in the color of the area being fed this extra blood—it became redder. Thus, the original paper reported changes in both cerebral blood volume and the oxygenation state of the blood (later called the BOLD effect) that are triggered by neural activity.

2. Functional MRI is difficult

Functional MRI is difficult because there are so many things one must understand to use it well. One must have a good working knowledge of:

- the physics and engineering underlying the imaging device;
- the temporal and spatial properties of the physiology and hemodynamic responses that we measure using that imaging device;
- the functional and structural anatomy of the human brain;
- the myriad of data analytic steps that must be considered (and, normally, applied) in going from a collection of MR images to a statistically meaningful map of an activation pattern;
- the way all of the above interact with each other and with psychology in the design of fMRI-based experiments;
- and perhaps most importantly, a feel for the way all these things interact when trying to interpret the results of experiments from your own or others’ laboratories.

Functional MRI is difficult because most of the easy questions have been asked. It is virtually impossible to think of basic psychological or clinical brain-related questions that have not been addressed – and often addressed in more than one paper – by the tools of functional brain imaging. Audiences frequently inquire of speakers: “What is the future direction of functional brain imaging in general, and functional MRI in particular?” My answer is: “Take a look at a photograph of an exploding star—functional brain imaging research is exploding in all directions at once.” Research using fMRI and other brain imaging tools is growing at an overwhelming rate. In 2001 there were over 800 peer-reviewed research articles published in over 200 journals [7]. This growth is in all subdivisions of cognitive and clinical psychology and medicine. It is rapidly expanding to social psychology and applied areas such as neuroeconomics.

Experiments in functional brain imaging (as with any experiment that involves human psychology and behavior) are influenced by many subtle factors. Experiments on cognition are that much more difficult because cognitive function is a moving target. Your subjects’ brains change all the time, and especially in response to cognitive challenges. Ask a person a question once, and it is a different person to whom you repeat that question the next minute or the next day. This complication has less practical impact when studying low-level sensory or motor processes. But most of the issues of importance in neuroeconomics are cognitive: related to choice behavior, attitudes, and emotions. Simply repeating the same thing over and over is not a way to get better signal-to-noise in your data. The tasks must have enough novelty, while staying focused on the core issue, to yield useful data.

And finally, designing, conducting, analyzing and interpreting fMRI-based experiments is difficult because there are so many ways to go wrong. The next section will review the fundamentals of designing fMRI-based experiments, with a few examples of good design, and discuss some of the ways in which the various components (the MRI machine, temporal aspects of hemodynamics, general issues in “difference imaging”, and the ubiquitous question of “thresholds for significance in statistical testing”) interact. The subsequent section will examine a few famous (or infamous) examples of published, peer-reviewed imaging studies that were later discovered to have misleading or erroneous elements. The purpose is not to condemn these examples. In fact, it is quite the opposite. In most of these cases the researchers were open and ultimately public in discussing the sources of error or misinterpretation associated with the original studies. Nonetheless, these examples and the general comments that follow may help to alert practitioners in the field to the many sources of potential error in designing, conducting, analyzing and interpreting such experiments.

3. Review of the basics of experimental design in functional MRI

In addition to providing exquisitely detailed structural images of the soft tissues of the body, MRI can be used to detect local changes in blood flow and the concentration of deoxyhemoglobin in the blood. Neural activity leads to changes in local cerebral blood flow (CBF) and the local concentration of deoxyhemoglobin in the human brain. We can use MRI to detect these changes and infer associated changes in neural activity. (Other technologies, including positron emission tomography (PET) and variants of MR Imaging using externally supplied contrast agents, are based on detecting associated changes in cerebral blood volume (CBV).)

At present, it is only changes that we can detect using functional MRI. That is, although one structural MR image is useful by itself, a single image of the sort collected in fMRI experiments is generally not very informative. The power of functional MR images is based on the analysis of many different brain images (at least two, and more commonly large collections of brain images). We cannot yet use fMRI to detect the absolute level of neural activity. Rather, we can detect the difference in activity associated with two different tasks, or
The physical apparatus for the test is not magnet compatible. Moreover, if the visual stimuli associated with that task were merely presented on an MR-compatible visual display system, the subjects would be engaging in numerous extraneous activities involving eye movements, hand movements, and arrangement strategies that would vary widely from subject to subject. In a purely behavioral setting, this is not a problem. But in a context where brain activation is being examined, all of that extraneous and poorly controlled activity will just generate irrelevant brain activity (irrelevant, that is, for the target issue of color discrimination). It is better to adapt the task in a novel way that still maintains the essentials of the color vision test (thereby changing the procedure from the standard that has been so thoroughly studied) to the magnet environment if one wants to minimize extraneous activity. That is, if one uses a relatively direct adaptation of the conventional task, subjects will be generating large eye-movement signals and motor signals from arm movement and have highly variable duration of activity associated with each trial. If, instead, an adaptation of the task is developed that minimizes these extraneous neural activations and focuses on the core issue of color discrimination (see, e.g., [2]), the resulting imaging data are simpler, clearer, and much easier to interpret. (It is true that a separate behavioral validation of the adapted procedure should be done to confirm the validity of the behavioral results in the adapted situation, perhaps in the less expensive context of a mock scanner.)

Some examples are chosen because they violate standard rules and guidelines in useful and instructive ways. For instance, experiments should always minimize head movement of the subject—but sometimes head movement can be studied using fMRI! In one study [10] subjects were asked to turn their heads in response to a stimulus that was briefly presented to the left or right of a visual fixation point. After the stimulus disappeared, subjects were to return their head to the position it originally had. (Pillows and padding of various kinds helped the subjects recover their original position.) The idea in this experiment was to take advantage of the time lag between neural activity and the resulting hemodynamic changes detectable using MRI. When the subjects moved their heads, the MR images collected were badly out of register with the other images of the run, and had to be discarded. However, the hemodynamic response associated with those body movements (the head turn) would lag behind the neural responses that triggered the head movements. By the time the MR scanner was collecting data about the hemodynamic response to those head movements, the subjects’ heads would, hopefully, be back where it started. Analyzing the data revealed that this was largely the case. Thus, the brain activations associated with these brief head movements could be studied.

There are many aspects of the physics of image generation, and the physiology of the hemodynamic changes that must be understood to design good experiments. Many references are available to give a brief summary (e.g. [13,14]) or a thorough discussion (e.g. [6,8]) of these issues.
one understands these things, the better one can design experiments. However, it is not essential to become an expert on everything to run or supervise fMRI-based research.

Arguably, one can avoid being an expert in the physics and engineering of magnetic resonance imaging and the physiology of hemodynamics. Having a good, basic knowledge of the technology of the machine, especially in a context where there is good support for the machine and maintenance of its performance in these experiments, is sufficient for designing and overseeing most experiments.

Having said that, it should be emphasized that having good technical support for the machines is essential. Functional MRI requires that the imaging device is operating with lower tolerances and variability than is acceptable for conventional structural imaging. In addition, these experiments typically require a kind of MR Imaging (T2*-weighted EPI) that induces a number of artifacts. This does not mean that every research installation must have its own physicist, but it does mean that the technicians operating the equipment must have special training to be aware of the techniques for minimizing artifacts, as well as performing daily tests to assure the required stability of the machine.

Similarly, one does not need to be an expert on the underlying blood-related physiology. It is generally sufficient to understand the temporal characteristics of the hemodynamic response in order to design experiments and analyze the associated data. Having said this, one should keep in mind the proviso that both the technology of imaging and the understanding of the physiology behind hemodynamics of these areas are constantly changing, and that some of these changes may be important for future aspects of experimental design. Therefore, while one may not need to be an expert in these areas, one must maintain a current working knowledge of the areas, with the emphasis being on noticing when a new technology or new understanding of the physiology is important for fMRI-based experiments. One recent example relevant to the technology is the development of multi-channel receiver coils that yield better contrast and spatial resolution of the cortex. A more speculative example is based on the idea that much of the BOLD signal may be driven by synchrony in neural firing across a population of neurons (e.g., [15]), rather than the absolute level of firing rate. This would suggest that tasks which are particularly good for EEG and MEG might be useful in selecting tasks for fMRI (because detection of the EEG and MEG signals requires the synchronous activity of neurons).

A good working knowledge of the brain’s structural and functional neuroanatomy is essential, and more knowledge about the areas that are particularly relevant to one’s own study is expected. But being an expert in neuroanatomy is not a requirement for conducting fMRI-based research.

However, in my view, the areas in which every investigator is obligated to be an expert are data analysis and interpretation of results. This is critical in order to protect the quality of your research, and in order to be a responsible supervisor of your students’ work. And becoming an expert in this area entails the commitment of a substantial amount of time and effort.

“Data analysis” in the context of fMRI refers to many different procedures. Some have to do with correcting for technical problems associated with magnetic field inhomogeneities, especially in high field (3 T and higher) scanners. Some have to do with quantifying the amount of head movement that occurred during each experiment. Some have to do with spatial and temporal smoothing of the data. And, critically, some have to do with the models used to represent a statistical fit to the data, as well as determining the appropriate thresholds to use when interpreting the relationship between the model and the fitted data. In addition, there is a growing literature on the use of multivariate analysis techniques (i.e., techniques that do not look only at the data from a single voxel in computing relevant statistics). The preceding is just a partial list.

Some aspects of data analysis (e.g., the use of algorithms to detect and quantify head movement) are universally agreed upon, in the sense that they would always be an expected part of any analysis. Other aspects of data analysis (e.g., the best use of the information obtained about motion: should it simply be a filter to discard data from subjects who move too much? How much is too much? Can the movement information be used to model the data, and reduce noise associated with movement? If there is a tiny component of movement that is stimulus-correlated, would modeling the movement result in loss of any expected signal due to the experimental stimuli?) do not have universal answers. The responsible investigator has to be expert at least with respect to the data analysis procedures they choose to use. They must also be aware of the new tools that are being developed.

In my experience, everyone makes errors when they start conducting fMRI-based experiments. We like to believe that most such errors are caught before reaching publication. The following examples illustrate a few of the problems or errors that reached publication.

Before describing the published examples, I should say that if there were a mechanism for including unpublished examples, a much larger – and probably very useful – compendium would result. In 1996, before the Second International Conference on Functional Mapping of the Human Brain, I sent an announcement to the pre-registered participants, asking if they would contribute “bad data”, i.e., data they had collected, but subsequently realized was incorrect for some interesting reason. Unfortunately, although I am sure every investigator has such data, only two people responded. A compendium of “bad data” from fMRI-based studies would surely be useful for the field.

4. Cautionary tales

One of the most famous cases of reported error in functional neuroimaging is from a study published in Science in 1989. The study was entitled “Neuroanatomical Correlates...
One of the perennial concerns about fMRI-based data is the variability of the hemodynamic response function (HDR) across individuals, across time for a given individual, and in different parts of the brain. These questions have been addressed in various published articles. Consider first a pair of papers that came from one laboratory at Stanford University [1,3]. The basic experimental technique employed in these papers required subjects to attend to a visual stimulus (a cross) that was constant for 15.5 s, and which then changed to a disk for 0.5 s. As soon as the change was detected, subjects made a simple motor response—pressing a button in each hand. This sequence repeated regularly, many times. These experiments supplied useful information about the reproducibility of the amplitude and general shape of the HDR for a given subject on a given day (over several runs), for a given subject across sessions on different days, across different subjects of similar ages, and across subjects with a wide range of ages. However, the data also appeared to demonstrate considerable variability in the delay between the stimulus trigger (i.e., the moment when the cross changed to a disk) and the initial rise of the HDR. In particular, some of the subjects had HDR functions that started to rise almost immediately after the stimulus changed. This was surprising because many other studies indicated a lag of approximately 2 s before the HDR (as measured by fMRI, using BOLD contrast) started to rise, with the peak being reached 4–6 s after the onset of the stimulus.

Perhaps there was a problem with using regular timing in the presentation of these stimuli. Even though 16 s sounds like a long time, perhaps some subjects (or, at least, some subjects’ brains) began to anticipate the change in the target stimulus, thus eliciting some sort of “expectancy-based” neural and hemodynamic responses. Because they were being “anticipated”, these stimuli could elicit signals from the brain that would appear sooner than predicted by assuming a fixed lag time.

When this possibility was brought to the attention of the investigators, they retested the subjects who had shown the unusually short lag in HDR [personal communication]. In the retest, the regular, periodic presentation of the target stimulus change was changed to a random intertrial interval, to minimize the effects of expectancy. The consequence of this change was that the early response disappeared. Thus, some of the data reported in the earlier studies could (and has been) interpreted as showing a variability in temporal onset of the BOLD effect that was much shorter than indicated by a host of other studies. The preceding work has been expanded [5], and the new work incorporates a variable intertrial interval, so this lesson was well-learned. On the other hand, there is no easy way to remove the erroneous parts of the figures from the literature, and, indeed, these figures are propagated in, for example, a recent text on fMRI (as will be discussed in more detail at the end of this section).

Another article dealt with variability in fMRI data in a different way [9]. Activation maps for a single subject were generated across multiple sessions. Three simple tasks, one...
problems like the above are just the tip of the iceberg. There are many other ways that the design and analysis of fMRI-based experiments can go wrong. The above examples illustrate the generation and publication of erroneous or misleading reports based on problems with intermodal image alignment, experimental design, and data analysis. In all cases the requisite work to track down the problem was done either explicitly by some of the authors of the original work, or with their cooperation. This is, of course, good, and how science works. But some readers (e.g., me) who happened to have had direct experience with some of these stimuli and tasks were immediately suspicious that there was a problem when they first saw the published data. So how did these studies get published in the form they did? My guess is that it was not because the reviewers were poor, but because, for many issues in functional brain imaging, it is only direct experience with similar or identical stimuli and tasks that leads to the right kind of skepticism. If you have done something very close to the indicated study, you may know that something is wrong. But the results were not, a priori, absurd or even unlikely, so it would be easy to accept them at face value. And this was in a context where the stimuli and tasks were extremely simple. As the phenomena being studied become more esoteric, and as they generate activity in complicated networks of brain areas, it is less likely that any reviewer would have the kind of experience necessary to be confidently suspicious as soon as they see new data. Given the flood of new studies, and the financial as well as time-consuming costs associated with replicating a published study, this problem is likely to get worse in the coming years.

5. Conclusion

It is not the aim of this essay to completely discourage new investigators from using functional MRI as a tool. Rather, the aim is to convince people considering the application of this tool in their own investigations that one cannot be a “dabbler” or “ dilettante” in this field. The tool is powerful, but complicated. There are many ways in which it is possible to make errors in reaching conclusions. Practice with these techniques in studying simple questions will help to teach the new practitioner something of the ways in which investigations can go wrong. But the appeal of the technique, especially for neuroeconomics, involves more subtle experimental tasks involving choice and attitude and emotional responses. These can be addressed with suitable experimental designs, but they are even more difficult to check because of the problem of repeating the same cognitive or emotional study on a given individual. In the end, we will have to rely on the usual mechanisms of science for correction: If a significant conclusion is found, and it impacts many other studies and stimulates more investigation, problems with early reports will be revealed. On the other hand, in a domain where the sheer number of studies (each costing thousands of dollars to conduct) is reaching almost 1000 per year, there is no way that they can
all be replicated and carefully studied. The wise reader will take all results with a large degree of skepticism, until replications have been obtained.

References