ANALYSIS OF FUNCTIONAL MRI FOR PRESURGICAL MAPPING: REPRODUCIBILITY, AUTOMATED THRESHOLDS, AND DIAGNOSTIC ACCURACY

by

M. Tynan R. Stevens

Submitted in partial fulfillment of the
requirements for the degree of
Master of Science

at

Dalhousie University
Halifax, Nova Scotia
August 2010

DALHOUSIE UNIVERSITY

DEPARTMENT OF PHYSICS AND ATMOSPHERIC SCIENCE

The undersigned hereby certify that they have read and recommend to the Faculty of Graduate Studies for acceptance a thesis entitled "ANALYSIS OF FUNCTIONAL MRI FOR PRESURGICAL MAPPING: REPRODUCIBILITY, AUTOMATED THRESHOLDS, AND DIAGNOSTIC ACCURACY" by M. Tynan R. Stevens in partial fulfillment of the requirements for the degree of Master of Science.

Dated: August 27th, 2010

Supervisors: _____

_____

Readers: _____

_____

DALHOUSIE UNIVERSITY

DATE: August 27th, 2010

AUTHOR: M. Tynan R. Stevens

TITLE: ANALYSIS OF FUNCTIONAL MRI FOR PRESURGICAL MAPPING: REPRODUCIBILITY, AUTOMATED THRESHOLDS, AND DIAGNOSTIC ACCURACY

DEPARTMENT OR SCHOOL: Department of Physics and Atmospheric Science

DEGREE: M.Sc. CONVOCATION: October YEAR: 2010

To my Daddats.

In whose honor I persevere.

Table of Contents

List of Tables

List of Figures

Abstract

Examination of functional brain anatomy is a crucial step in the process of surgical removal of many brain tumors. Functional magnetic resonance imaging (fMRI) is a promising technology capable of mapping brain function non-invasively. To be successfully applied to presurgical mapping, there are questions of diagnostic accuracy that remain to be addressed.

One of the greatest difficulties in implementing fMRI is the need to define an activation threshold for producing functional maps. There is as of yet no consensus on the best approach to this problem, and *a priori* statistical approaches are generally considered insufficient because they are not specific to individual patient data. Additionally, low signal to noise and sensitivity to magnetic susceptibility effects combine to make the production of activation maps technically demanding. This contributes to a wide range of estimates of reproducibility and validity for fMRI, as the results are sensitive to changes in acquisition and processing strategies.

Test-retest fMRI imaging at the individual level, and receiver operator characteristic (ROC) analysis of the results can address both of these concerns simultaneously. In this work, it is shown that the area under the ROC curve (AUC) can be used as an indicator of reproducibility, and that this is dependent on the image thresholds used. Production of AUC profiles can thus be used to optimize the selection of individual thresholds on the basis of detecting stable activation patterns, rather than *a priori* significance levels.

The ROC analysis framework developed provides a powerful tool for simultaneous control of protocol reproducibility and data driven threshold selection, at the individual level. This tool can be used to guide optimal acquisition and processing strategies, and as part of a quality assurance program for implementing presurgical fMRI.

List of Abbreviations and Symbols Used

Abbreviations

AUC	area under the curve
BOLD	blood oxygen level dependent
CBF	cerebral blood flow
CBV	cerebral blood volume
$CMRO_2$	cerebral metabolic rate of oxygen consumption
CS	cortical stimulation
ECoG	electrocorticography
EEG	electroencephalography
fMRI	functional magnetic resonance imaging
FDR	false discovery rate
FN	false negatives
FOV	field of view
FP	false positives
FPR	false positive rate
GLM	general linear model
GLT	general linear test
Hb	hemoglobin
Hct	hematocrit
HRF	hemodynamic response function
ICA	independent component analysis
ICC	intraclass correlation
IR	infra-red
MEG	magnetoencephalography
MP-FLASH	magnetization prepared fast low angle shot
MRI	magnetic resonance imaging
RF	radio frequency
ROC	receiver operator characteristic

ROI	region of interest
SNR	signal-to-noise ratio
SPGR	spoiled gradient recalled
TN	true negatives
TP	true positives
TPR	true positive rate
WHO	world health organization

Symbols

$\vec{B}$	magnetic field
$\vec{H}$	applied (magnetizing) field
$\vec{M}$	magnetization
$\vec{\mu}$	magnetic moment
μ_0	permeability of free space
χ_m	magnetic susceptibility
t^*	t-statistic (or t-value)
T_1	spin-lattice relaxation time
T_2	spin-spin relaxation time
T_2^*	effective spin-spin relaxation time
γ	gyromagnetic ratio ($\mathrm{radT^{-1}s^{-1}}$)
$\gamma\!\!\!^{-}$	gyromagnetic ratio ($\mathrm{MHzT^{-1}}$)
ω_0	Larmor precession frequency
T_R	repetition time
T_E	echo time
T_I	inversion time
$\hbar$	reduced Plank's constant (1.055×10^{-34} $\mathrm{m^2kgs^{-1}}$)

Units

T	tesla (magnetic field strength)
MHz	megahertz
μm	micrometers

Acknowledgements

This project would not have been possible without the gracious help of many. To start, I would like to extend a very special thanks to Dr. Gerhard Stroink, who introduced me to the lab at the NRC Institute for Biodiagnostics Atlantic, and whose consistent encouragement has pushed me to do my best. It has not gone unnoticed over the years.

To my supervisors Dr. Steven Beyea and Dr. Ryan D'Arcy, I owe everything I have accomplished in these years as a graduate student. I came to you as naive as a newborn, and you have done your best to shape me into a scientist. Maybe some day I will catch on.

To Dr. David Clarke, whom I hold in the highest esteem, I extend my gratitude. Your willingness to invite me into your world, to demonstrate, and to share knowledge has provided an unparalleled experience. And I doubt any of us will forget that luncheon I chose not to attend.

Thanks to all the support staff at the NRC. Janet Marshall for keeping my act together. Sujoy for help with the technical side of things, and for stepping in when I couldn't be there. And especially to Careesa for running the scans with me. To all the other students for your company and friendship. Couldn't ask for better lab-mates.

I would like to acknowledge the help I have received from Carl Helmick, who taught me a lot of what I know about AFNI, and helped set up my trusted MacBook, which has been a workhorse for me. And last, but not least to Ron Hill, who made the integration with the neuronavigation equipment possible, and without whom this project would never have gotten off the ground.

Chapter 1

Introduction

This thesis will address practical concerns regarding the application of functional MRI to presurgical mapping for brain tumor surgery. The problems investigated include controlling for the reproducibility of the functional mapping protocol, and setting of activation thresholds for viewing and interpreting the functional maps. To gain an appreciation of the problem at hand, this chapter will focus on the contextual background of presurgical mapping by fMRI, followed in the next chapter by the theoretical foundations.

1.1 Functional Neuronavigation

Neuroimaging has been playing an increasingly important role in the surgical treatment of brain tumors with the advancement of image-guided surgery technology [1–3]. Surgeons now routinely have access to high-resolution imaging in the operating room while removing a tumor, and can use a navigation wand to match locations on the exposed brain surface to precise locations in the images. The vast majority of neuronavigation uses magnetic resonance imaging (MRI). MRI is the modality of choice because it offers resolution on the order of 1mm^3 and soft tissue contrast that can be tailored to highlight differences between tumor and surrounding tissue [1].

While the goal of surgical treatment of brain tumors is complete removal of the abnormal tissue, this is often difficult to achieve in the case of infiltrative tumors, or tumors with poorly defined borders [2, 4]. The appropriateness of the surgical approach depends on many factors, including tumor type, location, and surrounding tissues, and often radiation and chemotherapies are used in conjunction with, or instead of, surgery. In any case, when the surgical option is chosen, the neurosurgeon faces the delicate tasks of choosing from possible access routes to tumors, and subsequently removing an infiltrative growth from surrounding healthy brain tissue. The surgical procedure therefore inherently involves balancing conflicting aims of total

resection and functional preservation [2].

Adequate preservation of functional zones cannot be based solely on anatomical landmarks. The morphological variation from individual to individual is pronounced, complicating precise identification of anatomical landmarks. This problem is amplified in individuals with brain tumors, which can cause significant compression and displacement of brain tissue, a phenomenon called the mass effect. A second caveat to functional preservation on the basis of anatomical landmarks is that functional organization in the brain differs between individuals, especially in higher-level cognitive functions, which involve more complex neural networks for completing the tasks [2,4].

Given that the anatomical regions of the brain for two people performing the same functions are not necessarily identical, a method to measure and localize brain function on the individual level is necessary to ensure functional preservation during brain tumor surgery. This procedure is interchangeably referred to as presurgical mapping, functional neuronavigation, and functionally guided surgery [1,2,4].

1.2 Direct Electro-cortical Methods

The first such methods to receive widespread scientific attention are still considered the gold standards by neurosurgeons today. These are cortical mapping by direct electrical stimulation (cortical stimulation, CS) and surface electrode recording (electrocorticography, ECoG) [5–8]. In cortical stimulation, an electrode is used to impart biphasic current to specific locations on the cortical surface, altering the excitability of the neural tissue in the vicinity. This allows critical functional regions to be identified by either interfering with the performance of a task, or by eliciting a sensorimotor response. Alternatively, ECoG uses an electrode array to measure the postsynaptic potentials related to task performance or spontaneous ictal activity.

These methods were investigated by renowned Canadian neurosurgeon Wilder Penfield in a series of influential studies from the 1930s to 1960s, mapping the functional organization of hundreds of patients on whom he performed neurosurgery [5,9,10]. CS and ECoG are highly invasive procedures, as they both require exposure of the cortex, limiting repeatability. Cortical stimulation also takes up valuable time in the operating room. An additional complication of CS is that the procedure requires the patient to be awake during the functional mapping stage, which is often

met with apprehension from the patient, who is already under significant stress related to coping with illness. ECoG in theory is capable of superb spatial resolution, capable of measuring the action potential of single neurons when placed within 50 µm [7]. In typical clinical practice, grids are not placed in close enough proximity to the neuron for single unit recording, and instead measure local field potentials of all neurons in the region. Grids may have electrode spacing on the order of 1-10 mm.

1.3 Non-invasive Methods

A non-invasive alternative to ECoG is electroencephalography (EEG), in which the electrical potentials from intracellular ionic currents are recorded from the scalp. Unfortunately, the EEG signal is blurred and attenuated by the large conductivity changes near the skull and scalp, resulting in lower SNR and spatial resolution, and making it challenging for functional neuronavigation [4]. A related technology that in part overcomes this limitation is magnetoencephalography (MEG) [2, 3, 11], achieving resolution in the range of 2-10 mm^3 [3, 11]. MEG measures the magnetic fields produced by synchronized neuronal current originating from post-synaptic currents. The fields produced are minute, and require sophisticated detection schemes.

Another promising complementary technology for presurgical mapping is functional MRI (fMRI). Functional MRI uses the same hardware as MRI, and therefore is widely available in the clinical setting [1,2,4]. It is also considered a safe, non-invasive technique, which means that it can be performed repeatedly, and at any time. Typical spatial resolution of clinical fMRI studies is 2-4 mm^3 [12–16]. Although fMRI studies have been performed with much higher resolution, the signal-to-noise ratio (SNR) is reduced, and therefore typically done only at higher magnetic fields [17]. This work will investigate functional MRI for presurgical mapping.

1.4 Methodological Challenges of Functional MRI

In the basic fMRI experiment, a task is presented following a known on/off timing structure, while rapid images are taken of the brain to observe the response. A model of the expected response to the task presentation is used as a predictor, and a statistical correlation is formed with the observed time series of each image element.

The resulting functional images are called statistical activation maps, and must have a threshold applied to determine the minimum level considered correlated with the task. The choice of threshold will have great impact on the accuracy of the fMRI map, and is especially critical in the context of presurgical mapping. This is particularly challenging given the low SNR typical of fMRI data [18].

In neuroscience research, this can often be overcome with averaging of multiple scanning sessions, or group level analysis. In the clinical setting, multiple tests are needed and inferences must be conducted at the individual level [4]. As a result, many of the procedures that are typically used to increase statistical power cannot be utilized for an individual patient fMRI scan. This leads to practical challenges associated with determining the optimal threshold settings, which will have a large effect on the diagnostic accuracy of the activation maps. The set of voxels above threshold would ideally not change within one subject from one scan to another, but in reality the construction of thresholded activation maps is susceptible to both type I and II errors (false positives and negatives, respectively). Error in the statistical processes used to construct the activation maps may be a result of any component of the data collection or processing strategy, and may represent inherent or accountable sources of variation [19]. Not only can voxels have erroneously supra or sub-threshold values, but the choice of threshold level is not objectively defined, and often based on heuristic motivation [20].

1.5 Ensuring the Validity of Presurgical fMRI

Researchers have taken a variety of approaches to validation of fMRI in both patients and healthy controls. This typically requires demonstrating convergent validity between two or more modalities. As such studies have been done comparing fMRI with CS [12–16, 20–30], EEG [31–34], MEG [32, 34–36], single-cell recordings [37], ECoG [7], and positron emission tomography [38]. General agreement is found between the methods, but it should not be surprising that the results vary between modalities as they are all subject to different limitations.

Results of comparisons between fMRI and cortical stimulation are of particular interest because CS is the gold standard for functionally guided brain tumor surgery. Kurth et al. [27] performed motor cortex mapping with fMRI and CS in 9 subjects,

finding an average separation between activation foci of 8.9 ± 6.0 mm. Pouratian et al. [28] used a battery of five tasks, and achieved 100% sensitivity, 66.7 % specificity in the frontal cortex, and 96.2% sensitivity, 69.8% specificity in the temporal cortex. Roux et al. [15] studied language localization in 14 right-handed patients with left hemisphere tumors. They found that a combination of two tasks could be used to achieve 59% sensitivity and 97% specificity compared to CS, whereas individually the tasks managed only 22% or 36% sensitivity. Rutten and coworkers [14] used a sample of 13 patients with temporal lobe epilepsy to compare the efficacy of four language tasks for presurgical mapping by fMRI. They found an overall sensitivity of 92% and specificity of 61% when all four tasks were used to compare with CS. More recently, using a verb generation task, Bizzi and coworkers [16] found 80% sensitivity and 78% specificity. Direct comparison of these results is complicated by the different proximity criteria used for calculating sensitivity and specificity of fMRI against CS.

While fMRI is the most commonly used non-invasive method for functional localization, there are a few caveats that are widely recognized. As will be discussed in chapter 2, the fMRI signal originates from a vascular effect, and is therefore an indirect measure of brain function [39]. Studying tumor patients with fMRI is especially challenging due to physiological complications, task performance difficulties, and patient availability [40–43].

1.6 Ensuring the Reliability of Presurgical fMRI

Variation in the extent of functional activation comes from many sources, making it difficult to identify the borders of eloquent tissue on the basis of fMRI alone [4, 44–50]. However, methods have been developed for accurately identifying critical functional zones [4,13,14,19,47,50–52]. Repeated imaging is commonly recommended, especially for higher-level cognitive functions. Typically this is done by repeating a task over multiple trials [4, 13, 19, 47], or performing single trials of multiple related tasks [14]. There are many approaches to quantifying fMRI reliability, including linear correlation [53], intraclass correlation (ICC) [54], dice similarity coefficient [55], and receiver operator characteristic (ROC) curve analyses [19].

The works of Genovese [19], and later Maitra [51] and Liou [47], indicate that ROC curve analysis of the repeatability of activation patterns can approximate the

true extent of activation. Their methods have yet to be applied extensively to clinical imaging [56], likely due to the increased imaging time required. Repeat imaging may also help in the identification and elimination of noisy data [4]. The mounting evidence has led several authors to advocate assessing reliability routinely for conducting presurgical fMRI [4, 13, 14, 48, 55].

The aim of this work is to investigate a novel method of ensuring reliability and accuracy of fMRI protocols for presurgical mapping at the individual level. The goal of this research will thus be to develop a tool capable of the following:

- Assessing reproducibility within the time constraints of a presurgical mapping protocol.
- Objectively determining thresholds for fMRI activation maps based on reproducibility.
- Producing automated thresholds that are diagnostically accurate when compared with CS.

Chapter 2

Functional MRI for Presurgical Mapping

Clinical functional MRI requires familiarity with imaging physics, physiology, and neuroscience principles, while maintaining perspective on their applicability in a clinical setting. This chapter will provide an introductory treatment of MR physics for the unfamiliar reader, as well as an overview of the relevant anatomy. The topic of the physical/physiological origin of the fMRI signal will be addressed, and followed by the theoretical considerations of fMRI data processing. Finally, the role of receiver operator characteristic (ROC) curves and test-retest imaging in analyzing individual functional MRI results will be explained, motivating the work undertaken herein.

2.1 Magnetic Resonance Imaging

2.1.1 Nuclear Magnetic Resonance

The basis of most MRI is the interaction of the nuclei in human tissue with a strong (1-10 Tesla) static magnetic field ($\vec{B_0}$):

$$\vec{B_0} \equiv B_0 \hat{z} \tag{2.1}$$

Species with an odd number of protons or neutrons have an unpaired nuclear spin, resulting in a non-zero net spin angular momentum ($\vec{I}$) and magnetic dipole moment ($\vec{\mu}$). While any nucleus with an odd number of protons or neutrons is suitable for MR, the spin-$\frac{1}{2}$ ^{1}H nucleus is by far the most commonly imaged species, because of its relative biological abundance and high sensitivity (table 2.1). The nuclear magnetic dipole experiences a torque ($\vec{\tau}$) given by the cross product of $\vec{\mu}$ and $\vec{B_0}$, tending to align it with the main magnetic field. Given that the nucleus has an existing angular momentum, the spin will precess about the main magnetic field, much like a top precesses about its axis of rotation because of the force of gravity.

The nature of this precession is elucidated by application of the Schrödinger equation. Recognizing the interaction energy of the dipole and magnetic field as the

negative dot product of $\vec{\mu}$ and $\vec{B_0}$, the Hamiltonian is:

$$U = -\vec{B_0} \cdot \vec{\mu} \tag{2.2}$$

$$= -\gamma B_0 I_z \tag{2.3}$$

$$= -\gamma \hbar B_0 S_z \tag{2.4}$$

where $\hbar$ is the reduced Planks constant (1.055×10^{-34} $\mathrm{m^2kgs^{-1}}$), and γ is the gyromagnetic ratio of the nuclei being imaged (the ratio $|\mu|/|I|$, in $\mathrm{rads^{-1}T^{-1}}$). I_z and S_z represent the components of angular momentum or nuclear spin parallel to the static field respectively. In this case, the time evolution operator is equivalent to a clockwise rotation about the static magnetic field, with an angular frequency of oscillation (ω_0) given by the Larmor equation:

$$\omega_0 = \gamma B_0 \tag{2.5}$$

The energy difference (Zeeman splitting) between $\vec{\mu}$ and B_0 in an aligned ($\uparrow\uparrow$) and anti-aligned ($\uparrow\downarrow$) state causes a slight polarization given by the Boltzmann distribution. It can easily be shown that the polarization fraction (P) is:

$$P \equiv \frac{n_{\uparrow\uparrow} - n_{\uparrow\downarrow}}{n_{\uparrow\uparrow} + n_{\uparrow\downarrow}} \simeq \frac{\gamma \hbar B_0}{2k_B T} \tag{2.6}$$

where k_B is Boltzmanns constant (1.23×10^{-23} $\mathrm{JK^{-1}}$), and T is the temperature. The polarization fraction is on the order of parts per million for conventional field strengths. As it is this polarization that creates the small net nuclear magnetization

Table 2.1: MR characteristics of the most commonly found elements in the human body. BA: Biological abundance (percent), Spin: Nuclear quantum spin number, γ: gyromagnetic ratio in $\mathrm{MHz}T^{-1}$ [57, 58]

Nucleus	BA	Spin	γ
^{1}H	63	1/2	42.6
^{16}O	24	0	0.0
^{12}C	12	0	0.0
^{14}N	0.58	1	3.1
^{31}P	0.14	1/2	17.2
^{13}C	0.12	1/2	10.7
^{17}O	0.094	5/2	-5.8
^{23}Na	0.037	3/2	11.3
^{15}N	0.0021	1/2	-4.3
^{19}F	0.0012	1/2	40.0

(M_0) we will exploit to produce the MR images, the MR signal is relatively weak. The hydrogen nucleus has a second advantage here, in that the large gyromagnetic ratio (42.6 MHzT^{-1}) leads to a large polarization fraction. Using higher field strengths will increase the signal to noise further [59].

2.1.2 Excitation and Detection

At this point we have a net longitudinal magnetization (M_z) of ^{1}H nuclei precessing throughout the body with the same frequency (ignoring the effects of local field variations). However, the ensemble of spins lacks phase coherence, and so there is no transverse magnetization component. This is not appropriate for imaging the ^{1}H nuclei, as the magnetization will not induce signal in the MRI receive coil. To induce a signal, we require a coherent component of the magnetization rotating in the transverse plane (M_{xy}) [59, 60].

Thus the second key ingredient in MRI is a mechanism for manipulating the spin system, to induce a coherence of the transverse magnetization, such that an appreciable oscillatory transverse field will exist at the Larmor frequency. For the field strengths used in typical human imaging systems, the Larmor frequency is in the RF range (60-400 MHz), and happily coincides with a window of electromagnetic transparency in the human body [59].

Consider the case of a circularly polarized RF field applied at the Larmor frequency, rotating in the same sense as the spin precession, creating a time-varying magnetic field ($\vec{B_1}$):

$$\vec{B_1} = B_1 \cos\omega_0 t\hat{\imath} - B_1 \sin\omega_0 t\hat{\jmath} \tag{2.7}$$

Furthermore, treating the ^{1}H nuclei as independent spin-$\frac{1}{2}$ particles, we may adopt a semi-classical ensemble description, replacing the magnetic dipole $\vec{\mu}$ with the magnetization $\vec{M}$. The macroscopic angular momentum becomes $\vec{M}/\gamma$. In this framework, the torque of the combined B_0 and B_1 fields causes a change in angular momentum governed by:

$$\vec{\tau} = \frac{1}{\gamma}\frac{d\vec{M}}{dt} \tag{2.8}$$

$$= \vec{M} \times \vec{B} \tag{2.9}$$

where $\vec{B}$ is the total magnetic field:

$$\vec{B} = \vec{B_0} + \vec{B_1} \tag{2.10}$$

Under the initial condition of magnetization along the polarizing axis, the solution to this differential equation is a simultaneous rotation about $\vec{B_0}$ and $\vec{B_1}$ at angular frequencies of γB_0 and γB_1 respectively (figure 2.1). In frame of reference rotating at the Larmor frequency, the magnetization vector simply precesses about the RF ($\vec{B_1}$) field. This process is referred to as RF excitation.

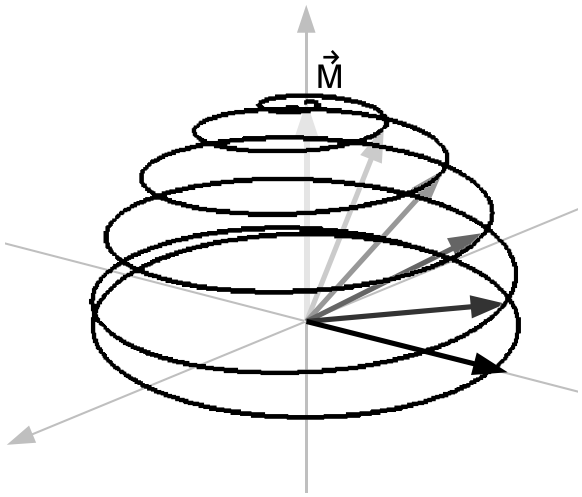


Figure 2.1: Excitation in the lab frame. The net magnetization vector is rotated away from the B_0 ($\hat{z}$) axis by the circularly polarized RF field B_1 while precessing rapidly at the Larmor frequency (in practice B_0 is so large in comparison with B_1 that several orders of magnitude more rotations about the $\hat{z}$ axis occur than shown).

By manipulating the magnitude, phase, and duration of a series of RF pulses, the spin system can be manipulated. The most elementary example is a single RF pulse of appropriate magnitude and duration (t_{pulse}) to cause a 90-degree rotation about the $\vec{B_1}$ axis ($\theta = \gamma B_1 t_{pulse}$). This results in the equilibrium longitudinal magnetization being rotated into the transverse plane coherently, where it will precess at the Larmor frequency. This transverse magnetization can be detected by suitable hardware (in fact, the hardware used to create the RF pulse is by design resonant at the Larmor frequency, and with suitable control circuitry can be used as both a transmit and receive device), and constitutes the MRI signal [59, 60].

2.1.3 Relaxation

The excited spin system will not stay in a state of coherent transverse magnetization indefinitely. Indeed, we started our description by postulating a thermal equilibrium polarization along the applied field axis, which is only possible if there is a mechanism for the spins in the system to return to the low-energy equilibrium state. The probability of this happening through spontaneous emission is far too small to provide a satisfactory explanation; relaxation in ^{1}H MR systems is described as a stimulated emission phenomenon [59]. Typically longitudinal magnetization recovery is called spin-lattice relaxation or T_1 decay, and the loss of transverse magnetization is called spin-spin relaxation or T_2 decay, or simply signal decay.

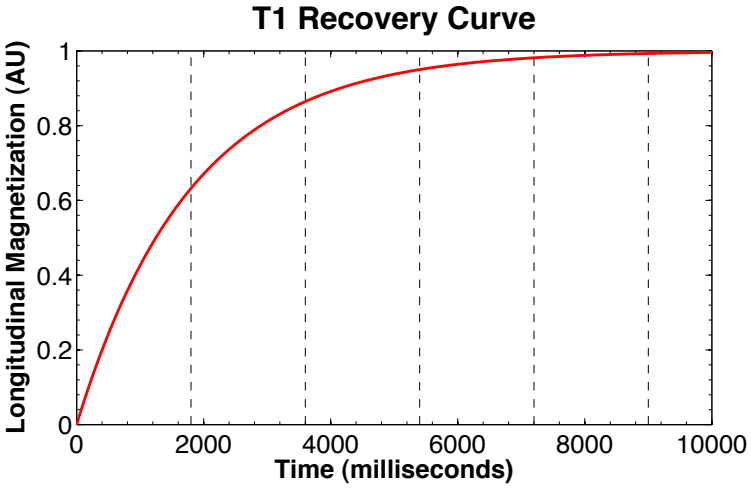


Figure 2.2: Longitudinal recovery is an exponential process (red), with a time constant T_1. Vertical dashed lines show multiples of T_1 (1800 ms in this example). Full longitudinal recovery can generally be assumed to occur at t = $5T_1$.

Spin-lattice relaxation is the exchange of energy between the spin system and its environment, which acts as a thermal reservoir. The rate of longitudinal relaxation can be found by treating the dipolar interaction with the lattice by perturbation theory. The phenomenological result is that the longitudinal magnetization recovers exponentially from its initial value ($M_z(0)$) to the equilibrium value (M_0) as:

$$M_z(t) = M_z(0)e^{-t/T_1} + M_0(1 - e^{-t/T_1}) \tag{2.11}$$

with a rate constant $1/T_1$ (figure 2.2). The rate of longitudinal relaxation is proportional to the spectral intensity of lattice fluctuations at the Larmor frequency, and

therefore depends on there being translational, vibrational, or rotational lattice modes at the correct frequencies [59]. As the primary mechanism for relaxation comes from dipolar interactions, any change in the interaction strength or characteristic frequency will affect the relaxation rates.

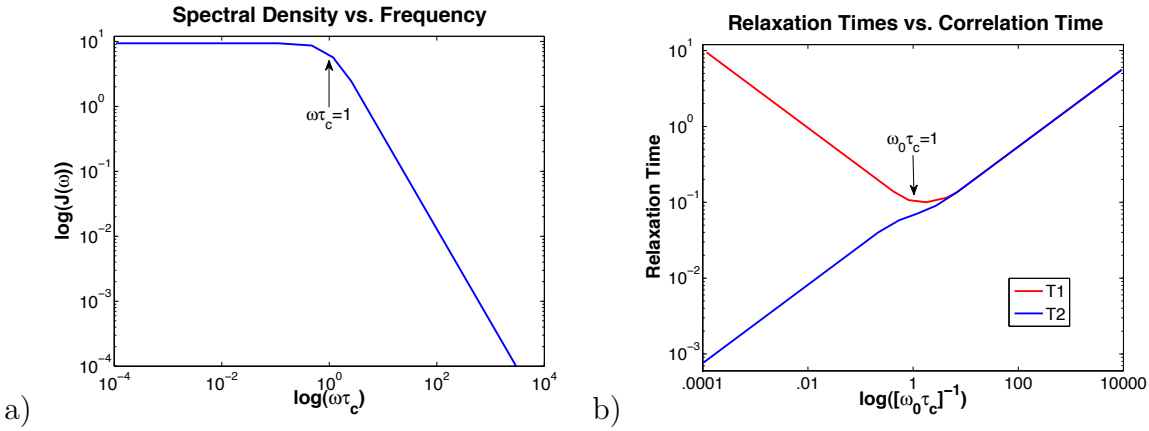


Figure 2.3: a) Spectral density as a function of frequency (ω), at a constant rotational correlation time (τ_C). b) T_1 and T_2 relaxation constants as a function of the rotational correlation time of the nuclei, at a constant Larmor frequency (ω_0). Reproduced from Callaghan [59].

The transverse magnetization certainly cannot exist longer than the longitudinal magnetization takes to recover, and in general the signal decays much more rapidly than this limit (figure 2.3). The reason is that the longitudinal decay is insensitive to phase coherence, whereas transverse magnetization exists only for as long as the contributing spins are precessing in phase. This process is called spin-spin relaxation because it is the process by which the spins come to thermal equilibrium with each other. Again, phenomenologically this is an exponential process, with a rate constant $1/T_2$ (figure 2.4):

$$M_{xy}(t) = M_{xy}(0)e^{-t/T_2} \tag{2.12}$$

Transverse relaxation is affected by the same factors as longitudinal relaxation. Additionally transverse relaxation is expedited by static field inhomogeneities within an imaging element, which cause spins to precess with different frequencies, destroying phase coherence. It is common to divide transverse relaxation caused by inherent random processes, and by static inhomogeneities. The observed relaxation time constant

$(1/T_2^*)$ is:

$$\frac{1}{T_2^*} = \frac{1}{T_2} + \frac{1}{T_2'} \tag{2.13}$$

where T_2' is the extra relaxation term due to longitudinal field offsets, and may be a result of main field inhomogeneity, local susceptibility effects, or both [59, 60].

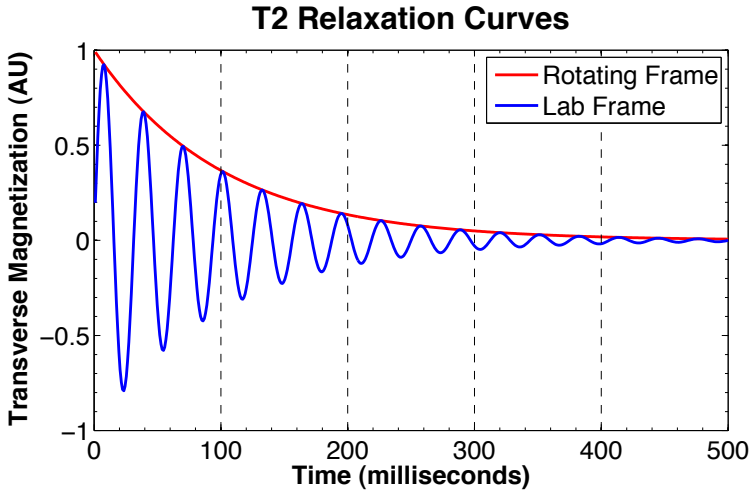


Figure 2.4: Transverse relaxation is an exponential process in the rotating frame (red), with a time constant T2. In the lab frame, the exponential decay of the signal is modulated by the Larmor frequency precession frequency (blue). For diagrammatic purposes, the Larmor frequency modulation is greatly reduced. Vertical dashed lines show multiples of T2 (100 ms in this example). Full transverse decay can generally be assumed to occur at t=5T2.

The local susceptibility effects are often the dominant source of transverse relaxation in biological systems, which are quite inhomogeneous by nature. The total magnetic field experienced by a nucleus ($\vec{B}$) is proportional to sum of the applied field ($\vec{H}$) and the sample magnetization ($\vec{M}$):

$$\vec{B} = \mu_0(\vec{H} + \vec{M}) \tag{2.14}$$

where μ_0 is the permeability of free space ($4\pi 10^{-7} NA^{-2}$), and not to be confused with the magnetic dipole moment. An alternative casting in terms of the magnetic susceptibility (χ_m) is:

$$\vec{B} = \mu_0(1 + \chi_m)\vec{H} \tag{2.15}$$

where the magnetic susceptibility can be interpreted as the proportionality constant between $\vec{M}$ and $\vec{H}$ for most non-ferromagnetic materials (in ferromagnetic materials,

hysteresis must be accounted for). Most biological tissue is diamagnetic or paramagnetic, where for diamagnetic materials $\chi_m < 0$, and for paramagnetic materials, $\chi_m > 0$. For spatially dependent χ_m, the field experienced by the spins in the sample is spatially dependent. This causes spins to precess with different frequencies, and leads to the phase dispersion that is responsible for the shortened coherence times. The complex dependence of transverse and longitudinal relaxation times on physical and physiological conditions gives rise to the wealth of contrast options available in MRI [59, 60].

2.1.4 Spatial Encoding (aka Imaging)

So far we have considered only the case of a relatively homogenous longitudinal field, in which case the Larmor frequency is approximately the same at all locations ($\omega(\vec{r})$=ω_0). In this case, the signal (dS) from a volume element in the image (dV) is simply an oscillatory function that depends on the density of hydrogen nuclei (ρ), and the particular image contrast employed (C):

$$dS(\vec{r},t) \propto C(\vec{(}r))\rho(\vec{(}r))e^{i\omega(\vec{r})t}dV \tag{2.16}$$

Without loss of generality, we will ignore the proportionality constant in 2.16, and call the product of C and ρ the contrasted spin density ρ'.

Spatial locations can be encoded in one dimension by introducing a spatially dependent component of the longitudinal field. This is generally accomplished by imposing a linear gradient ($\vec{G}$) in the encoding direction. The gradient introduces spatial dependence to the Larmor frequency:

$$\omega(\vec{r}) = \gamma(B_0 + \vec{G}\cdot\vec{r}) \tag{2.17}$$

Ignoring the spatially uniform oscillations of the main field term, and integrating over the imaging volume:

$$S(t) = \iiint_V \rho'(\vec{r})e^{i\gamma\vec{G}\cdot\vec{r}t}d\vec{r} \tag{2.18}$$

Sir Peter Mansfield realized that this bore a great deal of resemblance to the reciprocal lattice of X-ray diffraction (and would later win a Nobel prize for his efforts). Defining the reciprocal space vector ($\vec{k}$) as:

$$\vec{k} \equiv \frac{\gamma\vec{G}t}{2\pi} \tag{2.19}$$

we see that the signal is collected in reciprocal space, and forms a Fourier pair with the contrasted spin density:

$$S(\vec{k}) = \iiint \rho'(\vec{r}) e^{i2\pi\vec{k}\cdot\vec{r}t} d\vec{r} \tag{2.20}$$

$$\mathcal{FT} \Updownarrow \mathcal{FT}^{-1}$$

$$\rho'(\vec{r}) = \iiint S(\vec{k}) e^{-i2\pi\vec{k}\cdot\vec{r}t} d\vec{k} \tag{2.21}$$

The fact that the signal is collected directly in the Fourier conjugate space of the image has interesting ramifications. First of all, sampling in 3D k-space can be simply achieved by controlling the strength, direction, and duration of the linear gradients. Secondly, sampling enough discrete locations in k-space to reasonably approximate the integral of equation 2.20 will allow us to reconstruct the image by Fourier transform. Thirdly, equations 2.20 and 2.21 tell us how changing the sampling scheme in k-space will affect the images, through the properties of the Fourier transform.

While it is possible to collect an entire 3D image by traversing 3D k-space after a volumetric excitation, it is more common to excite only a thin slice at a time, and resolve the plane through 2D k-space encoding. How much k-space data must be collected, and what the best order (i.e. gradient-time waveform) of collection are both depend on the application. In general, the low frequency k-space data represent large structures, and provide image contrast, whereas high values in k-space encode the fine details (i.e. resolution). The modification of collection schemes, introduction of contrast agents, or changes in local tissue properties can change the contrast obtained, and is largely determined by the signal at the k-space origin [59]. One of the primary sources of contrast in MR imaging is the modulation of the signal intensity by T_1 and T_2 relaxation.

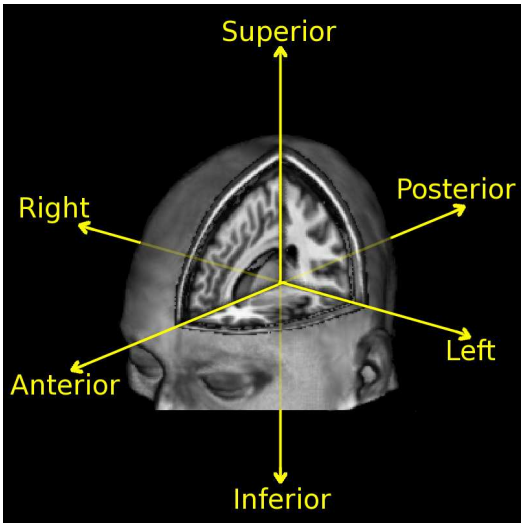
2.2 Functional Brain Anatomy

Discussing functional brain imaging is rather difficult without reference to various structures and locations in the brain. The next section of this chapter is therefore devoted to an introductory review of the relevant anatomical terms and concepts that arise when discussing the imaging methods used herein. This section is not intended

to provide an exhaustive treatment of the subject matter, and readers are directed towards devoted textbooks for further information [61, 62].

2.2.1 Directions, Axes and Planes

To describe the localization of function in the brain, it is useful to start by defining the basic anatomical directions (figure 2.5a), which give rise to the three planes in which diagnostic images are typically viewed (figure 2.5b). While the exact terms used to express anatomical directions will vary with morphological differences from species to species, or even body part to body part, the terms used for the human brain are relatively straightforward and consistent. The front of the head is referred to as the anterior direction, while the back of the head is posterior. The top of the head is in the superior direction, while inferior is towards the feet. Left and right are self-explanatory, but one should note that because of the near symmetry of this axis, lateral and medial are also often used, indicating whether the object being described is close to or far from the left-right midline. A few other terms that may be encountered are ipsilateral and contralateral, meaning on the same or opposite side as another structure respectively. Superficial and deep form another opposing pair, meaning near to or far from the surface.

a) 

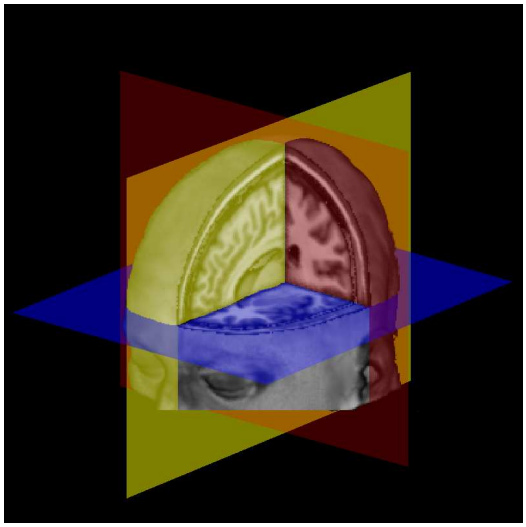
b) 

Figure 2.5: Anatomical directions and planes displayed on a three-dimensional render of a high-resolution T_1 weighted MRI. a) The three major axes are: left-right, anterior-posterior, and superior-inferior. b) The three major planes are: axial (blue), coronal (red) and sagital (yellow).

The main three planes in which brain images are formed are the axial, sagital, and coronal planes (figure 2.7). These are each defined in terms of an orthogonal pair of axes created by the opposing directional pairs given above. The axial plane is that formed by the anterior-posterior (AP) and left-right (LR) axes, the sagital plane by the AP and superior-inferior (SI) axes, and the coronal plane by the LR and SI axes. Any other plane can be defined by arbitrary combination of the orthogonal AP, LR, and SI axes, and such planes are commonly referred to as oblique. Neuroscientists typically view coronal and axial images with the left side of the anatomy displayed on the left side of the page (neurological view), while clinicians view these planes with the left side displayed on the right of the page (radiological view).

2.2.2 Cortical Organization

The brain consists of the two cerebral hemispheres, the cerebellum, and the brainstem. The cerebral cortex is by far the largest of these structures, and is particularly enlarged in humans as compared to other vertebrates. The cerebral cortex has a convoluted topology, formed by concave sulci (figure 2.6a) and convex gyri (figure 2.6b). The cerebrum is divided into four lobes: frontal, parietal, occipital, and temporal (figure 2.6c). Much finer organization can be described based on individual gyral folds, or alternatively on cytoarchitectural or functional delineations (figure 2.6d).

There are some regions of the cerebrum that demand special attention during functionally guided surgery, because of increased risk of postoperative functional deficits associated with their removal. The central sulcus divides the frontal and parietal lobes and also the primary motor area anteriorly from the primary sensory cortex posteriorly. The primary sensory and motor cortices are further organized from the medial to lateral extent, as shown in figure 2.7. Both the primary sensory and motor cortices are responsible for the contralateral side of the body. The calcarine sulcus at the occipital pole contains a representation of the visual field, damage to which can cause visual impairment.

While distributed in nature, critical aspects of higher cognitive functions can also be functionally localized. However, cognitive tasks frequently engage several cortical regions, and are considered to exhibit higher variability. Of note is language function, which shows a high degree of asymmetry to the left hemisphere. Two critical

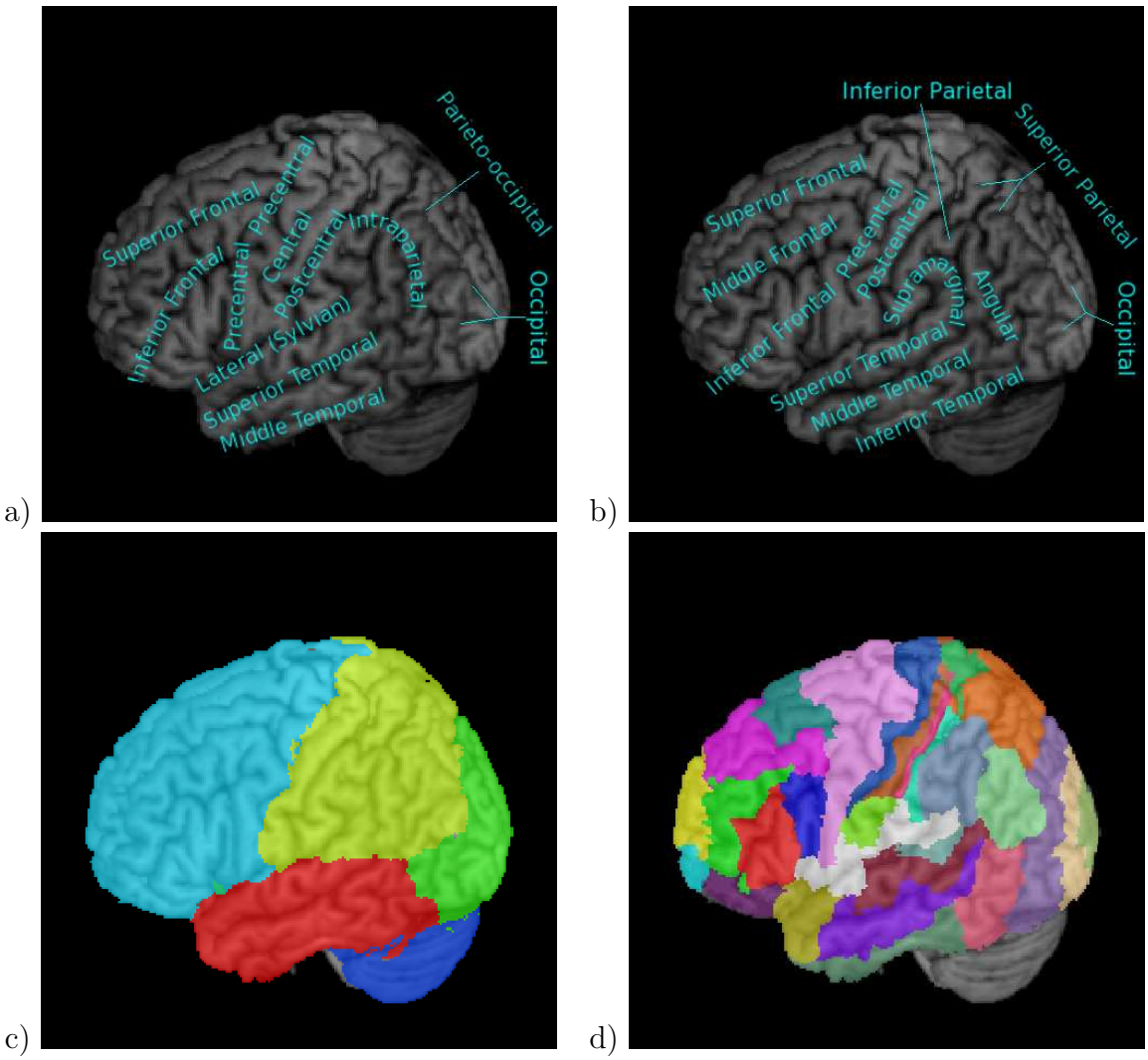


Figure 2.6: a) Major surface sulcal anatomy. b) Major surface gyral anatomy. c) The four main lobes of the cerebrum: frontal (cyan), parietal (yellow), occipital (green), and temporal (red). Also shown is the cerebellum (blue). d) The cerebrum can be divided according to cell type by the widely referenced Brodmann areas.

language-specialized regions are Broca's and Wernicke's, defined by the specific language deficits that arise form damage to these regions (or less commonly to their right hemisphere homologues). Wernicke's area is in the postero-superior temporal gyrus, and is involved in receptive comprehension for both input through primary visual cortex (i.e. reading) and auditory cortex (i.e. listening). Broca's area is located in the inferior frontal gyrus, and is responsible for articulating words. It is thus common to refer to Wernicke's area as the receptive, and Broca's area as the expressive, language zones [64].

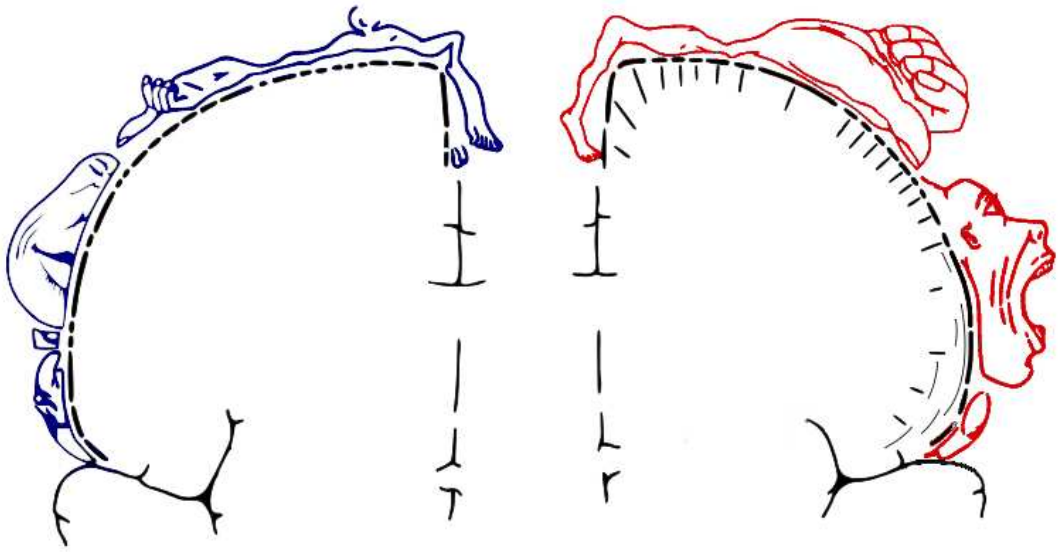


Figure 2.7: The Penfield sensory (blue) and motor (red) homunculi. These show the relative representations and locations of the major body parts along the primary motor and sensory cortices respectively [63].

2.3 Biophysical Origin of the BOLD fMRI signal

2.3.1 The Neural Basis of Brain Activation

The neuron is thought to form the basic unit of brain function, although even the simplest behaviors are the result of a concerted effort of networks composed of millions of neurons. A typical neuron is shown in figure 2.8, consisting of a cell body (soma), dendrites, an axon, and axon terminals. The exact form of a neuron is highly variable, expressing regional and functional differences [65].

At a basic level, the nerve cells function is to process and transmit information. The neuron has a cell membrane capable of controlling the flow of molecules, which allows it to establish an ionic potential difference across the cell wall, called the membrane potential. The nerve cell receives input signals from other cells at the dendrites in the form of neurotransmitters changing cellular regulation of specific ions (chiefly Na^+ and K^+), and causing deviations from the resting membrane potential. For instance, a typical ionotropic transmission may involve excitatory input increasing the inward flux of Na+ ions through voltage sensitive ion channels, thereby causing the trans-membrane potential to increase. If the membrane potential increases to

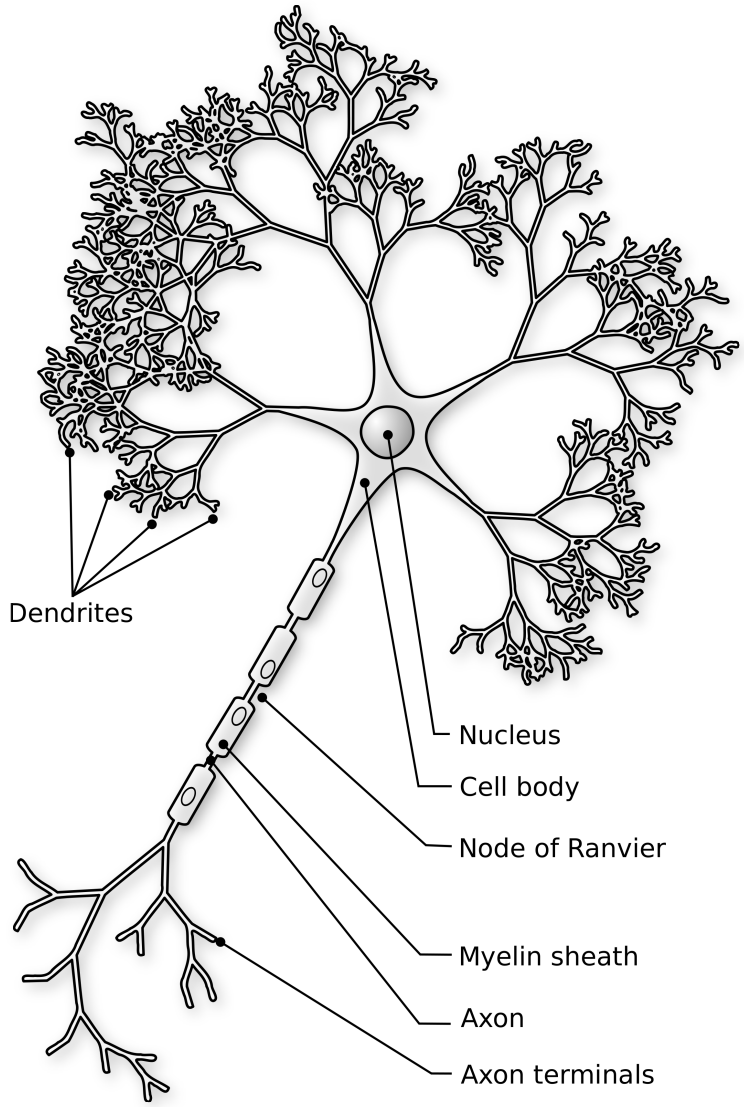


Figure 2.8: Structure of a typical neuron. Transmission of neuronal signals originates with input neurotransmitter at the dendrites, and culminates in an action potential down the axon, and release of neurotransmitter at the axon terminals [63].

a threshold, an action potential is triggered by the positive feedback mechanism of the voltage sensitive ion channels (figure 2.9). The action potential rapidly and dramatically increases the membrane potential, causing cascading depolarization down the axon to the axon terminals, where neurotransmitter substance is released to pass the signal on to other neurons or target cells. A negative feedback loop of voltage sensitive channels that remove K^+ from the cell ensure that the resting membrane potential is restored. The action potential is described as an all-or-none response

because the signal magnitude is independent of the stimulus magnitude so long as the threshold potential is reached [65].

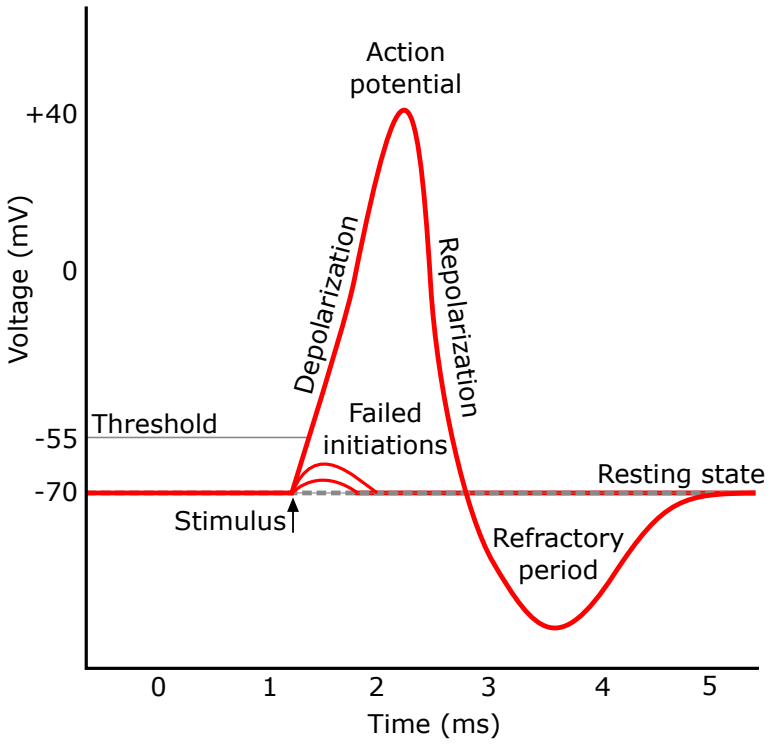


Figure 2.9: Schematic representation of an action potential. If the input stimulus is strong enough to reach the threshold potential, the action potential is initiated, and full depolarization occurs. After the repolarization phase, there is a refractory period during which no new action potentials can occur [63].

2.3.2 The BOLD Response

Functional MRI is not sensitive to individual action potentials. However, a region of cortex involved in excitatory and inhibitory activity experiences increased metabolic demand to facilitate events like restoration of resting potentials, and regulation of neurotransmitter substances [64]. This induces a vascular response, increasing cerebral blood flow (CBF) and cerebral blood volume (CBV) to provide oxygenated blood for the increase in metabolic rate of oxygen consumption ($CMRO_2$). Functional MRI is sensitive to these vascular changes, through the blood oxygen level dependent (BOLD) contrast mechanism.

To understand the origin of BOLD contrast, we must examine the vascular system.

The blood is made up primarily of plasma and red blood cells. The fraction of blood volume that is red blood cells is called the hematocrit (Hct). The red blood cells contain hemoglobin (Hb), which bind with oxygen by iron containing heme groups. When the Hb molecule is bound to oxygen, the iron binding sites are shielded, and the molecule is diamagnetic. When the oxygen is unbound, the molecule is paramagnetic because of the unbound electrons. We define Y as the fractional oxygenation of the red blood cells. The bulk susceptibility of the blood (χ_{blood}) can therefore be modeled as:

$$\chi_{blood} = Hct(Y\chi_{oxy} + (1-Y)\chi_{deoxy}) + (1-Hct)\chi_{plasma} \tag{2.22}$$

where χ_{oxy}, χ_{deoxy} and χ_{plasma} are the magnetic susceptibilities of oxygenated Hb, deoxygenated Hb, and blood plasma respectively.

During excitatory activity the relative concentration of oxygenated hemoglobin to deoxyhemoglobin can be modeled as:

$$\Delta Y = (1-Y)\left(\frac{\Delta CBF/CBF - \Delta CMRO_2/CMRO_2}{\Delta CBF/CBF + 1}\right) \tag{2.23}$$

The changes in cerebral blood flow far exceed those of oxygen utilization (i.e. there is an exaggerated change in supply relative to the change in demand):

$$\Delta Y = (1-Y)\left(\frac{\Delta CBF/CBF}{\Delta CBF/CBF + 1}\right) \tag{2.24}$$

leading to an increased fractional oxygenation of the blood. This comes with a concomitant decrease in blood susceptibility:

$$\chi_{blood} = \Delta Y(\chi_{oxy} - \chi_{deoxy})Hct \tag{2.25}$$

Recall that the effective transverse relaxation rate ($1/T_2^*$) is sensitive to magnetic susceptibility changes. It is therefore this change in blood susceptibility that will cause a change in the signal decay rate, and creates contrast between active and inactive brain states. The relationship between the neural processes and the vascular (hemodynamic) response is called neurovascular coupling. The hemodynamic response function (HRF) is slow compared to the time scale of neural events (figure 2.10). While individual action potentials may take only milliseconds, the hemodynamic response has a temporal scale of several seconds.

It has been suggested that the fundamental limit of spatial resolution for fMRI may be approximately 1 mm due to the granularity of the vascular system [4]. However, the precise nature of the physical-physiological contrast mechanism is complex, and depends on field strength, pulse sequence, local tissue properties, and analysis methods [59, 60]. The image resolution of most clinical fMRI studies is 2-4 mm, and so each image element can be expected to contain millions of neurons, and even several tissue types. The latter consideration is called the partial volume effect, and can reduce sensitivity. Studying tumor patients with fMRI presents unique challenges. Tumors can lead to alterations in vascularization patterns, reorganization of activity patterns, and decoupling of the neural activity from the blood flow response [40, 43]. Determining the extent to which BOLD response is spatially specific is crucial to the successful use of fMRI mapping for surgical planning.

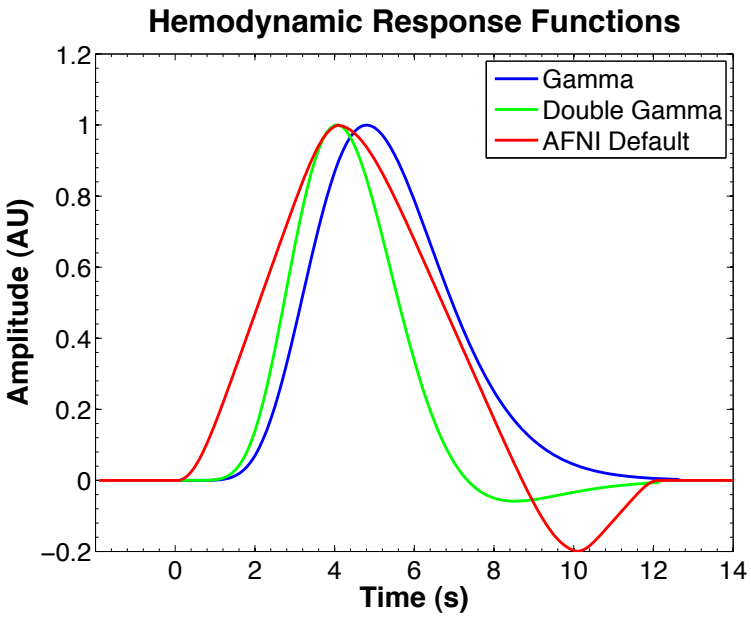


Figure 2.10: Example HRF used to model the neurovascular coupling. The gamma function response (form $t^n e^{-t/m}$) is available in many software packages (blue). The double gamma response includes a second gamma function to model the post stimulus undershoot, and is available in the popular FSL software (green). The default in the AFNI software package is a non-gamma function based HRF that somewhat resembles the form of the double gamma (red).

2.4 Functional MRI Processing

Functional MRI data are typically analyzed using either a model-based approach or a model-free approach. The discussion below address the more common of these two, the model-based approach.

2.4.1 The General Linear Model

The paradigm for most fMRI experiments is to measure the hemodynamic response to a time varied task stimulus (figure 2.11). Let f(t) represent the expected task response function (i.e. the task timing function convolved with the canonical hemodynamic response function), and y(t) the actual measured response of a voxel. The response of an active voxel can be modeled as the weighted sum of f(t), a polynomial to account for average signal intensity and baseline drift (2nd order is usually sufficient), and a random error term (ϵ(t)):

$$y(t) = \alpha_0 f(t) + a_0 + a_1 t + a_2 t^2 + \epsilon(t) \tag{2.26}$$

where α_0, a_0, a_1, and a_2 are the weights calculated by least squares fit. The response of an inactive voxel is expected to obey the baseline model:

$$y(t) = b_0 + b_1 t + b_2 t^2 + \epsilon(t) \tag{2.27}$$

different weights (b_i) are used in the fit of the baseline to reflect that these parameters are not necessarily identical.

Of course the actual measurement is performed at N discrete time points, separated by the repetition time, T_R. Equations 2.26 and 2.27 can thus be written as matrix equations relating the observed time series y to the modeled response in the form:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \tag{2.28}$$

where $\boldsymbol{X}$ is the design matrix, containing the predicted response, and baseline functions, $\boldsymbol{\beta}$ is the matrix of weights, and $\boldsymbol{\epsilon}$ is random Gaussian error matrix:

$$\boldsymbol{y} = \begin{bmatrix} y_0 \\ y_1 \\ \vdots \\ y_{N-1} \end{bmatrix}, \boldsymbol{X} = \begin{bmatrix} f_0 & 1 & 0 & 0 \\ f_1 & 1 & 1 & 1 \\ \vdots & \vdots & \vdots & \vdots \\ f_{N-1} & 1 & N-1 & (N-1)^2 \end{bmatrix}, \boldsymbol{\beta} = \begin{bmatrix} \alpha_0 \\ a_0 \\ a_1 \\ a_2 \end{bmatrix}, \boldsymbol{\epsilon} = \begin{bmatrix} \epsilon_0 \\ \epsilon_1 \\ \vdots \\ \epsilon_{N-1} \end{bmatrix} \tag{2.29}$$

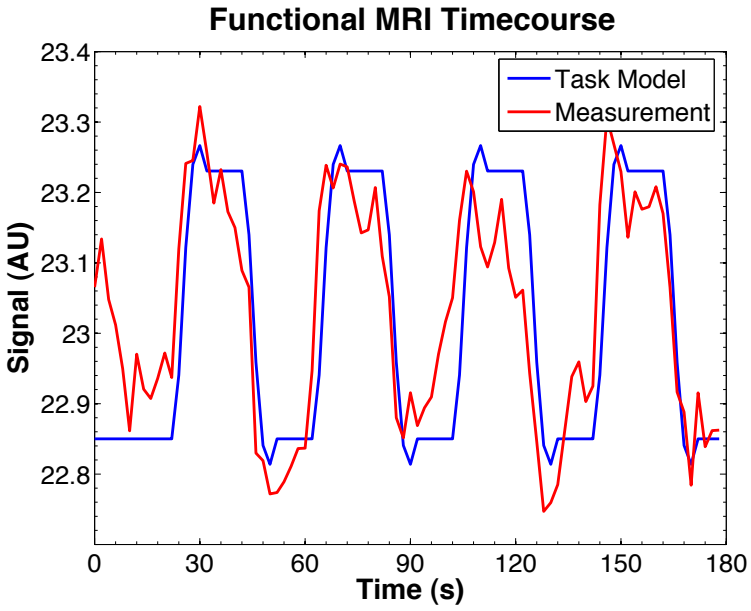


Figure 2.11: Example fMRI time-course modeled by convolution of the HRF with the stimulus function and addition of a DC baseline (blue), and a raw voxel time-course from the subsequent experiment (red).

This formulation is easily generalized for multiple independent stimuli by adding predictor task response functions (e.g. f(t), g(t), h(t), etc.) with independent weights (α_i). To fit the model, the time series is estimated as:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{b} \tag{2.30}$$

It can be shown that the estimate of $\boldsymbol{\beta}$ given by $\boldsymbol{b}$ that minimizes the sum of squares error with the measured data is given by:

$$\boldsymbol{b} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y} \tag{2.31}$$

Estimation of the regressor weights is thus reduced to linear algebra [66].

2.4.2 The Statistical Parametric Map

Once the best-fit parameters for the model are determined, an analysis of the goodness of fit is performed. Typical approaches to determining the goodness of fit include the t-test (t^*), the F-test (F^*), and the coefficient of determination (R^2). The t-test is a common measure of statistical significance, and was used throughout this work. For

large N (more than a few dozen time points), the t^* of a regression parameter (b_i) for an individual voxel is given by:

$$t^* = \frac{b_i}{\sigma(b_i)} \tag{2.32}$$

where $\sigma(b_i))$ is the sample standard deviation of the estimate for b_i, and is found from the corresponding diagonal element of the variance-covariance matrix, $\sigma^2(\boldsymbol{b})$:

$$\sigma^2(\boldsymbol{b}) = MSE \cdot (\boldsymbol{X}^T\boldsymbol{X})^{-1} \tag{2.33}$$

where MSE is the mean squared error of $\boldsymbol{b}$ [66].

A t^* value is calculated in such a manner for every voxel to create the statistical activation map. To mask out the inactive voxels and reduce the image to a comprehensible form, a minimum t^* value level for significant activation is chosen. As the t^* statistic is generated from a voxel-based analysis, setting significance threshold based on theoretical error rejection rates is inappropriate. Consider the case of using the t^* value corresponding to a p-value (i.e. type I error rate, or false positive rate) of 0.05. With a standard fMRI image of grid dimensions 64×64×20, there are 81920 voxels, and so under the null hypothesis we would expect 0.05×81920=4096 false positive voxels. This is clearly an unacceptable level of false alarms, so we need to turn to more conservative methods of setting significance levels.

2.4.3 Thresholding of Statistical Parametric Maps

Improvements on the voxel-wise error control can be divided into those that use a fixed significance level (like the voxel-wise method described above) and those that do not. Methods that use a fixed significance include the Bonferroni correction for family-wise error, and fixed cluster-level statistics. Family-wise error corrections adjust the significance threshold to account for the large number of voxels, and are the most commonly performed statistical correction in fMRI data analysis. The most basic family-wise error correction is the Bonferroni correction. The Bonferroni correction adjusts the acceptable error rate by the number of independent tests (ie. voxels), V:

$$p_B = \frac{p'}{V} \tag{2.34}$$

where p_B is the corrected voxel-wise error rate for a family-wise error rate of p' [67].

Bonferroni corrections have been criticized for being overly strict [67]. An improvement on the traditional Bonferroni correction can be made by correcting only for the number of voxels inside the brain, as defined by the high-resolution anatomical image. This approach is reasonable, as long as these voxels are masked out in the analysis process.

Another alternative is to use a more lenient correction scheme for the multiple comparison problem, such as the false discovery rate (FDR). The FDR method controls the expected ratio of false positives to total positives:

$$FDR = \frac{FP}{FP + TP} \tag{2.35}$$

The strategy for controlling the FDR suggested by Genovese et al. [68] is first to order the voxels from lowest to highest p-value (highest to lowest significance) in a 1D array 'p_i'. The corrected voxel-wise error rate (p_G) for an FDR of q is:

$$p_G \equiv max\left\{i : p_i \leq \frac{iq}{Vc(V)}\right\} \leq \frac{iq}{Vc(V)} \tag{2.36}$$

that is, it is the p_i corresponding to the largest value of 'i' which satisfies the inequality of 2.36. The c(V) term in the denominator is an optional correction for spatial correlations in the SPM that may be appropriate for fMRI:

$$c(V) = \sum_{i=1}^{V} \frac{1}{i} \simeq \ln V + 0.5772 \tag{2.37}$$

where the approximation is made for large V. This method controls the FDR at 'q' on average. The Genovese procedure can be viewed graphically as finding the intersection of the ordered data vector and the function (FDR(i)) given by the right of the inequality in 2.36 (figure 2.12) Given the large number of voxels, and the practically continuous distribution of p-values typical of fMRI images, the inequality will normally be approximated by equality.

A few interesting observations can be made for c(V)=1. Under this condition, the Genovese FDR procedure controls the family-wise error (p') as:

$$p' = p_G V = iq \tag{2.38}$$

Because i is strictly greater than or equal to 1, there is weaker control of family-wise error in the FDR method than the Bonferroni correction. The FDR method can in this case be considered to control family-wise error on the subset of active voxels.

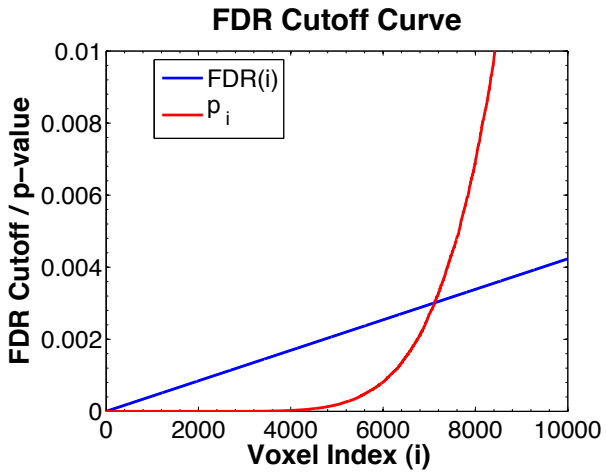


Figure 2.12: Graphical interpretation of the FDR procedure. Where the ordered p-values (red) intersect the FDR cutoff line (blue) is the corrected p-value for a given choice of q and c(V). In this example, q=0.05, and c(V)=log(V)+0.5772, giving a corrected p-value of 0.003.

There is as of yet no consensus on the best approach to setting the activation threshold. Previous research on presurgical fMRI has suggested that fixed confidence levels do not reflect the complexity of the neuroimaging process, and are inappropriate in the context of determining safe surgical boundaries [4, 30]. Data driven methods could easily overcome this if the desired signal were reliably separated from the background noise [19, 69]. Unfortunately the contrast to noise of fMRI experiments is low, and significant overlap of the signal and noise distributions exists (figure 2.13). The application of t^* value threshold inevitably involves a trade-off between allowing false positives and false negatives. The use of receiver operator characteristic (ROC) curves has been suggested as a method of determining appropriate thresholds based on optimization of true and false positive rates [19, 51, 70].

2.5 ROC Curves and Test-Retest Imaging

2.5.1 ROC Curve Basics

ROC curve analysis is a generalized method of examining the performance of a detection scheme. An ROC curve is a plot of the true positive rate (TPR or sensitivity)

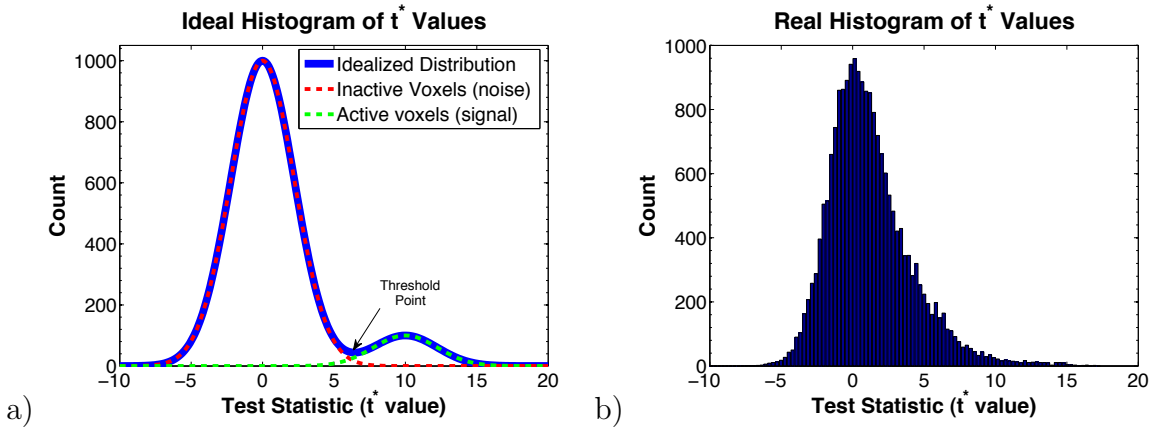


Figure 2.13: a) Schematic of idealized signal and noise distributions. The ideal threshold point can easily be found either manually, or by an automated procedure like the Otsu method [69]. b) A real fMRI histogram. Even in this relatively robust example dataset, the signal distribution is not easily distinguishable from the noise background, and significant tradeoffs between true and false positives are inevitable.

versus the false positive rate (FPR or 1-specificity) as a decision threshold is manipulated (figure 2.14). The true positive and false positive rates are defined as:

$$TPR = \frac{TP}{TP + FN} \tag{2.39}$$

$$FPR = \frac{FP}{FP + TN} \tag{2.40}$$

In the context of fMRI, the TP and FN are the number of truly active voxels identified correctly and incorrectly respectively, and TN and FP are the number of truly inactive voxels identified correctly and incorrectly. At very low thresholds, most true positives will be identified, and so TPR is high. However, this comes at the expense of many invalid supra-threshold voxels (high FPR). At very high thresholds, very few truly active voxels will survive the threshold mask, increasing the number of true and false negatives, and decreasing the number of both true and false positives.

ROC curves have several interesting features, but the most prominent in the context of appraising functional MRI images is the area under the curve (AUC). The AUC is numerically equivalent to the average sensitivity over all values of the specificity, or conversely the average specificity over all values of sensitivity, and can be used as an indicator of how well the measured pattern matches the true activation. The maximum possible value of the AUC is 1.00, at which point some threshold on

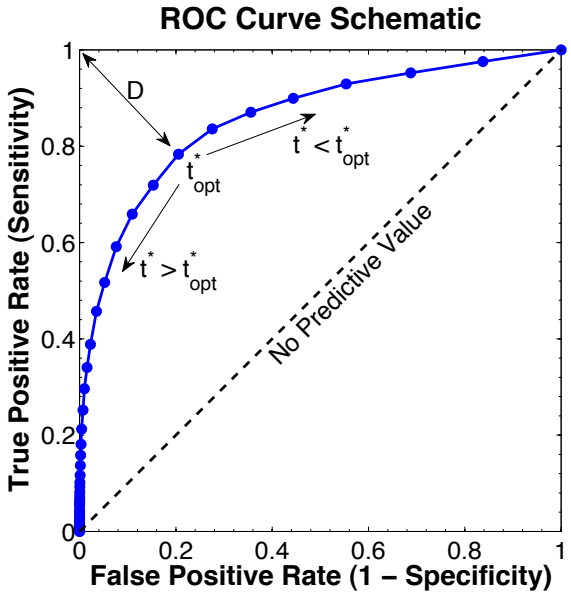


Figure 2.14: Schematic ROC curve, plotting true positive rate (TPR) against false positive rate (FPR), as the decision threshold is varied. A common choice for the optimal trade-off between true and false positives is the point that minimizes the distance to the top left corner (which represents perfect detection).

the fMRI image renders it identical to the estimate of the true activation pattern. An AUC of 0.5 corresponds to a test that has no discriminatory power, as the rate of false positives is equal to the rate of true positives on average, and thus a positive result is as likely to be correct as incorrect. Another salient feature of ROC curves is the optimal point. The top-left corner of an ROC plot corresponds to zero false positives, and 100% true positives. The closer the ROC curve approaches to this corner, the better. While weighting can be introduced to bias sensitivity or specificity, the simplest choice of optimal point is that which is nearest to the top-left corner in the Euclidean distance sense.

2.5.2 ROC Curves for fMRI: Reproducibility and Optimal Thresholds

How to measure the voxel-based TPR and FPR from experimental fMRI data is a subject of ongoing discussion [56]. The gold standards in functional brain mapping are sensitive to different aspects of physiology [3], typically unavailable preoperatively, and restricted to a small field of view within the craniotomy. Lack of an a priori knowledge of the correct activation pattern prohibits a single fMRI session from being

capable of generating effective estimates of the TPR and FPR. Within a test-retest framework, multiple replications of the same scanning experiment can be conducted, and the activated regions can be compared to infer the TPR and FPR [19,51,56,71]. One possible approach is to threshold one of the images very stringently, and to use this as a true positive template [71]. Like all methods of estimating ROC curves from test-retest data, this is fundamentally related to image reproducibility. This method is limited in that there is no basis on which to decide which image should be used as the template, and it discards sub-threshold data from the template that may reflect true activation.

An improved method was developed to address this issue by Genovese et al [19], by extending the test-retest framework to larger numbers of replications (M). For a given threshold, the probability of observing a voxel is above threshold m times (P_m) is modeled as a mixture of two binomials:

$$P_m = \lambda Binomial(M, TPR) + (1-\lambda) Binomial(M, FPR) \quad (2.41)$$

$$= \binom{M}{m} \left(\lambda TPR^m (1-TPR)^{M-m} + (1-\lambda) FPR^m (1-FPR)^{M-m}\right) \quad (2.42)$$

where λ is the proportion of voxels that are truly active (an unknown parameter), and the first bracketed term is the binomial coefficient, which accounts for the fact that m above threshold voxels can arise from M replications in M choose m combinations. The first term in 2.41 represents the probability of the TPR contributing to the m active classifications, and the second term the contribution of the FPR.

In the simplest form, the voxel probabilities are treated as independent to compute a likelihood function of observing the whole dataset. The likelihood function can be thought of as roughly the converse of a probability. Rather than predicting unobserved data based on known probabilities, the likelihood function estimates unknown probabilities based on observed data. Under the assumption of independent voxels, the natural log of the likelihood function (ℓ) is:

$$\ell(TPR, FPR, \lambda | \boldsymbol{n}) = \sum_{m=0}^{M} \ln \left(P_m\right)^{n_m} \quad (2.43)$$

$$= \sum_{m=0}^{M} n_m \ln \left(P_m\right) \quad (2.44)$$

where $\boldsymbol{n}$=$(n_0, n_1, \ldots, n_M)$, that is a vector of length M+1, where the m^{th} entry denotes how many voxels are above threshold m times (n_m). By maximizing the likelihood as a function of TPR, FPR, and λ, the parameters which best fit the model are found. Methods of accounting for dependence between voxels have been developed [51]. Genovese et al [19] point out that repeating this process for different threshold levels can produce an ROC curve. Following the general observations of the preceding section, this curve can be used as a data driven method of determining the optimal threshold. The drawback of the Genovese method is that it requires three or more replications of the task, significantly extending the duration of scanning procedures. Additionally, the threshold used to compute the vector n is fixed across replications, which may not be ideal.

The test-retest sensitivity and specificity of an fMRI experiment depend on a number of factors including scanner noise, choice of tasks [3], task performance [48], pulse sequence design [72], intersubject variability [42, 73], neurological status [56] and choice of test statistic used to construct activation maps [19]. The sources of variation in fMRI data include both purely random effects and systematic biases, resulting in a complex interaction between signal, noise, and artifact. In an ROC analysis performed on simulated data, Skudlarski et al. [70] explored the effects of a number of data processing steps on sensitivity and specificity. They found that temporal high-pass filtering, spatial smoothing of raw data, and drift removal all had beneficial results on the TPR and FPR. Averaging of multiple scanning sessions was also shown to improve the ROC analysis results. To estimate the TPR and FPR from only two replications of an fMRI task, a novel ROC generation scheme was developed. This routine will be used to assess the reproducibility of fMRI data, and to determine data driven thresholds for improved individual functional mapping.

Chapter 3

Experimental Details

The data for this project was obtained in a two-stage approach. In the first stage, healthy control data from two different MRI scanners (with different field strengths) were used to develop the presurgical mapping protocol, and the ROC routines. The second stage consisted of collecting data from volunteers with operable brain lesions (tumors or vascular malformations), who were identified as candidates for awake surgery and cortical stimulation. The purposes of this stage were a) to demonstrate the analysis procedures developed in the first stage, and b) to perform post-surgical comparison with the gold standard to investigate accuracy. In both stages, participants had the study explained to them in detail prior to obtaining informed consent. Ethics review boards of the National Research Council, Capital District Health Authority, and Izaak Walton Killam Health Centre reviewed and provided approval for both stages of the study.

3.1 ROC Generation Algorithm

The ROC generation algorithm created for this project is suitable for both assessing the test-retest reliability of individual mapping results, and automating threshold selection on the basis of optimizing true and false positive rates. Like in the method of Le and Hu [71], true and false positive rates are determined by assuming one of the images to be the true activation pattern, and comparing the other image to it. The routine is referred to in our lab as the Procedure for ROc CREATion and Evaluation or simply pROCreate.

3.1.1 ROC Reliability Analysis

Let the first image of the test-retest pair be A, and the second B. Let A_t and B_t be the set of voxels in each image whose value is above thresholds t_A^* and t_B^* respectively. With A as the template (reference) image, and for a given t_A^* and $^*_B$, the true and

false positive rates can be expressed in terms of set theory as:

$$TPR(t_A^*, t_B^*) = \frac{num(A_t \cap B_t)}{num(A_t \cap B_t) + num(A_t \cap B_t^C)} \quad (3.1)$$

$$FPR(t_A^*, t_B^*) = \frac{num(A_t^C \cap B_t)}{num(A_t^C \cap B_t) + num(A_t \cup B_t)^C} \quad (3.2)$$

where $\cup$ and $\cap$ represent the union and intersection sets respectively, and the superscript 'C' denotes the complement.

Caclulating the TPR and FPR for all values of t_B^* will produce an ROC curve for that value of t_A^*. The deviation from the methods proposed by Le and Hu [71] is to repeat for all values of t_A^* to generate a family of ROC curves that characterize the test-retest pair (figure 3.1). A selection of these ROC curves can be plotted to inspect the sensitivity-specificity tradeoffs. The area under the curve (AUC) is calculated for each value of t_A^* and plotted, to assess general reproducibility (figure 3.1). This whole procedure is repeated using B as the image template, maximizing the use of information in the test-retest pair. The average AUC over all template thresholds (and over both images as the true positive template) is output as an overall indicator of reliable activation patterns.

3.1.2 ROC Threshold Selection

The ROC threshold algorithm is based on a few simple observations of the data provided in the reliability analysis. Assume that the reproducible voxels are those with high t^* values, and that low t^* values correspond chiefly to random noise. In this case, increasing the threshold on the template image should make the activation pattern more easily detectable by the retest image, and the AUC will increase. As was discussed in chapter 2, the AUC has an upper bound of 1.0 for a perfect detection and classification of active voxels, and a lower bound of 0.5 for chance classification. The AUC is expected to increase monotonically until the majority of background noise is below threshold. At high thresholds, the AUC should asymptote towards 1.0, as stronger correlations will be more spatially reproducible. If the approach to an AUC of 1.0 is rapid, there will naturally come a point where increasing the template threshold further is only eliminating true positives.

The rate at which the AUC increases per unit increase in t^* threshold can be

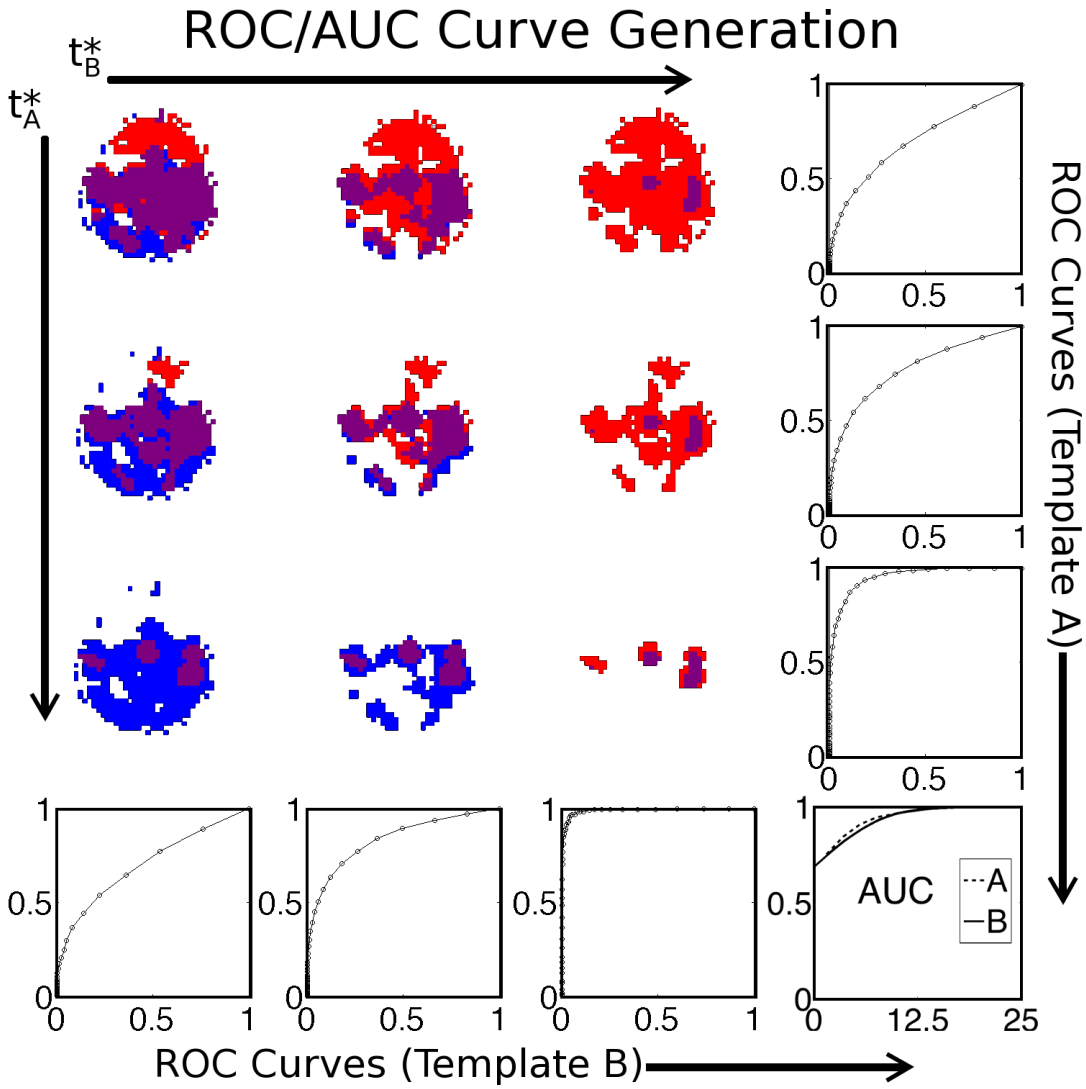


Figure 3.1: Example of ROC curve and AUC profile generation from test-retest data. The first image (A, red) threshold is increased down the columns, and the second image (B, blue) threshold increases across the rows. Overlap of the two activation maps at a given threshold pair is shown in purple. Each value of the t^* threshold on image A produces an ROC curve (right-most column), and likewise for each template threshold on B (bottom row). The area under the curve is calculated from each ROC curve, for both templates, and plotted as a function of the template threshold used to derive it (bottom right).

approximated by fitting the AUC data by a cubic spline interpolant, and evaluating the derivative. A cutoff value for the derivative is assigned (d), above which a unit increase in template threshold results in marginal improvement of the AUC (figure 3.2). This template threshold is used only to select the ideal template. Once the

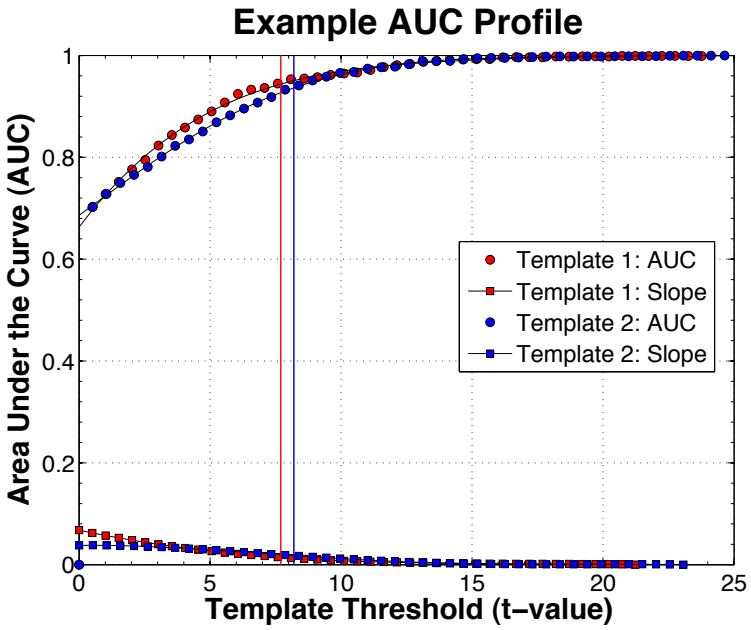


Figure 3.2: An example AUC profile generated by the pROCreate method. The average AUC from each template (circles) is used to measure reproducibility, and the slope of the AUC profile (squares) is used to determine the optimal template (vertical lines) and retest thresholds (figure 2.14). In this example the derivative cutoff d=0.02.

ideal template is determined, the corresponding retest ROC curve is examined. The optimal point of the ROC curve is found as the point closest to the top-left corner in ROC space, and the corresponding retest threshold is applied to the retest image (figure 2.14). Each image is appraised in the role as the template to determine retest thresholds for both images. To run pROCreate requires as arguments only the test-retest images in NIfTI file format, the number of threshold points to evaluate for each image, and the derivative cutoff for threshold selection. Optionally pROCreate can be run for a restricted region of interest (ROI), in which case a binary mask file must be supplied (NIfTI format).

3.2 Functional MRI Data Acquisition

3.2.1 Healthy Controls - Subject Info

Eight healthy control volunteers with no known neurological impairments (4 males, 4 females; 24.4 $\pm$ 3.5 years old) were recruited for this phase of the study. All healthy controls were right handed according to the Edinburgh Handedness Inventory [74].

The participants all spoke English as their first language, and had either normal or corrected to normal vision.

Figure 3.3: The 4 tesla MR scanner at the National Research Council Institute for Biodiagnostics (Atlantic), used in both stages of this study.

3.2.2 Healthy Controls - Acquisition Details

The healthy control volunteers were scanned twice with a 1.5 T MRI (GE Signa) and twice with a 4 T scanner (Varian INOVA, figure 3.3), for a total of four scanning sessions each. During each session, both structural and functional images were acquired. The structural images were collected with an MP-FLASH (figure 3.4a) sequence at 4 T (T_I = 500 ms, T_R = 10 ms, T_E = 5 ms, $\alpha = 11°$), and a SPGR (figure 3.4b) sequence at 1.5 T (T_R = 25 ms, T_E = 5 ms, $\alpha = 40°$), with a 256×256 matrix, 64 slices, and 0.94×0.94×3 mm voxels (FOV = 24×24×19 cm). Functional images were collected with a single-shot spiral out (figure 3.4c) sequence, using T_R = 2 s, α=90°,

64×64 matrix, 22 slices, and 3.75×3.75×5 mm voxels, with a 0.5 mm gap (FOV = 24×24×12 cm). The echo time used was different at 1.5 T (40 ms) and 4 T (15 ms), to compensate for the shortening of T_2^* at higher field strengths. Other imaging parameters were matched to maintain a fair comparison between the two scanners.

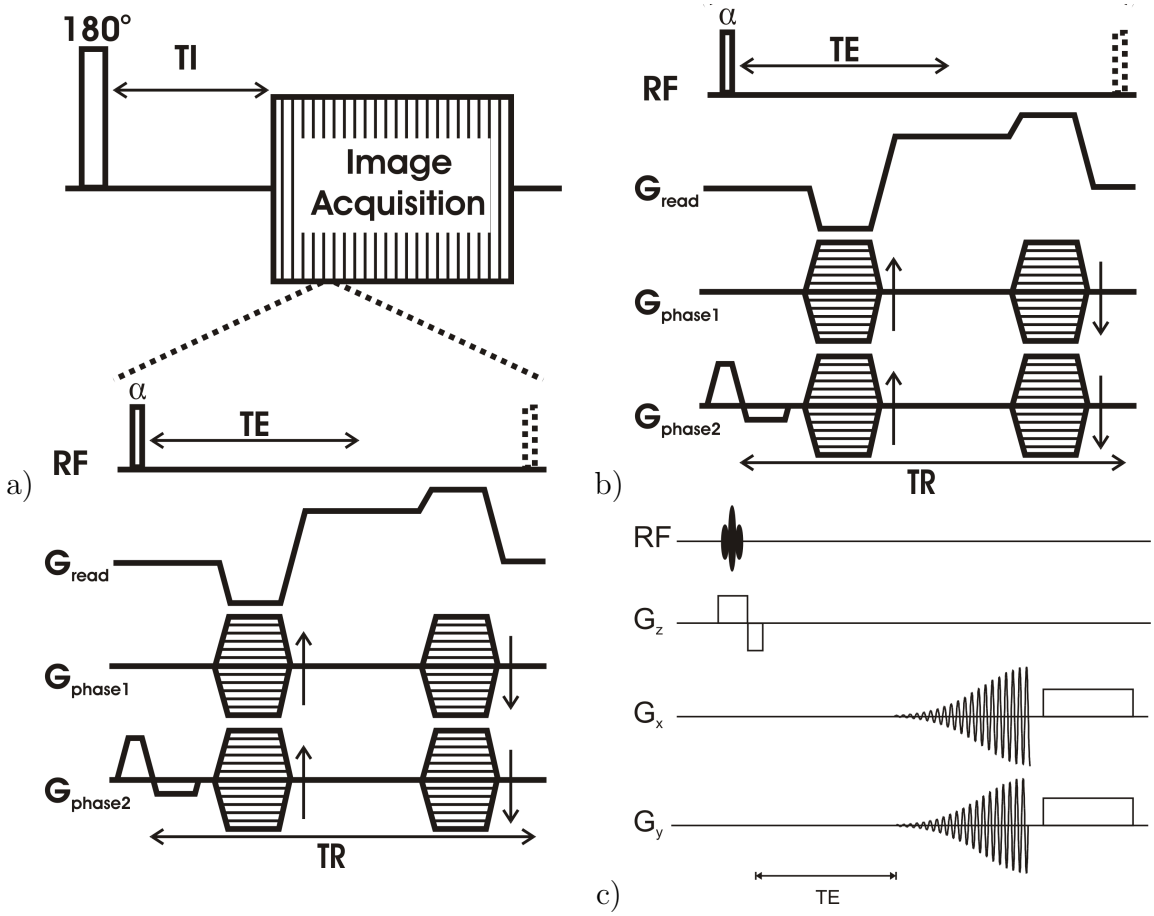


Figure 3.4: a) MP-FLASH pulse sequence diagram used for T_1 anatomical images as 4 tesla. b) SPGR pulse sequence used for anatomicals at 1.5 tesla. c) Spiral-out pulse sequence used for T_2^* weighted functional imaging at both field strengths.

3.2.3 Healthy Controls - Functional MRI Tasks

The healthy volunteers performed one cognitive (sentence/math comprehension) and one sensorimotor task (paced finger tapping). Task order was counterbalanced between subjects to ensure the order of performance did not affect the results. All tasks for the healthy controls used a block design, consisting of interspersed 20 s blocks of stimuli and rest. The stimuli were presented using E-Prime [75] by a projector in

the console room, projecting onto a screen fixed to the end of the magnet bore. The subjects could view the screen by looking ahead through a mirror mounted on the head coil. Subjects were given an opportunity to practice each task prior to entering the MRI to mitigate compliance and performance related issues.

Sensorimotor Task

The sensorimotor task had two active finger-tapping conditions (one for the left and the other for the right hand), and one rest condition. The left and right hand condition each appeared four times, and consisted of simple thumb-to-digit tapping in ascending/descending order, paced by visually presented cues (figure 3.5). For the rest condition, a fixation cross was displayed, and the subjects were instructed to relax. The active block order was randomized, and a fixation block was placed before and after each active block.

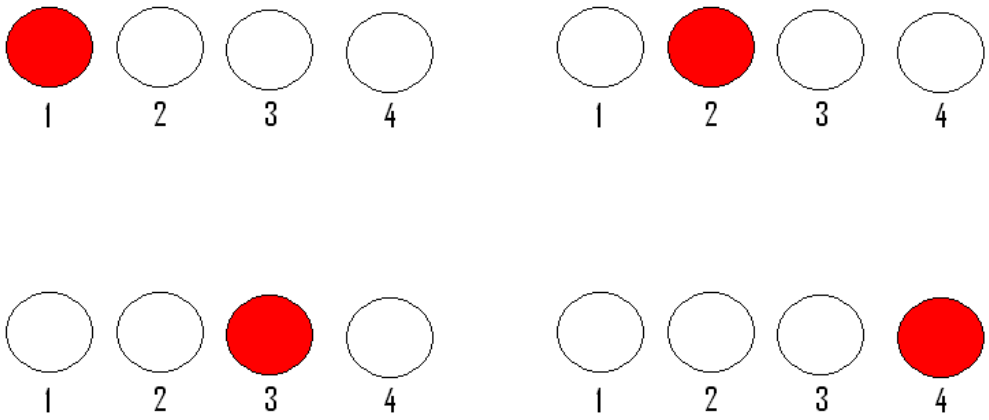


Figure 3.5: The visual stimuli presented to cue the finger tapping task. The red circle cyclically progressed from digit one to four, and then in reverse, changing at a rate of 1 Hz.

Cognitive Task

The cognitive task consisted of two active conditions and one passive rest condition. The rest condition was fixation. For the active condition, either four English sentences or four single digit addition/subtraction math statements

were displayed one at a time for 5 seconds each. The sentences contained 50 percent semantically correct ("He stroked her face with a feather."), and 50 percent semantic violation ("I drank some ability from California.") stimuli. The math stimuli also consisted of 50 percent correct (4+2=6) and 50 percent incorrect (9-4=4) stimuli. The participants were asked to indicate via a button press which stimuli were correct, and which contained a violation. The block order was randomized, and a fixation block was placed before and after each active block.

3.2.4 Patient Cases - Subject Info

Nine patients (table 3.1) with cerebral lesions volunteered to take part in an fMRI study on the 4 tesla scanner prior to neurosurgical intervention (2 males, 7 females; 45.0 ± 20.2 years old). Of the nine patients, five were right handed, three left handed, and one of mixed handedness according to the Edinburgh Handedness Inventory [74]. All patients spoke English as a first language, and were tested prior to entering the MRI to ensure that they could resolve the stimuli. Prescription MR safe goggles were available on site if needed.

The nine patients had a variety of lesion locations and types, including two oligoastrocytomas, one oligodendroglioma, two glioblastoma, three meningiomas, and one cavernous angioma. A rather heterogeneous sample is inevitable, as the practical difficulties of finding suitable patients and obtaining all the required data limits sample size. The lesions were located mostly in the left hemisphere (n=7), and of those, four were located in the frontal lobe, and three in the temporal lobe. One of the two right hemisphere cerebral tumors was in the frontal lobe, and the other was located in the parietal lobe. Eight of the nine patients (all except patient 9) received cortical stimulation. One patient was unable to complete the procedure (patient 8). Eight of the nine patients performed repeat imaging, allowing the ROC analysis to be performed. Six patients received both test-retest imaging and cortical stimulation mapping (patients 1-6).

Table 3.1: Summary of the patient volunteers for the second stage of this study. Handedness was assessed by the Edinburgh Handedness Inventory [74]. Tumor types according to the World Health Organization classification system.

Patient	Age/Gender	Hand	Tumor Type (W.H.O. grade)	Location	CS	Test-Retest
1	35 / F	R	Anaplastic Oligoastrocytoma (Gr. 3)	L. Inferior Frontal	Yes	Yes
2	62 / M	L	Oligodendroglioma (Gr. 2)	L. Middle Frontal	Yes	Yes
3	20 / M	R	Cavernous Angioma	L. Inferior Parietal	Yes	Yes
4	24 / F	L / R	Glioblastoma Multiforme (Gr. 4)	L. Anterior Temporal	Yes	Yes
5	26 / F	R	Mixed Oligoastrocytoma (Gr. 2)	R. Inferior Frontal	Yes	Yes
6	45 / F	R	Glioblastoma Multiforme (Gr. 4)	L. Inferior Temporal	Yes	Yes
7	48 / F	L	Metaplastic Meningioma (Gr. 1)	L. Inferior Frontal	Yes	No
8	81 / F	L	Meningioma (Gr. 2)	L. Medial Temporal	No	Yes
9	41 / F	R	Fibrous Meningioma (Gr. 1)	R. Superior Parietal	No	Yes

3.2.5 Patient Cases - Acquisition Details

Structural and functional MR imaging for the patient cases were performed on the 4T Varian INOVA MRI (figure 3.3). Structural T_1 weighted MR images were collected using an MP-FLASH sequence (figure 3.4a), T_I = 500 ms, T_R = 10 ms, T_E= 5 ms, $\alpha = 11°$, 256×256 matrix, resolution of 0.94×0.94×1 mm, no gap, 170 axial slices (FOV = 24×24×17 cm). For functional imaging, a two-shot T_2^* sequence with spiral readout weighted for BOLD contrast was used (figure 3.4c with T_R = 2 s (1 s per excitation), T_E = 15 ms, $\alpha = 60°$, 64×64 matrix, voxel resolution of 3.75×3.75×4.0 mm, 0.5 mm between slices, 22 slices (FOV = 24×24×10 cm). The patient scans used a 4 mm slice thickness to improve the voxel resolution. Further reduction in slice thickness would result in unacceptable reduction of the FOV. The flip angle was reduced to 60 degrees, as this is near the Ernst angle. Imaging with this flip angle provides optimal signal to noise, and simultaneously decreases the RF power deposited during the scan. A two shot spiral was used to reduce distortion of the T_2^* weighted images.

3.2.6 Patient Cases - Functional MRI Tasks

Each patient received a unique task battery, based on the location of the tumor and a-priori knowledge of functional organization of the brain. For patients that presented with specific functional deficits, tasks reflecting those functional systems as closely as possible were chosen. For example, a patient with seizures and spasms of the left leg would receive a task battery focused on motor function, including left and right foot flexion, and left and right hand movement. The following tasks were available for motor investigations in the patient studies:

Finger Tapping

The finger-tapping task used the same stimuli as that for healthy controls. To reduce scan time requirements, the left and right hand conditions were separated into individual tasks, each of four active and five fixation rest 20 second blocks. The active blocks consisted of simple thumb-to-digit tapping in ascending/descending order, paced by visually presented cues (figure 3.5). For the rest condition, a fixation cross was displayed.

Foot Flexion

The foot flexion task consisted of a paced active condition and a fixation rest condition. The subject was prompted by a circle that alternated between white and red, changing color every second (figure 3.6). The patient was instructed to flex and relax one of their feet each time the circle changed color. Four active blocks and five rest blocks were performed, and the task could be done with either the left or right foot. All blocks were 20 seconds long.

Tongue Movement

The tongue movement task consisted of a paced active condition and a fixation rest condition. The subject was prompted by a circle that alternated between white and red, changing color every second (figure 3.6). The patient was instructed to move their tongue to the opposite side of their mouth (alternating left and right) each time the circle changed color. Four active blocks and five rest blocks were performed. All blocks were 20 seconds long.

Figure 3.6: The visual stimuli presented to cue the foot/tongue motor tasks. The red circle changed color at a rate of 1 Hz, to pace the motor task.

Additionally, the following cognitive tasks were available for testing both expressive and receptive language:

Sentence/Math Comprehension

The sentence and math comprehension task used the same stimuli as in the healthy controls. Block length was reduced to 18 seconds, and stimuli presentation time reduced to 4.5 seconds each accordingly. The block order was altered as well to fit more active condition blocks in roughly the same total scanning time. Six active blocks of each task condition were randomly ordered, with fixation rest before and after every second active block.

Overt Object Naming

The object naming used two active conditions and one fixation rest condition. For the 'objects' active condition, 3D images were displayed of objects, and the patient was asked to name aloud what appeared in the image (figure 3.7a). The images were controlled across blocks and trials for familiarity, word length, and average naming errors, based on a published image database. In the second active condition, the subjects were shown pictures of 3D non-objects (figure 3.7b), and were asked to passively view the images without attempting to discern meaning from them. This task used 16-second blocks of activation and rest, during which 8 images of objects or non-objects were shown. Six active blocks of each task condition were randomly ordered, with fixation rest before and after every second active block.

Semantic Decision Task

The semantic decision task consisted of two active and one fixation rest block. For first active condition, 2D images were displayed of animate (50 percent) or inanimate items (50 percent), while scrambled images were presented for the second active condition (figure 3.7c-d). The participants were asked to indicate via a button press, which stimuli were animate and which were inanimate objects. This task used 16-second blocks of activation and rest, during which 8 images of objects or non-objects were shown. Six active blocks of each task condition were randomly ordered, with fixation rest before and after every second active block.

3.3 Functional MRI Data Analysis

Functional MRI data analysis was done using the AFNI software package [76]. The AFNI software package can be used through a graphical interface, but the real flexibility of this software package is the command-line interface. Creating custom scripts to execute the series of processing steps allowed much more control over processing choices. These scripts also ensured both standardization and automation of the pipeline. Where custom solutions were needed, the IDL programming language [77]

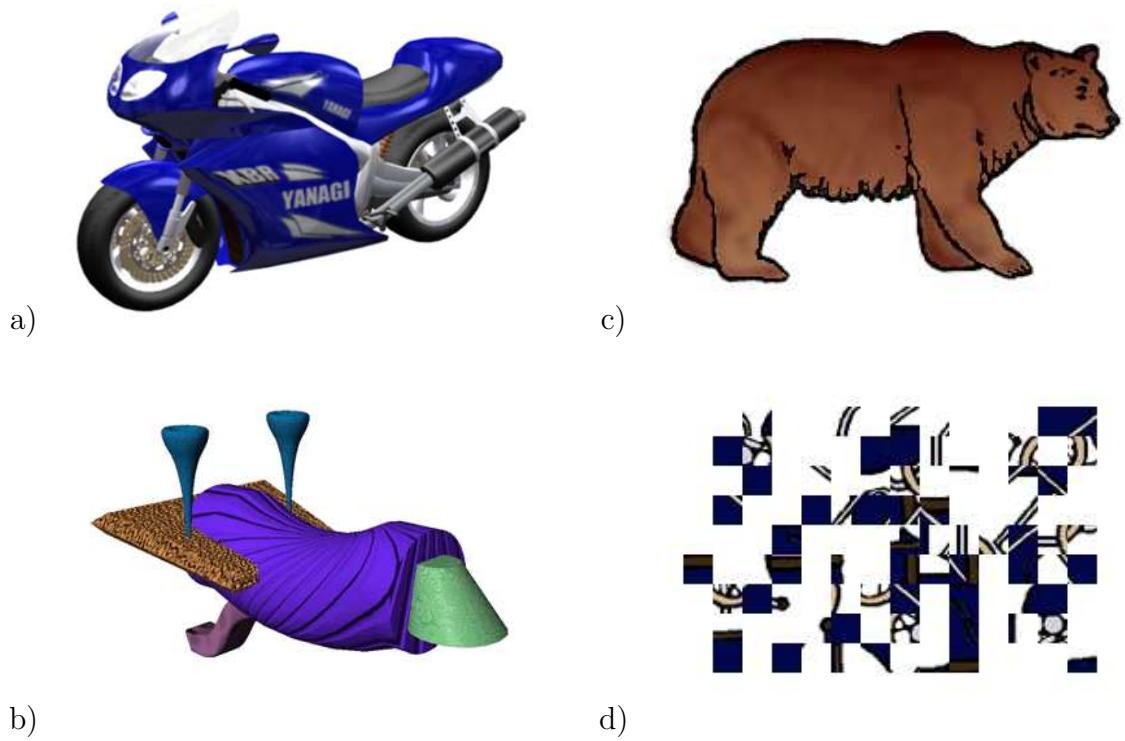


Figure 3.7: Examples of a) 3D objects for the first active condition of the object-naming task. b) 3D non-objects for the second condition. c) 2D objects for the first active condition of the semantic-decision task. d) 2D scrambled objects for the second active condition.

was used to develop the necessary tools. The same processing pipeline was applied to the healthy control and patient data.

Functional MRI data is first motion corrected to align all images in each time series with the first image of that time series. The motion correction parameters are saved and plotted to inspect the data for excessive motion. Furthermore, the motion parameter time series can be used in the GLM as part of the baseline model to reduce the sensitivity to motion artifact if needed. Both the functional and anatomical images are segmented to create masks of the brain, after which the masked anatomical is down-sampled to the resolution of the functional image. The masked low-res datasets are used to calculate registration parameters using a general (12 parameter) affine transformation, allowing shift, rotation, scale, and shear of the images. Functional scans from different imaging sessions were all registered to the first anatomical scan at that field strength to permit voxel-to-voxel comparisons. A Gaussian spatial

smoothing kernel was applied to the data prior to application of the GLM, with a 6mm full width at half maximum.

For the functional analysis, every condition, trial, and task was first parameterized by a 1D box car function, and each of these were convolved with the hemodynamic response function. Constant, linear, and quadratic terms were included as parameters of the baseline model to account for low frequency drifts. For each unique linear combination of the task regressors, the 3dDeconvolve program in AFNI estimated the significance as a predictor of the observed time-series, and output a t-statistic map. Whole-brain analysis was performed rather than region-of-interest based analysis, to ensure that we did not overlook reliable activation in unexpected locations.

3.4 Cortical Stimulation

CS mapping was performed by Dr. David Clarke using an Ojemann OCS-1 cortical stimulator (figure 3.8), with a bipolar probe, 5 mm spacing between electrodes. Stimulation was performed using 0.5 ms duration pulses. This was repeated at increasing current levels, from 4 mA peak-to-peak, in steps of 2-6 mA to a maximum of 20 mA, or until a response was elicited. The pulse frequency used for motor mapping (5 Hz) differ from those used for language investigations (60 Hz).

For motor investigations, the patient was observed for involuntary movements, and was asked to report any unusual sensations that occurred during the stimulation session. For language function, the patient was asked to count slowly, or to recite the days of the week, or months of the year. The surgeon would tell the patient which number, day, or month to start with, and to go forward or backward. Additionally, a laptop and monitor was brought into the operating room to present the fMRI object naming stimuli during cortical stimulation when this task was used presurgically. Locations that consistently caused errors in the cognitive tasks, or had an effect on the sensorimotor system were labeled, recorded, and annotated with the specific effect.

3.5 Quantitative Spatial Comparison of fMRI and CS

In order to make quantitative spatial comparisons between fMRI and CS, each location that had an effect on sensorimotor or cognitive function when stimulated by CS was digitized using a Medtronic TREON neuronavigation system (figure 3.9b). The TREON uses a system of infrared (IR) emitters, reflectors, and cameras to triangulate positions in the operating room. The patients head is fixed to the surgical table by a Mayfield clamp, to which a set of IR reflectors is attached (figure 3.9a). The TREON uses these reflectors to establish the coordinate system. A separate wand is equipped with IR reflectors that have a known geometrical relationship to the tip of the wand. This is used to trace the skull of the patient, and the trace is registered to a high resolution anatomical MRI of the patient. After this initialization, the tip of the wand will show up as crosshairs on the MR image, and can be used to navigate and record locations. The setup in action in the operating room is shown in figure 3.9b.

The 3D coordinates that are supplied by the TREON use an axis and origin defined by the first image loaded during the session. To make accurate comparisons to the images used for functional mapping, we developed the following routine to transform this coordinate system into the coordinates of the presurgical fMRI. First the leftmost, anterior and superior corner of the first image loaded is designated as the origin of the TREON coordinate system. Before loading a second image, one navigates to the rightmost, posterior, and inferior limits of the image box, and records the coordinates of these voxels. This way, coordinate points can easily be converted into voxel locations on the TREON reference image. It is important to do this before registering any secondary images to the reference image, as the registered images may extend the total image box, obfuscating the relationship between the coordinate system and the reference image. To convert coordinates in the TREON reference image to that of the presurgical mapping image, registration software is used to compute the transformation matrix that accurately co-registers the two images, and the transformation is applied to the CS data. For a given fMRI threshold, the fMRI to CS distance can then be calculated for each task and CS data point simply by locating the nearest active voxel, and calculating the Euclidean distance.

Figure 3.8: Ojemann OCS-1 bipolar stimulator. The Ojemann OCS-1 is capable of peak-to-peak currents of 0-22 mA, in pulses of 0.1-2 ms, at frequencies of 5-100 Hz.

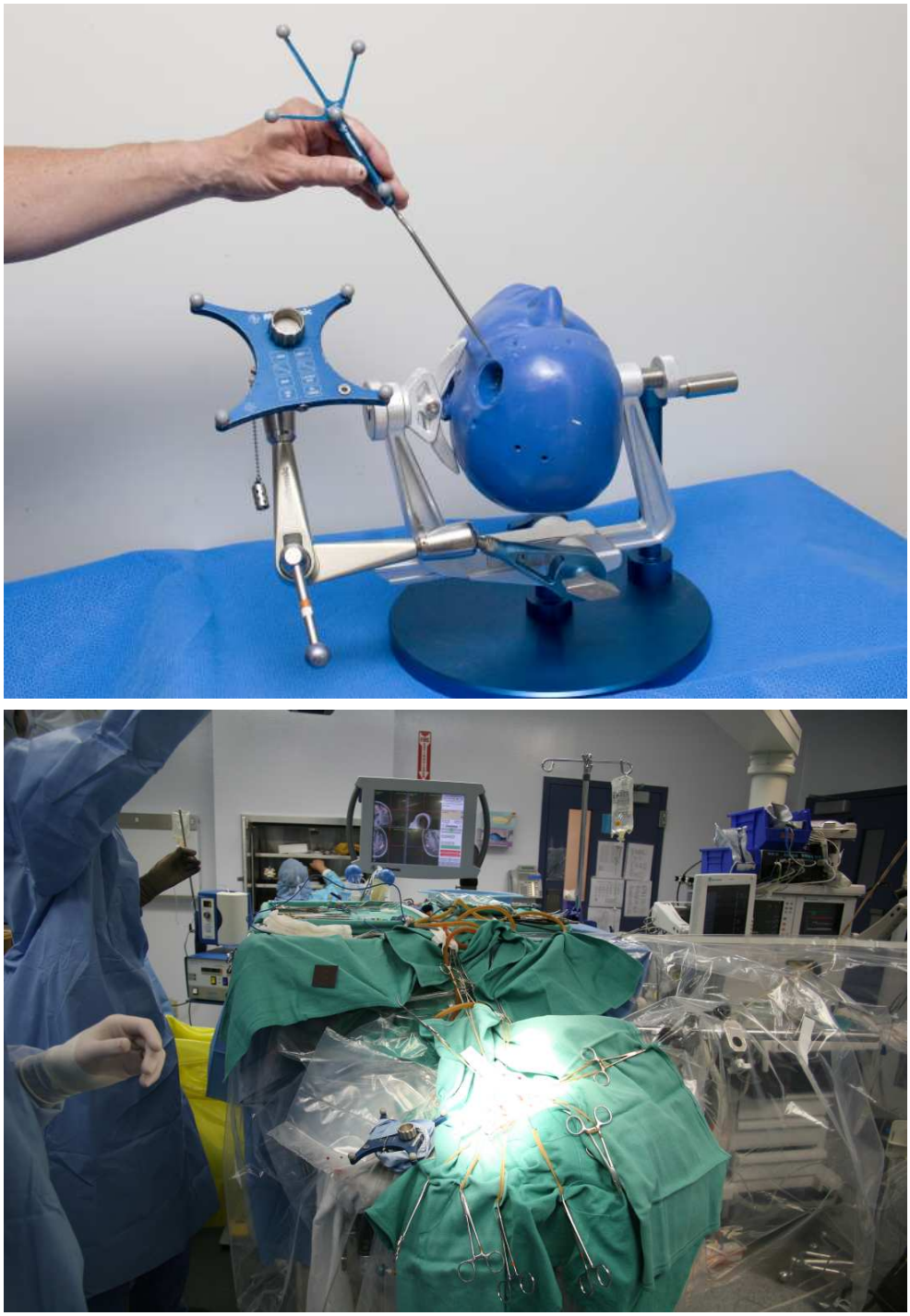


Figure 3.9: a) Mockup of the patient positioning system, including the Mayfield clamp (to which is attached an IR reflective coordinate frame), and an IR reflective navigation wand. b) Intra-operative setup. Foreground: the patient is under surgical drapes, and the IR reflective coordinate frame can be seen. Background: the TREON display screen can be seen at the foot of the patient bed. Not shown: the TREON IR emitter and camera.

Chapter 4

Test-Retest Reproducibility of fMRI Activation Maps

The reproducibility of functional mapping must be assessed to assure that the results are reliable. This chapter explores the reproducibility of the data collected from healthy control volunteers in the first phase of this study. Two different methods of ranking reproducibility are presented: the linear regression method, and the ROC/AUC analysis described in section 3.1.1. The results of each analysis are presented separately and compared with one another. The available test-retest data are further divided according to characteristics such as field strength, functional task, and subject, to assess the effect of these factors on reproducibility. Finally, some comments are made on the overall performance of the two reproducibility analyses, and the implications for presurgical mapping.

4.1 Methodological Details

Each volunteer performed both the finger tapping and comprehension tasks. Linear regression and ROC analyses are performed on each healthy control's test-retest pair to assess reproducibility. The linear regression analysis consists of plotting the voxel-by-voxel values of each image against each other and fitting with a least-squares linear regression. The correlation coefficient is used as a general indicator of reproducibility. The ROC analysis is performed by the pROCreate algorithm described in section 3.1.1. There are 8 subjects, 2 field strengths, 2 tasks, and 2 conditions per task. Labelling the two conditions as 'C_1', and 'C_2', four general linear tests (GLT) are assessed: C_1 greater than baseline, C_2 greater than baseline, C_1+C_2 greater than baseline, and C_1-C_2 greater than baseline. The last option is the method specified by AFNI for generating contrasts between two active conditions. Identical activation maps of opposite polarity are generated by assessing C_2-C_1 greater than baseline, and so this test is not included. There are thus 128 independent test-retest pairs of images for analysis of reproducibility.

4.2 Linear Regression Results

The distribution of regression coefficients is shown in figure 4.1, and example plots are shown in figure 4.2. The example plots are all taken from the motor task on the 4 tesla scanner, using contrasts C_1 greater than baseline or C_2 greater than baseline (right or left hand). These test-retest pairs are chosen to represent poor, average, and excellent datasets, as compared with the distribution of regression coefficients. The magnitude of the regression coefficients range from 0.059 to 0.840, with a mean (SD) of 0.584 (0.159). The distribution of correlation coefficients is skewed, with more test-retest pairs exhibiting a high degree of reprodubility than low. The magnitude of the regression slopes range from 0.039 to 1.069, mean (SD) of 0.583 (0.199). A plot of the regression slope and correlation coefficients is shown in figure 4.3. The line of best fit (R^2=0.685) is determined with the intercept set to 0. The slope of this line is 1.00.

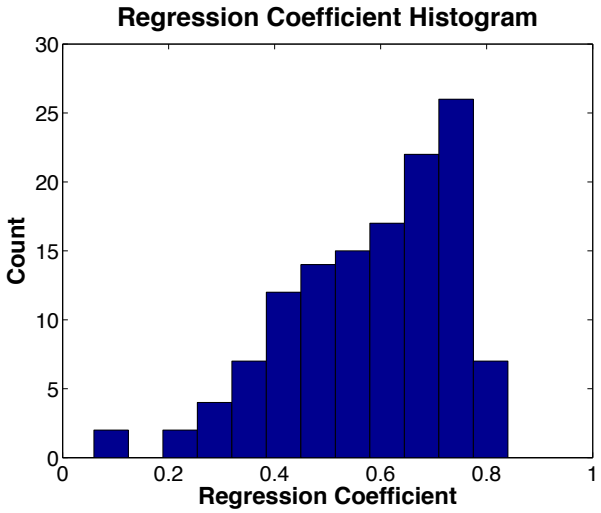


Figure 4.1: Histogram of the correlation coefficients from the linear regression analysis. The histogram shows that the test-retest pairs are concentrated at the upper end of this reproducibility scale, with mode at r=0.75, and mean of r=0.584.

4.3 ROC/AUC Reproducibility Results

The same 128 test-retest pairs are analyzed by the ROC reproducibility algorithm. Example AUC plots and their ROC generators are shown in figure 4.4 for the same

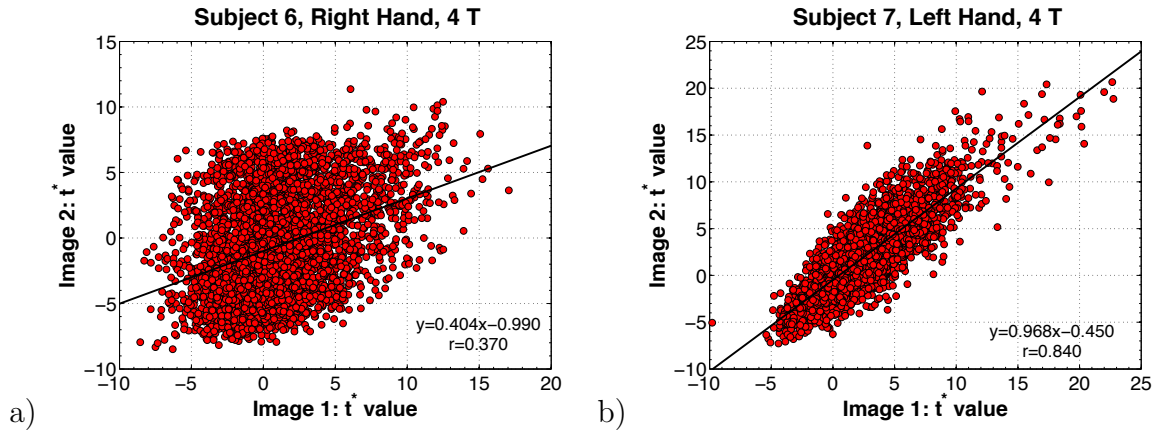


Figure 4.2: Example datasets from the linear regression analysis. Correlation coefficient and slope are shown for: a) A poorly reproduced test-retest pair. b) A well reproduced dataset

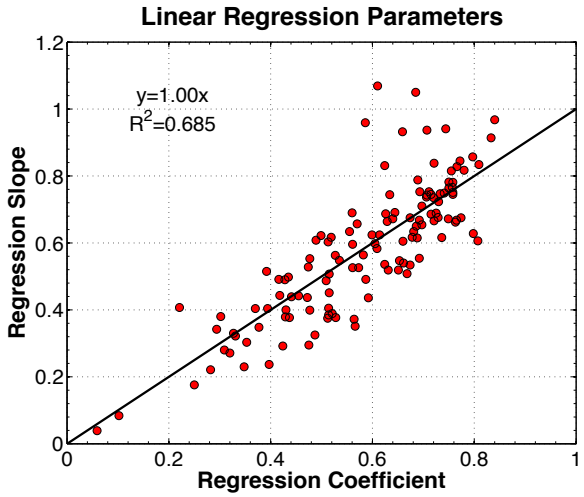


Figure 4.3: This figure shows that the correlation coefficient and slope of the linear fit in the regression analysis are themselves highly correlated

test-retest pairings as the example regression plots of figure 4.2. The distribution of average AUC values obtained when using either the first or the second image as a template are shown in figure 4.5. The mode of these histograms are also skewed towards the high end of the scale, indicating that the majority of the test-retest pairs are quite reproducible. With the first image as the template, the max (min) and mean (SD) of the average AUC are 0.937 (0.495) and 0.821 (0.089) repsectively. With the second image as the template, these values are 0.939 (0.405) and 0.830 (0.086). The average AUC for each template is plotted in figure 4.6, and fit with a

linear relationship (R^2=0.716). The intercept was again set to 0, which resulted in a slope of 1.01.

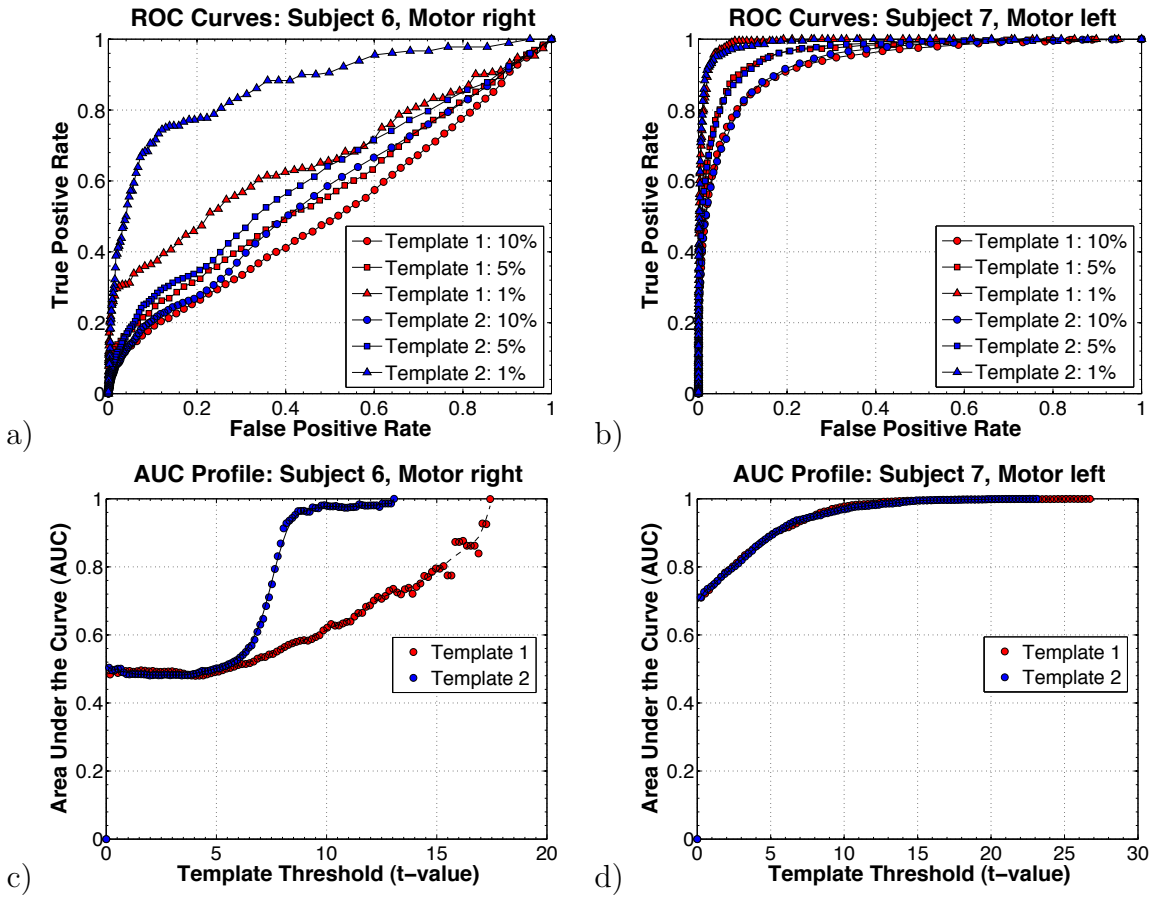


Figure 4.4: a) and b) ROC curves are shown corresponding to template thresholds for which 1, 5, and 10 percent of the brain volume is declared active, for the same example datasets as figure 4.2. The resulting AUC profiles from testing the full range of template thresholds are shown in c) and d). In all plots image 1 as the template is represented by red, and image 2 in blue. Note that both the initial value of the AUC and the average value over the range are indicative of reproducible datasets.

4.4 Comparison of ROC/AUC and Linear Regression

The test-retest pairs with good test-retest sensitivity and specificity (i.e. large AUC) tend to correspond with those that performed well on the regression analysis. This is exemplified by comparing the plots of figures 4.2 and 4.4. To test this relationship, the correlation coefficient from the linear regression analysis is plotted against the average AUC obtained in the ROC analysis in figure 4.7. The average AUC from both

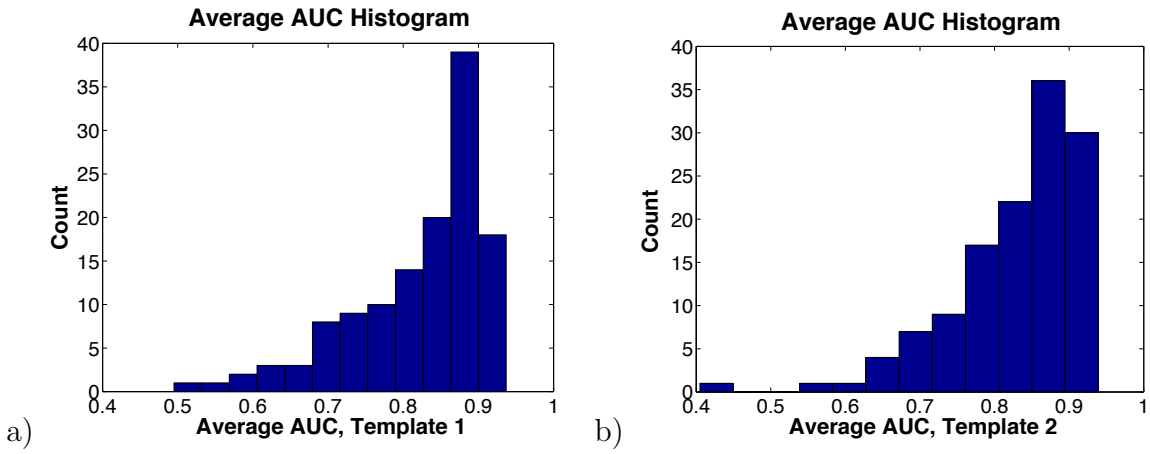


Figure 4.5: Histograms of the average AUC, for a) the first image of the test-retest pair as template and b) the second. Like the histograms of the correlation coefficient on the linear regression analysis, these distributions are skewed, with modes very near the upper end of the range (0.88 and 0.87), and means of 0.82 and 0.83.

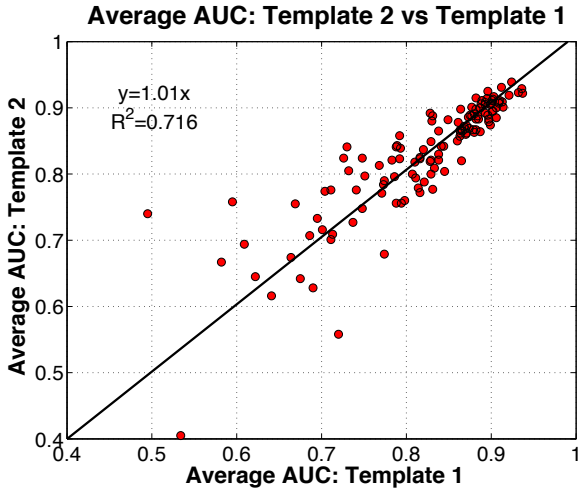


Figure 4.6: The average AUC of template 1 and template 2 for each test-retest pair is plotted in this figure, and fit with a linear relationship. The two average AUC values are highly similar, especially when the average AUC is large.

templates are shown, as well as the mean value, which is fit by a linear relationship (R^2=0.683). In this case, the intercept was chosen to be 0.5, to demonstrate that a test-retest pair that performed at chance on the ROC analysis could also be expected to have a correlation coefficient of nearly zero. The slope of this fit is 0.549. As in the previous case, the fit is better for more reproducible image pairs, whereas less reproducible pairs display greater deviation from the linear relationship.

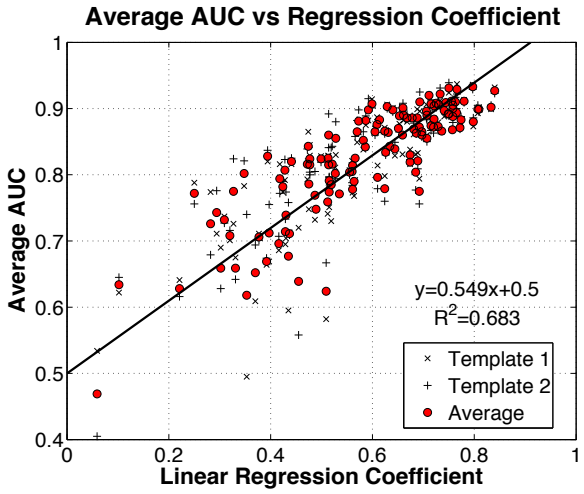


Figure 4.7: The average AUC as a function of correlation coefficient is shown for template 1 (x's) and template 2 (+'s) as well as the two averaged together (red circles). One can see that there is general agreement between the two methods, as both require the images to be reproducible. However, the methods are inherently different, and thus there is significant spread about the linear fit.

4.5 Factors Affecting Reproducibilty

Field strengths, tasks, contrasts, and subjects are examined here for significant between groups effects on reproducibility of the activation maps. Both mean regression coefficients and AUC (averaged over all template thresholds, for both template options) are plotted with the standard error of the mean for each between groups analysis. Statistical significance between pairs of groups is assessed using the two-sampled t-test with unequal variances.

4.5.1 Field Strength

No significant differences in reproducibility are found by either regression or ROC analysis (figure 4.8).

4.5.2 Functional Task

The between task analysis yielded significant differences only on the AUC analysis. Both the left hand tapping task and the sentences conditions have statistically greater

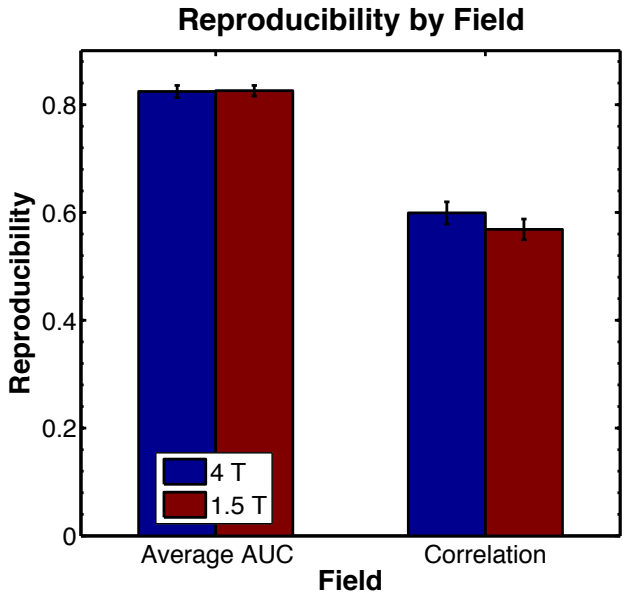


Figure 4.8: Bar plot of the average (with standard error of the mean) of both reproducibility measures sorted by field strength. No significant differences are found between 4 tesla (blue) and 1.5 tesla (red).

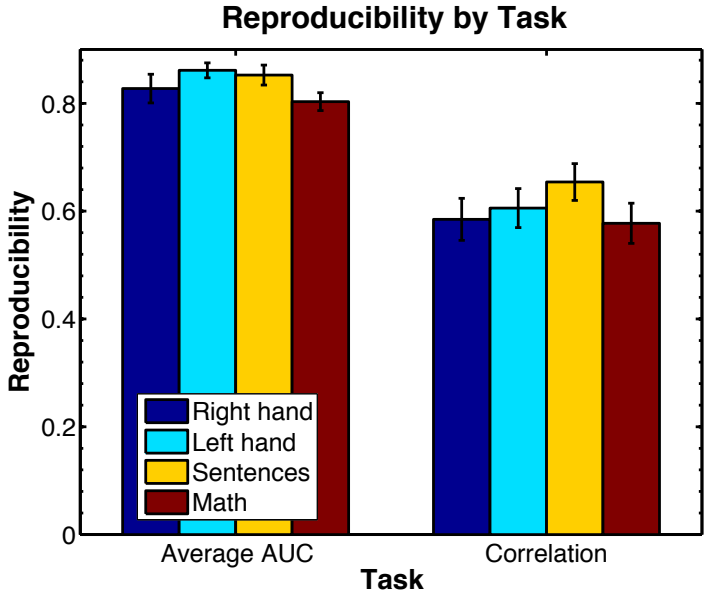


Figure 4.9: Bar plot of the average (with standard error of the mean) of both reproducibility measures sorted by task. The sentences (yellow) and left hand tapping (cyan) are significantly more reproducible than the mathematics (red) task, but only by the AUC metric.

average AUC than the mathematics task at the α=0.05 level (one-tailed) when contrasted to baseline (figure 4.9). The motor task additionally has significantly greater average AUC than the cognitive task when the two within task conditions are contrasted against each other (α=0.05, one-tailed, not shown), but not when the within task conditions are contrasted together against rest (C_1+C_2 >0). No pairings were found to be statistically significant on the linear regression analysis.

4.5.3 Contrast Generation

A much more pronounced effect on reproducibility is seen when the datasets are divided by contrast generated, regardless of task (figure 4.10). The reproducibility of the C_1-C_2 contrast is significantly less than all other contrast options at the α=0.01 level (one tailed). This level of significance is found on both the ROC and regression analyses.

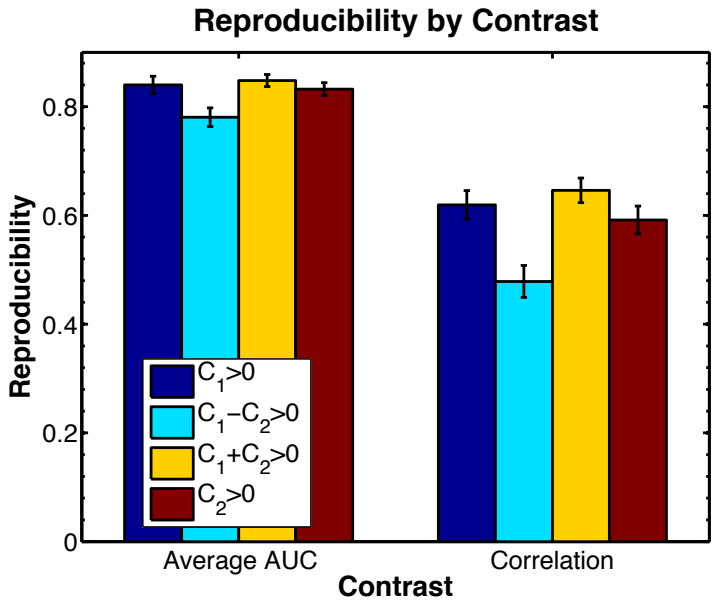


Figure 4.10: Bar plot of the average (with standard error of the mean) of both reproducibility measures sorted by contrast. Both the AUC and correlation measures of reproducibility find that the C_1-C_2 >0 contrast (cyan) is significantly less reproducible than the other contrasts.

4.5.4 Subjects

The greatest variability is seen between subjects (figure 4.11). There are many significant reproducibility differences between subjects, which are summarized in table 4.1 for the ROC analysis, and table 4.2 for the regression analysis. The significant differences are all related to subjects 4 and 6, who have lower reproducibility scores than all other subjects, to varying degrees of significance. Subject 2 exhibits a trend to lower reliability scores, but this does not reach significance at the 0.05 level. Subject 4 is found to perform significantly worse than other subjects on the ROC analysis more frequently than on the regression analysis (n=5 versus n=3 respectively), and vice versa for subject number 6 (n=5 versus n=7).

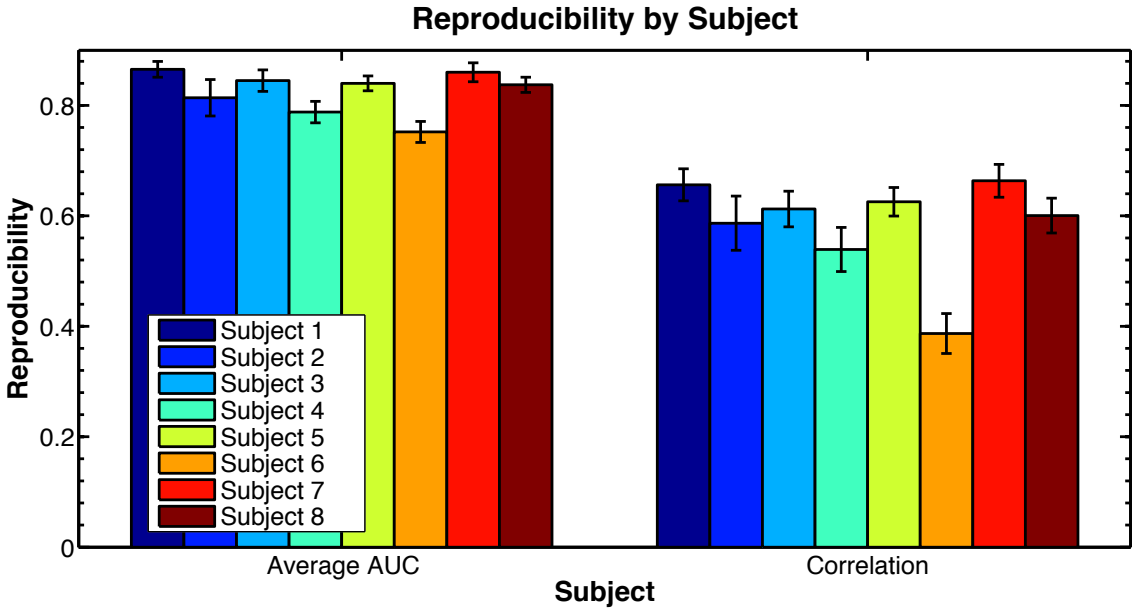


Figure 4.11: Bar plot of the average (with standard error of the mean) of both reproducibility measures sorted by subject. Refer to tables 4.1 and 4.2 for a summary of the significance tests.

4.5.5 Motion

Motion is a major sources of degradation of functional MRI image quality. In this work, the motion correction preprocessing steps are monitored, and the resulting realignments plotted. Figure 4.12 shows such a plot for the example datasets from figures 4.2 and 4.4. Subject 6 exhibits several large motion events, some approaching

Table 4.1: Differences by subject in average AUC (two-sample t-test, unequal variance, one-tailed). Entries represent the reproducibility of the subject indicated by the row being greater than that indicated by the column at: *: α <0.05, **: α <0.01, ***:α <0.001

Subject	1	2	3	4	5	6	7	8
1	X			**		***		
2		X						
3			X	*		***		
4				X				
5				*	X	***		
6						X		
7				**		***	X	
8				*		***		X

Table 4.2: Differences by subject in linear regression coefficient (two-sample t-test, unequal variance, one-tailed). Entries represent the reproducibility of the subject indicated by the row being greater than that indicated by the column at: *: α <0.05, **: α <0.01, ***:α <0.001

Subject	1	2	3	4	5	6	7	8
1	X			*		***		
2		X				**		
3			X			***		
4				X		**		
5				*	X	***		
6						X		
7				**		***	X	
8						***		X

the voxel size of the functional images. This dataset is very typical of subject 6, who moved by 1 millimeter or more in at least one image of every test-retest dataset. Subject 7 on the other hand, does not require motion correction greater than fractions of a milimeter, and shows no single motion events greater than roughly one to two tenths of a millimeter. The corresponding histograms of the relative abundance of t-values are shown in figure 4.13 (i.e. the number of voxels in each bin divided by the number in the largest bin), to give an idea of the severity of the effect of motion on theoretical statistical distributions.

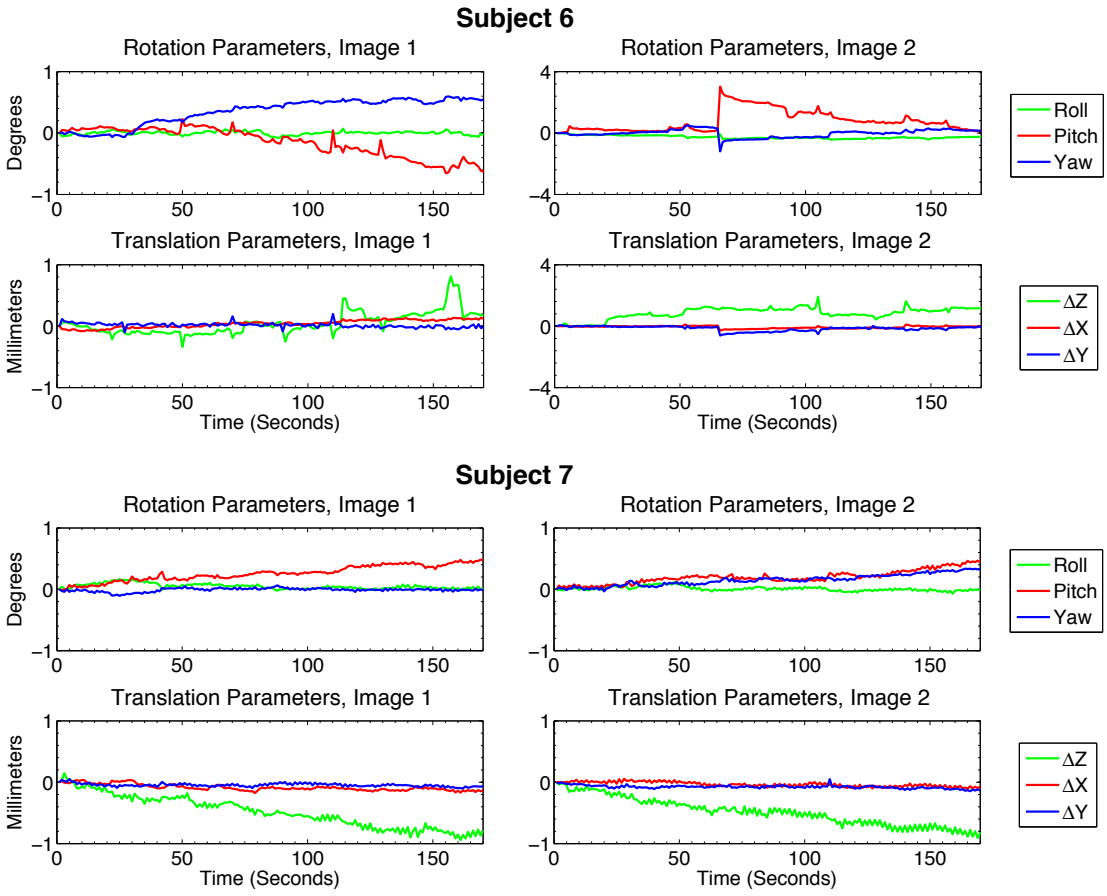


Figure 4.12: Motion correction plots for the example datasets of subject 6 (top two rows) and 7 (bottom two rows). Subject 6 moves considerably during all functional tasks, and by nearly a whole voxel in the seond trial of the motor task at 4 tesla, as shown. Subject 7, on the other hand, remains relatively still in all functional tasks, and also demonstrates significantly more image reproducibility.

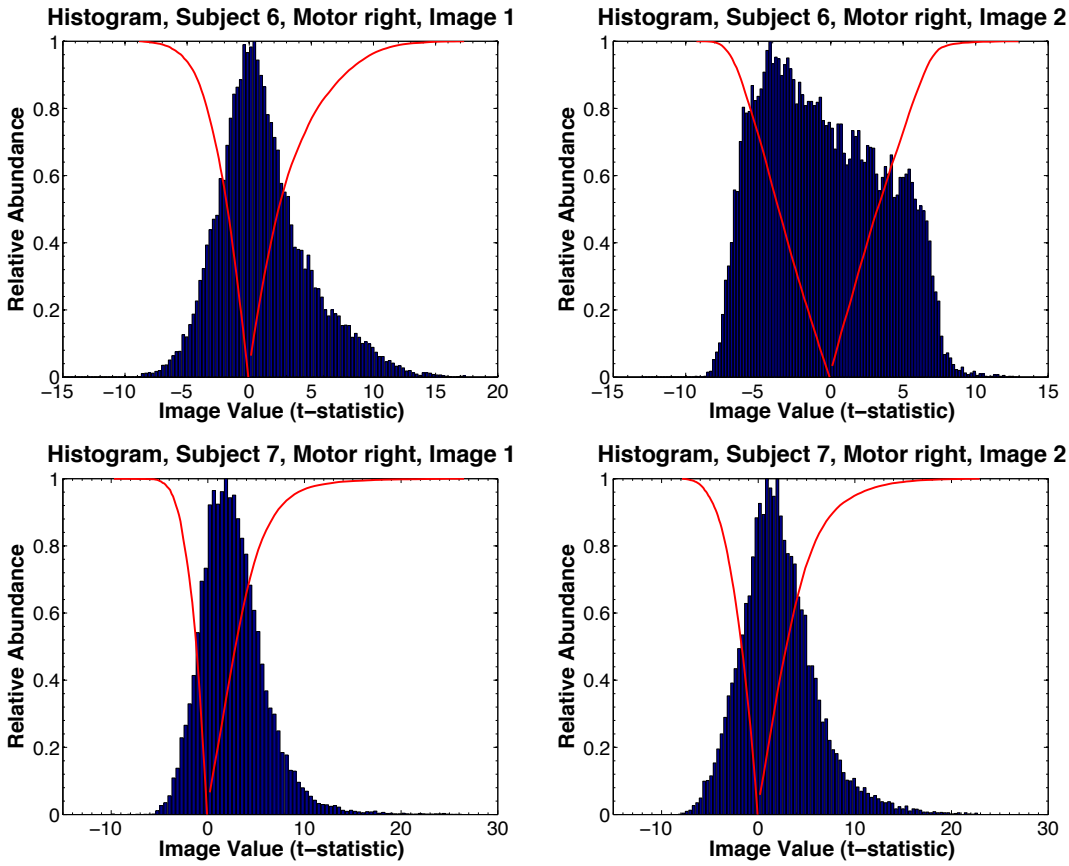


Figure 4.13: Histograms (blue) corresponding to the same fMRI images as the motion correction plots above for subjects 6 (top row) and 7 (bottom row). The cumulative distribution function is also shown as red lines for the positive and negative halves of the histogram separately. Note that the image with severe motion (top right) has a very unusual histogram. The t-values are symetrically spread over a significant range, with a broad rather than sharply peaked shape.

4.6 Discussion

Reproducibility of functional MRI results is found to be quite high by both measures employed in this study. The literature is quite sparse when it comes to individual level functional MRI, as most applications of fMRI attempt to make inferences at the group or population level. Many investigators have asserted the importance of reproducibilty analysis in this context, but the method of such an analysis is widely left up to the individual institution. While the linear regression method has been reported on previously, the method of reproducibility assessment by average AUC is novel to the knowledge of the author. Generation of both of these metrics requires

only that a second replicant of each functional task is performed, and so the two can be performed in concert quite easily.

4.6.1 Linear Regression Analysis

The strong relationship between the regression slope and regression coefficient shown in figure 4.3 is hypothetically explained in terms of subject fatigue, attention, and habituation. These three factors all have the potential to reduce the neural recruitment for a given task over time, which theoretically reduces the metabolic load, and therefore functionally related signal changes. As the sources of statistical noise cannot be expected to exhibit the same attenuation, the distribution of t-values associated with active voxels becomes more embedded in the noise background, and the predictive value of the first image decreases.

4.6.2 ROC/AUC Analysis

Recalling that an area under the curve of 0.5 indicates that the retest dataset discriminates at chance, an average AUC of 0.5 over template thresholds indicates that the test-retest pair performs at chance detection. The fact that there are AUC values below 0.5 (and in one case of template choice, an average AUC below 0.5) likely reflects that there are sources of structured noise, which lead to sub-chance predictive power. The most prevalent example of such artifact is likely subject motion.

With the spiral image acquisition used here, motion artifact appears in the raw images as long streaks of signal mislocation in the image, causing bright and dark bands. These are often caused by patient movement in response to the task, and are therefore correlated in time with the task. Thus the shape of these artifacts persists on the functional activation maps, as strongly positively and negatively correlated regions. As the spatial location of these artifacts is not necessarily consistent (depending on many factors such as direction and magnitude of the motion, and timing in relation to the imaging sequence), it degrades the test-retest sensitivity and specificity, thus reducing the AUC dramatically. This would also explain the tendency for datasets with lower average AUC deviate from the line of best fit more than the datasets with large average AUC, as structured noise in one of the two images would introduce bias into template performance.

4.6.3 Comparison of ROC/AUC and Linear Regression Analyses

Both the linear regression and average AUC measures of reproducibility found that the test-retest pairs were skewed towards the upper end of the respective scales. When compared directly (figure 4.7), it is obvious that while the relationship between the two is complex, there is a general trend of agreement. That the relationship could be fit relatively well with a line of fit that had a specified intercept of 0.5 is encouraging, as this intercept point was hypothesized to reflect zero reproducibility by both measures. If the intercept was left as a free variable, the intercept and correlation coefficient both increased by approximately 5 percent. Thus the ROC/AUC and linear regression analyses are certainly both reflective of the reproducibility of a test-retest pair, but use different levels of information to arrive at a quantitative result.

The regression analysis analyzes information at the voxel level. Test statistic values are compared at each voxel, regardless of the magnitude at neighboring voxels in space. It is well known that active voxels are likely to exist in contiguous regions (clusters), and therefore this may not be the optimal level of analysis for reproducibility. Additionally, the regression analysis is ignorant to the process of thresholding, a routine aspect of processing functional activation maps. For this reason, the linear fit of the regression analysis may be biased by the noise distribution centered about t-value of 0, especially when the number of truly active voxels is small. However, the linear regression method is fast, simple, and widely familiar, three characteristics that are certainly important when it comes to practical application.

The ROC analysis on the other hand calculates AUC by comparing distributed activation patterns. In other words, it is sensitive to the overlap of clusters of activation, topographies defined at a given threshold magnitude. In this process, at a given test-retest threshold pair, it does not matter how much above threshold the voxels within an active cluster are; the extent of the clusters determines the true and false positive rates. At a given template threshold, if there is some value of the retest threshold that makes the two spatial patterns highly similar, then the AUC will be large. If this can be said for all template thresholds, then the images are very similar, which will be reflected by a large average AUC. There are alternative metrics that could be used as indicators of reproducibility in the ROC analysis. For example, the shortest distance from the ROC curve to the optimal point of 100% sensitivity,

100% specificity displays similar assymptotic behaviour to the AUC as a function of template threshold. However, the AUC is calculated from all retest thresholds, whereas the optimal point is a single data point, and therefore the AUC may be a more representative metric of reliability.

The differences between the ROC and regression analyses make them inherently sensitive to different aspects of the processing pipeline. For example, the ROC analysis is likely to be less sensitive to different smoothing kernels used in the preprocessing stage, as this will only slightly alter the activation patterns. This is especially true at low thresholds, when regions of activation are large. However, smoothing may have a more significant impact on the linear regression, as it alters the precise test statistic values at all voxel locations.

The ROC and regression analyses do not always agree on which groupings are statistically different from one another. In the by task comparison, ROC analysis is the only method to detect significant differences. Additionally, the ROC analysis finds the left hand condition to be the most reproducible, while the regression analysis finds the sentences condition to be most reproducible (although these were not statistically different from one another in either case).

4.6.4 Factors Affecting Reproducibility

Increasing field strength is known to improve signal to noise and contrast to noise of functional MRI data, and so it was expected that reproducibility would also increase due to enhanced separation of signal and noise distributions. There are some image artifacts that are more problematic at higher field (e.g. field inhomogeneities), and some that are more prevalent at lower field (e.g. draining veins). If these artifacts are repeatable and consistent, they will appear like true activation on a reproducibility analysis. This could introduce erroneous sources of enhanced or degraded reproducibility that does not reflect the reliability of neural activation signals.

Sensory-motor tasks are widely considered more robust than cognitive tasks, but the finger tapping task requires more subject movement, which could be responsible for suppressing the difference between the tasks. That the sentences condition was found to be more reliable than the mathematics condition by ROC analysis could be explained by the fact that the sentence stimuli are more complex. Therefore, this

condition would require greater concentration and effort on behalf of the subject, increasing the activation strength.

The low reproducibility of the C_1-C_2 >0 contrast compared to the other contrasts explored is to be expected because testing for a difference in response by condition is more subtle than testing for the presence or absence or response of any kind. This contrast will also result in fewer active voxels, and so the bias of the linear fit in the regression analysis by the noise distribution centered about t-values of zero may be worsened compared to other contrasts.

Subject-to-subject differences are known to be a major source of variability in activation patterns, often overshadowing the contribution from between sessions or between scanners. It is natural that this would also be a dominant source of variation in reproducibility, as task performance, vigilance, and compliance are key to the validity of fMRI results. It is worth noting that the between subjects analysis is the least sensitive of the analyses performed on factors affecting reproducibility, because there are more groups (n=8), and therefore fewer test-retest pairs per group (n=16). It is thus likely that more significant differences could be detected by increasing the number of test-retest pairs per subject.

Motion during the scan significantly degrades the image quality of functional MRI scans by producing error in the spatial encoding procedures. Additionally, motion changes the distribution of inhomogeneities in the imaging volume, which reduces accuracy of the shimming procedures used to achieve a homogeneous magnetic field. Signal changes associated with motion events can be equal to or greater than the functional changes of interest, and as such can cause false activation. This is an even greater problem if the motion is correlated with the task, as is often the case for tasks requiring a subject response. In this case, the correlation based statistical procedures used to generate activation maps will very easily mistake motion induced signal changes for the BOLD effect.

4.7 Conclusion

While reproducibility is generally high, it varies considerably from subject to subject, and dataset to dataset. One test-retest pair has such poor reproducibility that predicitve value of the first image is less than that dictated by pure chance. It is

clear that even a validated protocol of excellent general reproducibility is subject to case-by-case variability. For example, the sentence task has the highest average correlation coefficients but performed as poorly as r=0.353 in one case. Similarly, the left hand tapping task was best on the ROC analysis, but exhibited an average AUC of just 0.732 in one case. Both of these are far from the group means for these tasks of 0.654 and 0.861 respectively. Thus it is not sufficient to assess reproducibilty of a protocol at the group level for applications to individual level functional MRI mapping. Individual level reproducibility analysis should be performed, so that unreliable results can be detected and treated accordingly.

The ROC analysis is more information rich than the regression analysis, as it better displays the intricacies of the process of applying a threshold to the activation maps. This comes at the price of increased computational demand, and conceptual complexity. This price may be justified as the inherent sensitivity to activation threshold can be exploited to guide or automate the choice of threshold on a case-by-case basis.

Chapter 5

Setting Activation Thresholds for Functional MRI

Setting thresholds on functional MRI images remains a significant challenge obstructing standardization of the processing pipeline. Typically decision thresholds are set in terms of acceptable error rates, such as the false positive rate in the Bonferroni threshold, or the false discovery rate in the FDR method. In this section, we will compare these traditional threshold methods with a novel data-driven method, based off the ROC and AUC analysis of reproducibility. In this scheme, the active/inactive decision cutoff is the slope of the AUC as a function of template threshold. This chapter will present the results of applying these three threshold schemes to test-retest data from 8 healthy controls subjects. The effect of factors such as field strength, task, and contrast generation will be examined for effects on the activation thresholds, and resulting activation maps. Finally the relationship between reproducibility and the automated threshold procedure will be addressed.

5.1 Methodological Details

As discussed in section 2.4.3, the use of a fixed false positive rate (voxel wise error rate) is inappropriate in functional MRI analysis because of the large number of comparisons (i.e. voxels). The two most common schemes for addressing this problem are the Bonferroni and False Discovery Rate (FDR) corrections. The Bonferroni correction uses a desired image-wide false positive rate (p-value), while the FDR method requires specification of a q-value, which reflects the proportion of positive results that are incorrect. These values are typically set in the range of 0.01-0.05, to parallel uncorrected error rates of α=0.01-0.05 used in single significance tests (uncorrected t-values of 2-3 for typical fMRI data sets).

By contrast, the derivative-based cutoff (d-value) of the AUC analysis has no direct interpretation in terms of error rates. Indeed, as described in section 3.1.1, the process is best illustrated graphically. Therefore, the example ROC/AUC plots

from section 4.3 are reproduced in figure 5.1, altered to illustrate example results of applying the pROCreate threshold algorithm. The derivative of the cubic spline fit is now plotted for each curve, and the upper and lower bounds of the template image t-value thresholds tested in the ensuing chapter are indicated (corresponding to d=0.01 and d=0.05). The ROC curves from which the retest thresholds are chosen are shown for three choices of d-value, representing the upper, lower, and middle regions of the template cutoff range.

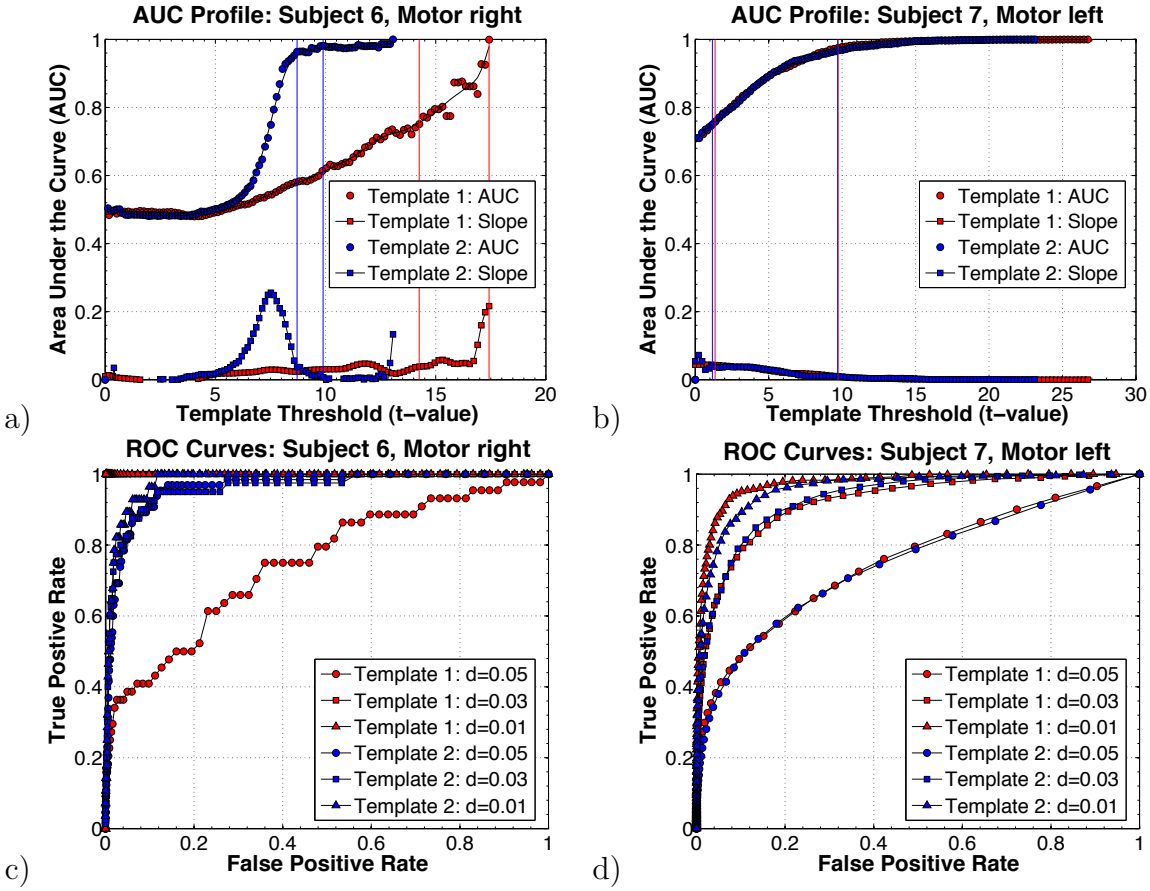


Figure 5.1: a) and b) AUC profiles from testing the full range of template thresholds for the same example datasets as figure 4.4. The vertical lines on the AUC plots are the d=0.05 (leftmost) and d=0.01 (rightmost) cutoff for the first (red) and second (blue) image as template. The ROC curves corresponding to template thresholds for d=0.01, d=0.02, and d=0.05 are shown in c) and d).

It is observed that the slope may begin below desired cut-off values, and so the algorithm must be made resilient to applying thresholds in this region. This can be accomplished by requiring the slope to cross the cutoff value from above, or by

imposing a minimum AUC value for template threshold detection. Due to fluctuations in the derivative calculation, especially near the endpoints, the latter may be more robust, although in practice the two produced nearly identical results. An additional caveat of using the cross-from-above method is that it can cause the less restrictive d-value cutoffs to produce thresholds greater than the more restrictive d-value cutoffs, as would be the case for the test-retest pair of subject 7, shown in figure 5.1. Because the AUC derivative is never above 0.05, this cutoff value would fail to detect a suitable threshold template, in which case the highest t-value possible is applied. Therefore in this work, an AUC value of 0.75 is required to assert that the asymptote had been detected, as this method maintains consistent ordering of d-value and t-value thresholds.

Fortuitously, the derivatives tend to reach a maximum value 0.1-0.2, and so suitable cutoff derivatives for detecting the asymptotic plateau fall in the same range as the error rates described above (0.01-0.05). This makes it easy to compare the results of the various threshold procedures on a single graph. The focus of this chapter will thus be plots of the image threshold as a function of decision cutoff, for all the threshold methods investigated (i.e. p, q, or d values). The primary effect of increasing the threshold is to reduce the number of voxels in the image, so the resulting number of active voxels in each image after application of the threshold will also be presented. The Bonferroni correction will be employed using the number of voxels containing brain tissue only for the multiple comparison correction. The False Discovery Rate correction will be performed both with and without the c(V) correlation adjustment term. The pROCreate method will be used for setting automated, data driven thresholds, as described above.

5.2 Comparison of Threshold Techniques on Test-Retest Images

The results of applying the three threshold methods to both the first and second trials of the test-retest pairs are shown in figure 5.2. The Bonferroni correction is the most restrictive, even when only the set of voxels containing brain tissue are included in the correction factor. The FDR correction is more strict with the c(V) term included, and the FDR threshold without the c(V) term is the least restrictive of all the traditional methods. The ROC threshold shows much more variation over the

corresponding range of d-value cutoffs. For d=0.05, the ROC automated thresholds typically lie between the two FDR corrections, while at the more restrictive end near d=0.01, the ROC thresholds tend to be between the Bonferroni correction, and the strict FDR method. There is also markedly more variation in the ROC threshold at a given cutoff than in any of the other methods tested.

One can observe on this same figure that the ROC method is the only method of the three to differentiate between the first (solid line) and second (dashed line) image of the test-retest pair. At nearly all d-value cutoffs tested, the first image threshold is on average higher than the second image, when all available test-retest datasets are considered. The corresponding plots of the number of active voxels as a function of decision cutoff reveals that the ordering of the different methods is reversed, as higher thresholds imply lower voxel counts. The variability in the ROC threshold t-value does not translate into increased variability of the number of active voxels. In fact, the opposite trend is observed and the variability in number of active voxels lies somewhere between the Bonferroni and FDR methods, and decreases with decreasing d-value cutoff (i.e. increasing threshold). While the numbers of active voxels in the first and second images are significantly different for the Bonferroni and FDR methods, in the ROC method, the two images have indistinguishable voxel counts.

5.3 Factors Affecting Threshold Levels

In this section, some of the factors affecting reproducibility identified in chapter 4 are revisited to examine the effect on activation threshold. Because of the variability of the ROC method, there is not sufficient power to detect differences between subjects, and this factor is omitted from the proceeding section.

5.3.1 Field Strength

The Bonferroni threshold is unaffected by field strength (figure 5.3), leading to a large discrepancy in the number of active voxels between 1.5 and 4 tesla. The greater number of active voxels is found at 4 tesla, indicative of the elevated t-values found at the higher field strength. The FDR method, by contrast, does result in a slight, but significant difference in t-value thresholds between field strengths, as shown in figure 5.4. The t-value threshold is larger at 1.5 tesla than 4 tesla, by roughly 0.1-0.2

at all q-value cutoffs, and for both the c(V) corrected and uncorrected method. This increases the difference between the numbers of voxels declared active at the two different field strengths.

The ROC method shows the opposite dependence of t-value threshold on field strength (figure 5.5) from the FDR method. The 4 tesla activation map thresholds are higher than the 1.5 tesla maps, although this difference is only present for cutoffs of d<=0.03. For AUC slope cutoffs more strict than this (d<0.03), the increasing difference in t-value threshold produces activation maps with the same number of active voxels, whereas for less strict cutoffs, the number of active voxels is greater at 4 tesla.

5.3.2 Functional Task

To examine the dependence on task, the two simple motor or cognitive contrasts (C1 >0 and C2 >0) are grouped together to increase detection power. Neither the FDR methods nor the Bonferroni method show any difference between the motor and cognitive tasks in either threshold value or extent of activation (figures 5.6 and 5.7 respectively). However, the pROCreate algorithm does produce significant differences, as shown in figure 5.8. At all tested derivative cutoffs, the automated threshold is larger for the motor task, and this results in fewer voxels being declared active (albeit of larger average significance).

5.3.3 Contrast Generation

The Bonferroni method characteristically displays no dependence on the choice of contrast (figure 5.9). The False Discovery Rate methods both threshold the contrast C_1-C_2 >0 at a higher t-value than the other contrast options, resulting in far fewer voxels being declared active for this contrast (figure 5.10). The C_1+C_2 >0 contrast is trending towards lower thresholds, and more active voxels by both the c(V) corrected and uncorrected FDR methods. The ROC/AUC threshold method assigns the lowest thresholds to the C_1-C_2 >0 contrast, for d<=0.02 (figure 5.11). The C_2 >0 contrast is given t-value thresholds greater than the C_1-C_2 >0 contrast, but lesser than the other two contrast options for d<=0.02. In terms of number of active voxels, for the C_1-C_2 >0 contrast produces significantly fewer active voxels than the other contrasts

despite the lower activation thresholds. The $C_1+C_2 > 0$ contrast generates the most active voxels, although the difference between it and the single condition vs. rest contrasts is not significant for d<=0.025.

5.4 Automated Thresholds and Reproducibility

No clear relationships are observed for any of the threshold methods between t-value threshold and either measure of reproducibility. However, there is a dependence of the number of voxels included in the template image by the pROCreate algorithm and the average AUC. The ROC/AUC threshold method uses many more voxels in the template image for reproducible test-retest pairs than for the pairs with low reproducibility scores. This relationship is shown in figure 5.12 for d=0.02, and is roughly exponential, so as to necessitate a logarithmic plot.

On the other hand, very little dependence on reproducibility can be observed for the retest threshold or number of active retest voxels. The optimal threshold that is applied to the retest image does not produce a reduced activation extent for the less reliable images. For unreliable test-retest pairs, the retest image thus often contains tens to hundreds of times more voxels than the template image it is matched to. The dependence is again approximately exponential, and requires a logarithmic plot (figure 5.13).

5.5 Discussion

5.5.1 Comparison of Threshold Methods

The three threshold techniques behave very differently, both in terms of their dependence on cutoff choice, and in terms of the variability of threshold levels and activation extent. In particular, the Bonferroni method shows little variability in threshold, as it only depends on the number of voxels in the brain and decision cutoff, which is relatively constant at a given resolution. While subject-to-subject differences will certainly be present, any differences from test to retest will be primarily a result of instability in the automated tissue segmentation routine. The FDR method on the other hand is data driven, depending on the strength of the statistical correlations in the data, and therefore results in inherently more variable thresholds. Although the

increased variability in thresholds for the FDR methods leads to more variability in activation extent, the ROC method uses an increased range of image thresholds to maintain more consistent numbers of active voxels. This is consistent with the fact that the automated threshold procedure works on matching patterns of activation, rather than a priori assumptions about statistical significance. The ROC method is also the most dependent on decision cutoff, and is the only method that does not maintain a consistent ordering with respect to the other methods in terms of threshold and number of active voxels across the range of cutoffs tested.

That the ROC method on average set higher thresholds on the first image of the test-retest pairs than the second is especially interesting in light of the linear regression analysis results of section 4.2, which shows that the slope of best fit to test-retest data is normally less than 1. Thus, if the images have reproducible distributions of activity (i.e. activation patterns or neural networks), but diminished magnitude, we expect that the threshold could be adjusted downward on the second image to produce comparable results. This is precisely what the present results of the pROCreate algorithm seem to suggest, and is further supported by the number of active voxels being equal in the test and retest images by this method.

5.5.2 Factors Affecting Threshold Levels

It is seen from the results of the Bonferroni threshold method, that when divided by field strength, equal threshold levels result in significantly more voxels being declared active at 4 tesla. However, the previous chapter showed no difference in the correlation coefficient or average AUC of the test and retest datasets between field strengths. Indeed, the ROC algorithm compensates at strict d-value cutoffs by using higher t-value thresholds at 4 tesla than 1.5 tesla, resulting in a reduced difference between the numbers of active voxels at the two field strengths. Strangely, the FDR method shows the opposite trend, applying stricter t-value thresholds to the lower field strength. This is because the elevated t-values at higher field correspond to smaller uncorrected p-values. The maximum value of 'i' satisfying the inequality of equation 2.36 will thus be larger, and the corrected p-value will be larger. A larger corrected p-value in turn equates to a smaller t-value, and thus whenever the t-values in an image are increased, the FDR corrected threshold will decrease. It is questionable as to whether or not

this behavior is desirable for functional MRI, especially at the individual level.

Only the pROCreate method showed significant differences between the motor and cognitive task for both t-statistic threshold and number of active voxels. The threshold applied for the motor task was generally higher by this method, resulting in fewer active voxels. This is sensible from a neural perspective, as cognitive tasks are known to recruit a more distributed network of functional nodes.

The results by contrast once again demonstrate that there are far fewer significant voxels in the C_1-C_2 contrast, regardless of choice of threshold method. However, in this case we once again see the ROC and FDR methods producing opposite results. Because of the lower t-values in this contrast, the FDR method applies a larger t-value threshold, whereas the ROC method applies a lower t-statistic threshold than to the other contrasts. It is not immediately obvious which of these is more appropriate.

5.5.3 Reproducibility and Automated Thresholds

The relationship between the number of voxels included in the template by the pROCreate routine and reproducibility is shown in figure 5.12, but is also suggested by the panels of figure 5.1. In the latter figure, it is observed that the AUC for subject 6 remains near 0.5 until template thresholds of t=5 or t=6 for either template. The AUC for template 2 then rapidly increases and plateaus, while template 1 does not asymptote at all. In the latter case, the algorithm fails to find a suitable template cutoff, and only a single voxel is included in the template. The algorithm thus naturally suppresses the activation template for unreliable image pairs because the AUC remains low for most of the template threshold range. Only the most strongly task correlated voxels in these datasets are reliable, and the reduction of active voxels in the template reflects this.

It is not entirely obvious why this does not result in reduced activation extent in the retest image, but reflects that, for less reproducible pairs, the algorithm must include comparatively many retest voxels to match a pattern of relatively few template voxels. For average AUC values above 0.85, the ratio of active voxels in the retest to template images tends not to exceed an order of magnitude difference, and this range appears most suitable for application of the pROCreate threshold algorithm.

5.6 Conclusions

Of the three threshold methods tested in this chapter, the ROC method shows by far the most flexibility, while the Bonferroni method is the most rigid. The pROCreate algorithm seems very promising for setting data-driven thresholds, as it was able to adapt to different inputs easily, such as setting different thresholds for the first and second image of the test-retest pair, for different field strengths, and for different tasks. Furthermore, the adjustments in threshold often produced sensible results in terms of activation extent, and theoretical properties of the fMRI images. The routine appears best suited for datasets that are highly reproducible, which is not surprising, as these will naturally contain easier patterns to match. Further investigation of the performance of the pROCreate method is warranted, and in particular, cluster level analysis would be a natural extension of this work. The next chapter will demonstrate the use of the pROCreate algorithm in a series of patient cases, including a comparison with the gold standard in functional mapping, cortical stimulation.

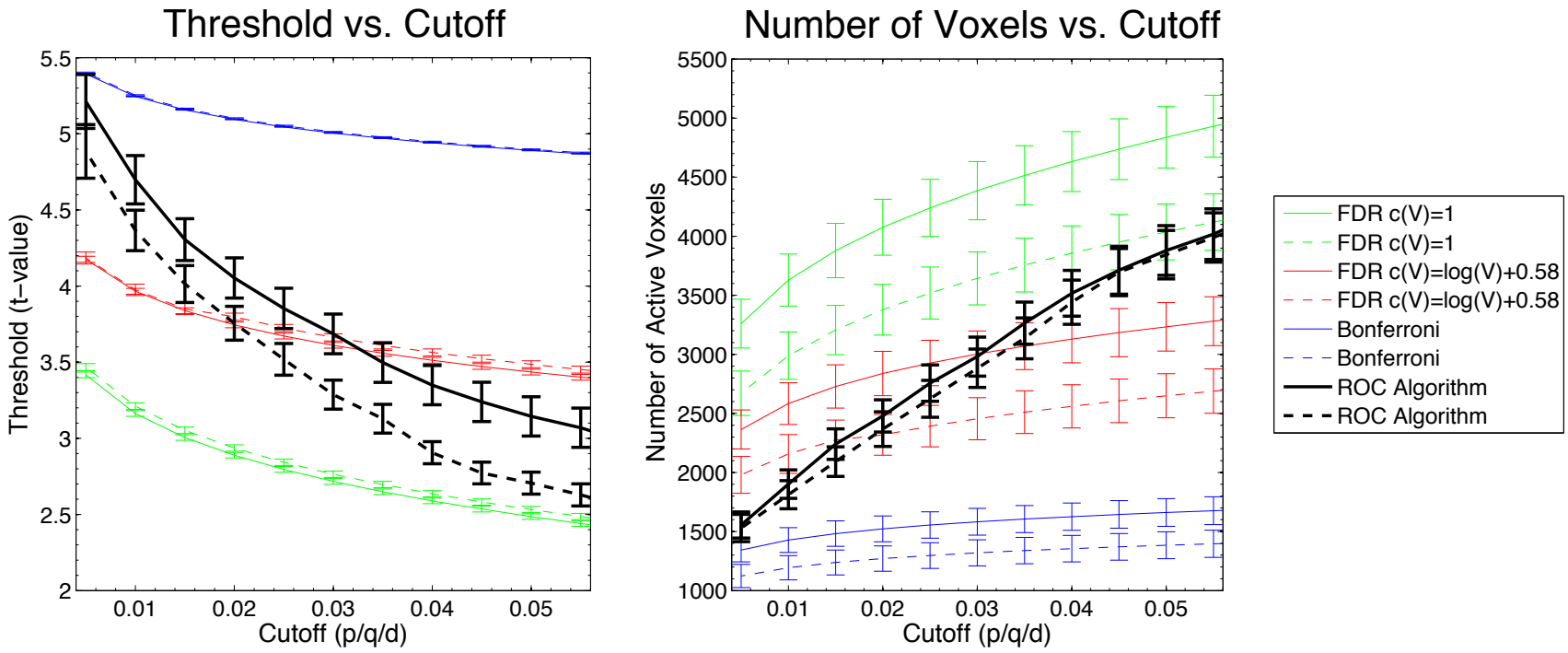


Figure 5.2: The t-value threshold as a function of decision cutoff for all threshold methods investigated (i.e. p, q, or d values). b) The number of (retest) voxels declared active as a function of decision cutoff. The heavy black line shows the novel ROC threshold method, while the thin colored lines are the traditional methods. The solid lines represent the threshold applied to the first image (i.e. the second image acting as template), and the dashed lines to the second image of the test-retest pair.

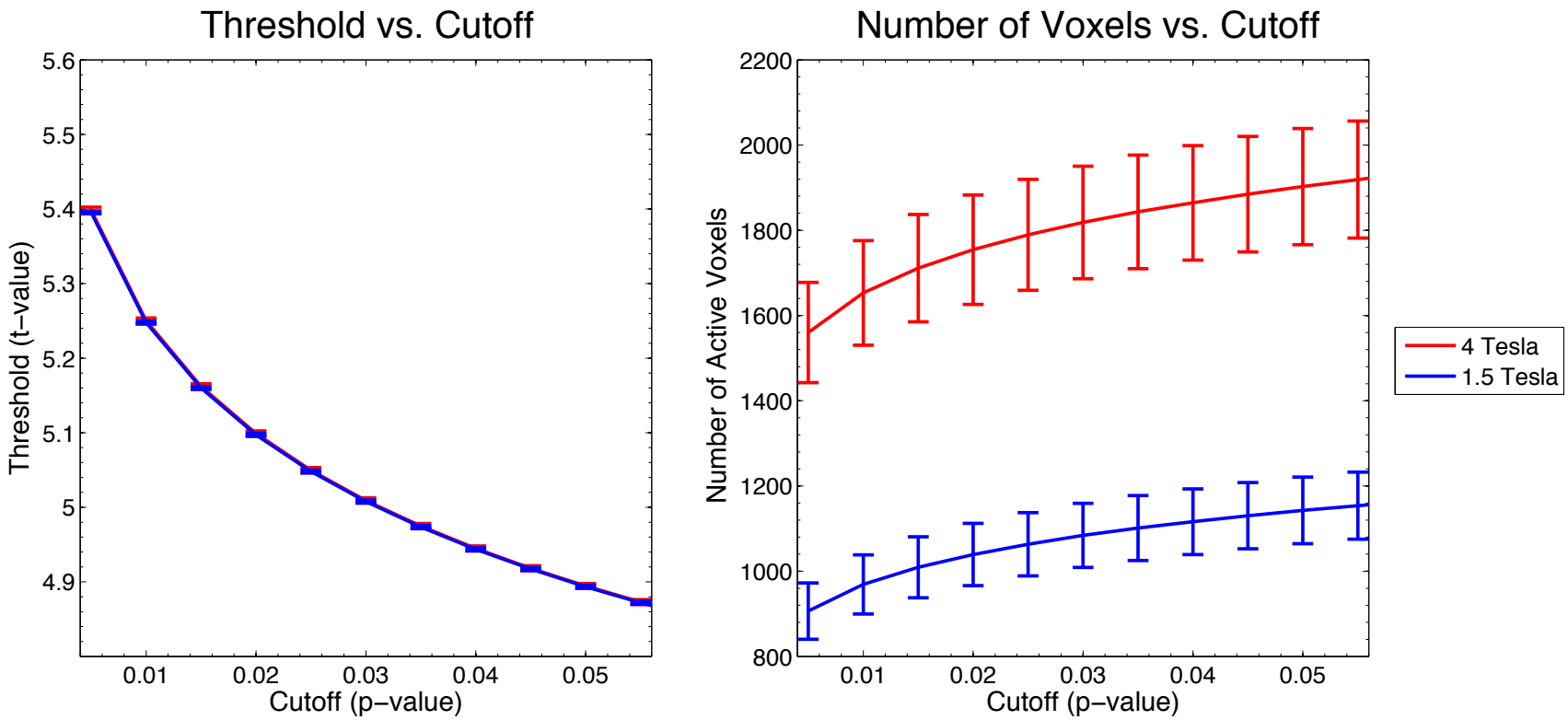


Figure 5.3: Bonferroni threshold results by field strength. a) The t-value threshold and b) The number of (retest) voxels declared active as a function of decision cutoff (p-values).

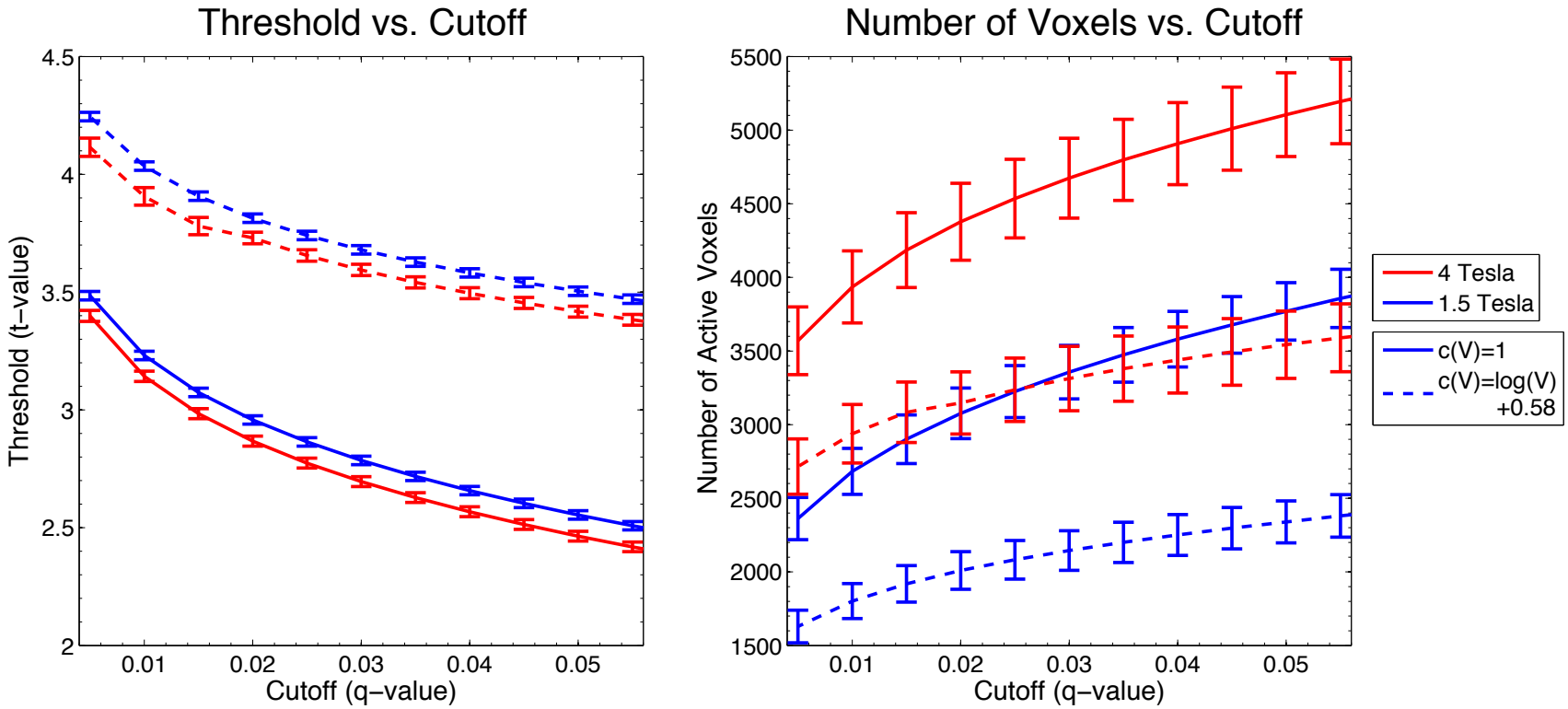


Figure 5.4: False discovery rate threshold results by field strength. a) The t-value threshold and b) The number of (retest) voxels declared active as a function of decision cutoff (q-values).

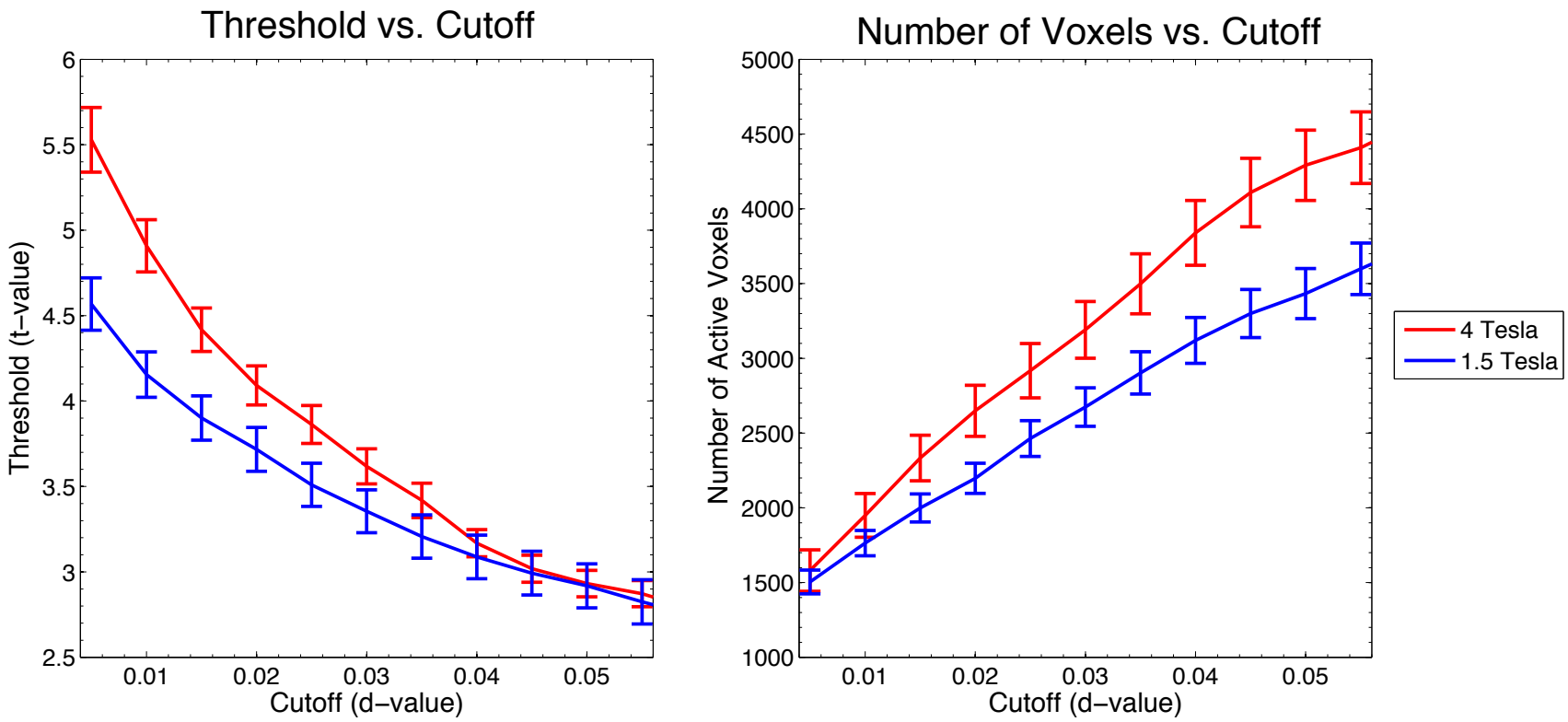


Figure 5.5: ROC threshold results by field strength. a) The t-value threshold and b) The number of (retest) voxels declared active as a function of decision cutoff (d-values).

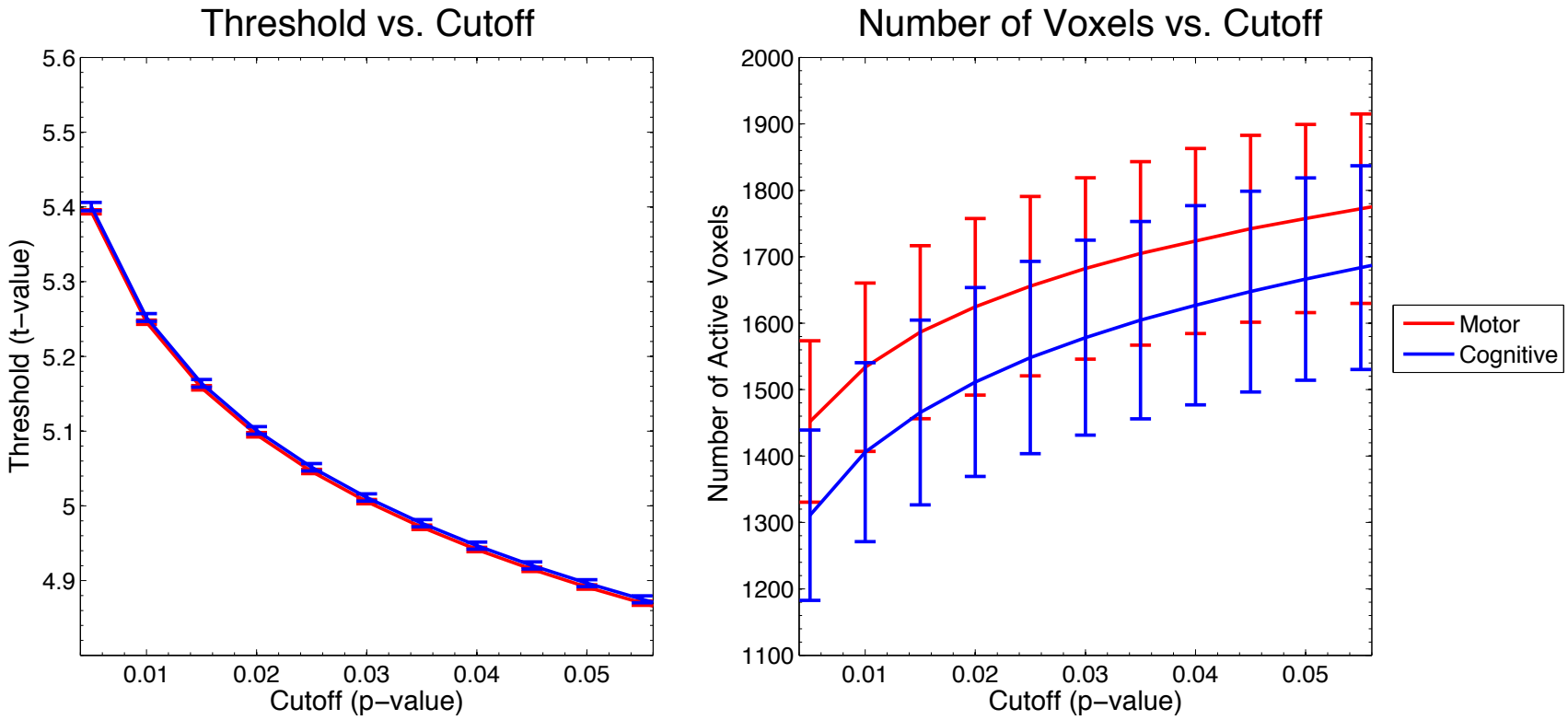


Figure 5.6: Bonferroni threshold results by task. a) The t-value threshold and b) The number of (retest) voxels declared active as a function of decision cutoff (p-values).

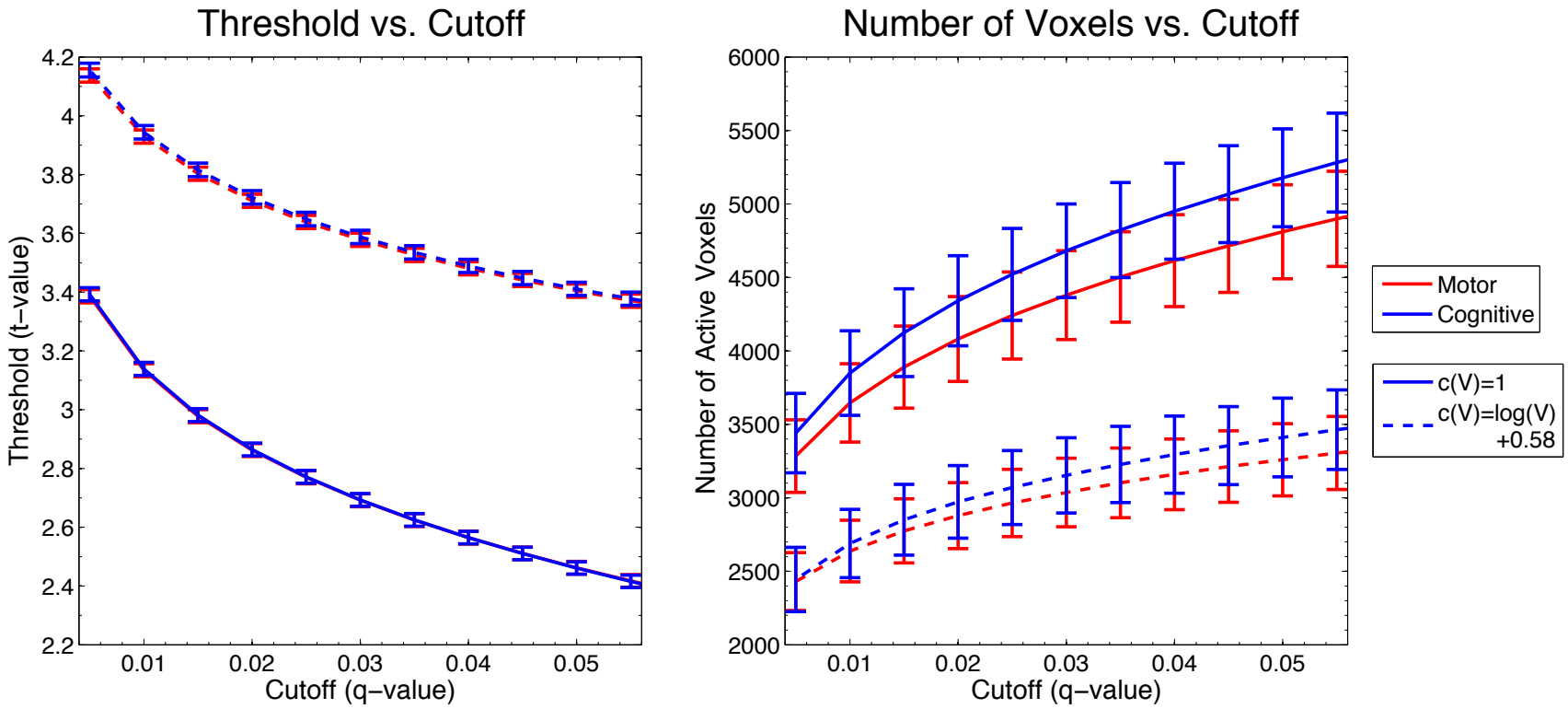


Figure 5.7: False discovery rate threshold results by task. a) The t-value threshold and b) The number of (retest) voxels declared active as a function of decision cutoff (q-values).

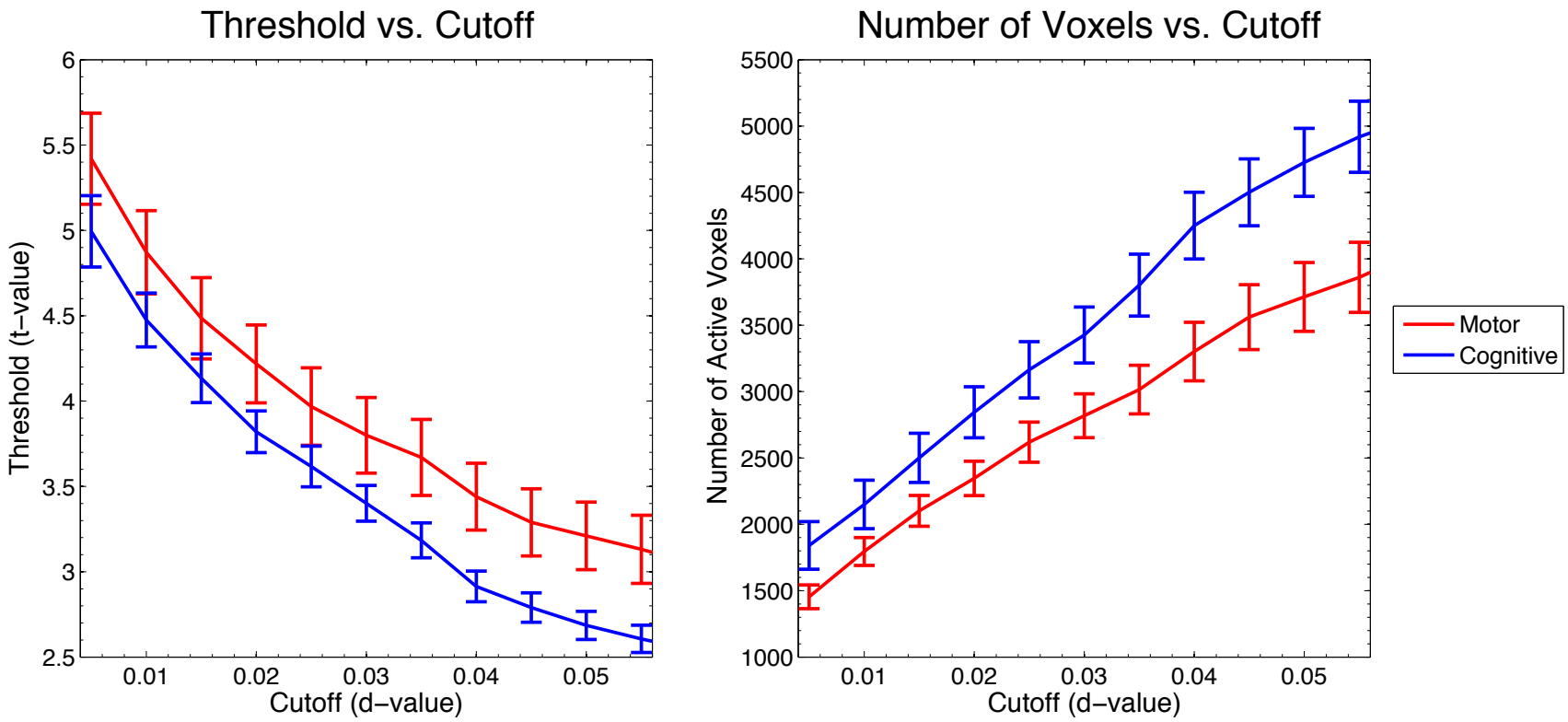


Figure 5.8: ROC threshold results by task. a) The t-value threshold and b) The number of (retest) voxels declared active as a function of decision cutoff (d-values).

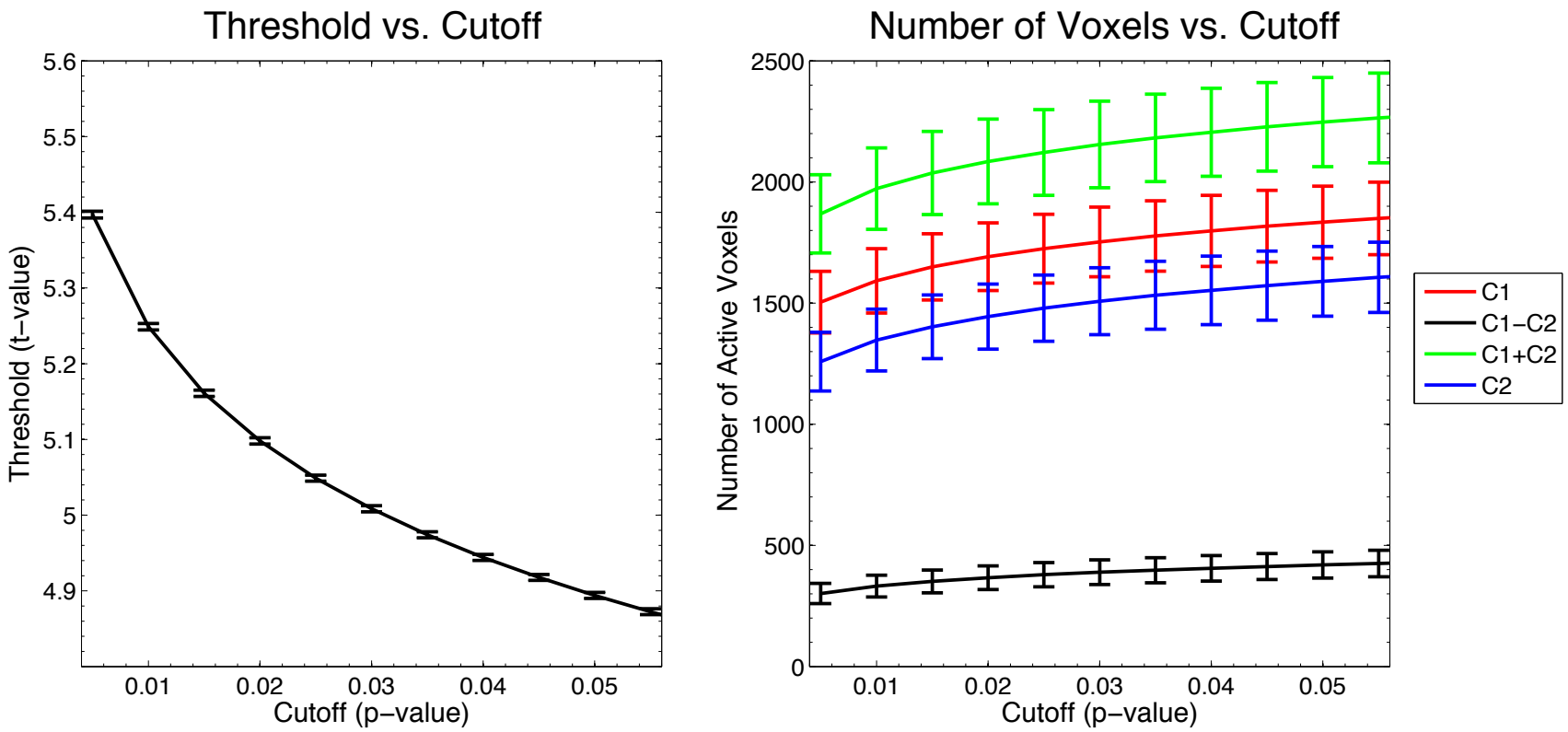


Figure 5.9: Bonferroni threshold results by contrast. a) The t-value threshold and b) The number of (retest) voxels declared active as a function of decision cutoff (p-values).

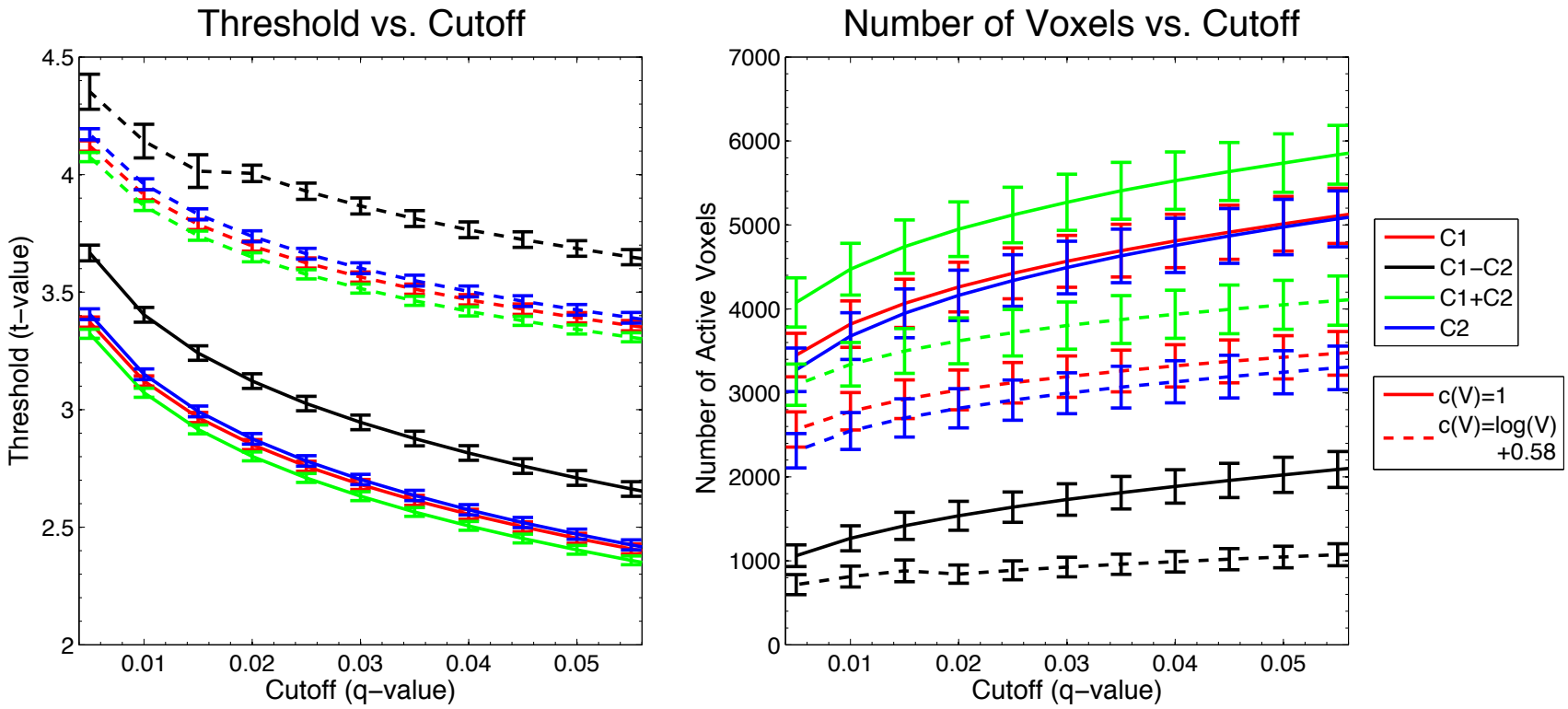


Figure 5.10: False discovery rate threshold results by contrast. a) The t-value threshold and b) The number of (retest) voxels declared active as a function of decision cutoff (q-values).

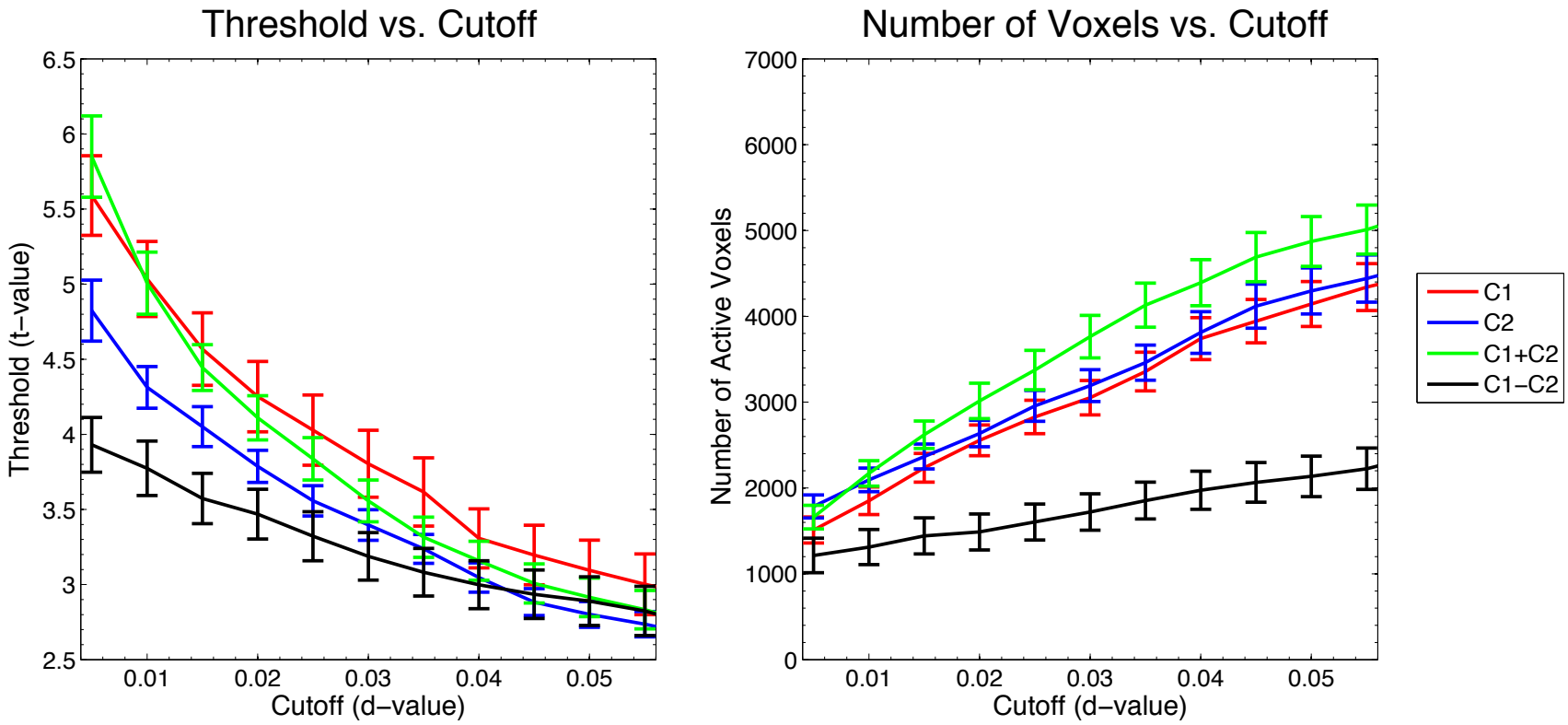


Figure 5.11: ROC threshold results by contrast. a) The t-value threshold and b) The number of (retest) voxels declared active as a function of decision cutoff (d-values).

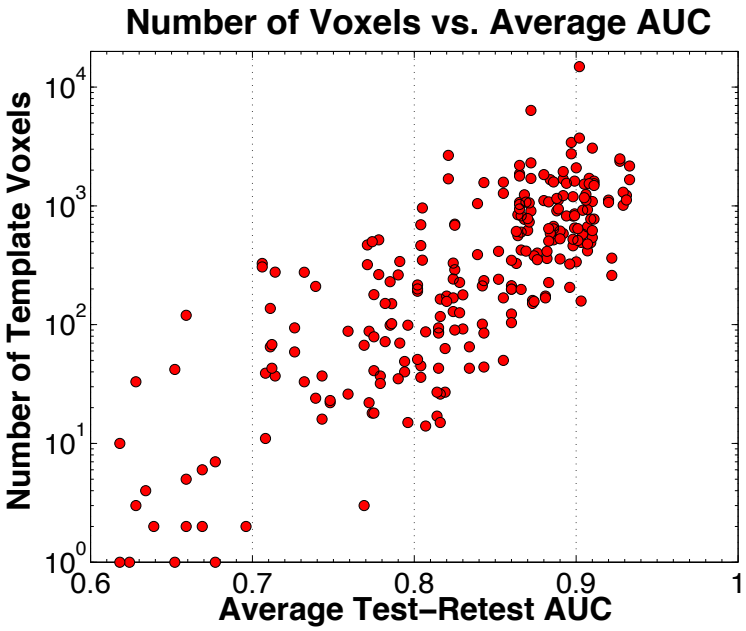


Figure 5.12: Plot of the number of template voxels used by the automated threshold routine versus test-retest AUC for both templates 1 and 2.This data is for d=0.02, but a similar relationship is seen at other d-value cutoffs. Test-retest pairs with lower reproducibility use far fewer template voxels.

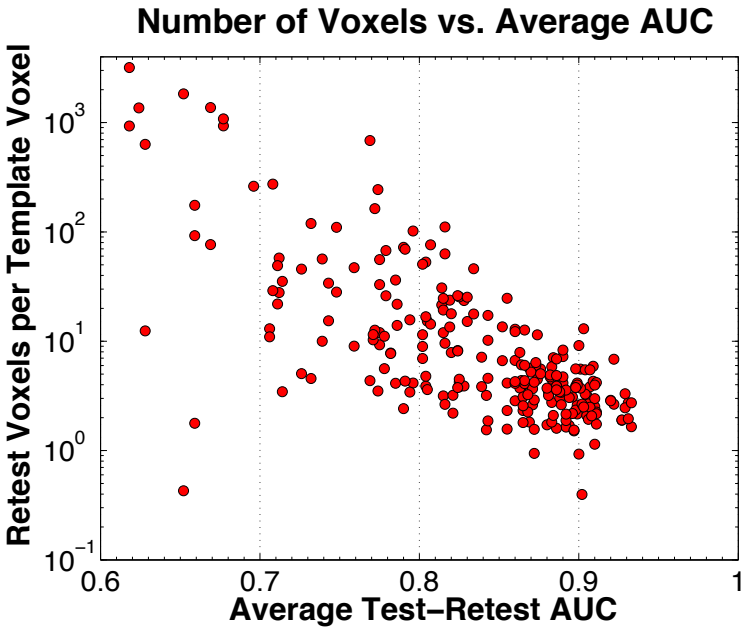


Figure 5.13: Plot of the ratio of voxels declared active in the retest to template image versus average AUC. This data is for d=0.02, but a similar relationship is seen at other d-value cutoffs. Test-retest pairs with lower reproducibility produce tens to thousands of voxels in the retest image per voxel included in the template image.

Chapter 6

Patient Cases and Comparison with Cortical Stimulation

In this chapter, the results of the patient studies undertaken in the second phase of this research will be presented. In particular, focus will be placed on the six patients who performed test-retest functional MRI as well as cortical stimulation mapping. The cases will be presented one-by-one, starting with a summary of the presurgical mapping including both the reproducibility tests discussed in chapter 4, and the automated threshold routine presented in chapter 5. This will be followed, for each case, by a description of the intraoperative cortical stimulation mapping results, and a quantitative analysis of the spatial relationship between the non-invasive functional MRI results and the invasive gold standard.

6.1 Methodological Details

For all patient cases, both the ROC and linear regression reproducibility analyses are performed on each available test-retest dataset. The pROCreate algorithm is used to set automated, case-specific threshold levels, using a derivative cutoff of 0.01. The single condition contrasts to rest condition are used unless otherwise specified (i.e. $C_1 > 0$ or $C_2 > 0$), as these are more reproducible than the contrast C_1-$C_2 > 0$, and contain fewer voxels than the contrast C_1+$C_2 > 0$. The three-dimensional renders of the functional MRI presented in the ensuing text displaying any fMRI activation within approximately 10 mm of the surface. Cortical stimulation procedures were observed in the operating room. The results of these sessions are reported, along with the Euclidean distances to the nearest fMRI activations.

6.2 Patient 1

Patient 1 is a 35-year-old, strongly right hand dominant female (laterality quotient 80). She presented with headaches, nausea, and vomiting. Clinical imaging revealed

a left frontal tumor, suspected to be a low-grade glioma or astrocytoma (figure 6.1). The planned surgical approach would include brain regions adjacent to the suspected region of Broca's area, in order to perform a full resection of the tumor. The patient was asked to perform both a right hand finger-tapping task, and the overt object-naming task for presurgical functional MRI. The same overt naming task was used in the operating room during cortical stimulation to test for language.

6.2.1 Functional MRI

Both the object-naming and finger-tapping tasks produced widespread activation, for even the strictest traditional threshold levels. The test-retest linear correlation coefficients (slopes) were 0.722 (0.698) and 0.647 (0.532) for the finger-tapping and object-naming tasks respectively. The average AUC for the first (second) image as the template was 0.915 (0.925) for the motor task and 0.834 (0.813) for the language task, indicating that the motor task was more reproducible.

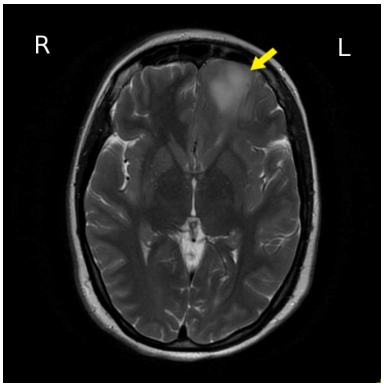
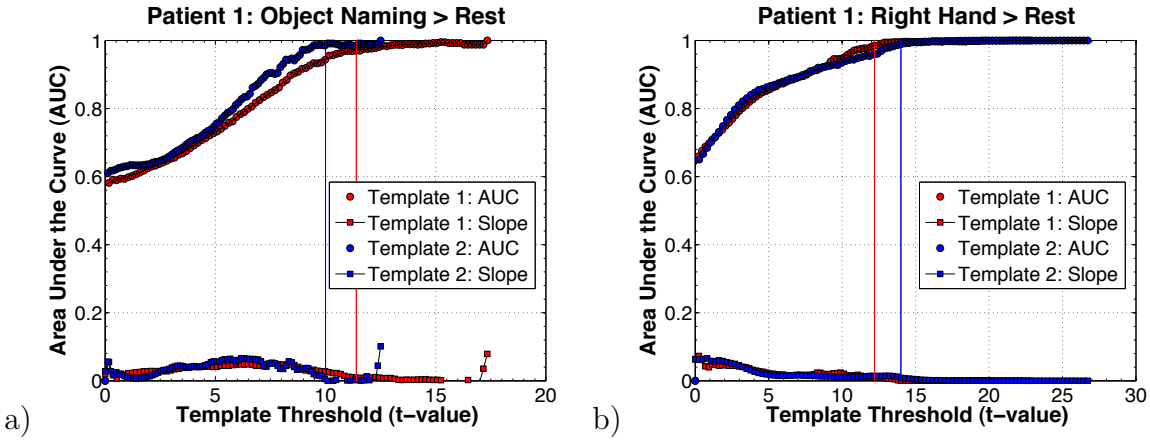


Figure 6.1: Patient 1: T_2 weighted anatomical image showing left inferior frontal tumor (arrow).

The AUC profile and template thresholds are shown for the object-naming and finger tapping tasks in figure 6.2. Application of the automated threshold routine produced first (second) retest image thresholds of 7.13 (7.01) for the motor task, and 7.71 (4.42) for the language task. At these thresholds there were 1283 (1300) active voxels in the motor task and 792 (1033) voxels in the language task.

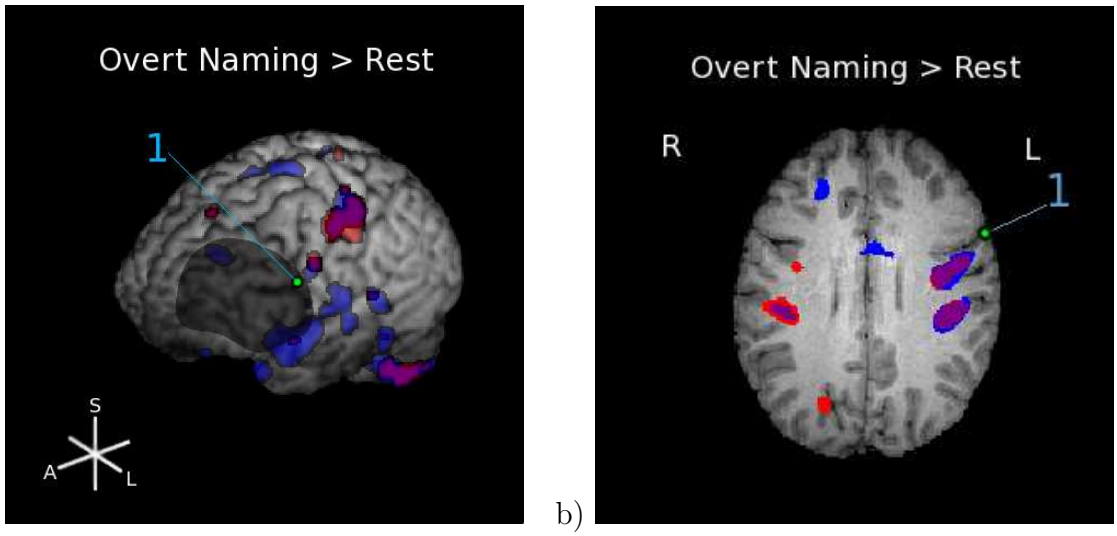
6.2.2 Cortical Stimulation

Both the overt naming and right hand motor tasks were imported to the neuronavigation. Cortical stimulation produced a single distinct location of speech arrest and speech errors at 4 mA, and further investigation at this current level produced a convulsive seizure. At a current of 8 mA, a second seizure was observed, at which point the surgeon ceased cortical stimulation. Biopsy revealed the tumor to be an anaplastic oligoastrocytoma (WHO grade 3).



a) b)

Figure 6.2: Patient 1: AUC profiles. a) Overt naming contrast to rest. b) Finger tapping contrasted to rest. Vertical lines represent the AUC cutoff points for image 1 (red) and image 2 (blue).



a) b)

Figure 6.3: Patient 1: a) 3D render and b) 2D slice depicting the fMRI and cortical stimulation results. The shaded region represents the exposed cortex. Green label (1) is the CS site causing speech arrest. The fMRI results for the overt naming task are in red (image 1) and blue (image 2). Regions of overlap of the two test-retest fMRI images are shown in purple.

6.2.3 Quantitative Comparison

The functional MRI and cortical stimulation results are shown overlaid on the anatomical image collected during the presurgical evaluation in figure 6.3. Only the overt object-naming task is used for comparison, as the sole intra-operative testing result affected speech production. At the thresholds set by the pROCreate algorithm, the

distance from the positive cortical stimulation site to the first and second trial of the overt naming task is 10.6 mm and 8.4 mm respectively. In both cases, this fMRI activation site was on the same gyrus as the cortical stimulation point, approximately 1 cm deep (figure 6.3).

6.2.4 Discussion

One useful aspect of test-retest imaging is the potential to discriminate between reproducible activation clusters and spurious activations. In this case, a large active region near the inferior border of the craniotomy was present in only one of the two images, and cortical stimulation produced no result. On the other hand, a region just outside of the craniotomy with a small surface projection, but active on both images, seems to be related to the CS speech arrest site.

6.3 Patient 2

Patient 2 was 62 years old, male, and left-handed (laterality quotient -80). This patient presented with a generalized convulsive seizure, but no post-ictal functional deficits. Patient 2 had a contrast-enhancing tumor in the left frontal lobe, anterior to the precentral gyrus, and superior to the presumed expressive speech area (figure 6.4). The tumor was characterized as a possible low-grade glioma, with some calcification present. The patient was seen for a presurgical fMRI session the day before surgery, at which time both overt object-naming and right-hand finger tapping were performed.

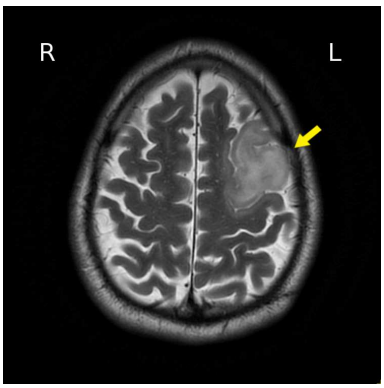


Figure 6.4: Patient 2: T_2 weighted anatomical showing the low-grade glioma in the left frontal lobe.

6.3.1 Functional MRI

Patient 2 demonstrated highly reproducible fMRI activity, with linear correlation coefficients (slopes) of 0.732 (0.643) and 0.822 (0.860)

for the language and motor task respectively. The average AUC for the first (second) image as template was 0.948 (0.951) for the right hand finger tapping contrasted to rest, and 0.888 (0.886) for the overt naming condition contrasted to rest.

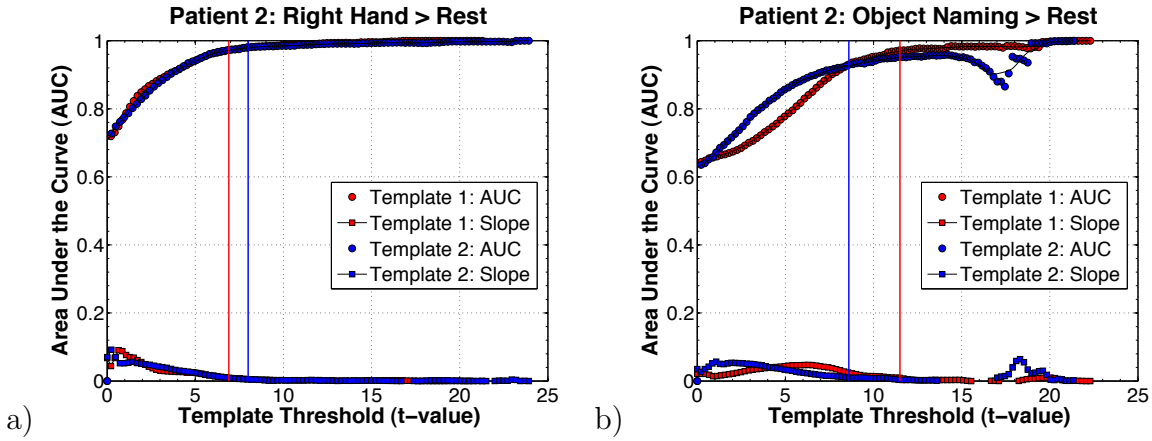


Figure 6.5: Patient 2: AUC profiles. a) Right finger tapping contrast to rest. The AUC cutoffs are nearly identical for both image templates. b) Object-naming contrast against rest.

The pROCreate threshold algorithm was used to determine test-retest thresholds at the d=0.01 cutoff (figure 6.5), and resulted in t-value thresholds of 6.98 (6.03) for the first (second) image of the overt naming task, and 4.73 (4.83) for the first (second) image of the finger tapping test-retest pair. At these thresholds, 3043 (2644) and 3529 (3199) voxels were included in the first (second) image of the language and motor tasks respectively.

6.3.2 Cortical Stimulation

Cortical stimulation of the left hemisphere produced motor responses at four locations of the exposed cortex, as shown in figure 6.6. The first location produced twitches in the right hand at currents as low as 7 mA. Locations two through four produced right upper extremity motor responses at current levels of 8 mA or greater. Cortical stimulation site number 2 produced the most consistent response, resulting in flexion at the right wrist. The craniotomy did not expose the inferior frontal gyrus, and no language effects were observed. Biopsy revealed the tumor to be a grade 2 oligodendroglioma.

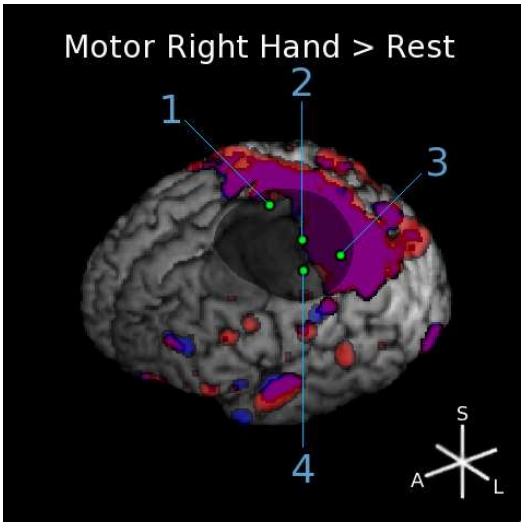


Figure 6.6: Patient 2: 3D render of the functional mapping results. CS labels (green) 1, 3, and 4 reliably caused subtle hand twitches. Point 2 caused more pronounced wrist flexion. The red overlay is fMRI image 1, blue image 2, and purple is the overlap, for the right hand tapping task. The shaded region represents the exposed cortex.

6.3.3 Quantitative Comparison

The functional MRI results for right hand finger tapping are shown in relation to the cortical stimulation results in figure 6.9. The distance from the cortical stimulation points to the nearest fMRI activation in the first (second) image of the finger-tapping test-retest pair was: 13.9 (13.9) mm for CS site 1, 5.9 (9.8) mm for site 2, 5.9 (9.8) mm for site 3, and 12.1 (12.1) mm for site 4.

6.3.4 Discussion

Patient 2 demonstrated highly reproducible finger tapping results, which resulted in widespread activation, as determined by the ROC threshold algorithm. The activation coincided with the location of the cortical stimulation results. Referring to figure 6.6, the functional MRI results appear to coincide more closely than the distances reported above for some of the CS labels. In cases like this, CS points appear to be directly within a region of fMRI activity, but the fMRI activations are deep to the surface. It is possible that the CS and fMRI results are indicating the same cortical location; however, the restriction of CS to surface measurements prevents direct testing of the fMRI activation. This is a fundamental difficulty of comparing cortical

stimulation, which is performed only on the exposed cortical surface, with functional MRI, which is by nature a volumetric measurement.

6.4 Patient 3

Patient 3 was a male 20 years of age, and right hand dominant (laterality quotient 90). The patient experienced seizures that were characterized by loss of consciousness, facial spasms, and what was described as a numb, thick feeling of the tongue. Figure 6.7 shows the lesion discovered in the inferior parietal lobe (postcentral sulcus). This location is consistent with the clinical symptoms and a-priori functional organization of the primary sensory cortex. The lesion was suspected to be a vascular malformation known as a cavernous angioma, and the surgical approach would expose parts of the sensory/motor cortex, as well as the receptive language zone. The patient was given three tasks to complete: right handed finger tapping, tongue movement, and sentence/math comprehension.

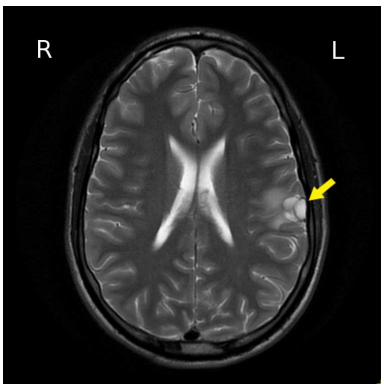


Figure 6.7: Patient 3: T_2 weighted image of left parietal cavernous angioma.

6.4.1 Functional MRI

The test-retest linear regression analysis produced correlation coefficients (slopes) of 0.650 (0.632) for the tongue movement task, 0.782 (0.875) for the finger movement task, and 0.788 (0.708) for the language task. The corresponding average AUC values for the first (second) image acting as the activation template were 0.848 (0.874), 0.880 (0.878) and 0.915 (0.906) for the respective tasks (figure 6.8). Both the regression coefficients and the average areas under the curve are above the group average observed in healthy controls.

The pROCreate algorithm resulted in t-value thresholds for the first (second) image of the finger-tapping task of 6.31 (6.31). The thresholds for the tongue movement

task were 4.82 and 6.61 for the first and second image of the test-retest pair respectively. The sentences versus rest condition resulted in automated thresholds of 6.83 and 5.53 for the two images. At these thresholds, the number of active voxels in the first (second) image was 915 (1162) for the right hand motor task, 386 (336) for the tongue movement task, and 2004 (1854) for the sentence comprehension task.

6.4.2 Cortical Stimulation

Despite testing of the pre- and post-central gyri, cortical stimulation did not produce any functional response up to the maximum current of 10 mA. The testing was repeated at 5 and 50 Hz stimulation frequencies (i.e. for language and motor investigations), to no avail. Therefore no quantitative comparison was possible.

6.4.3 Discussion

This case included relatively well-reproduced functional MRI activation that spanned the superior to inferior-anterior regions of the exposed cortical surface. The activation results follow typical functional organization patterns, with finger movements eliciting superior central region activation, tongue movement producing inferior central region results, and sentence comprehension producing inferior frontal and posterior temporal lobe activation (as well as superior central region because button-pressing is used for making responses). However, none of these could be substantiated by cortical stimulation.

This case illustrates the typical difference between functional MRI and cortical stimulation sensitivity. Functional MRI often identifies many cortical regions correlated with a given functional task, however often CS will corroborate only very few of these. This does not render these fMRI activations invalid, but rather indicates that the two techniques are different in both the physical and physiological principles involved. In this case in particular, the functional MRI results were very reproducible, and used automated thresholds as high or higher than the traditionally restrictive Bonferroni method, thus suggesting a high degree of reliability of the presurgical mapping results. It is unknown why no effects of cortical stimulation were observed, but it may reflect that other cortical regions were able to compensate for some of the intra-operative tasks, although this does not explain the absence of a sensory response

from the post-central gyrus.

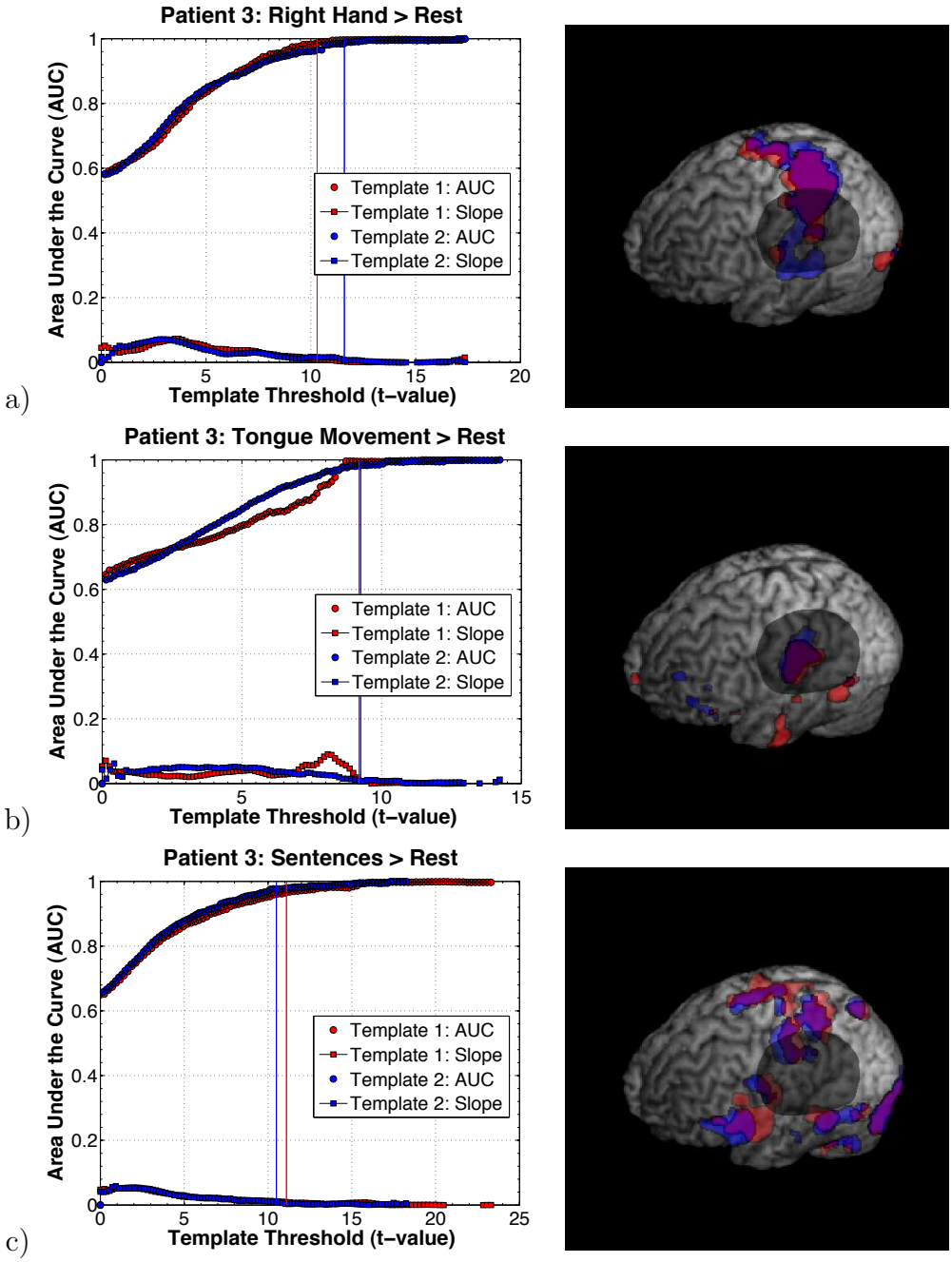


Figure 6.8: Patient 3: The AUC profiles for each of the three administered tasks, and corresponding 3D renders are shown: a) right hand tapping, b) tongue movement and c) sentence comprehension. The fMRI results (red - image 1, blue - image 2, purple - overlap) closely reflect the *a priori* cortical organization. No cortical stimulation results were obtained. The shaded region is the exposed cortex.

6.5 Patient 4

Patient 4 was a 24-year-old female, and reported ambidexterity (laterality quotient 40). The patient presented with generalized seizures, and memory difficulties that were affecting daily life. A suspected high-grade glioma was discovered in the medial temporal lobe of the left hemisphere (figure 6.9). The patient was brought in for a presurgical functional MRI, at which time the overt object-naming and sentence/math comprehension tasks were carried out a test-retest fashion.

6.5.1 Functional MRI

Object naming contrasted to rest resulted in a test-retest linear correlation coefficient (slope) of 0.644 (0.968). The sentences condition contrasted to rest produced a linear regression correlation coefficient and slope of 0.715 and 0.344 respectively. The average AUC of the first (second) template was 0.882 (0.800) for the object naming task (figure 6.13) and 0.913 (0.932) for the sentence comprehension task (figure 6.10).

Figure 6.9: Patient 4: T_2 weighted MR image showing the tumor in the left temporal lobe, believed to be a high-grade glioma.

It is observed that the AUC profiles for the two different templates of the object-naming task are quite disparate. In particular, there is an early plateau in the AUC of the second template. It was hypothesized that this plateau was caused a cluster of activation in the medial frontal lobe that was characteristic of motion artifact. This cluster was prevalent only the second image. Inclusion of the motion parameters as regressors decreased the average AUC values to 0.812 (0.748), and so this tactic was not employed. Instead, the region was manually masked in both images, reducing the number of tested brain voxels by approximately 2000 out of 20000. The pROCreate algorithm was run again, resulting in an average AUC of the first (second) template of 0.884 (0.827). It can be seen in figure 6.11 that the early plateau region is no longer present, and the template image threshold has shifted upwards. The threshold

determined for the first (second) image of the overt naming task was 3.86 (6.26) and for the sentence comprehension task 9.41 (3.00). The statistical maps produced had 1981 (1417) and 1819 (2744) active voxels respectively.

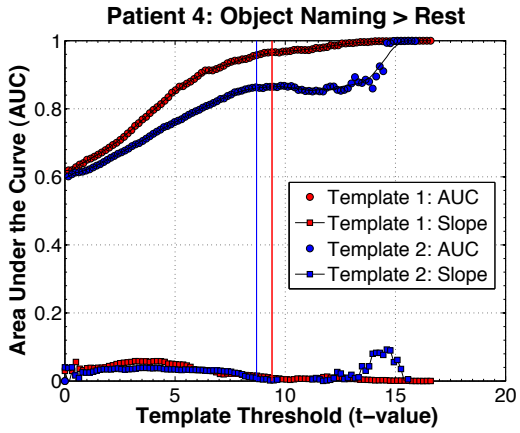


Figure 6.10: Patient 4: AUC profile for the overt naming task before removal of the region of motion artifact. The AUC cutoff for the second image is applied early because of the false plateau.

6.5.2 Cortical Stimulation

Cortical stimulation was carried out both for investigation of language and sensory/motor representations in the region of the craniotomy. A variety of effects were observed at both 5 Hz and 50 Hz (figure 6.11). At 5 Hz, two locations causing speech impairment were identified using a current of 8 mA (sites 1 and 2). These impairments ranged from slurred and distorted speech (dysphasia - site 1) to full speech arrest (aphasia - site 2). At this same current, using 50 Hz stimulation, another nearby location causing speech arrest was identified (site 3). Additionally, stimulation at this frequency and current level caused sensation in the right pharyngeal region when site 4 was stimulated. Increasing the stimulation current to 10 mA produced counting errors when sites 5 and 6 were tested.

6.5.3 Quantitative Comparison

The cortical stimulation results are shown in relation to the overt object-naming and sentence comprehension results in figure 6.11. The distances from each of the 6 CS

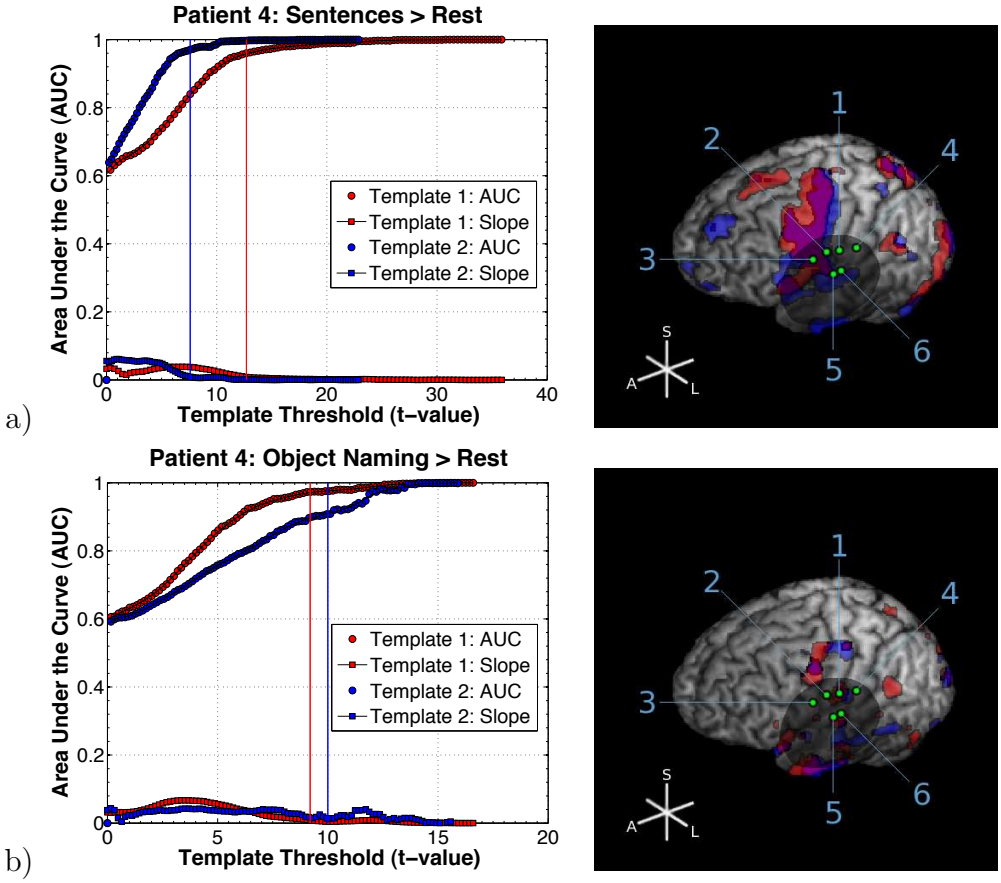


Figure 6.11: Patient 4: The final AUC profiles for a) the sentence and b) the object-naming tasks, and the corresponding rendered activation maps (red - image 1, blue - image 2, purple - overlap). The six labeled CS sites (green) produced the following effects: (1) dysphasia (2) aphasia (3) aphasia (4) pharyngeal sensation (5) number counting errors (6) number counting errors. The shaded region represents the approximate area of exposed cortex.

labels to the nearest fMRI activation on the first (second) image of the overt-naming task are: 5.3 (3.8) mm, 3.8 (7.5) mm, 3.8 (15.0) mm, 3.8 (5.9) mm, 5.9 (7.5) mm, and 5.9 (5.3) mm. The distances to the first (second) image of the sentence comprehension test-retest pair are: 14.3 (5.3) mm, 8.7 (0.0) mm, 7.0 (0.0 mm), 19.1 (12.7) mm, 8.7 (4.5) mm, 11.3 (7.0) mm.

6.5.4 Discussion

The motion artifact in the frontal lobe was resistant to the use of motion correction parameters as regressors of no interest in the fMRI model. It is possible that this

artifact was caused by one or two motion events, in which case it would not be well modeled by the full time-series of motion correction parameters. A more sophisticated approach, such as an independent component analysis (ICA), may be more successful. This was not deemed necessary as the region was well outside of the planned craniotomy, and recovery of the AUC profile was observed with the simple masking procedure used. Drawbacks of both the masking and ICA procedures are that they require manual intervention, increasing analysis time and subjectivity.

The thresholds applied to the sentence task are quite different from first to second image. The high reproducibility and low linear regression slope seem to support this outcome. However, it was observed that this patient had strong activation in the cerebellum, which most certainly affects the output of the pROCreate algorithm. The concept of ROI based ROC analysis has been explored previously, and may be suitable in the present context given the limited window of the craniotomy. Indeed, the aforementioned use of an artifact exclusion mask is an example of this kind of analysis.

The posterior-superior cortical stimulation sites (labels 1 and 4) were more closely matched by the overt language task than the covert language task. The most posterior of these elicited a sensory response, while the more anterior caused slurred vocalizations. Combined with the location on the anatomically identified pre- and post-central gyri, these CS sites likely form a part of the primary sensory/motor representation of the speech anatomy. Thus there are task-specific effects that will enhance, or conversely limit, the accuracy of fMRI for presurgical mapping. No single task is suitable for all cases, especially for higher cognitive functions such as language. Employing several presurgical tasks will enhance the predictive capacity of functional MRI mapping under these circumstances.

6.6 Patient 5

Patient 5 is 26 years old, female, and right handed (laterality quotient 100). She presented clinically with a grand mal seizure, which led to the discovery of a right inferior frontal lobe tumor (figure 6.12). The location of the tumor was near the right hemisphere homologue to Brocas area. Presurgical functional mapping was carried out on the 4 Tesla scanner, employing the overt object naming and tongue movement

paradigms.

6.6.1 Functional MRI

Functional MRI results were analyzed for reproducibility by both the linear correlation and ROC methods. The average AUC for the first (second) template was 0.616 (0.444) for the tongue movement task, and 0.926 (0.932) for the overt object-naming task (figure 6.13). The correlation coefficients (slopes) were 0.607 (0.677) for the tongue movement task, and 0.826 (0.651) for the overt naming paradigm. The motion correction parameters for the tongue movement task revealed small, but highly task-correlated head motion. Use of the motion parameters as regressors in this case increased the average AUC for the tongue movement task to 0.636 and 0.751 for the first and second template respectively.

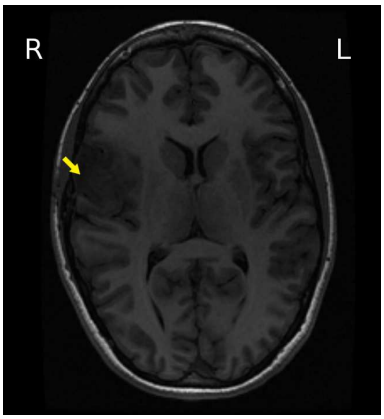


Figure 6.12: Patient 5: T_1 image of right frontal tumor.

The pROCreate algorithm for determining image thresholds was therefore carried out on the tongue movement and object naming maps derived with and without motion parameters as regressors respectively. The t-value thresholds for the first and second image of the tongue movement task were 3.17 and 2.89, resulting in active voxel counts of 532 and 1438 respectively. The thresholds applied to the object-naming task were 4.85 and 3.93 for the first and second image, producing 1925 and 1908 voxels respectively.

6.6.2 Cortical Stimulation

Cortical stimulation was performed using both 5 Hz and 50 Hz stimulation frequencies. No results were obtained at 5 Hz. Number counting errors were produced at two locations using 50 Hz stimulation, and 7 mA current (figure 6.13). Stimulation at these same sites produced speech arrests in reciting the days of the week and months of the year. The patient reported that she was unable to make the words. Continued investigation produced a seizure, and cortical stimulation was concluded. Biopsy revealed the tumor to be of the mixed oligoastrocytoma type, WHO grade 2.

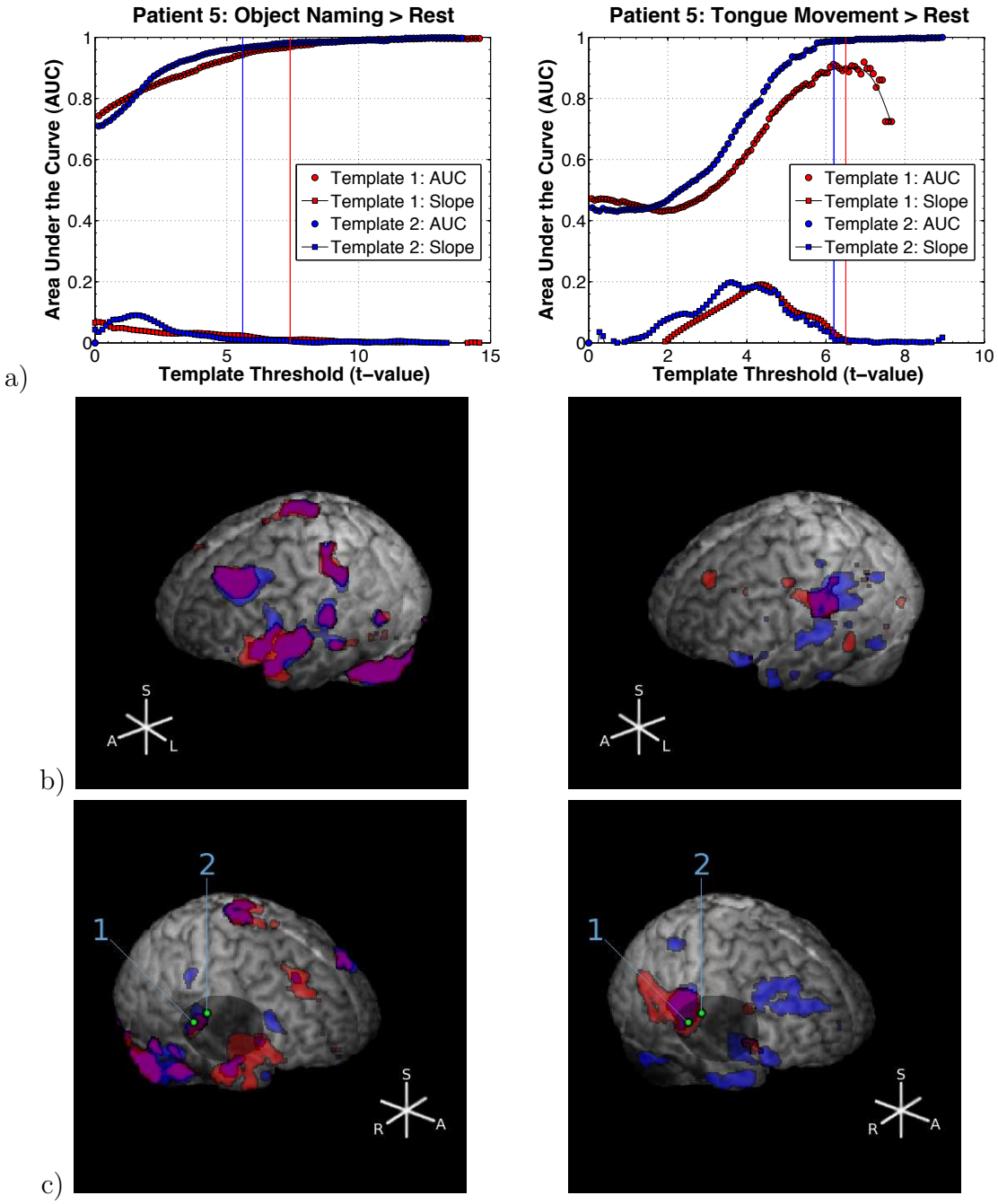


Figure 6.13: Patient 5: Functional MRI and CS results. The left column is from the object naming task, and the right column from the tongue movement task. Top: AUC profiles and automated thresholds for the two fMRI tasks employed. Middle: Left hemisphere activation patterns after application of the pROCreate algorithm. Bottom: Right hemisphere activation (red image 1, blue image 2, purple overlap), cortical stimulation results (green), and exposed cortical region (shaded). CS labels (1) and (2) both produced speech arrest.

6.6.3 Quantitative Comparison

Both the tongue movement and object-naming activation maps were compared to the cortical stimulation sites recorder intra-operatively. The first (second) image of the tongue movement task contained activation a Euclidean distance of 3.8 (5.3) mm from the first labeled CS location, and 9.5 (11.5) mm from the second. The first (second) image of the object-naming task contained activation 3.8 (5.3) mm and 9.5 (10.6) mm from the two CS locations.

6.6.4 Discussion

Contrary to the case of patient 4, in this example the inclusion of motion correction parameters in the baseline model for fMRI analysis did a good job of reducing image artifact. Although the low average AUC scores for the tongue movement task persisted after the use of the motion regressors, the AUC profiles do exhibit the desired rapid increase and plateau features. The average AUC is suppressed by the early part of the curve for both templates, where the AUC is below 0.5, and in the case of the first template, by the drop in AUC near the upper end of the range. It is believed that these two characteristics are both caused by remaining motion artifact, which, in the case of the first image template, appears to persist at high threshold levels. The thresholds produced by applying the pROCreate method to this test-retest pair were quite low, and produced widely differing activation extents. This is indicative of the difficulty the algorithm has in dealing with poor quality images. By contrast, the object naming test-retest pair is among the more reproducible datasets studied, and the near identity of the number of voxels declared active in the test-retest pair is characteristic.

For both functional tasks used in this case, the nearest cluster to the CS sites was bridging the inferior portion of the central sulcus, the a-priori location of the tongue and mouth representation on the sensory-motor strip. This cluster was much larger for the tongue movement task than the object-naming task. There are several explanations for this; one being that the object-naming task requires less frequent movement of the speech anatomy. Another is that the object-naming task employs a more widespread cortical network. Likely both of these factors contribute. Despite both the larger extent of this cluster on the tongue movement task and the lower

reproducibility of the test-retest pair, the two paradigms performed nearly identically when compared to cortical stimulation. Cortical stimulation of the inferior right frontal lobe produced no language effects, despite the presence of an fMRI activation cluster elicited by overt naming in the region (Brocas homologue). The negative result of CS was no surprising because to the larger, more reproducible representation of Brocas area in the left hemisphere (figure 6.13).

6.7 Patient 6

Patient 6 is a 45-year-old right-handed woman (laterality quotient 100). The patient presented with general drowsiness, nausea, and headaches. The headaches would most frequently occur whilst reading. Imaging revealed a large left medial temporal tumor, suspected to be a high-grade glioma (figure 6.14). The prepared task battery included finger tapping, object naming, and sentence/math comprehension.

6.7.1 Functional MRI

The patient began to suffer from a strong headache and neck pain during the fMRI scan, and so the session was cut short. Thus only the sentence comprehension task was completed in test-retest fashion, and will be the focus of this discussion. The two images of the sentences contrasted to rest condition produced a test-retest correlation coefficient (slope) of 0.728 (0.741). The average AUC of the first (second) image was 0.719 (0.766). The two images of the mathematics contrasted to rest condition had a test-retest correlation (slope) of 0.745 (0.686), and a first (second) image template average AUC of 0.825 (0.845). In this case, the conditional contrast of mathematics minus sentences was found to be most reproducible by the ROC method, with average AUC of 0.839 (0.877) for the first (second) template, but least reproducible by the linear regression analysis (correlation coefficient and slope of 0.614 and 0.695 respectively).

Figure 6.14: Patient 6: T_2 weighted anatomical MRI displaying the left temporal lobe tumor.

The pROCreate algorithm produced retest image thresholds for the first (second) image of 3.95 (3.07) for the sentences contrast, 6.12 (4.23) for the mathematics contrast, and 2.94 (2.76) for the math minus sentences contrast (figure 6.15). The resulting first (second) statistical maps had 1670 (1053), 1467 (1193), and 1350 (1262) active voxels for the three respective contrasts. In the interest of isolating the mathematics greater than sentences contrast, the mathematics minus sentences statistical map was masked for voxels greater than or equal to zero.

6.7.2 Cortical Stimulation

Cortical stimulation was carried out at 50 Hz, using both number counting, and the object-naming task. At 9 mA current, stimulation produced speech interruption at a single location for both language tasks (figure 6.15).

6.7.3 Quantitative Comparison

The three contrasts generated all resulted in significant activation as determined by the pROCreate algorithm. The sentences-to-rest contrast was a minimum distance of 4.5 (5.9) mm from the CS location for the first (second) image of the test-retest pair. The math-to-rest and math-to-sentences contrasts were 4.5 (4.5) mm and 3.8 (3.8) mm from the CS site respectively for the first (second) image.

6.7.4 Discussion

For this case, the two measures of reproducibility are not in good agreement. For the sentences to rest condition contrast, the AUC was low, while the linear correlation was high; in the mathematics contrasted to sentences condition, the converse was true. The sentences contrast to rest exhibited strange behavior for the AUC of the first image template. Above template thresholds of approximately 9.5, the AUC suddenly drops. There was no pronounced motion correction needed for either replication of the task, so motion artifact is not likely the cause. For the math contrasted to sentences condition, the low linear regression coefficient may be due to the fact that this contrast contained many voxels with near zero t-value, biasing the linear regression towards lower correlation values.

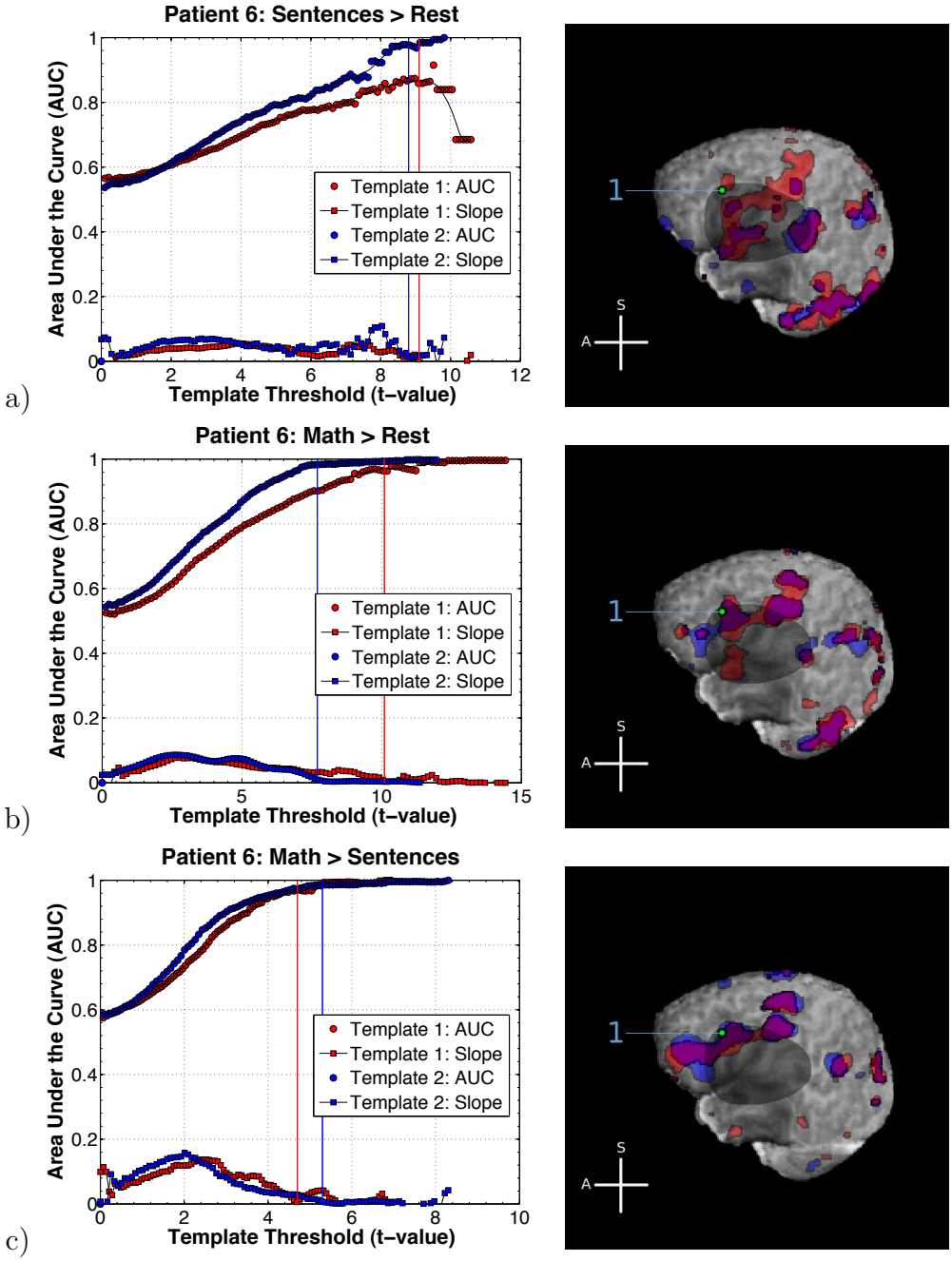


Figure 6.15: Patient 6: AUC profiles (left) and 3D renders (right) for the sentence comprehension task performed by patient 6. Three different contrasts are shown: a) C_1 >0, b) C_2 >0, and c) C_2-C_1 >0. Contrary to the trends discussed in chapter 4, the contrast math>sentences performed best on the ROC analysis, and was the most specific to CS in this case as well. The fMRI overlays are shown in red (image 1), blue (image 2), and purple (overlap). The green point is the speech arrest CS location, and the shaded region is the exposed cortex.

While little difference is seen in the accuracy of the three contrasts when compared to cortical stimulation, the math greater than sentences contrast has far fewer active voxels in the region of the craniotomy. Contrasts between multiple active conditions may thus be more specific than active conditions contrasted to rest. In general one must be weary of sacrificed test-retest reproducibility of these contrasts (see chapter 4), although as demonstrated in this patient, this is not always the case. It is unknown why this region would cause speech arrest and yet be more strongly activated by mathematics than sentence comprehension. The object-naming task would possibly have provided further evidence on the nature of this region, but as mentioned before, the fMRI session had to be cut short due to patient discomfort.

6.8 Conclusion

The usefulness of ROC analysis in the application to presurgical mapping has been demonstrated in this chapter. The novel component of ROC analysis for test-retest fMRI images in the pROCreate method is the generation of AUC profile curves, which allow detailed inspection of data quality. Certain image artifacts are apparent on such profiles, and processing options for dealing with these artifacts can be compared quantitatively using the average AUC metric. Additionally, the automated threshold procedure by combined asymptote detection (from the AUC profile) and optimal point threshold (from the corresponding ROC curve) has been demonstrated to perform well at predicting the location of critical functional zones. On average, the nearest fMRI result to each cortical stimulation site was (mean $\pm$ standard error) 5.8 ± 1.1 mm by the pROCreate threshold method (table 6.1).

Figure 6.16 shows the performance of the pROCreate threshold method compared with the other threshold techniques in terms of mean distance to cortical stimulation. Both the case of the best matching fMRI image (i.e. the data from table 6.1) and all fMRI images of similar function to the CS effect are considered. In both cases, the pROCreate threshold method is approximately equal in accuracy to the FDR correction with c(V) term included. The Bonferroni method is the least accurate by this measure, and the FDR without c(V) term is the most accurate. However, these results should be interpreted carefully, as lower thresholds will naturally result in smaller CS to fMRI distances, at the expense of specificity. In this context, the

pROCreate method performs particularly favorably, as it achieves the same average accuracy as the FDR with c(V) threshold scheme, but on average includes fewer active voxels.

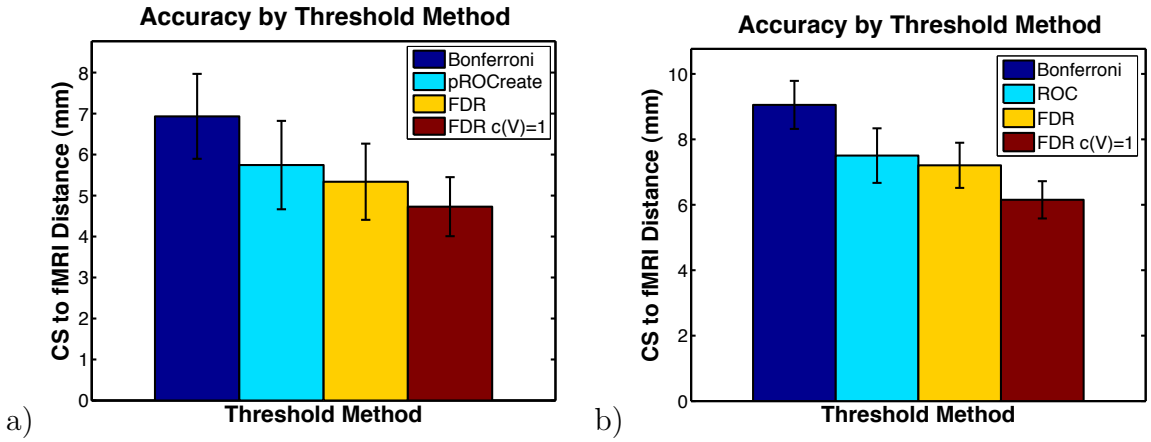


Figure 6.16: Mean Euclidean distance between cortical stimulation data points and the nearest fMRI activation by threshold method. a) Only the nearest fMRI image is used in the average. b) All fMRI images of similar function to the CS effects are included in the average.

There are limitations to the accuracy of calculating the Euclidean distance from fMRI to CS. Brain shift during surgery is likely a possible source of error in this

Table 6.1: List of cortical stimulation sites, the nearest suprathreshold fMRI voxel according to the pROCreate method, and the distance between them.

Patient	CS Site	Effect	Distance (mm)	fMRI Task
1	1	Aphasia	8.4	Overt naming
2	1	Hand twitch	13.9	Finger tapping
2	2	Wrist flexion	5.9	Finger tapping
2	3	Hand twitch	5.9	Finger tapping
2	4	Hand twitch	12.1	Finger tapping
4	1	Dysphasia	3.8	Overt naming
4	2	Aphasia	0	Sentences
4	3	Aphasia	0	Sentences
4	4	Pharynx sensation	3.8	Overt naming
4	5	Counting errors	4.5	Sentences
4	6	Counting errors	5.3	Overt naming
5	1	Aphasia	3.8	Overt naming
5	2	Aphasia	9.5	Overt naming
6	1	Aphasia	3.8	Math>Sentences

calculation, most significantly in the left-right direction as the patient is positioned on their side, making this the direction of gravitational force. Making the craniotomy and opening the dura can bring about cortical shifts of several millimeters, as the soft brain tissue responds to changing pressures. This will in general cause either an increase in the distance to fMRI if the shift is outwards (for example due to mass effect of a large tumor), or decrease the distance, if the shift is inwards (for instance due to loss of CSF). There exists the possibility of performing intra-operative MRI to obtain anatomical images during the operation, and to use the registration of these to presurgical images to improve the accuracy of the CS data registration.

There are additional sources of error from the TREON coordinate frame and wand, and registration of the image guidance MRI to the presurgical MRI. These are likely less than 1mm due to the accuracy of modern coregistration techniques. The low resolution of functional MRI means that the comparison is made on a coarse grid to begin with, and so such small errors in the localization of CS are unlikely to impact the distance calculations.

The use of fMRI for presurgical mapping is not without caveats. Firstly, the data must be inspected carefully for the presence of image degrading artifacts, especially if considerable patient motion is detected in the motion correction process. Secondly, the results of fMRI mapping are highly sensitive, but not specific, when measured against cortical stimulation. This suggests that fMRI may in its current state be best suited in an augmentative role to cortical stimulation. In particular, depending on fMRI alone for determining resection boundaries is not considered safe, as the spatial extent of the active regions depends strongly on image thresholds. While the pROCreate algorithm attempts to alleviate this problem, performance is still case specific.

Additionally, dependence on highly sensitive fMRI mapping would lead to conservative resection strategies, which may be disadvantageous from the standpoint of disease control. However, there is little doubt that the fMRI mapping process adds valuable information to the presurgical workup that is not available by the invasive methods. For example, when considering alternative access pathways to a surgery, the knowledge provided by fMRI mapping could allow the surgeon to avoid potential complications. Additionally, the knowledge of fMRI results could expedite

intra-operative mapping, by providing the surgeon with knowledge of the potentially viable, and unadvisable surgical routes.

The comparisons with cortical stimulation illuminate the need for multiple functional tasks to be used in each case, especially when higher cognitive functions are involved. Clear differences have been observed in the results of different language tasks (e.g. object naming and sentence comprehension in case 4), or in the different contrasts of a single task (e.g. the math>rest and math>sentences in case 6). Of course the tradeoff is that more tasks take more time, and doing each twice to use the ROC analysis even more so. However, this is already a considerable improvement over the previous methods of using ROC curves for determining image thresholds, which required at least 3 replicates of each task. Potential savings in time could be achieved by shortening the duration of each functional task. The ROC analysis could be used to perform an examination of the test-retest reproducibility as a function of task duration, to find the optimal trade-off between scan time and image quality.

Chapter 7

Conclusions

The primary goal of this work was to develop a standardized fMRI processing pipeline to ensure reliability and accuracy. To this end, a novel method of assessing fMRI reproducibility and setting activation thresholds has been presented, based on the area under ROC curves derived from test-retest sensitivity-specificity analysis. The method developed is easy to implement, and requires less extension of the scanning session compared to existing methods for determining image thresholds from test-retest ROC curves [19]. The performance of this method was evaluated on a database of healthy control test-retest imaging collected using multiple fMRI tasks, scanners, and individuals, as well as in a sample of brain tumor patients undergoing surgical intervention. In the latter group, this technique was compared with the gold standard in functional mapping, demonstrating the accuracy of the fMRI images produced.

7.1 Test-retest Reproducibility

Reproducibility of the fMRI mapping in healthy controls was generally high, although slightly less than that reported by Fernandez et al [53], who found a mean linear regression coefficient for a language task of 0.737, with a range from 0.608 to 0.890. The most prevalent factors affecting reproducibility were choice of image contrast, and variation between subjects. The subject with the least reproducible test-retest datasets also moved the most during the scanning sessions. Analysis of the raw images and statistical maps showed the impact of motion on image reproducibility. Many other factors could contribute to the inter-subject variability observed, such as those related to task performance (e.g. attention, ability), physiology (e.g. neurovascular coupling), and physical limitations (e.g. susceptibility boundaries). The high degree of inter-subject variability indicates the need to control reproducibility at the individual level, rather than the group level, when implementing a presurgical mapping protocol.

The contrasts comparing task conditions to rest were generally more reproducible than the conditions comparing task conditions to one another. Intuitively this makes sense, as the motivation behind making contrasts between tasks is to elucidate the subtle differences between the task requirements, trading sensitivity for specificity. Given that the task conditions used are typically quite similar in many regards, they recruit much of the same cortical networks to varying degrees. The metabolic demand in these regions, and therefore the T_2^* decay changes measured by the BOLD technique, will be less different than when compared to rest. Many of the sources of noise from both the scanner and physiology will be similar under all conditions, and thus detection sensitivity is compromised. An additional caveat is that motion during either task condition will affect the contrast of the two conditions, whereas the non-contaminated condition may still successfully be contrasted to rest. Special care should be taken to examine the test-retest reproducibility when using contrasts between active conditions for presurgical mapping, to ensure that the activation maps are reliable.

7.2 Activation Thresholds and Accuracy

It was shown that test-retest AUC depends on the significance threshold applied to the functional images, and that this could be exploited to set automated, data-driven significance thresholds. These thresholds are implicitly related to detecting stable patterns of activation rather than using a priori levels of statistical significance, as evidenced by the equal number of active voxels on the test and retest images. In the context of the sources of variation in fMRI data identified in section 2.5.2, it should be pointed out that threshold selection on the basis of reproducibility is most appropriate when random noise dominates. As has been demonstrated, the presence of structured noise sources can cause suppression of reproducibility indices which will affect the data-driven threshold levels. The worst case scenario is perhaps systematic error, as the noise source will appear as true activation according to the test-retest criteria, and reproducibility optimization will be biased towards inclusion in the thresholded image.

The automated threshold algorithm produced different threshold levels for data from the different field strengths, resulting in less difference in the number of active

voxels between the two scanners than traditional threshold methods. This may be more suitable for individual level functional imaging than the existing statistically driven methods, as the underlying neural activation patterns should not change from site to site [50], whereas statistical significance values obtained generally do. By contrast, when applied to different tasks the ROC method was the only threshold method studied to produce a different number of active voxels. This is consistent with the understood differences in the cortical network engaged [4, 64].

When compared to cortical stimulation, the fMRI maps produced by the ROC method were accurate to within less than 6 mm on average, although in some cases the discrepancy is as great as 14 mm. This is similar to the results of earlier studies [14, 16, 21, 22, 27], and supports the currently accepted notion that fMRI boundaries are not sufficiently accurate for determining safe resection limits [4, 30]. However, this level of accuracy is sufficient for directing intra-operative examinations, and for informing choices between possible alternatives for surgical approach. Accuracy will likely depend on the function under investigation [14], imaging hardware [54], software strategies [14, 15], and patient compliance [4].

7.3 Future Directions

There are many avenues available for continued investigation of the pROCreate algorithm. Modern threshold techniques include a variety of cluster-based criteria for determining activation patterns [Logan, 2004], as well as dynamic thresholding [Stippich, 2007]. The characteristics of these methods should be compared to the ROC method, and a more complete analysis of the images produced, including lateralization and cluster size distribution should be performed. While previous within versus between site test-retest reproducibility analyses have shown that within site data is more consistent [54], it would be interesting to see the performance of ROC data-driven threshold algorithm applied to data from the same individual and task but different scanners. There are many modulating factors that affect the imaging results between sites (e.g. pulse sequences, head coils). Some of these may affect images in ways more subtle than reflected by the image wide reproducibility, significance threshold, and number of active voxels measures used in this study (e.g. local sensitivity changes near susceptibility boundaries, characteristic artifacts).

It has been shown in this work that the optimal threshold need not be identical for both the test and retest image. In fact, it is not necessarily true that the reproducibility, and therefore optimal threshold, is identical in different regions of the brain [48, 51]. Different parts of a neural network may exhibit different degrees of activation, and yet be reliably identified. A possible extension of the ROC threshold algorithm is to set regional significance thresholds, by running the algorithm several times using different masks. For instance, there are standard masks available to define the different lobes of the brain, and these could be used to define ROIs.

The pROCreate algorithm could potentially be implemented without the need for test-retest imaging, by splitting a single acquisition in half, and using the two halves as a simulated test-retest set [53]. This idea is particularly attractive, as it could be implemented in a real-time analysis package, with the benefit of assuring suitable data quality at time of acquisition [4]. Such an extension of the pROCreate software could be demonstrated offline in a simulation of the real-time acquisition process, without the need for real-time fMRI console software.

In conclusion, the pROCreate software allows quick and quantitative inspection of reproducibility, and can easily be adapted to any imaging or parametric mapping modality that uses a decision threshold for interpreting results. This reproducibility assessment could be used as part of a quality assurance routine, ensuring reliability of test results at the individual level. Other potential uses include comparing the performance of data processing strategies or functional task paradigms, optimization of acquisition parameters, and piloting of study protocols. In the context of presurgical fMRI mapping, the pROCreate algorithm provides a valuable tool, as it determines automated image thresholds according to a universal criteria, which adapt on a case-by-case basis to detect reliable activation patterns. Such software could be valuable in obtaining more widespread acceptance of fMRI in the clinical setting.

Bibliography

[1] Y. Enchev. Neuronavigation: Geneology, reality, and prospects. *Neurosurg. Focus*, 27(3E11):1–18, 2009.

[2] A.R. Asthagiri, N. Pouratian, J. Sherman, G. Ahmed, and M.E. Shaffrey. Advances in brain tumor surgery. *Neurol. Clin.*, 25:975–1003, 2007.

[3] S. Tharin and A. Golby. Functional brain mapping and its applications to neurosurgery. *Neurosurgery*, 60(ONS2):185–202, 2007.

[4] Chrisoph Stippich. *Cinical Functional MRI: Presurgical Functional Neuroimaging*. Springer, Heidelberg, Germany, 1st edition, 2007.

[5] W. Penfield and E. Boldrey. Somatic motor and sensory representation in the cerebral cortex of man as studied by electrical stimulation. *Brain*, 60:389–443, 1937.

[6] G. Ojemann, J. Ojemann, E. Lettich, and M. Berger. Cortical language localization in left, dominant hemisphere: An electrical stimulation mapping investigation in 117 patients. *J Neurosurg*, 71:316–326, 1989.

[7] N. Logothetis. The underpinnings of the BOLD functional magnetic resonance imaging signal. *J. Neurosci.*, 23(10):3963–3971, 2003.

[8] H. Duffau. Contribution of cortical and subcortical electrostimulation in brain glioma surgery: Methodological and functional considerations. *Clin. Neurosci.*, 37:373=382, 2007.

[9] W. Penfield and T. Rasmussen. *The Cerebral Cortex of Man.* Macmillan, New York, 1950.

[10] W. Penfield and P. Perot. The brain's record of auditory and visual experience. *Brain*, 86(4):595–696, 1963.

[11] M. Hämäläinen, R. Hari, J. Knuutila, and O. Lounasmaa. Magnetoencephalography - theory, instrumentation, and applications to noninvasive studies of the working human brain. *Rev. Mod. Phys.*, 65(2):413–595, 1993.

[12] C.R. Jack, R.M. Thompson, R.K. Butts, F.W. Sharbrough, P.J. Kelly, D.P. Hanson, F.W. Riederer, R.L. Ehman, N.J. Hangiandreou, and G.D. Cascino. Sensory motor cortex: Correlation of presurgical mapping with functional MR imaging and invasive cortical mapping. *Neuroradiology*, 190:85–92, 1994.

[13] R. Beisteiner, R. Lanzenberger, K. Novak, V. Edward, C. Windischberger, M. Erdler, R. Cunnington, A. Gartus, B. Streibl, E. Moser, T. Czech, and L. Deecke. Improvement of presurgical patient evaluation by generation of functional magnetic resonance risk maps. *Neurosci. Lett.*, 290:13–16, 2000.

[14] G.J.M. Rutten, N.F. Ramsey, P.C. van Rijen, H.J. Noordmans, and C.W.M. val Veelen. Development of a functional magnetic imaging protocol for intraoperative localization of critical temporoparietal language areas. *Ann. Neurol.*, 51:350–360, 2002.

[15] F.E. Roux, K. Boulanouar, J.A. Lotterie, M. Mejdoubi, J.P. LeSage, and I. Berry. Language functional magnetic resonance imaging in preoperative assessment of language areas: Correlation with direct cortical stimulation. *Neurosurgery*, 52:1335–1347, 2003.

[16] A. Bizzi, V. Blasi, A. Falini, P. Ferroli, M. Cadiolo, U. Danesi, D. Aquino, C. Marras, D. Caldrioli, and G. Broggi. Presurgical functional MR imaging of language and motor functions: Validation with intraoperative electrocortical mapping. *Radiology*, 248(2):579–589, 2008.

[17] A. Shmuel, D. Chaimow, G. Raddatz, K. Ugurbil, and E. Yacoub. Mechanisms underlying decoding at 7T: Ocular dominance columns, broad structures, and macroscopic blood vessels in V1 convey information on the stimulated eye. *NeuroImage*, 49:1957–1964, 2010.

[18] K.J. Friston, A. Holmes, J.B. Poline, C.J. Price, and C.D. Frith. Detecting activations in PET and fMRI: Levels of inference and power. *NeuroImage*, 40:223–235, 1996.

[19] C.R. Genovese, D.C. Noll, and W.F. Eddy. Estimating test-retest reliability in functional MR imaging I: Statistical methodology. *Magn. Reson. Med.*, 38:497–507, 1997.

[20] F.E. Roux, D. Ibarrola, M. Tremoulet, Y. Lazorthes, P. Henry, J.C. Sol, and I. Berry. Methodological and technical issues for integrating functional magnetic resonance imaging data in a neuronavigational system. *Neurosurgery*, 49(5):1145–1157, 2001.

[21] F.Z. Yetkin, W.M. Mueller, G.L. Morris, T.L. McAuliffe, J.L. Ulmer, R.W. Cox, D.L. Daniels, and V.M. Haughton. Functional MR activation correlated with intraoperative cortical mapping. *Am. J. Neuroradiol.*, 18:1311–1315, 1997.

[22] D.B. FitzGerald, G.R. Cosgrove, S. Ronner, H. Jiang, B.R. Buchbinder, J.W. Belliveau, B.R. Rosen, and R.R. Benson. Location of language in the cortex: A comparison between functional MR imaging and electrocortical stimulation. *Am. J. Neuroradiol.*, 18:1529–1539, 1997.

[23] J. Fandino, S.S. Kollias, H.G. Wieser, A. Valavanis, and Y. Yonekawa. Intraoperative validation of functional magnetic resonance imaging and cortical reorganization patterns in patients with brain tumors involving the primary motor cortex. *J. Neurosurg.*, 91:238–250, 1999.

[24] R.J. Tomczak, A.P. Wunderlich, Y. Wang, V. Braun, G. Antoniadis, J. Görich, H.P. Richter, and H.J. Brambs. fMRI for preoperative neurosurgical mapping of motor cortex and language in a clinical setting. *J. Comput. Assist. Tomo.*, 24(6):927–934, 2000.

[25] F.E. Roux, K. Boulanouar, D. Ibarrola, M. Tremoulet, F. Chollet, and I. Berry. Functional MRI and intraoperative brain mapping to evaluate brain plasticity in patients with brain tumours and hemiparesis. *J. Neurol. Neurosurg. Psychiatry*, 69:453–463, 2000.

[26] S. Lehéricy, H. Duffau, P. Cornu, L. Capelle, B. Pidoux, A. Carpentier, S. Auliac, S. Clemenceau, J.P. Sichez, A. Bitar, C.A. Valery, R. van Effenterre, T. Faillot, A. Srour, D. Fohanno, J. Philippon, D. le Bihan, and C. Marsault. Correspondence between functional magnetic resonance imaging somatotopy and individual brain anatomy of the central region: comparison with intraoperative stimulation in patients with brain tumors. *J. Neurosurg.*, 92:589–598, 2000.

[27] R. Kurth, K. Villringer, T. Kombos, O. Süss, M. Brock, A. Villringer, and K.J. Wolf. Correlation of preoperative fMRI of the primary motor cortex and intraoperative monopolar electrostimulation during neuronavigated-surgery of intracerebral tumours. *NeuroImage*, 13(6):S1208, 2001.

[28] N. Pouratian, S.Y. Bookheimer, D.E. Rex, N.A. Martin, and A.W. Toga. Utility of preoperative functional magnetic resonance imaging for identifying language cortices in patients with vascular malformations. *J. Neurosurg.*, 97:21–32, 2002.

[29] N.M. Petrovich Brennan, S. Whalen, D. de Morales Branco, J.P. O'Shea, and A.J. Norton, I.H. aand Golby. Object naming is a more sensitive measure of speech localization than number counting: Converging evidence from direct cortical stimulation and fMRI. *NeuroImage*, 37:S100–S108, 2007.

[30] C. Giussani, F.E. Roux, J. Ojemann, E.P. Sganzerla, D. Pirillo, and C. Papagno. Is preoperative functional magnetic resonance imaging reliable for language areas mapping in brain tumor surgery? review of language functional magnetic resonance imaging and direct cortical stimulation correlation studies. *Neurosurgery*, 66(1):113–120, 2010.

[31] R.C.N. D'Arcy. *Determining the Functional Neuroanatomy of Speech Perception and Reading*. PhD thesis, Dalhousie University, 2002.

[32] Z. Liu, L. Ding, and B. He. Integration of EEG/MEG with MRI and fMRI. *IEEE Eng. Med. Biol.*, 25(4):46–53, 2006.

[33] K. Whittingstall, G. Stroink, and M. Schmidt. Evaluating the spatial relationship of event-related potential and functional MRI sources in the primary visual cortex. *Hum. Brain Mapp.*, 28:134–142, 2007.

[34] W.J. Freeman, S.P. Ahlfors, and V. Menon. Combining fMRI with EEG and MEG in order to relate patterns of brain activity to cognition. *Int. J. Psychhophysiol.*, 73:43–52, 2009.

[35] P. Grummich, C. Nimsky, E. Pauli, M. Buchfelder, and O. Ganslandt. Combining fMRI and MEG increases the reliability of presurgical language localization: A clinical study on the difference between and congruence of both modalities. *NeuroImage*, 32:1793–1803, 2006.

[36] M. Liljeström, A. Hultén, L. Parkkonen, and R. Salmelin. Comparing MEG and fMRI views to naming actions and objects. *Hum. Brain Mapp.*, 30:1845–1856, 2009.

[37] D.S. Kim, I. Ronen, C. Olman, S.G. Kim, K. Ugurbil, and L.J. Toth. Spatial relationship between neuronal activity and BOLD functional MRI. *NeuroImage*, 21:876–885, 2004.

[38] D.J. Veltman, K.J. Friston, G. Sanders, and C.J. Price. Regionally specic sensitivity differences in fMRI and PET: Where do they come from? *NeuroImage*, 11:575–588, 2000.

[39] R.L. Savoy. *Encyclopedia of the Human Brain*, chapter Functional Magnetic Resonance Imaging (fMRI), pages 1–21. Academic Press, 2007.

[40] A. Schreiber, U. Hubbe, S. Ziyeh, and J. Hennig. The influence of gliomas and nonglial space-occupying lesions on blood-oxygen-leveldependent contrast enhancement. *Am. J. Neuroradiol.*, 21:1055–1063, 2000.

[41] A. Aubert, R. Costalat, H. Duffau, and H. Benali. Modeling of pathophysiological coupling between brain electrical activation, energy metabolism and hemodynamics: Insights for the interpretation of intracerebral tumor imaging. *Acta Biotheor.*, 50:281–295, 2002.

[42] D.A. Handwerker, J.M. Ollinger, and M. D'Esposito. Variation of BOLD hemodynamic responses across subjects and brain regions and their effects on statistical analyses. *NeuroImage*, 21:1639–1652, 2004.

[43] B.L. Hou, M. Bradbury, K.K. Peck, N.M. Petrovich, P.H. Gutin, and A.I. Holodny. Effect of brain tumor neovasculature defined by rCBV on BOLD fMRI activation volume in the primary motor cortex. *NeuroImage*, 32:489–497, 2006.

[44] C. Tegeler, S.C. Strother, J.R. Anderson, and S.G. Kim. Reproducibility of BOLD-based functional MRI obtained at 4 T. *Hum. Brain Mapp.*, 7:267–283, 1999.

[45] J.A. Maldjian, P.J. Laurienti, L. Driskill, and J.H. Burdette. Multiple reproducibility indices for evaluation of cognitive functional MR imaging paradigms. *Am. J. Neuroradiol.*, 23:1030–1037, 2002.

[46] D.J. McGonigle, A.M. Howseman, B.S. Athwal, K.J. Friston, R.S.J. Frackowiak, and A.P. Holmes. Variability in fMRI: An examination of lntersession differences. *NeuroImage*, 11:708–734, 2000.

[47] M. Liou, H.R. Su, J.D. Lee, P. Cheng, C.C. Huang, and C.H. Tsai. Bridging functional MR images and scientific inference: Reproducibility maps. *J. Cognitive Neurosci.*, 15(7):935–945, 2003.

[48] H. Otzenberger, D. Gounot, C. Marrer, I.J. Namer, and M.N. Metz-Lutz. Reliability of individual functional MRI brain mapping of language. *Neuropsychology*, 19(4):484–493, 2005.

[49] G.S. Harrington, S.T. Farias, M. Buonocore, and A.P. Yonelinas. The intersubject and intrasubject reproducibility of FMRI activation during three encoding tasks: implications for clinical applications. *Neuroradiology*, 48:495–505, 2006.

[50] J.T. Voyvodic. Activation mapping as a percentage of local excitation: fMRI stability within scans, between scans and across field strengths. *Magn. Reson. Imaging*, 24:1249–1261, 2006.

[51] R. Maitra, S.R. Roys, and R.P. Gullapalli. Test-retest reliability estimation of functional MRI data. *Magn. Reson. Med.*, 48:62–70, 2002.

[52] G. Fesl, B. Braun, S. Rau, M. Wiesmann, M. Ruge, P. Bruhns, J. Linn, T. Stephan, J. Ilmberger, J.C. Tonn, and H. Brüchmann. Is the center of mass (COM) a reliable parameter for the localization of brain function in fMRI? *Eur. Radiol.*, 18:1031–1037, 2008.

[53] G. Fernandez, K. Specht, S. Weis, I. Tendolkar, M. Reuber, J. Fell, P. Klaver, J. Ruhlmann, J. Reul, and C.E. Elger. Intrasubject reproducibility of presurgical language lateralization and mapping using fMRI. *Neurology*, 60:969–975, 2003.

[54] R. Bosnell, C. Wegner, Z.T. Kincses, F. Korteweg, T. Agosta, O. Ciccarelli, N. de Stefano, A. Gass, J. Hirsch, H. Johansen-Berg, L. Kappos, F. Barkhof, L. Mancini, F. Manfredonia, S. Marino, D.H. Miller, X. Montalban, J. Palace, M. Rocca, C. Enzinger, S. Ropele, A. Rovira, S. Smith, A. Thompson, J. Thornton, T. Yousry, B. Whitcher, M. Filippi, and P.M. Matthews. Reproducibility of fMRI in the clinical setting: Implications for trial designs. *NeuroImage*, 42:603–610, 2008.

[55] C.M. Bennett and M.B. Miller. How reliable are the results from functional magnetic resonance imaging? *Ann. N.Y. Acad. Sci.*, 1191:133–155, 2010.

[56] E.E. Chen and S.L. Small. Test-retest reliability in fMRI of language: Group and task effects. *Brain Lang.*, 102:176–185, 2007.

[57] S. Stoll. Nuclear isotope database. `http://www.easyspin.org/documentation/isotopetable.html`, August 2010.

[58] R.A. Frietas Jr. *Nanomedicine*. Landes Bioscience, Austin, 1998.

[59] P.T. Callaghan. *Principles of Nuclear Magnetic Resonance Microscopy*. Clarendon Press, Oxford, 1st edition, 1993.

[60] E.M. Haacke, R.W. Brown, M.R. Thompson, and R. Venkatesan. *Magnetic Resonance Imaging: Physical Principles and Sequence Design*. Wiley-Liss, New York, 1st edition, 1999.

[61] H. Gray. *Gray's Anatomy: The Anatomical Basis of Clinical Practice, Expert Consult*. Churchill Livingston, New York, 40th edition, 2008.

[62] H. Blumenfeld. *Neuroanatomy Through Clinical Cases*. Sinauer Associates, Sunderland, 1st edition, 2002.

[63] Various. Wikimedia commons. `http://commons.wikimedia.org/wiki/Main_Page`, August 2010.

[64] B. Kolb and I.Q. Whishaw. *Fundamentals of Human Neuropsychology*. Worth Publishers, New york, 5th edition, 2003.

[65] P.E. Roland. *Brain Activation*. Wiley-Liss, New York, 1st edition, 1997.

[66] D.B. Ward. *Deconvolution Analysis of FMRI Time Series Data*, 2006.

[67] B.R. Logan and D.B. Rowe. An evaluation of thresholding techniques in fMRI analysis. *NeuroImage*, 22:95–108, 2004.

[68] C.R. Genovese, N.A. Lazar, and T. Nichols. Thresholding of statistical maps in functional neuroimaging using the false discovery rate. *NeuroImage*, 15:870–878, 2002.

[69] N. Otsu. A threshold selection method from gray-level histograms. *IEEE T. Syst. Man. Cyb.*, SMC-9:62–66, 1979.

[70] P. Skudlarski, R.T. Constable, and J.C. Gore. ROC analysis of statistical methods used in functional MRI: Individual subjects. *NeuroImage*, 9:311–329, 1999.

[71] T.H. Le and X. Hu. Methods for assessing accuracy and reliability in functional MRI. *NMR Biomed.*, 10:160–164, 1997.

[72] G.K. Aguirre, J.A. Detre, E. Zarahn, and D.C. Alsop. Experimental design and the relative sensitivity of BOLD and perfusion fMRI. *NeuroImage*, 15:488–500, 2002.

[73] S.C. Strother, N. Lange, J.R. Anderson, K.A. Schaper, K. Rehm, L.K. Hansen, and D.A. Rottenberg. Activation pattern reproducibility: Measuring the effects of group size and data analysis models. *Hum. Brain Mapp.*, 5:312–316, 1997.

[74] R.C. Oldfield. The assessment and analysis of handedness: The edinburgh inventory. *Neuropsychologia*, 9(1):97–113, 1971.

[75] Psychology Software Tools Inc. E-prime v2.0. `http://www.pstnet.com/eprime.cfm`, 2009.

[76] National Institute of Mental Health. Analysis of functional neuroimages (afni). `http://afni.nimh.nih.gov/afni/download/afni`, 2008.

[77] ITT Visual Information Solutions. Interactive data language (idl) v7.1. `http://www.ittvis.com/ProductServices/IDL.aspx`, 2008.