USING FUNCTIONAL MAGNETIC RESONANCE IMAGING TO PLAN SURGICAL RESECTIONS OF BRAIN TUMOURS

KRZYSZTOF J. GORGOLEWSKI

Doctor of Philosophy
School of Informatics
University of Edinburgh
2012

ABSTRACT

Brain tumours, even though rare, are one of the deadliest types of cancer. The five year survival rate for the most malignant type of brain tumours is below 5%. Modern medicine provides many options for treating brain cancer such as radiotherapy and chemotherapy. However, one of the most effective ways of fighting the disease is surgical resection. During such a procedure the tumour is partially or completely removed.

Unfortunately, even after a complete resection some tumourous tissue is left behind and can grow back or metastasise to a different location in the brain. It has been shown, however, that more aggressive resections lead to longer life expectancy. This does not come without risks. Depending on tumour location, extensive resections can lead to transient or permanent post-operative neurological deficits. Therefore, when planning a procedure, the neurosurgeon needs to find balance between extending patients life and maintaining its quality.

Recent developments in Magnetic Resonance Imaging (MRI) fueled by the field of human cognitive neuroscience have led to improved methods of non-invasive imaging of the brain function. Such methods allow the creation of functional brain maps of populations or individual subjects. Adapting this technique to the clinical environment enables the assessment of the risks and to plan surgical procedures. The following work aims at improving the use of functional MRI with a specific clinical goal in mind.

The thesis begins with description of etiology, epidemiology and treatment options for brain tumours. This is followed by a description of MRI and related data processing methods, which leads to introduction of a new technique for thresholding statistical maps which improves upon existing solutions by adapting to the nature of the problem at hand. In contrast to methods used in cognitive neuroscience our approach is optimized to work on single subjects and maintain a balance between false positive and false negative errors. This balance is crucial for accurate assessment of the risk of a surgical procedure.

Using this method a test-retest reliability study was performed to assess five different behavioural paradigms and scanning parameters. This experiment was performed on healthy controls and was aimed at selecting which paradigms produce reliable results and therefore can be used for presurgical planning.

This allowed the creation of a battery of task that was applied to glioma patients. Functional maps created before the surgeries were compared with electrocortical stimulation performed during the surgeries.

The final contribution of this work focuses on technical aspects of performing neuroimaging data analysis. A novel data processing framework which provides means for rapid prototyping and easy translation and adaptation of already existing methods taken from cognitive neuroscience field is introduced. The framework enables fully automatic processing of patient data and therefore greatly reduced costs while maintaining quality control. A discussion of future directions and challenges in using functional MRI for presurgical planning concludes the thesis.

ACKNOWLEDGEMENTS

Many thanks to my supervisors: Cyril Pernet, Mark Bastin, and Amos Storkey. Without their guidance, patience, and wisdom this work would not be possible.

I would like to also thank Satrajit Ghosh for his advice and support.

DECLARATION

I declare that this thesis was composed by myself, that the work contained herein is my own except where explicitly stated otherwise in the text, and that this work has not been submitted for any other degree or professional qualification except as specified.

Edinburgh, 2012

Krzysztof J. Gorgolewski,
March 26, 2013

GLOSSARY

ADP Adenosino-di-phosporates. 16

AMPLE Activity Mapping as Percentage of Local Excitement. 48, 49

AT Adaptive Threshold. 38–49, 51, 71, 73–81, 93, 107, 112

ATP Adenosino-tri-phosporates. 16

B-H Benjamin-Hochberg. 39, 45, 82

BA Brodmann Area. 65

BIC Bayesian Information Criterion. 30, 38

CT Computer Tomography. 8, 9, 11, 12

DTI Diffusion Tensor Imaging. 121

ECS Electrocortical Stimulation. 12, 13, 19–22, 53, 113, 115–122

EPI Echo Planar Imaging. 16, 18, 61, 95

FDR False Discovery Rate. 23, 25, 26, 29–35, 37–39, 41, 43–45, 48, 62, 82

FMRI functional Magnetic Resonance Imaging. 13, 15–31, 35, 38, 47–52, 54, 55, 59, 63, 91, 92, 108–111, 113–122

FNR False Negative Rate. 31–34

FT Fixed Threshold. 38–40, 42–49, 51, 62, 71, 73–79, 107, 112

FWE Family Wise Error. 22, 23, 25, 26, 29, 30, 34, 38, 42–45, 47, 62

GBM Glioblatoma Multiforme. 2, 10, 113

GLM General Linear Model. 19, 23, 27, 31, 32, 37, 38, 61

H-D Harrell-Davies. 39, 43, 45, 46, 62, 106

HRF Hemodynamic Response Function. 17, 19, 93

ICC Intraclass Correlation Coefficient. 93

IPC Inferior Parietal Cortex. 65

IPS Intra–Parietal Sulcus. 54

LPC Lateral Peristriate Cortex. 54

MRF Markov Random Field. 30, 32, 48

MRI Magnetic Resonance Imaging. 8, 9, 11, 12, 14, 15, 22, 55, 63, 115

NMR Nuclear Magnetic Resonance. 14

RCBF regional Crebral Blood Flow. 17

RCMR regional Crebral Metabolic Rate. 17

RFT Random Field Theory. 23, 25, 26, 29–31, 61

ROI Region of Interest. 48, 72

RTMS repetitive Transcranial Magnetic Stimulation. 13

SMA Supplementary Motor Area. 65

SNR Signal to Noise Ratio. 27, 31–34, 40–47, 49, 53, 61, 110

SPM Statistical Parametric Map. 18, 19, 27, 28, 30, 31, 35, 37, 38

TR Time of Repetition. 52, 112

TSNR temporal Signal to Noise Ratio. 94, 95

CONTENTS

1 BRAIN TUMOURS

1.1 INTRODUCTION

Brain tumours are abnormal tissue growth inside the head infiltrating the brain or the central spinal canal (see Figure 1.1). In general tumours can also grow in other parts of human body, but the following work will focus on those that affect the central nervous system. Brain tumours can be divided into two major categories, namely benign and malignant. Benign tumours resemble healthy tissue and grow only up to a certain size. Unless causing damage by pressing on other tissue (the so called "mass effect") they are not dangerous and can often be left untreated. Malignant tumours on the other hand are a type of cancer —- an uncontrollable tissue growth. Untreated they will keep growing and infiltrating more of the surrounding tissue thus presenting a substantial risk to the patient.

The mechanism behind the transformation of a healthy cell into a cancer cell is complex. It requires a series of mutations to promote growth and inhibit tumour suppressants. Tumours affecting the brain can arise from different tissue types such as retinal cells, glia cells, meninges, cerebellar stem cells, or cells of the retina.

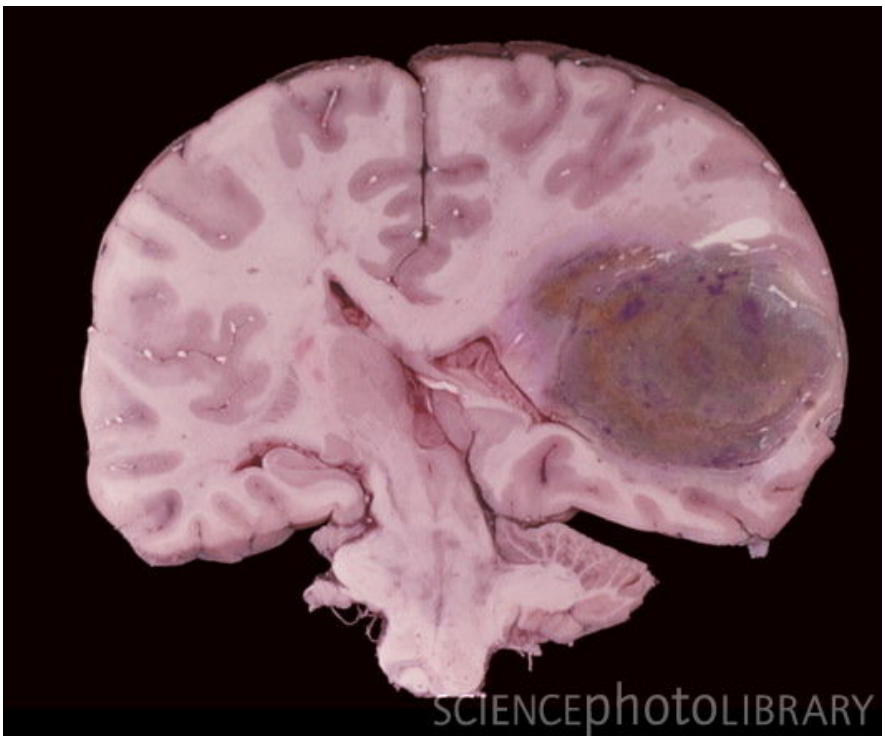

Figure 1.1: Post mortem brain section. The dark mass in the temporal cortex is a glioblastoma — one of the most common types of brain tumour. Source: Science Photo Library.

1.2 TYPES OF BRAIN TUMOURS

There are many types of brain tumours. The most common include the following.

1.2.1 *Gliomas*

Gliomas are the second most common type of tumour and at the same time they are most likely to be malignant and infiltrate surrounding tissue. Among gliomas the most common are:

ASTROCYTOMAS
: they originate from astrocytes, the most common cells in the brain. Highly malignant astrocytoma, Glioblatoma Multiforme (GBM), is the deadliest of all brain tumours having a median survival time of 17 to 37 weeks depending on treatment.

EPENDYMOMAS
: tumours originating from ependymal cells compromising the membrane of ventricles. Ependymomas inside the brain are most common in children. In adults they tend to develop in the spinal cord.

OLIGODENDROGLIOMAS
: a malignant transformation of oligodendrocytes, glial cells responsible for myelinating neurons.

1.2.2 *Meningiomas*

Meningiomas are the most common type of brain tumour. They originate from the meninges, a sheet of protective tissue between the brain and the skull. 90% of meningiomas are benign and often require no treatment apart from periodic observation. They are also typically asymptomatic since in many cases they do not infiltrate the surrounding brain tissue. However, a fast growing meningioma can cause increased intracranial pressure.

1.2.3 *Pituitary tumours*

The third most common type of brain tumour. These are located in the pituitary gland, a part of the brain responsible for releasing hormones such as the growth hormone, prolactin, thyroid-stimulating hormone, adrenocorticotropic hormone, melanocyte-stimulating hormone, follicle stimulating hormone , and luteinizing hormone. Pituitary tumours can disrupt the hormonal balance in the body and produce unique symptoms. Even though many pituitary tumours are benign they are often surgically removed because of the proximity of other vital brain regions (e.g. brain stem).

1.3 EPIDEMIOLOGY AND ETIOLOGY

1.3.1 *Incidence*

Cancer is relatively rarely found in the brain. In a registry of cancer patients in England available at http://www.ons.gov.uk, cancer of the brain accounted for only 1.5% of all cases (see Figure 1.2). This data, however, does not include benign tumours which, depending on location, can also be a health hazard and can potentially turn into a malignant form. The fact that brain tumours are just a small fraction of all cancers is true when averaged over all ages, but the situation changes dramatically when we focus just on the youngest patients. In the US, brain tumours are the second most common form of cancer among children, with leukemias being the most common (Gurney et al., 1999). Overall incidence of brain tumours (malignant and non-malignant) is estimated at a level of 19.89 per 100,000 person-years. This translates into an estimated 12500 new cases a year in UK and 62500 in the US.

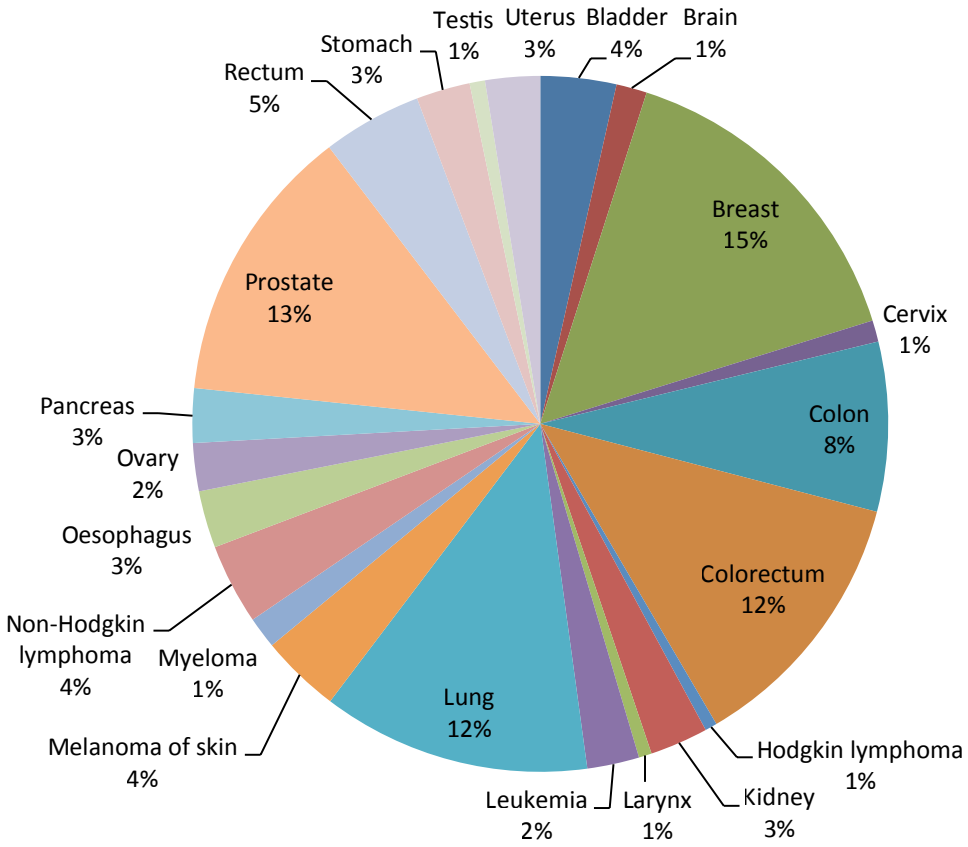

Figure 1.2: Distribution of cancer patients diagnosed in England in 2005-2009. Brain cancer accounts for only 1.5% of all cancer cases. Source http://www.ons.gov.uk.

Brain tumours are also more common in older people and children than in young adults. In England, most patients are diagnosed at age 50 years and above (see Figure 1.3). In the US, young adults account for only 9% of all malignant and non-malignant tumours (Gurney et al., 1999).

The interaction between gender and brain tumour incidence is not clear. According to statistics collected in England between 2005–2009, males are more likely to be diagnosed with brain tumours. However, a similar study carried out in the US showed the opposite trend (Gurney et al., 1999). It is worth noting that the US registry included malignant and non-malignant tumours whereas the English study focused only on malignant neoplasms.

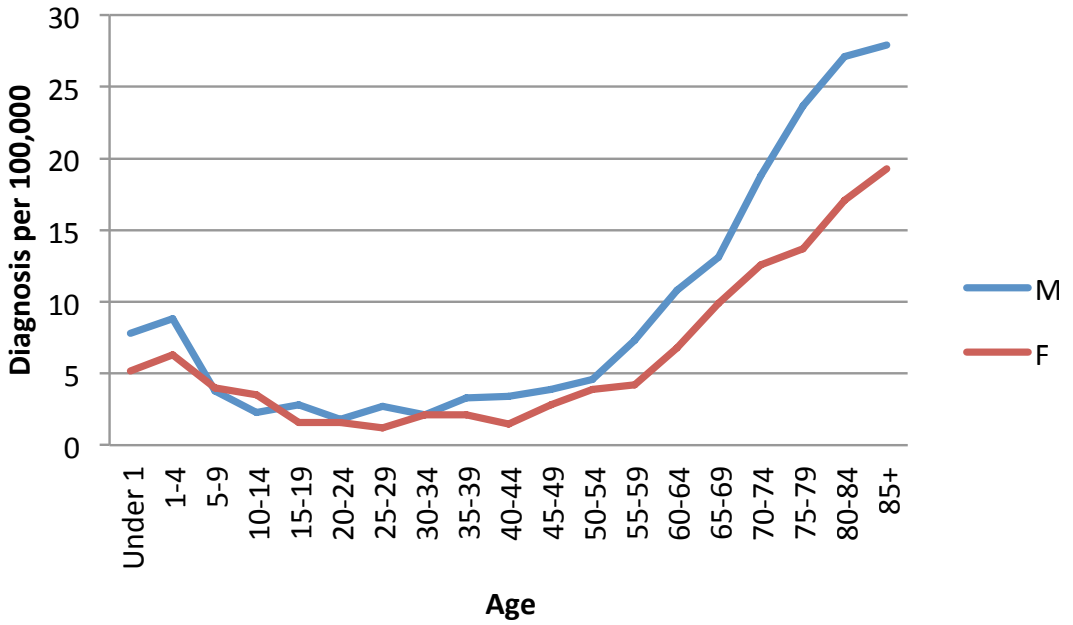

Figure 1.3: Age standardized incidence of brain cancer across ages for males and females. Source http://www.ons.gov.uk.

According to US cancer registries of malignant and non-malignant tumours, the most common histological type is meningioma (34%) followed by glioblastoma (16.3%) and then pituitary tumours (13.5%). All gliomas taken together account for 30% of tumours and 80% of malignant tumours (see Figure 1.4).

The distribution of histological tumour types changes across ages. Children (ages 0–14 years) are most likely to develop medulloblastomas and pilocytic astrocytomas. Youths and young adults (15–34 years) are more likely to be diagnosed with pituitary tumours. From age 35 onwards meningliomas and glioblastomas are the most common tumours, with exception of very old age (85+) people where the incidence of gliomas start to decrease.

1.3.2 *Survival*

Even though cancer of the brain is relatively rare it remains one of the deadliest of all cancers. As mentioned before it is the second deadliest cancer among children. In terms of five year survival rate (as estimated in England between 2005–2009) cancer of the brain comes fourth after lung, larynx, and colon (see Figure 1.5). Females show slightly better survival than males and in general younger patients manage to fight the disease longer than older patients.

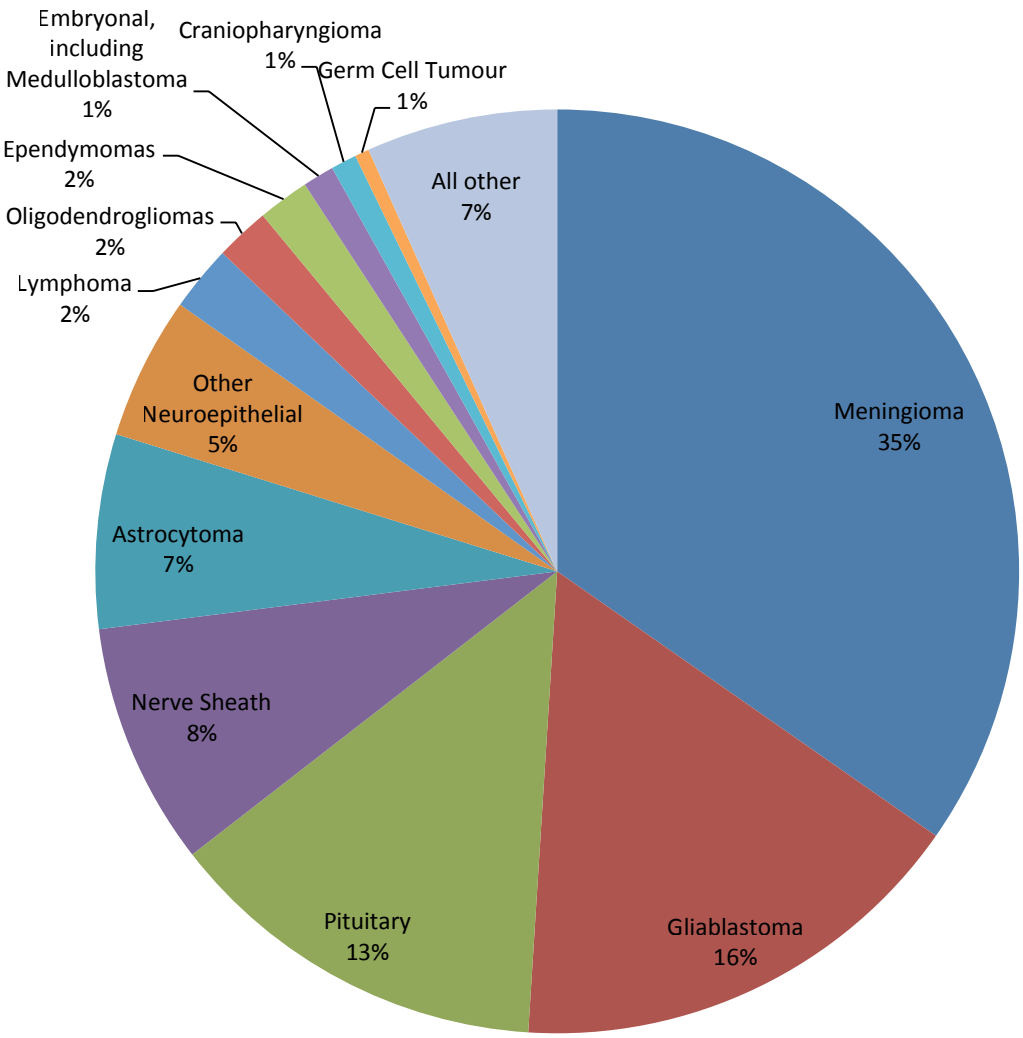

Figure 1.4: Distribution of brain tumour types. Gliomas account for 30% of all tumours and 80% of malignant tumours. Data collected between 2005-2008. Source: "CBTRUS Statistical Report: Primary Brain and Central Nervous System Tumors Diagnosed in the United States in 2004-2008".

Prognosis for different types of tumours can be drastically different. The benign tumours usually do not present a significant health risk, but in case of malignant tumours the life expenctancy after diagnosis can range from months to years. Out of all the malignant tumours, gliablastoma is the quickest in its progression with 75% of patients not surviving past one year after diagnosis. On the other side of the spectrum, 91% of the patients with pilocytic astrocytoma live beyond ten years after being diagnosed (see Figure 1.6). For a more extensive review of brain tumour epidemiology see Ohgaki and Kleihues (2005) and Ohgaki (2009).

1.3.3 *Risk factors*

Many risk factors have been investigated in the context of brain tumours. Certain occupations have been reported to coincide with a higher rate of brain tumour diagnoses, but many of these studies lack confirmation or have been contradicted by

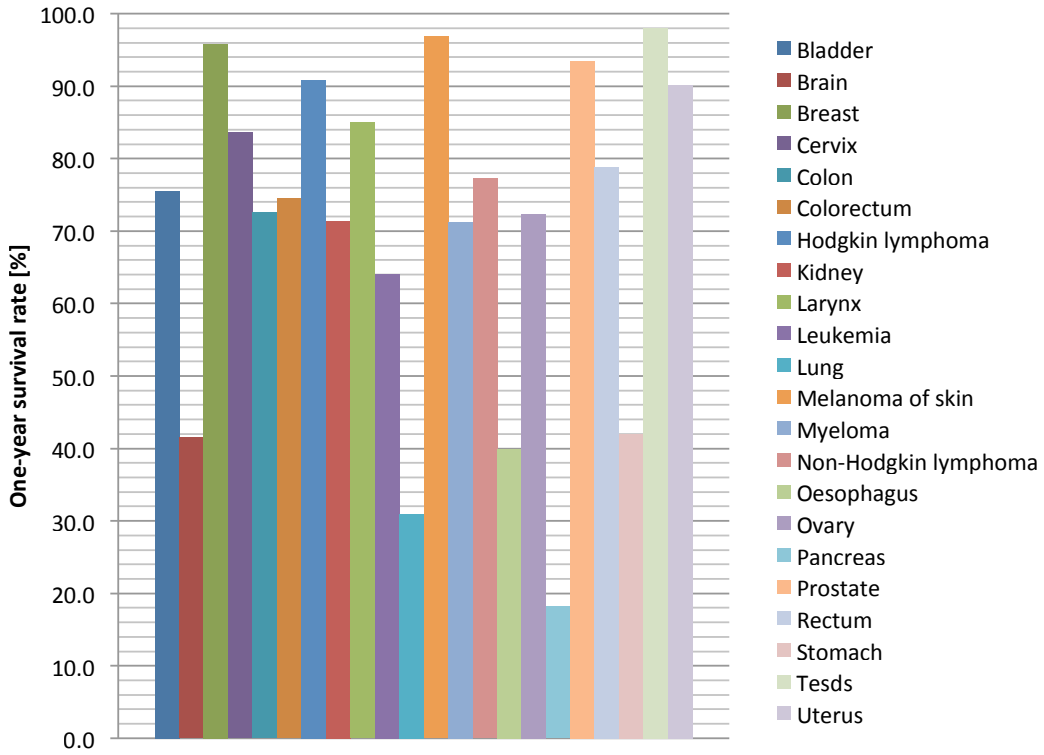

Figure 1.5: Age-normalized survival rate for patients diagnosed in England in 2005-2009, split between different types of cancer. Prognosis for brain cancer is relatively poor. Almost 60% of diagnosed patients do not live longer than a year after the diagnosis making it the 4th most deadly cancer. Source `http://www.ons.gov.uk`.

other investigations based on different samples. For example, it has been reported that jobs with a higher exposure to lead can cause brain tumours (Anttila et al., 1996; Cocco et al., 1998). These findings come from analysing occupation information of deceased brain tumour patient in 24 US states (Cocco et al., 1998), as well as monitoring lead levels in the blood of workers with high exposure risk (Anttila et al., 1996). However, another study of male US workers (Wong and Harris, 2000) did not find such a correlation despite finding a significantly increased risk developing cancer of stomach, lungs and endocrine organs.

Relationships between parental occupation and brain cancer risk have been also investigated. Children of parents working in the chemical industry (McKean-Cowdin et al., 1998), as farmers, drivers and mechanics, and in the textile industry (Cordier et al., 2001) have increased risk for developing brain tumours. Although those relationships appear plausible, no trend connected to a specific chemical environmental component has yet been found.

Other potential risk factors include electromagnetic field exposure, cell phone usage, diet rich in nitrates, and genetic factors, although very little actual evidence has been found to support the impact of electromagnetic fields and cell phone usage. However, diets based on cured meat and ham (which are rich in nitrates) have been shown to elevate the risk of brain cancer (Blowers et al., 1997). As in case of occupational risk factors not all studies confirm this diet risk (Chen et al., 2002). Genetic and

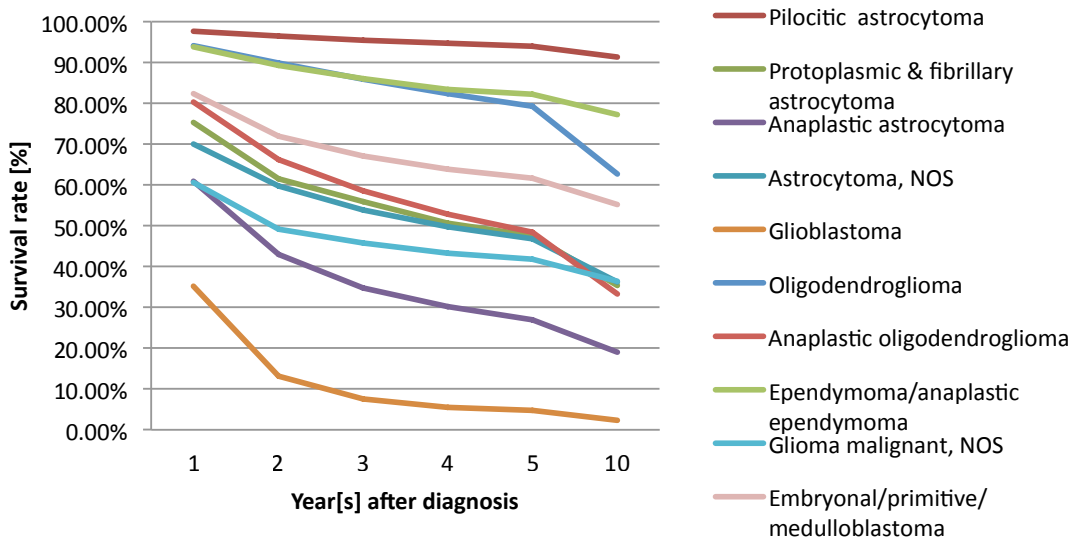

Figure 1.6: Survival rate of selected brain tumours. Data collected in the US, between 1995-2008. Source http://www.cbtrus.org/.

hereditary clustering focused research also does not show a clear causal effect of certain gene mutation. This might be due the complicated interaction between genetic and environmental factors (Inskip et al., 1995).

However, there is one factor that has been successfully recognized as creating significant risk of developing a brain tumour it is ionizing radiation in children (Brüstle et al., 1992). Exposure to therapeutic radiation at a young age increases the chances of developing gliomas 7–9 years later. This is true only for dosages of radiation used for therapy, 70–250 rad (see Ron et al., 1988). The influence of lower radiation exposure used for example in dental X-rays is not clear (Rodvall et al., 1998). For a further review of the etiology of brain tumours see Inskip et al. (1995) and Ohgaki and Kleihues (2005).

1.4 DIAGNOSIS

The diagnosis of brain tumours is a multistep process. In most cases the patient starts experiencing neurological symptoms such as, but not limited to, headaches, partial paralysis, speech problems, loss of memory, visual impairment, or seizures. Symptoms are highly dependent on tumour location and can be similar to other neurological issues, such as traumatic brain injuries or strokes. It usually takes years from the onset of the disease to diagnosis. This is because the tumour has to reach a certain size to start causing symptoms. In contrast to traumatic brain injuries and strokes, most tumours grow slowly giving the plasticity mechanisms in the brain the time to adjust to the change and maintain normal functioning. This is, however, only possible up to the point where the tumour infiltration is not too invasive.

Because the symptoms of brain tumours are overlapping with other neurological illnesses, diagnosis is never only based on symptoms. When a doctor of first contact suspects a brain tumour, the patient is sent to a neurologist or neurosurgeon who will most likely order a brain scan. Various brain scanning modalities and techniques

can reveal brain tumours. The simplest of them is Computer Tomography (CT). It reconstructs volumetric images of the scanned object from a series of X-ray pictures taken from different angles. Most tumours appear darker than the healthy tissue on a CT image (see Figure 1.7). Another imaging technique useful in brain tumour diagnosis is Magnetic Resonance Imaging (MRI). The principles of MRI are more complicated than CT and rely on the excitation of atomic nuclei, in particular water protons (see e.g. Huettel et al., 2008 for detailed explanation on MRI). Depending on the MRI sequence used, tumours can appear darker (see Figure 1.8a) or brighter than the surrounding brain tissue (see Figure 1.8b). Contrast agents such as iodine for CT and gadolinium for MRI administered intravenously can enhance the signal from some tumours.

Figure 1.7: A CT scan of the head. The darker region on the left indicates a diffuse glioma.

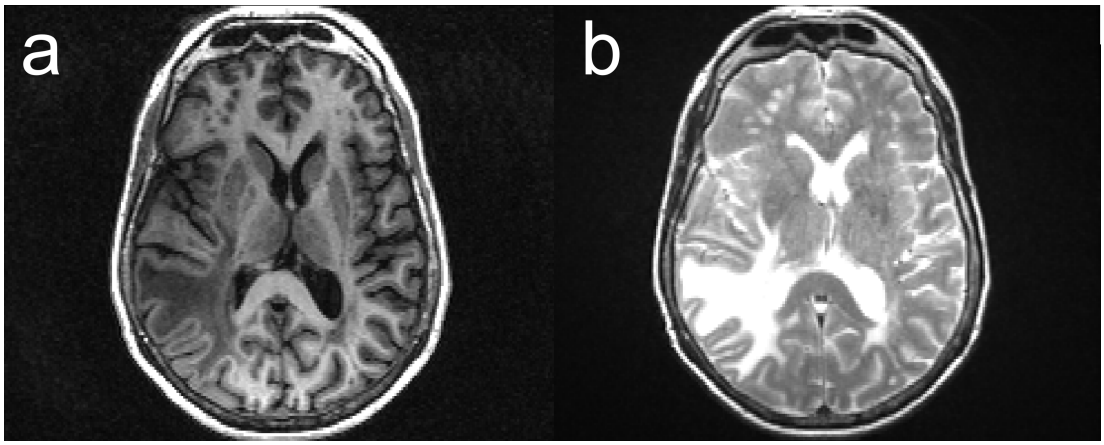

Figure 1.8: An MRI scan of the head. T1-weighted contrast shows a diffuse glioma as a darker region (a), whereas T2-weighted contrast reveals the tumour as a brighter region (b). Both scans are of the same patient.

Brain scans performed for different reasons, for example after a head trauma during a car accident, can also reveal asymptomatic brain tumours. These accidental finding give more time to counteract and choose the optimal course of treatment. Additionally if a patient is already suffering from cancer in another part of the body there is a chance of metastatic tumours (also called secondary brain tumours). Cancer can spread through the lymphatic system and blood to distant organs. The blood-

brain barrier protecting the brain does not to prevent this spread. Metastatic brain tumours are most common in patients who suffer from lung cancer, breast cancer, malignant melanoma, kidney cancer, and colon cancer. In such cases a brain scan or full body scan can be ordered and reveal a brain tumour before it will show any symptoms.

A definitive way of diagnosing a brain tumour is microscopic investigation of tissue samples (histology). These samples can be obtained through open skull surgery or (more commonly) biopsy. Brain biopsies are relatively simple procedures involving drilling a small hole in the skull and a small piece of the brain is removed using a needle. Often biopsy procedures are guided using previously acquired CT or MRI images to precisely target the potential tumour site. Tumour tissue can express various changes on the molecular level which are visible under the microscope (see Figure 1.9). Additionally staining can be used to increase the contrast between cellular structures. Hematoxylin and eosin staining for example renders nuclei dark blue and all other structures in different shades of red, orange, and pink. The following cellular changes can be observed in brain tumours:

NEOPLASIA
: an uncontrolled division of cells.

ANAPLASIA
: also known as dedifferentiation of cells. Cells lose structural and functional differentiation of normal cells. Often cells are unnaturally shaped and can have abnormally big nuclei. Ratio of nuclei to cytoplasm size can exceed the normal 1:6 and reach 1:1. Anaplastic tumours can arise from dedifferentiation of neoplastic tumours or grow from cancer stem cells. A grade between I and IV is given to each tumour depending on how dedifferentiated the cells are.

NECROSIS
: Cell death. Usually dead cells are disposed of by phagocytes and will not be visible in a tissue sample, but necrotic cells have disrupted cell signalling which makes them invisible to phagocytes. This not only leads to build up of dead cells but also harmful toxins.

ATYPIA
: abnormality of cell shape not included in any of the categories included above.

Additionally cancerous changes can disrupt blood supply by competing for access to ventricles with healthy cells. Functional abnormality can also change the balance of neurotransmitters in the extracellular space.

1.5 TREATMENT

There are three major courses of action to treat malignant brain tumours: chemotherapy, radiation therapy, and surgery. Even though they can be combined they differ in their mode of action, advantages, risks and side effects; therefore they will be discussed separately.

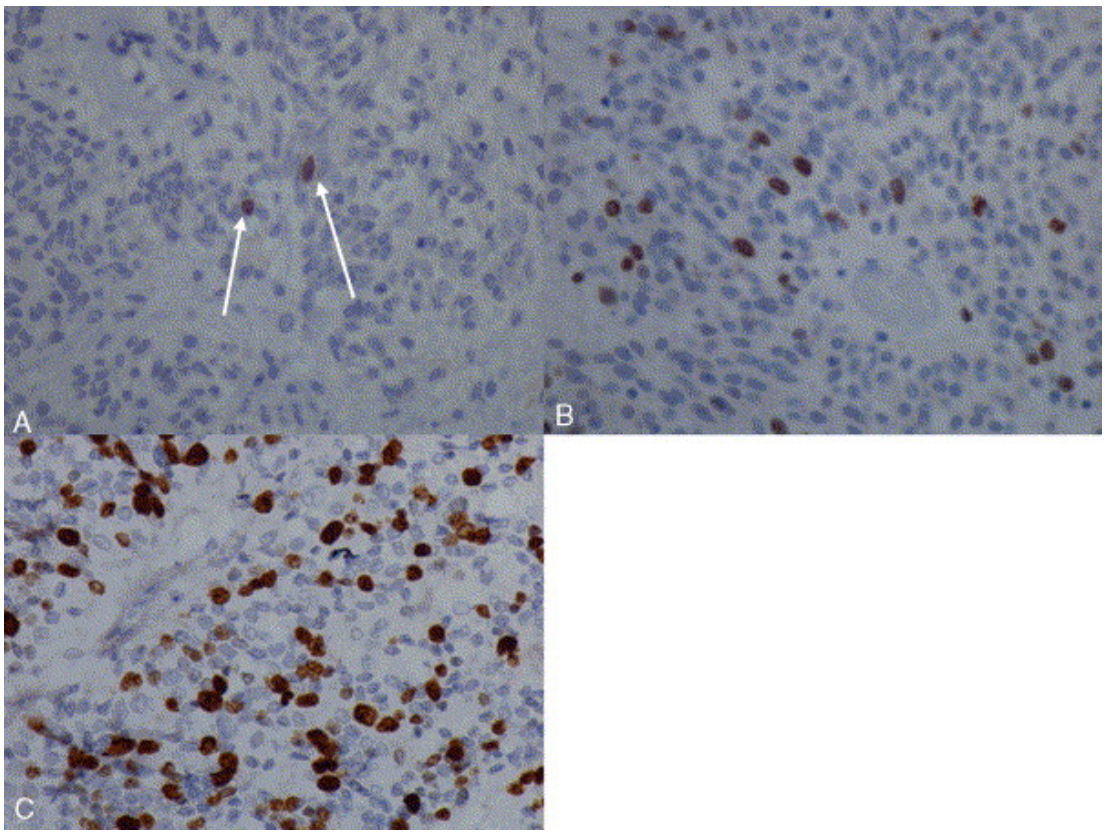

Figure 1.9: Histopathology of a brain tumour. Ki-67 staining shows proliferation (cell growth) in red. Specimens come from tumours at different stages of development: initial biopsy (A), six (B) and ten (C) years after presentation. Figure reproduced from (Tarnaris et al., 2006).

1.5.1 *Chemotherapy*

The defining characteristic of any cancer including malignant brain tumours is rapid cell growth. Chemotherapy consists of administering drugs that specifically target fast dividing cells. Administration is usually intravenous, but sometimes oral. Because of the mode of administration, drugs act globally targeting all fast dividing cells. Those not only include cancer tissue, but also blood cells, bone marrow, digestive tract, and hair follicles. Therefore chemotherapy can cause many side effects such as skin and nail loss, apathy, weight loss, and damage to the immune system. The latter is probably the most serious of the side effects since it makes the patients prone to infections. Chemotherapy can also damage reproductive system and fertility preservation methods are recommended. In the case of pregnant women risk of damaging the fetus is so high that abortion is recommended.

Because chemotherapy targets fast dividing cells, it is most effective on fast and/or young tumours. For the same reason the core of the tumour is usually left spared and requires additional treatment such as radiotherapy or surgery. Chemotherapy is also less efficient in brain tumours than cancer of other body parts because of the blood brain barrier which prevents some of the drugs from reaching the tumour. Therefore the improvement after chemotherapy in brain tumour patients is modest. It is usually prescribed for the most aggressive tumours such as GBM. Additionally it is used as a

replacement for radiation in children to avoid its detrimental effect on the developing brain.

1.5.2 *Radiation therapy*

Another common way of fighting cancer is using ionized radiation. In this treatment radiation is used to damage cell DNA and therefore slow down tissue growth and/or cause cell death. As with chemotherapy this method is most effective on fast dividing and undifferentiated cells (such as cancer cells but also stem cells) because their DNA repair mechanisms are least effective. Dividing cells with damaged DNA will cause accumulation of errors and cease growth. Unfortunately radiation has a detrimental effect on healthy tissue as well. Therefore narrow beams of radiation targeting only the tumour or potential tumour sites (lymph nodes) are used. Additionally to mitigate the effect of radiation on healthy tissue between the tumour and the source (skin, skull, grey matter) multiple angles are used. In this technique, the target (tumour) is exposed to a beam of radiation from multiple angles. Since all the beams cross at one point, the tumour will receive the biggest dose of radiation sparing the surrounding tissue. Additionally, dosages can be spread in time. This allows healthy cells to recover (since there are able to repair their DNA more efficiently than cancer cells).

Radiotherapy is often applied in conjunction with chemotherapy or surgery (see below). Because smaller tumours respond better to radiation it is often optimal to first remove the bulk of the tumour (in a safer non-radical resection) and then treat the remaining parts with radiation. Additionally radiation can also be delivered intraoperatively straight after removing the bulk of the tumour.

Modern, high resolution CT and MR imaging have revolutionized radiotherapy. Acquiring precise images of the tumour allows more precise targeting of the radiation beams. When combined with using the same external point of reference for both the imaging and the radiation (stereotactic procedures), radiotherapy can target just the tumour sparing healthy tissue to some extent. Procedures based on CT or MRI are called stereotactic radiosurgeries and are executed in a fully computerized way. The amount, rate, angle and width of radiation can be optimized using a computer system to minimize collateral damage. Despite many advantages of this approach, due to the nature of radiation it can be only successfully used on small tumours. In the case of bigger neoplasms, surgery can be the only option. Unfortunately, due to the plastic nature of the brain, tumours often stay unnoticed until they grow to a size that excludes the option of radiosurgery.

Despite the fact that radiation is mostly painless, it can cause some side effects. Whole brain radiation often leads to cognitive decline. This involves deficits in learning, memory and spatial information processing. This is commonly linked with impaired neurogenesis in hippocampus (Rola et al., 2004). Additionally (especially in case of children), radiation increases chances of getting a new malignant brain tumour 7–9 years after the treatment. However, in most cases this risk does not outweigh the benefits of radiation as a treatment.

1.5.3 *Surgery*

The third treatment option is surgical removal of the tumour. This procedure involves opening the patient's skull, cutting through the meninges, whilst avoiding any damage to the vascular system, and removing the tumourous tissue (see Figure 1.10). It is a serious procedure involving general anaesthesia. The surgery is planned using preoperative CT and MRI to decide the location and size of the craniotomy (opening the skull) and extent of the resection. Surgery comes with substantial risk not only from anaesthesia but also from the fact that parts of the removed brain tissue might be still involved, and crucial for, some cognitive and mental capabilities. At the same time it has been shown that the size of the resection correlates with life expectancy after the procedure (Patchell et al., 1990; Bindal et al., 1993; Woodward et al., 1996). Therefore surgeons are faced with a dilemma choosing between radical resection and a risk of post-operative neurological deficit, or a liberal resection and shorter lifespan. In other words it is a balance between the length of life and its quality.

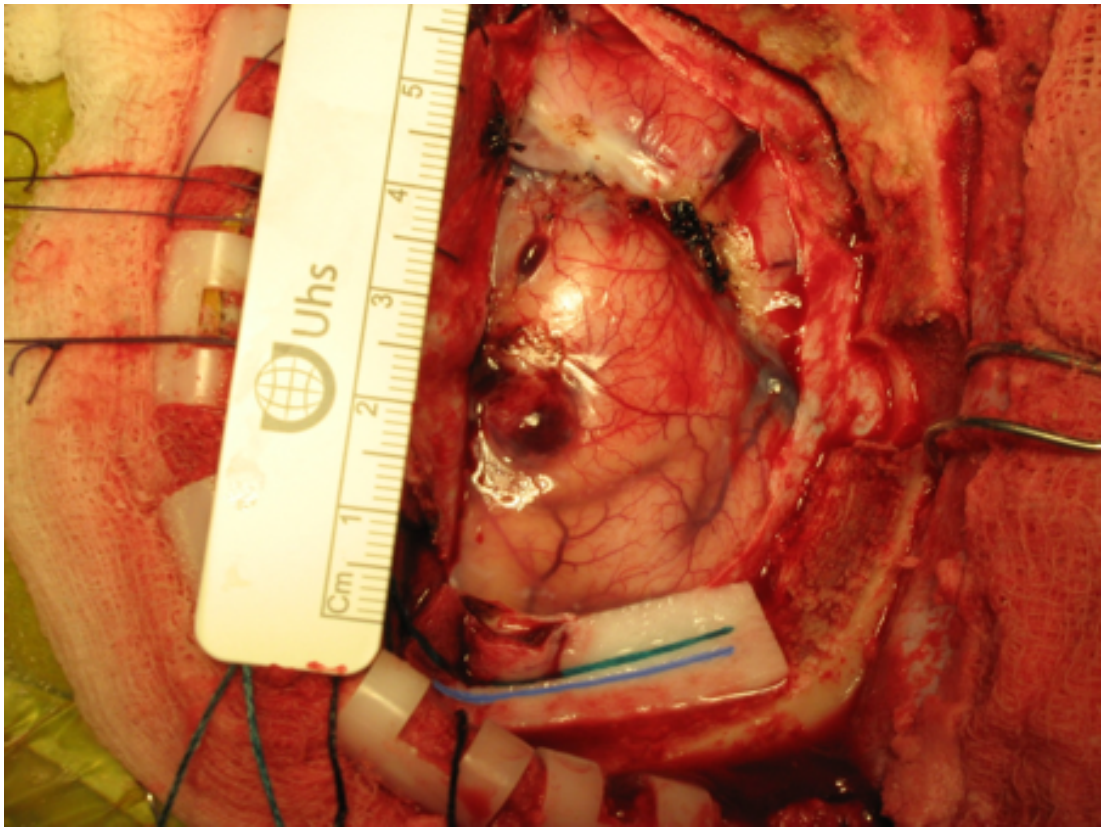

Figure 1.10: Brain surgery. The surface of the brain has been exposed after pulling away the skin, removing a bone flap from the skull and cutting through the meninges.

Some attempts have been made to alleviate the risks of post surgical deficits. These include using anatomical landmarks and functional anatomy knowledge gathered from lesion studies to localize and avoid functional areas. This techniques has, however, the obvious limitation of not taking into account the between subject variability. The anatomical landmark method has been superseded by Electrocortical Stimulation (ECS). This technique involves applying small electrical currents to the exposed

surface of the brain (Penfield and Boldrey, 1937; Berger et al., 1989). Electrical current can interferes with normal functioning of neurons which can be observed in the behaviour of the patient. The reaction depends on the stimulated region and cognitive function related to it. For example, stimulating primary motor cortex causes involuntary movement. In case of Broca's area patient can experience speech arrest. To fully observe the influence of the stimulation on patient's behaviour, he or she has to be temporarily woken up from the anaesthesia. Today ECS is carried out during most of the procedures and its usefulness in terms of reducing the severity of postoperative neurological deficits has been well established. However, it does not come without flaws. First of all it extends the length of the procedure which translates to increased risk and cost. Secondly, subjects woken from general anaesthesia can be confused and fail to perform the behavioural tests properly. Finally since the cortex is mapped just before the procedure, any decision about the extent of the resection has to be made ad hoc by the surgeon. As mentioned before this decision is a tradeoff between the potential extension of patient's life and risks of neurological deficits. Cortical mapping can be performed as a separate procedure allowing the patient to recover and make an informed decision about the treatment. This approach, however, adds the risks and costs of an extra procedure.

There have been attempts to use other techniques to map the cortex that would avoid the aforementioned issues. repetitive Transcranial Magnetic Stimulation (rTMS) has been suggested as a potential replacement for ECS (Krings et al., 2001a). This technique is based on a similar idea as ECS; both methods aim at influencing the neuronal dynamics to find areas that are related to certain behaviour. rTMS has a different mode of action. Fast switching electromagnets are used to focally induce current. The main advantage of this method is that the electromagnetic field can penetrate the skin and bone avoiding the need for opening the skull. Mapping procedure can therefore be performed, before tumour extraction. However, rTMS does not come without limitations. Because the magnetic field evoked by the rTMS magnets is very local it has limited range. This becomes an issue when trying to stimulate areas far from the surface of the skull.

Another approach involves using functional Magnetic Resonance Imaging (fMRI). This technique allows the mapping of a wide variety of cognitive skills, is not invasive and provides information about cortical and subcortical areas. The principles behind fMRI and its use for presurgical mapping is the focus of the next chapter.

1.6 SUMMARY

Brain tumour is a serious disease that often leads to death within months or years from the initial diagnosis. Among other treatments surgical removal of the tumourous tissue has proven to extend the survival rate of patients. However, the procedure carries a risk of neurological impairments caused by damaging functional tissue. In the next chapter we will discuss how fMRI can be used to asses and minimize that risk, by localizing the functionally eloquent areas before the surgery.

2

FMRI AND ITS USE FOR PRESURGICAL PLANNING

2.1 NUCLEAR MAGNETIC RESONANCE PRIMER

Even though MRI images have a spatial resolution in the range of millimetres, the underlying physical phenomena occurs on the level of atomic particles. During late 1930s and 1940s Isidor Rabi, Felix Bloch, and Edward Purcell discovered that protons when put in a static magnetic field can absorb energy in the form of an electromagnetic pulse and release it when the pulse is turned off. Their research (later culminating in a Nobel Prize) showed that energy can be absorbed only if the electromagnetic pulse had a particular frequency. This frequency matched the spin frequency of the atomic nucleus, hence the name Nuclear Magnetic Resonance (NMR).

Because 70% of the mass of a human body consists of water H_2O, hydrogen is by far the most abundant element. Hydrogen has only one proton and is positively charged which makes it particularly interesting in the context of NMR. Since each positively charged proton spins, it generates a magnetic field perpendicular to the plane of spinning. This field is infinitesimally small, and even for a large number of protons (an averaged sized person contains 5×10^{27} protons) the net magnetization will be close to zero. This is because without external interference, each proton points in a different direction and the magnetic fields created by them cancel each other out.

However, when put into an external static magnetic field (B_0) all protons align with that field. They can either align in parallel (along) or antiparallel to the direction of the field. The parallel state is also called the low-energy state and more protons reach it. The antiparallel state is called the high-energy state and is less common. Therefore the number of low-energy protons is higher than the high-energy protons. Because protons are not randomly scattered the net magnetization is not close to zero any more. It in fact depends on the ratio of low and high-energy state protons, since their magnetic fields point in opposite directions.

The net magnetization caused by unequal number of low and high-energy state protons remains constant for a given field strength. However, low-energy state protons can switch to the high-energy state if given an appropriate amount of energy at the correct frequency. This energy difference between the two states equals to:

$$\Delta E = h\nu_0 \tag{2.1}$$

where h is the Planck's constant and ν_0 is the resonance frequency given by:

$$\nu_0 = \frac{\gamma B_0}{2\pi} \tag{2.2}$$

where γ is the gyromagnetic ratio depending on the charge and mass of the atomic nucleus (proton in our case) and B_0 is the strength of the magnetic field. After applying the electromagnetic pulse a subset of the low–energy state protons will "jump" into the high-energy state. The exact number depends on the length of the pulse. A 90° pulse will equalize the number of low and high state protons yielding a zero net

magnetization. A 180° pulse will reverse the ratio leading to a net magnetization of the same strength as the stable state (before excitation) but in the opposite direction. When energy is no longer being supplied through an electromagnetic pulse, protons begin to recover to their stable state. Transitions from a high–energy state back to low–energy state is accompanied by a release of energy. The sum of that energy is equal to energy put into the system by the electromagnetic pulse. This energy is released in the form of photons and thanks to the fact that their frequency is the same as the excitation pulse, the same coils can be used for transmitting and receiving the signal. The change in net magnetization in the longitudinal direction (B_0 field direction) is called T1 recovery and depends on the tissue type. This allows different anatomical features and tissue types to be distinguished and is a key basis of MRI. T1 also influences how quickly the next electromagnetic pulse can be applied so protons can fully recover after the previous excitation.

So far we have only mentioned net magnetization in the longitudinal direction. Since protons are actually precessing in the B_0 field, they also have transverse (perpendicular) components. If the spins are not in sync this component will be zero, because each proton's precession will cancel the others out. However, an electromagnetic pulse synchronizes the spins causing a non-zero net magnetization in the transverse plane. After the electromagnetic pulse is switched off this coherence is gradually lost. This is called T2 decay. Despite a similar nature (returning to a stable, initial state) T1 and T2 relaxation rates are not the same. Both of them are exponential in form, but the time constant of T1 is usually an order of magnitude larger. In practice many tissue types can be defined only by a combination of T1 and T2 signals.

A decrease in coherence of the frequency and phase of the spins which leads to decreasing net magnetization in the transverse plane can be caused by two phenomena: spin-spin interactions and spatial inhomogeneity of the magnetic field. The first is caused by interference between spins of protons that are close to each other. This decay is what we described above as T2. The second cause, spatial inhomogeneity of the magnetic field, is additive to T2 and in literature is referred to as T2*. This magnetic field inhomogeneity and the T2* signal are the basis of the functional MRI signal.[1]

2.2 FUNCTIONAL MAGNETIC RESONANCE

2.2.1 *Physical principles*

fMRI builds on the basis of structural MRI, but focuses on the dynamical signal changes instead of a single point in time. To be more precise, fMRI acquires a series

1 This brief introduction to NMR is in no way exhaustive. MRI was possible due to major breakthroughs in physics, engineering, and mathematics. These advancements have been reflected by three Nobel Prizes: in 1944 to Isidor Isaac Rabi "for his resonance method for recording the magnetic properties of atomic nuclei"; in 1952 to Felix Bloch and Edward Mills Purcell "for their development of new methods for nuclear magnetic precision measurements and discoveries in connection therewith"; and in 2003 to Paul Lauterbur and Sir Peter Mansfield "for their discoveries concerning magnetic resonance imaging". Without the work of these and many other scientists the following research would not be possible. However, explaining in detail the physical principles and engineering solutions that made MRI possible is beyond the scope of this work. We refer the curious reader to the excellent handbook by Huettel et al., 2008.

of images of the brain in rapid (2–3 seconds apart) succession. One would expect that each of the images in such series would be more or less the same after accounting for random noise contributions. However, two physiological phenomena in the human brain make Echo Planar Imaging (EPI) sequences sensitive to local brain activation. Firstly oxygenated and non-oxygenated blood has different magnetic properties. These differences can be picked up by gradient-echo MRI. Blood flow issues seem somewhat removed from the issue of brain signals propagated by spiking neurons. One has to remember that recovering the ionic balance between the outside and inside of the cell is an active process requiring energy. The molecular ion pumps located on the cell membranes use Adenosino-tri-phosporates (ATP) which has to be delivered to fuel the process. This is where blood supply system comes into play. An intricate system of capillary vessels delivers oxygen and other substances crucial for smooth running of the neuronal cells.

This by itself would not allow us to visualise populations of spiking neurons. After all, the blood supply could be constant. In this way the varying demand for nutrients would have been met on average. This is, however not the case. Due to a phenomenon of neuro-vascular coupling, the vascular system adapts to local demands of nutrients. In other words, vessels expand increasing the blood flow and allowing more blood to be delivered to the region that needs it. This, with the combination of higher use of oxygen, leads to a local change of magnetic properties that can be picked up by EPI sequences. The local aspect of this process is extremely important because otherwise MRI would not be able to perform the task of mapping behaviour and cognition to certain populations of neurons.

The difference in magnetic properties of oxygenated and deoxygenated haemoglobin was first noticed by Linux Pauling and Charles Coryell (Pauling and Coryell, 1936). They discovered that oxygenated blood is diamagnetic (has zero magnetic moment) whereas deoxygenated blood is paramagnetic (has significant magnetic moment). This difference in magnetic moments leads to 20% greater magnetic susceptibility of deoxygenated blood. Later, Seiji Ogawa discovered that deoxygenated blood attenuates T2*-weighted signal (Ogawa et al., 1990). In his experiments using rats he noticed dark lines resembling blood vessels that disappear when rats were breathing pure oxygen instead of a normal air mixture (rich in CO_2). The source of this signal loss was confirmed by scanning test-tubes with oxygenated and deoxygenated blood. The presence of deoxygenated blood therefore distorted the T2*-weighted signal. This discovery was a cornerstone of non-invasive in vivo human brain fMRI.

As mentioned earlier, oxygen is an important component of the energy management processes inside the human body. Even though energy is delivered to cells in the form of glucose, it is mainly used to restore membrane potential via active ion pumps. These molecular mechanisms are fuelled by the reaction of decoupling ATP into Adenosino-di-phosporates (ADP). Therefore glucose has to be converted to ATP to be of use to neurons. There are two chemical reactions that can turn glucose into ATP: aerobic and anaerobic glycosis. The variant without oxygen produces 2 molecules of ATP for every molecule of glucose. Addition of oxygen glycosis boosts its performance to 36 molecules of ATP.

Oxygen is therefore crucial for restoration of membrane potential after neuronal spiking. Demand for oxygen and nutrients triggers a reaction of the vascular system: blood vessels expand in width increasing the blood flow. As mentioned above,

this happens on a local level. In other words blood supply will change only within a millimetre or so from the spiking neuron. How does this change influence the T2* signal? One would think that increased activation will lead to bigger oxygen consumption, more deoxygenated blood and decrease of the T2* signal. However, it turns out that the vascular system overcompensates for the oxygen demand and the observed change is an increase of T2* signal due to the increased blood flow and deoxygenated blood being temporarily displaced by oxygenated blood. PET studies using H_2O^{15} and fluorodeoxyglucose have shown that the regional Cerebral Blood Flow (rCBF) and the regional Cerebral Metabolic Rate (rCMR) increase during activation phases (neuronal stimulation). However, rCBF and rCMR are not completely coupled. During activations, the rCMR increase by 5% but the rCBF increases by 29% (Fox and Raichle, 1986).

The neurovascular coupling is not fully understood, but the results obtained from fMRI have been shown to be both reliable (Smith et al., 2005) and in agreement with other measures of neuronal activity (Logothetis et al., 2001). Therefore, there must be a form of signalling mechanism that would inform the vascular system of a local demand for nutrients. This signal would have to originate from spiking neurons. The current most widely believed theory about this pheonomenon involves mediation of astrocytes. A side product of glycosis, namely lactate, is hypothesised to trigger astrocytes which interact with the vascular system causing vessel dilation (Pellerin and Magistretti, 1994). The neuro-vascular relation has been studied extensively and has lead to the discovery of the Hemodynamic Response Function (HRF). It is a function between the neuronal activation (e.g. driven by a stimulus or behaviour) and the observed fMRI signal. The most pronounced characteristic of this relation is the delay. We observe reaction of the vascular system only 2–3 seconds after the onset of the stimulus, with the signal reaching a plateau in 5–8 seconds. In addition the HRF is wide which leads to temporal smoothing and can be a limiting factor in terms of temporal resolution.

Nonetheless, with appropriately designed stimuli one can elicit a measurable change in the brain. A goal of a typical fMRI experiment is to find neuronal correlates of cognitive functions, mental states or interaction between mental illnesses and behaviour. Depending on what the question is, subjects are asked to perform different tasks inside the scanner. It is worth remembering that fMRI does not measure absolute values of activation. It only allows differences between states of the brain to be identified. Therefore all of the fMRI experiments are optimized to contrast one behaviour/task with another. For example, a simple task might consist of alternating between looking at a rotating checkerboard and looking at a cross. In such a set-up, differences between the visual stimulus during the task (rotating checkerboard) and rest phase (static cross) should elicit activation in the visual cortex.

2.2.2 *Data processing*

Images produced by an fMRI scanner are three-dimensional cubes of uniformly sampled space. Every point (also called a voxel) is assigned a value, which on its own does not directly correspond to any physical property of the tissue it is depicting. As mentioned before, fMRI studies the dynamics of a process and, because of that, many images (volumes) are acquired in rapid succession. This allows a timeseries

signal to be extracted for each voxel in the brain. If one takes the timecourse of the stimulus (for example, when the checkerboard appeared on the screen and for how long) and convolve it with the HRF, one obtains the expected timecourse of the activation. Brain mapping using fMRI is therefore essentially finding which parts of the brain elicit this expected temporal pattern.

However, before statistical analysis of the timecourse can be performed several preprocessing steps are necessary. The typical preprocessing pipeline involves slice time correction, realignment possibly artefact removal, and finally coregistration with an anatomical image. After, discarding the first few volumes to allow for magnetic saturation effects to reach a steady state, one has to deal with inaccuracies and artefacts that might arise from the different acquisition times for different slices within the same volume. Most modern clinical scanners are not able to take a "snapshot" of the whole brain at once, and multiple two dimensional slices are acquired separately instead. Therefore the fMRI signal is sampled at different timepoints depending on the slice being imaged. Slice time correction shifts and interpolates the signal so all slices from the same volume are sampled at the same point in time. This correction depends on the order in which slices are acquired which can be different for different scanners, and usually slice order is not sequential to avoid signal cross-talk effects.

The next stage of preprocessing (realignment) deals with potential head movement of the patient. Even if the patient's head is fixed to the table and/or tightly cushioned some movement is unavoidable. This is especially true for paradigms involving overt speech and for children or patients experiencing micro tremors or seizures. If the patients head moves between time point t_1 and t_2 during the scan it means that for any given location (x,y,z) within the volume, the signal value at t_1 does not correspond to the same part of the brain as the signal measured at t_2. To correct for this, an affine transformation (translations and rotations in the x,y,z directions) is applied to minimize the difference between volumes. Various cost functions exist in order to quantify this difference, and their choice depends mainly on the software used, but differences between techniques are minimal.

Based on the realignment and intensity information one can try to find volumes that most likely contain artefacts and therefore cannot be trusted. This can be caused by extreme movement or temporary scanner coil failure during the EPI sequence. Volumes detected this way can be discarded or replaced by an average of the previous and the next volume. Volumes affected by "large" motion or shift in global intensity are also often 'marked' during the statistical analysis.

Finally to be able to visualise the result on a clinically relevant T1- or T2-weighted volume one has to transform the fMRI volumes to the dimensions of the T1 or T2-weighted volumes and realign them. This process (coregistration) is similar to realignment. It consists of finding an affine transformation (again with 6 degrees of freedom) that minimizes differences between two volumes: T1 or T2 weighted images on one hand and the mean of the fMRI timeseries on the other hand. The main difference from realignment lies in the cost function. While realignment can be performed using linear functions, because images have all the same scales, non-linear functions like mutual information must be used for coregistration since images have often different scales (e.g. ventricles appear white on T2* fMRI data but dark on T1, thus a linear distance cannot be used).

After data preprocessing, the signal is analyzed using statistical procedures, leading to a Statistical Parametric Map (SPM). There are manny different procedures achieving this goal but we will focus here on the most common approach, namely the fitting of a General Linear Model (GLM) (Friston et al., 1994). First, a design matrix X is constructed. It consists of the stimuli/task timecourse convolved with the HRF. Additionally it can include motion (i.e. parameters estimated during motion correction) and artefact regressors. It is expressed in the timescale of the volume acquisition (in other words X has as many rows as there are volumes acquired). With the design matrix X and the timecourse Y from one voxel one can see that:

$$Y = \beta\, X + \epsilon \tag{2.3}$$

where ϵ is the normally distributed noise. Thus one can estimate parameters using the least squares method by simple inversion:

$$\hat{\beta} = (XX^T)^{-1}X^TY \tag{2.4}$$

These parameters ($\hat{\beta}$) show how strong the relationship is between the signal and regressors. Their linear combination c (also called a contrast) divided by the noise ($\sigma^2 = (\epsilon^T\epsilon)/df$) leads to a t statistic:

$$T = \frac{c^T\hat{\beta}}{\sqrt{\sigma^2 c^T(X^TX)^{-1}c}} \tag{2.5}$$

This is calculated separately for every voxel and leads to a three dimensional SPM[2].

2.3 FMRI AND PLANNING SURGICAL PROCEDURES

The clinical potential of using fMRI was investigated soon after its conception (Ogawa et al., 1990). Jack et al. first showed that fMRI could be used to map motor cotices of tumour patients (Jack et al., 1994). In their proof of principle study using two subjects who were experiencing seizures caused by tumours, they showed the potential of the new imaging technique to improve the safety of surgical procedures. Many others studies soon followed confirming these findings. Even though this study of tumour-affected sensory-motor activation provided an early benchmark, it was soon shown that fMRI could also be used for language mapping. The first attempts at this investigated the use of fMRI to replace an invasive intracarotid amobarbital procedure also known as the Wada test. In this test, amobarbital is injected in one carotid artery at time, which causes impairment of neuronal processes in the ipsilateral hemisphere. This, coupled with behavioural tests can be used to establish the dominant hemisphere in the context of language and memory. In a novel study on seven epilepsy patients, Desmond et al. (1995) have shown strong agreement of fMRI data with the Wada test. FitzGerald et al. were one of the first to map Wernicke and Broca areas using five different paradigms tested on 13 subjects (FitzGerald et al., 1997). Their findings were confirmed in the operating room using ECS.

2 For detailed overview of fMRI data processing see Friston, 2007; Ogawa and Sung, 2007; Poldrack et al., 2011.

With time, studies investigating the use of fMRI for presurgical planning improved in terms of the number of subjects. Schlosser et al. (1997) compared fMRI maps for 24 patients with arteriovenous malformation and concluded that it provided useful and reliable information on the location of the eloquent cortex. Krings et al. (2001b) also addressed the question of how reliable fMRI is terms of finding eloquent cortex. In their extensive study of 103 patients they managed to find motor cortex in 85% of the cases.

fMRI has been proven to show reliable information about cortical activation, but the most important question from a clinical point of view is how this information could change and improve clinical practice. Lee et al. (1999) attempted to tackle this difficult question in their retrospective study of 32 tumour and epilepsy patients. In 89% of cases, they found that presurgical fMRI was useful for either checking the feasibility of the procedure, planning the resection or selecting patients for awake ECS. The same study has shown that fMRI was also found to be useful for assessing whether patients should undergo ECS in 52% of tumour cases and 78% of epilepsy patients.

Another important issue is how to interpret activation maps produced using fMRI. From a neurosurgeon's point of view this fundamentally relates to the margin of safety. In other words, this is the minimal distance between the resected area and the eloquent cortex that does not lead to postoperative neurological deficits. Initial studies have shown that if the tumour is 10 mm or further from fMRI activation the risk of neurological deficits is significantly lower (Hå berg et al., 2004). This was further confirmed by Krishnan et al. (2004). They have found that there were significantly more neurological deficits in patients with partial resections when the distance between the tumour and fMRI activation was lower than 5 mm. They concluded that in case where the tumour was 10 mm or closer to the activated region ECS should be performed. This, however, shows that in cases where the tumour margin is far enough away, ECS is not necessary and fMRI can be a valid replacement. This potentially risky claim was validated in a retrospective study on 25 patients with low grade gliomas (Hall et al., 2005). In this study patients did not undergo ECS and resection planning and execution was performed purely using fMRI data. Tumour progression was also slowed down: there were no signs of it within 25 months after surgery. Despite the fact that ECS was not used to direct the surgery it did not impair the outcome of the procedure. Only two patient experience temporary deficits and those were only temporary.

fMRI based cortical mapping was also validated against the well established ECS technique. In general, studies found good agreement between ECS and fMRI (Hirsch et al., 2000). Studies report that between 83% (Majos et al., 2005) and 92% (Lehéricy et al., 2000) of the ECS stimulation points are also found using fMRI based on sample size of 33 and 60 patients respectively. However, comparison between fMRI and ECS is very challenging. First and foremost the two techniques are mapping different areas. ECS is showing only the essential brain areas for certain behaviour and fMRI is mapping all brain areas involved, even the non-essential ones. Additionally, both techniques have intrinsic thresholds. ECS might not elicit a reaction with too low amplitude, while, as discussed in the next section, the fMRI inferred activation area changes border location based on the statistical threshold used. Depending on the selection of these two thresholds one can find good or poor agreement between fMRI

and ECS. Lastly, one has to be certain about the locations of the ECS stimulations in the fMRI scans. This is a non-trivial problem because of the non-linear shift of the brain due to craniotomy and brain resection.

Despite methodological difficulties in direct comparison with ECS, fMRI provides many advantages. It is able to map subcortical areas in a non-invasive way; ECS requires partial resection to uncover subcortical tissue. Additionally, because fMRI mapping can be done before the procedure, it provides valuable information for planning the craniotomy location and resection extent. The later is mostly related to the location and distance of the tumour from the eloquent cortex. As mentioned before, radical resections are more likely to give better results in terms of life expectancy, but at the same time they provide higher risk of postoperative deficits. fMRI mapping can be used to asses this risk for different resection scenarios and allows the neurosurgeon to discuss different options with the patient before the procedure. ECS is usually performed just before the resection and does not give such an option. Additionally, ECS gives a picture of the eloquent cortex, which is limited to the size of the craniotomy. It also makes the procedure longer and involves waking the patient. It is very rare, but the experience of waking up in an operating room can cause shock to the patient and lead to premature termination of the procedure. ECS can also cause seizures in 5–20% of cases (Sartorius and Wright, 1997). These seizures have to be carefully monitored and can be chemically managed, but nonetheless provide another risk factor.

Patients performing behavioural tasks during ECS are often confused and still under some influence of the global anaesthesia. This can lead to failure to follow the instructions or alteration of the normal neuronal response. Admittedly this issue is most likely not a problem for basic tasks such as finger tapping, but when cortical mapping moves towards more complicated paradigms, which aim to identify regions performing higher cognitive functions, it might become a bigger problem.

2.4 THRESHOLDING STATISTICAL MAPS IN THE CONTEXT OF PRESURGICAL PLANNING

The following work was done in collaboration with A. Golby, L. Soleiman, and L. Rigolo and was presented at ISMRM 2011 (Gorgolewski et al., 2011a).

As mentioned above, fMRI measures how much the signal is similar to the expected one. The expected signal is estimated using experimental design (i.e. the onset and duration of the tasks performed by the subject) and assumptions about the HRF. This correspondence between the measured and expected signal is determined for each voxel leading to a statistical map of the brain. This map essentially shows how much each voxel was involved in a particular mental/behavioural task.

However, to establish "safety margin" and make the interpretation of the imaging data easier for clinicians, one has to transform a continuous statistical map into a binary map. This essentially means giving each voxel one of two labels: "active" (i.e. "not safe to remove") or "not active" (i.e. "safe to remove"). Such process is known as thresholding, however, one has to keep in mind that it can take a more elaborate form than labelling voxels just based on their statistical value. The location of each voxel and the statistical values of its neighbours also has to be taken into account.

As one can imagine, given the same statistical map, two different thresholding methods (or even the same thresholding method with different parameters) can lead to different binary maps. This in turn can lead to a different estimation of the distance between the tumour and the eloquent cortex. Such difference can lead to different clinical decisions (see Figure 2.1). For example, depending on the threshold used, patient might or might not get ECS mapping during the surgery.

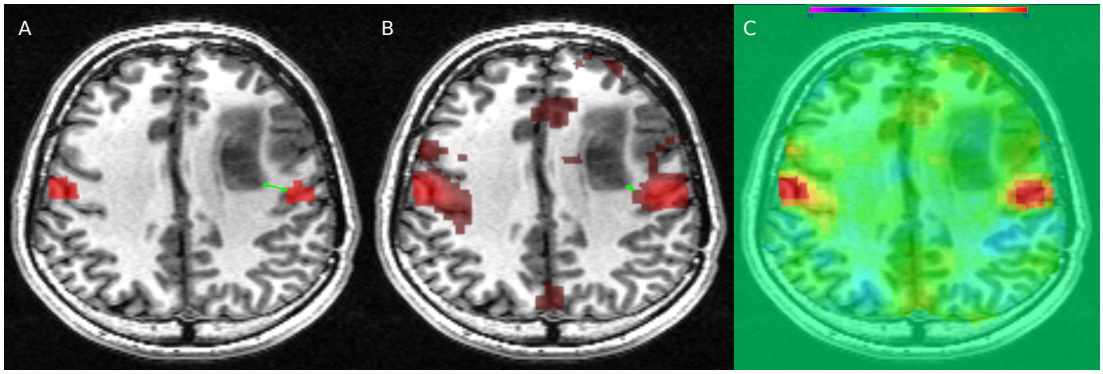

Figure 2.1: Influence of thresholding on the distance between the eloquent cortex and the tumour. A low threshold (A) results in bigger activation area and smaller distance from the tumour. A high threshold (B) results in small region and a bigger distance from the tumour. Panel C shows an semi transparent unthresholded map overlaid on the structural image.

One of the most discussed issues in thresholding statistical maps is the multiple comparison problem. Each voxel individually and independently gets assigned a statistical value based on its timecourse. However, when the number of measurements increases the chances of finding at least one accidental finding, also called Family Wise Error (FWE), increase. This is equivalent to the question, what is more likely: getting at least one head in 10 coin flips or in 100? If you are going to test the same variable multiple times, it is more likely you are going to find a certain value (even if its rare) at least once than if you would test it only once. The fact that variables (voxels) are samples from the same distribution means that you are testing the same quantity multiple times.

In the past two decades, this issue has been acknowledge by the neuroimaging community and many methods for correcting for multiple comparisons have been introduced (Nichols, 2012). This is, however, true only for the use of neuroimaging for basic science (i.e. making inferences how the brain works in healthy or diseased populations). When it comes to using fMRI for presurgical planning the situation looks different.

There is significant variability in the selection of methods used to threshold fMRI activation maps acquired for brain tumour presurgical planning. In a review of 50 recent papers (Gorgolewski et al., 2011a), only 12% of studies claimed to use some kind of FWE correction for multiple comparison testing. 42% of these studies used a p-value threshold lower than the standard 0.05 in an attempt to minimize the number of false positives, while 22% did not use the same threshold for all of the subjects, with the threshold being manually adjusted on per subject basis. What is more, most of the studies used simple thresholds without taking the spatial properties of the statistical maps into account; only 4% used cluster size as an additional threshold

(see Figure 2.2). Although thresholding methods used in neuroscience fMRI studies have greatly improved in the last few years, there still remains a lack of consensus in the clinical literature about how to identify activation boundaries accurately and objectively. As a first step towards developing robust frameworks for assessing thresholding methods for tumour resection, we investigated how several automated thresholding methods affect the distance between the activation areas determined from fMRI experiments and the tumour boundary defined on structural MRI, and how this data might potentially change surgical practice.

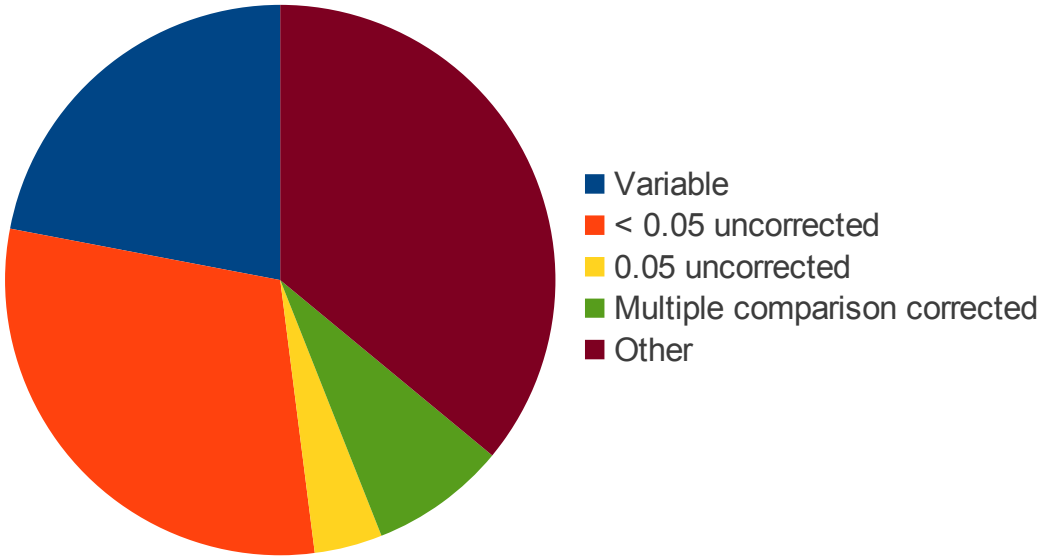

Figure 2.2: Thresholding methods used in studies concerning the use of fMRI for presurgical planning.

11 patients with primary brain tumours situated near the motor cortex underwent a hand clenching fMRI task. The hand used was always contralateral to the tumour location. After fitting a GLM, T-values were calculated for every voxel. These T-maps were thresholded in three ways: (i) manually by an expert rater with additional cluster extent threshold of 10 voxels, (ii) using SPM8 with a cluster forming threshold of 0.05 FWE corrected and False Discovery Rate (FDR) corrected[3] of clusters extent probability (assessed using Random Field Theory (RFT)) lower than 0.05 (SPM)(Chumbley and Friston, 2009), and (iii) using FSL and Spatially Regularized Mixture Models with a 0.5 probability threshold of belonging to the activation class (FSL)(Woolrich et al., 2005). For each subject, the distance between the activation area and the manually segmented tumour was determined by an image analyst and verified by a neurologist. In the case where the tumour margin and the activation area overlapped, a Dice's similarity coefficient was calculated (see Table 2.1).

Determining what the appropriate distance from the tumour margin is for surgical resection based on fMRI activation maps is not straightforward (Hå berg et al., 2004). We therefore determined how a hypothetical 'safety margin' of 5, 10, 15 and 20 mm would influence the clinical procedure. Specifically, if the activation region is

3 A more detailed review of existing thresholding techniques can be found in Chapter 3.

Table 2.1: Distance between the edge of the activation areas and the tumour margin. Dice's similarity measures are calculated in case of an overlap.

	SPM	FSL	Manual
Patient 1	0.23% overlap	0.10% overlap	8.4 mm
Patient 2	2.46% overlap	0.18% overlap	3.07% overlap
Patient 3	2.82% overlap	0.16% overlap	2.39% overlap
Patient 4	3.58% overlap	0.01% overlap	0.08% overlap
Patient 5	11.87 mm	25.61 mm	29.18 mm
Patient 6	13.52% overlap	0.82% overlap	24.91% overlap
Patient 7	1.68% overlap	0.27% overlap	0.48% overlap
Patient 8	1.83% overlap	12.4 mm	18.31 mm
Patient 9	2.38% overlap	0.18% overlap	.02 mm
Patient 10	7.14 mm	0.00% overlap	7.28 mm
Patient 11	0.01% overlap	0.00% overlap	7.17 mm

further from the tumour than the safety margin then full resection is recommended, otherwise a partial resection should be performed. Additionally, we calculated the theoretical statistical properties of the thresholds generated by the expert rater.

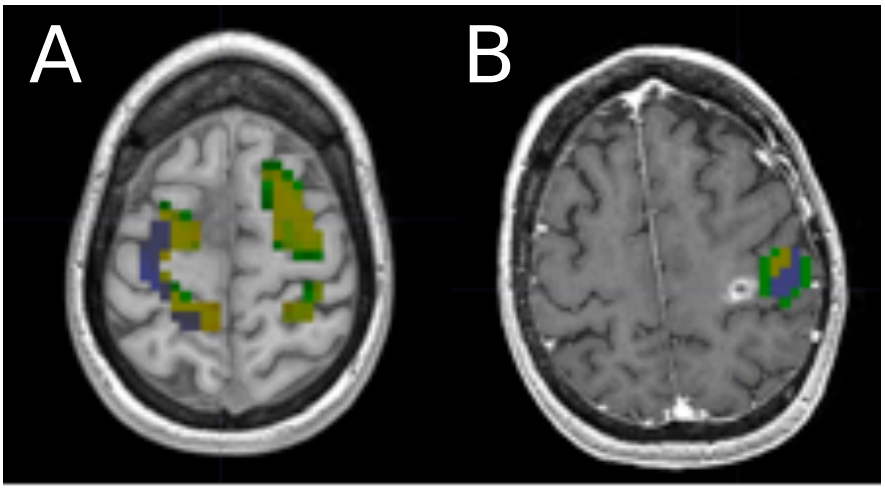

Figure 2.3: Patient 1 (A) and 10 (B). Voxels are colour coded by the thresholding method they were included by: FSL (green), FSL+SPM (yellow), FSL+SPM+manual (purple).

Figure 2.3 shows two example cases. Table 2.2 shows tumour distance and overlap for all subjects. Figure 2.4 shows that if the safety margin is large enough (20 mm) both automated thresholding methods perform similarly to manual thresholding, which we assume is the "gold standard". However, if the safety margin is reduced to 5 mm almost one third of the cases are classified differently; automated methods tend to produce larger activation regions leading to a partial resection recommendation. Additionally statistical analysis of theoretical properties of the manual thresholds shows that the rater chooses thresholds to minimise the number of false

positive voxels, i.e. FWE corrected p-values lower than 0.002 and FDR lower than 0.0003. However, the clusterwise statistics show that the FDR values in two cases are surprisingly high (Table 2.2). This is due to a fixed cluster size threshold of 10 voxels which does not take into account the height of the cluster forming threshold. This problem can be addressed using modern thresholding methods based on RFT which estimate the expected cluster size for a given cluster forming threshold (Friston et al., 1993).

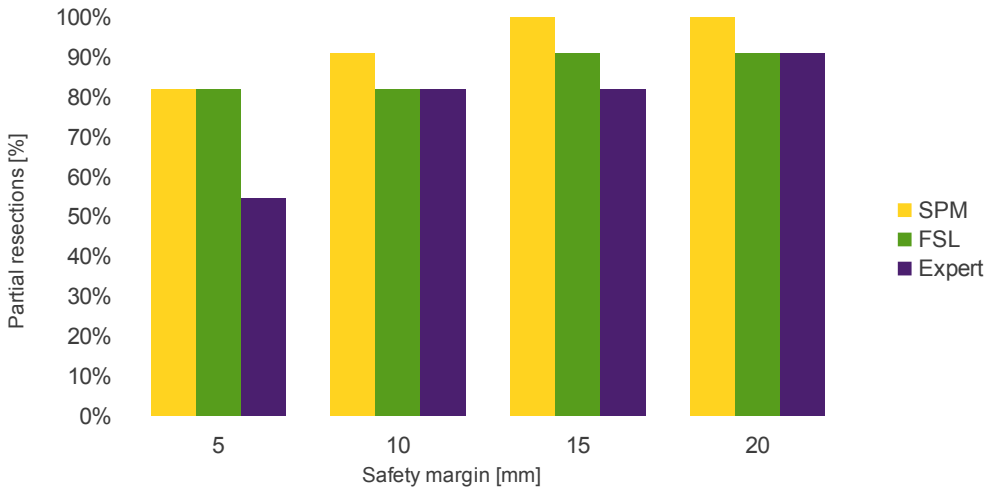

Figure 2.4: Percentage of cases with partial resection recommendation using different safety margins.

The conservative approach of using very high p-values, both in the reviewed papers and presented data, results in very few if any false positives. This, however, means that the number of false negatives (falsely claiming that a piece of tissue is not involved in the particular cognitive task and can be safely removed) will also be very high. Neuroscience approaches focus on controlling the number of false positives to ensure that the reported findings, if any, are true. In neurosurgery we are in fact interested in the opposite question —- which parts of the brain are not involved in a particular task and can be safely removed. Further research is needed to investigate the use of equality testing in presurgical mapping using fMRI data to create "safety maps", rather than mapping functional brain regions. It is also important to note that as long as the person making clinical decisions based on thresholded maps, presumably the neurosurgeon, understands how they were prepared he or she can adjust the safety margins appropriately. In case of manual thresholding even if it is very conservative it can be a successful base for procedure planning as long as the expert thresholding it is consistent.

The problem becomes more significant when the fMRI cortical mapping is prepared by an outside centre. Such situations call for standardization in the way thresholded fMRI activation maps are prepared and data to inform the neurosurgeon about

Table 2.2: Statistical properties of the manual thresholds. FWE corrected voxelwise P — using Bonferronni correction (P_{Bonf}) or RFT based correction (P_{RFT}), voxelwise FDR (FDR), and clusterwise FDR (cFDR).

	thr	P_{Bonf}	P_{RFT}	FDR	cFDR
Patient 1	10	0	0	0	0
Patient 2	9	0	0	0	0
Patient 3	10.3	0	0	0	5×10^{-6}
Patient 4	10	0	0	0	0
Patient 5	6	0.0006	0.002	4×10^{-5}	0.005
Patient 6	13.1	0	0	0	0.803
Patient 7	17.5	0	0	0	0.789
Patient 8	7	6×10^{-6}	4×10^{-5}	0	10^{-6}
Patient 9	10	0	0	0	0
Patient 10	5.7	0.0018	0.005	2×10^{-4}	0.003
Patient 11	9	0	0	0	0

the statistical properties of the presented data, such as the expected number of false positives and false negatives.

2.5 SUMMARY

In this chapter we have explained the physical principles of MRI. We have also briefly discussed the physiological basis of the fMRI signal, namely the magnetic properties of oxygenated versus non-oxygenated blood. Acquiring signal using an MRI scanner is only the beginning of the journey towards usable information. In the second part of this chapter we discussed the state of the art for using fMRI for presurgical planning. We put special focus on the last step of data processing, namely thresholding of the statistical maps. As we show in the literature review there is a lack of a consensus how to perform this step when using fMRI for presurgical planning. In the following chapter we introduce a novel technique for thresholding single subject fMRI statistical maps.

3

ADAPTIVE THRESHOLDING

3.1 INTRODUCTION

This work has been presented at OHBM 2011 (Gorgolewski et al., 2011c), ICML Workshop on Statistics, Machine Learning and Neuroscience 2012 (Gorgolewski et al., 2012c) and published in Frontiers of Human Neuroscience (Gorgolewski et al., 2012d).

After appropriate data pre-processing, a GLM is fitted to the measured signal and a T-test looking for differences between conditions or between a given condition versus rest is performed. As presented previously in Chapter 2, the final outcome from this analysis is a SPM, that is a 3D volume of T-values. Given these T-values, each voxel is labelled as being "active" (involved in the task) or "not-active" (not involved in the task) based on an *ad-hoc* threshold. This procedure has been successfully used in the context of cognitive neuroscience group studies for population inference. However, three major problems need to be addressed in order to improve inference at the subject level when used for clinical decision making (such as presurgical mapping), namely: (i) the impact of Signal to Noise Ratio (SNR) on thresholding, (ii) the relative importance of Type I versus Type II error rates, and (iii) the spatial accuracy of the thresholded maps. In this chapter, we investigate how these issues affect statistical maps and describe a new adaptive thresholding method which improves cluster detection and delineation.

SNR is usually higher in group studies than in single subject fMRI. In group studies, one averages the effect (beta parameters of the GLM) observed in multiple subjects, which usually leads to a stronger signal than that obtained for just one subject. In addition, statistical significance is assessed in comparison to the between subject variance, which is less dependent on scanner related noise than within subject variance. In single subject analyses, the effects are usually estimated on a single set of scans with comparison to the between scan variance. In this context, the SNR can be low due to scanner noise with potentially high between scan variance. This is particularly true in the clinical context in which patients are often advanced in age or impaired by medical conditions (Stippich et al., 2007), resulting in reduced scanning time (less signal) or increased motion (more noise). In consequence, researchers often threshold single subject maps manually based on prior anatomo-functional knowledge and expectations (O'Donnell et al., 2011) rather than using the signal properties or the statistical values. Such a liberal approach is problematic as it may prevent reliable (reproducible) results. Depending on the researcher, clinician, or radiologist, different thresholds will be used leading to different inferences. Single subject fMRI analyses thus require a thresholding method that gives more reliable results.

Cognitive neuroscience group studies have focused on avoiding false positives, whereas in the clinical context, false negatives are also an issue. The biggest concern of the researcher or clinician using fMRI is validity, i.e. is the brain activation that is observed real or an artefact? Statistical methods reflect this point of view by control-

ling for the probability of a false positive error, i.e. reporting an activation that is not present. By contrast, the goal of a surgical procedure such as tumour resection is to remove as much diseased tissue as possible while preserving mental and cognitive capabilities. In this context, surgeons are not only interested in delineating eloquent cortical areas, but also in delineating the tissue that is not involved with a particular cognitive skill. Therefore, the error of reporting an area as not active, and safe to cut out, when in fact it is active (a false negative error) has more profound consequences than a false positive error. In single subject fMRI which is used for clinical decision making, it is thus more important to have a method that provides a good balance between the two expected error rates rather than one that controls perfectly for only one of the two error rates.

The spatial extent of active areas is also of greater importance in the clinical context than in cognitive neuroscience studies. In the latter, it is often sufficient to answer the question of where certain neuronal processes take place in an average brain. As a consequence, many publications report only the peak coordinates of activation. However, in the clinical context, the precise location matters. In the case of presurgical planning for example, decisions about the safety of the procedure and extent of the resection are made based on the distance between a tumour and the eloquent cortex as revealed by fMRI. The statistical threshold used influences this distance by changing the spatial extent of activated areas, whilst it usually doesn't impact on the peak location. Therefore, the thresholding method used in single subject analyses must allow a good delineation of the true underlying signal extent.

3.2 REVIEW OF EXISTING THRESHOLDING METHODS

3.2.1 *Voxelwise thresholding, Family Wise Error and False Discovery Rate*

As already mentioned, a glsGLM is fitted to every voxel separately resulting in an SPM. Each voxel now contains e.g. a T-value (though F- or Z-values are also possible) that can be compared to a theoretical distribution (given the degrees of freedom) and thus, each voxel can be assigned with a corresponding probability value. This probability, often called the p-value, is the probability of the statistical value (T, F or Z) coming from the null distribution (in other words the probability of a false positive). Therefore the lower the p-value (or higher the statistical-value) the more likely it is that the observed effect (relation between the stimuli and observed timecourse) is significant. Following the Neyman-Pearson lemma, one can consider a voxel as being active (significant) if it has a p-value below an ad-hoc threshold, otherwise this voxel is said to be inactive (non-significant). For historical reasons, voxel-wise thresholding has been performed at $p < 0.001$.

Although this approach is simple, it only works if one considers a single voxel. In the typical volume sizes used in neuroimaging, there are thousands of voxels, leading to thousands of tests. Due to random noise each test has a small but non-zero probability of giving a false result. Chances of making such mistake in terms of one test are acceptable, but the more tests we make the bigger is the probability of finding some extremely high (or low) values driven only by random noise. Because of these issues the probability of obtaining at least one false positive is not 0.001 but much higher. To avoid such inaccuracies a series of corrections have been developed.

The first and probably most known correction for multiple comparisons is the Bonferroni correction (Dunn, 1961). This procedure controls for FWE, which is the chances of making at least one false positive error in a family of many tests/comparisons. The correction works by defining the p-value threshold as α/n where α is the desired FWE level and n is the number of tests. Because in neuroimaging context, this would be the number of voxels, this correction tends to be overly conservative. The notion of FWE does not, however, appeal to all users. Value of probability of at least one false positive among all tests is hard to use when inspecting one of the tests; having used FWE correction what is the chance that this particular voxel is a false positive? To address this issue another approach to multiple comparison correction was introduced, namely FDR. This is the ratio of false positive tests to the sum of false positive and true positive tests. There are various procedures to control for FDR, but the most popular is the top-down approach introduced by Benjamini and Hochberg (1995). This technique was later introduced to neuroimaging by Genovese et al. (2002).

3.2.2 *Clusterwise thresholding, topological inference and Random Field Theory*

All of the approaches mentioned above were focused on voxels. When we talk about false positive errors it means voxels that should have been labelled inactive but were not. However, fMRI data rarely produce activation maps with sparse pattern of a few separate voxels. This can be explained by the intrinsic smoothness of the data. If one voxel is active it is very likely that voxels surrounding it will be active as well. This lead to the idea that maybe it should not be the voxels what we should be concerned with, but groups of connected (neighbouring) voxels otherwise known as clusters. To address this issue a technique that uses a field of Gaussian processes to establish the probability of a cluster of given size has been developed (Worsley et al., 1992). This technique, also known as RFT, estimates how probable it is by chance that a cluster of suprathreshold voxels are obtained given the cluster-forming threshold (i.e. the voxelwise threshold that has been used to define suprathreshold voxels which can form clusters) and the smoothness of the Gaussian field under the null hypothesis of no activation. The smoothness is estimated from the residuals of the linear model (Kiebel et al., 1999). Alternatively, one can apply an FDR correction to clusters trying to control for the number of false positive clusters (Chumbley and Friston, 2009). The procedure consists in thresholding data at the voxel level without any correction, assigning a p-value to each cluster using RFT and finally deciding which of them to label as active and which to discard by applying the FDR probability threshold. This method is known as topological FDR.

The biggest contribution of these approaches is the change of perspective from single voxels to clusters of voxels. This level is indeed more intuitive to look at since in mapping the brain we are interested in active regions which are usually continuous. There are two things worth mentioning about topological inference techniques. First of all the cluster-forming threshold is still a parameter and even though it does not influence the significance of clusters (since in RFT for higher cluster-forming thresholds smaller clusters are expected) it will influence their size, and as we have shown in Chapter 3 that means according to the threshold the distance between the active area and the tumour will change. Secondly, the null distribution in RFT concerns

lack of activation and the technique is controlling for false positive clusters, but as mentioned in the Introduction, the false negative rate is also of concern in the clinical context.

At this point it is also worth mentioning that cluster probability can be also assigned without using a parametric distribution. The cluster size distribution can be estimated by permutations (Petersson et al., 1999). This, however, is a viable option only for group studies where the assumption of exchangeability of subject labels holds. In case of single subject studies one would have to permute volumes which can be problematic due to temporal correlations.

Additionally RFT can also be used to control for FWE rate on a voxel level similar to Bonferroni correction. This is done by reversing the inference and asking the question: what is the cluster-forming threshold that will yield the probability of a cluster of size with the desired FWE level.

3.2.3 *Mixture models*

From a logical perspective, since in most paradigms used it is expected that some signal is present in the brain, it seems reasonable to analyze the data assuming a signal model (Turkheimer et al., 2004). Mixture models represent the entire distribution of statistical values for a given space as a mixture of the "active" and "noise" distributions. The first application of mixture models to fMRI data was proposed by Everitt and Bullmore (1999). In this initial work, voxels values of an SPM were modelled as a mixture of central and non-central chi-square distributions thus producing a distribution corresponding to no activation, and another distribution corresponding to the presence of some (either positive or negative) activation. After fitting the model, posterior probabilities were obtained for each voxel to be active or inactive and this SPM was thresholded to reveal significantly activated areas. At the heart of such approach is the assumption that signal and noise, in particular the null distribution, can be separated via modelling. This idea was later adopted by others (Hartvig and Jensen, 2000; Woolrich et al., 2005) who incorporated spatial priors to account for the correlation between voxels. Both Hartvig and Jensen and Woolrich et al. used Markov Random Field (MRF) to spatially regularize labelling of the statistical map, although Woolrich et al. were the first to train parameters of the MRF from the data in a Bayesian way. More recently, Pendse et al. (2009) considered a mixture of Gaussians to model the null distribution in an attempt to improve voxelwise FDR. They used Bayesian Information Criterion (BIC) to choose how many mixture components were required to accurately describe the data. However, in this method the inference was carried out on the voxel level without taking into account spatial characteristics of the signal such as cluster size. It also suffers from problems with interpreting which Gaussians correspond to either noise or activation classes.

3.2.4 *Performance comparison*

3.2.4.1 *Methods*

To get an overview of a selection of existing thresholding methods we have performed a series of exploratory toy simulations. Series of eight 2D images of size 128×128 pix-

els and intensity zero were created and 'activations' inserted in the last four images. The activation pattern was systematically varied across the simulations by changing its size and the SNR. The activation was defined as a square of size 16×16, 32×32, 64×64, 80×80, 88×88, 92×92 or 96×96 pixels and intensity of 0.5, 0.75, 1.25 or 1.75. Normally identically and independently distributed noise ($\mu = 0$, $\sigma = 1$) was then added to each pixel of each images and volumes were finally smoothed by convolving with a Gaussian of full width at half maximum (FWHM) of 7 pixels. The SPM software package (`http://www.fil.ion.ucl.ac.uk/spm`) was used to estimate the β parameters of the GLM and obtain T-maps. Note that traditional steps included in fMRI data analysis like convolution with the hemodynamic response function, autocorrelation, high pass filter were omitted because our data did not vary with respect to the properties these steps are designed to account for. Every combination of size and intensity (SNR) of the activation area was run 100 times resulting in 2800 different T-maps.

The T-maps generated using the procedure described above were used to evaluate the following thresholding methods (see Figure 3.1):

1. Voxelwise:
   a) Voxel-wise FDR controlling method (Genovese et al., 2002) with the desired voxel-wise FDR set to q = 0.05.
2. Topological:
   a) Topological FDR on cluster extent with a desired cluster-wise FDR set to 0.05 (Chumbley and Friston, 2009). Two cluster-forming threshold variants were used:
      i. Low: using a p-value of 0.05 (uncorrected).
      ii. High: using a minimum between voxelwise Bonferroni and RFT.
3. Mixture models:
   a) Gaussian mixture model using two Gaussian distributions. Similar to the approach presented by Pendse et al. (2009) but with the threshold set by the crossing point between the two distributions.
   b) Gamma-Gaussian mixture model using one Gamma distribution. As presented by Beckmann et al. (2003), and identical to the approach described by Woolrich et al. (2005), but without the spatial regularization.
   c) Spatial mixture model using a Gamma-Gaussian mixture model with only one Gamma distribution with spatial regularisation based on MRF (Woolrich et al., 2005)

The Gamma-Gaussian mixture model with spatial regularisation was fitted using the implementation included in the FSL software package (`http://www.fmrib.ox.ac.uk/fsl`). Topological FDR related calculations were conducted using the SPM software package. Results obtained by these seven thresholding methods were examined in terms of voxel FDR, cluster FDR and voxel False Negative Rate (FNR):

- Voxel FDR : ratio of false positive voxels to all voxels labelled as active (false positive and true positive)

- Voxel FNR: ratio of false negative voxels to all voxels labelled as not active (false negative and true negative)
- Cluster FDR: ratio of false positive clusters[1] to all clusters

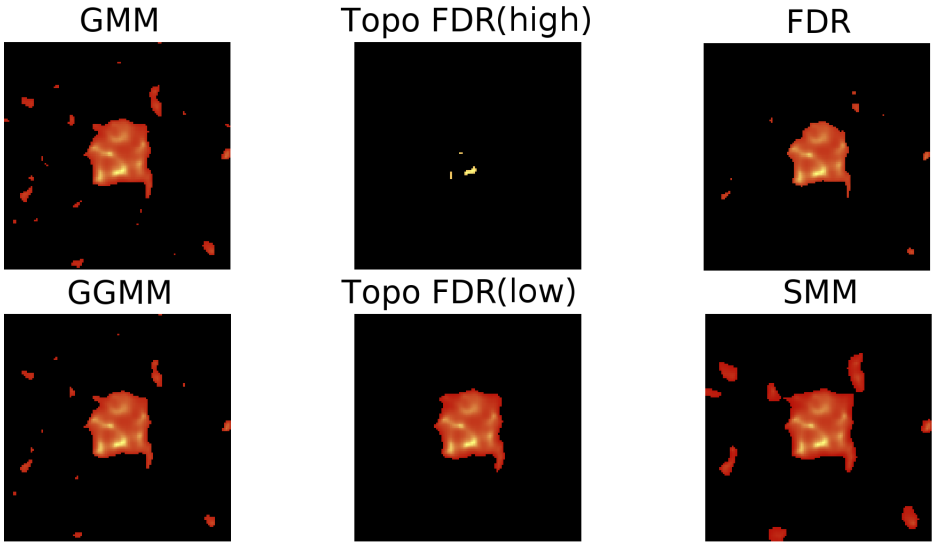

Figure 3.1: Example simulation run (SNR 0.5) thresholded using six evaluated methods.

3.2.4.2 *Results*

Overall, topological methods outperformed the other methods. In terms of voxel FDR, only topological methods gave satisfactory results. Looking at the cluster FDR, again topological methods performed better giving similar results whatever the cluster forming threshold, especially for bigger cluster sizes and SNR.

Among voxel-wise methods, mixture models (Gaussian, Gamma-Gaussian, spatial) performed better than the voxel-wise FDR procedure. The Gamma-Gaussian mixture model performed better than the Gaussian for many cluster sizes and SNR values suggesting a better fit using Gamma rather than Gaussian distributions. The spatial model showed a strong decrease of cluster FDR as a function of the cluster size, as a direct effect of accounting for spatial dependency via MRF.

Looking at the voxel FNR, topological FDR high and the spatial mixture model performed worse than the other methods (Table 3.1), which overall gave satisfactory results.

Because as in (Chumbley and Friston, 2009) we have assumed intrinsic spatial dependency of the underlying signal, the artificial binary maps were smoothed before applying the GLM. This resulted in smooth borders of the activation area, which were the main source of false positive voxels. The voxel-wise FDR and cluster-wise FDR decreased for all of the methods with increasing size of the true activation patch (see

1 A false positive cluster is a cluster of which at least half of the voxel do not belong to a true activation patch.

Table 3.1: Averaged (over sizes and SNR) results in terms of voxel and cluster FDR and FNR for the six methods tested. GMM: Gaussians Mixture Model, GGMM: Gamma-Gaussian Mixture Model, voxel FDR: FDR control over voxels, SMM: spatial mixture model (i.e. Gamma-Gaussian mixture model with regularisation), Topo FDR low and high: FDR control over clusters with low and high cluster-forming threshold. Performances below 0.1 are highlighted in bold.

Method	Voxel FDR	Cluster FDR	Voxel FNR
GMM	0.213	0.8873	**0.0011**
GGMM	0.1769	0.7877	**0.0033**
Voxel FDR	0.1754	0.8708	**0.0024**
SMM	0.2141	0.8708	0.0592
Topo FDR (low)	0.2232	**0.0868**	**0.0008**
Topo FDR (high)	**0.015**	**0.022**	0.1926

Figure 3.2). The decrease in the number of false positive voxels is easily explained by the fact that it is a function of the ratio of the border length ($4 \times a$) and the area (a^2). In other words for larger areas, false positive voxels on the borders have less influence on the FDR than for the small areas. This effect was also reflected on the voxel-wise FNR which increased with increasing size of the true activation patch (see Figure 3.2).

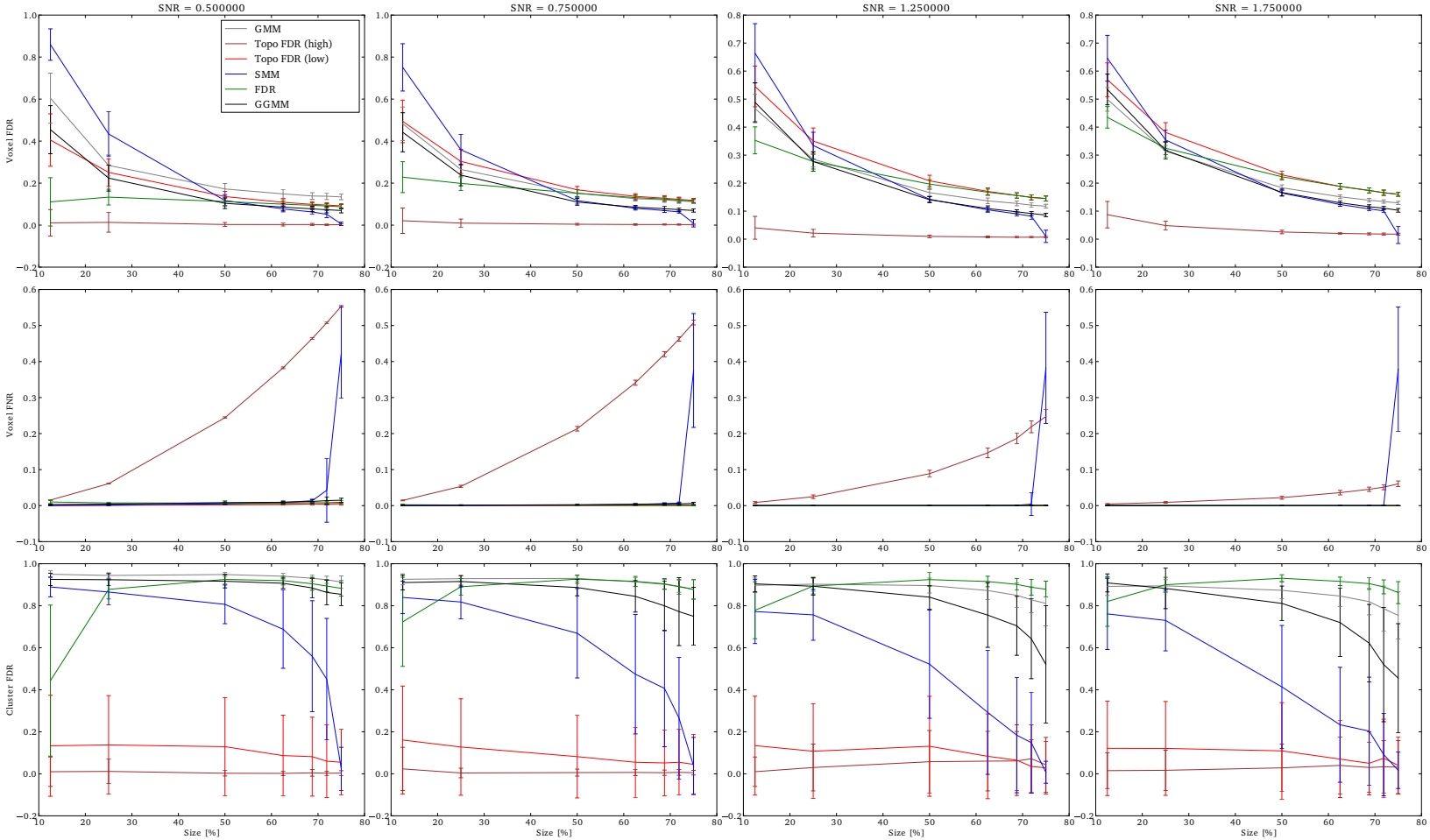

Figure 3.2: Simulation results. Performance (average data over simulations) in terms of voxel FDR, voxel FNR, and cluster FDR for four SNR levels and seven cluster size levels. Whiskers represent standard deviation around the mean. GMM: Gaussians Mixture Model, GGMM: Gamma-Gaussian Mixture Model, voxel FDR: FDR control over voxels, SMM: spatial mixture model (i.e. Gamma-Gaussian mixture model with regularisation), Topo FDR: FDR control over clusters whith low (uncorrected) and high (FWE corrected) cluster-forming threshold.

3.3 ADAPTIVE THRESHOLDING

3.3.1 *Motivation*

Our preliminary simulations show that mixture models excel in finding the right balance between false positive and false negative voxels, but fail in discarding small false positive clusters. Topological methods on the other hand perform well in terms of keeping the number of false positive clusters low, but depending on the cluster-forming threshold they can create a number of false negative voxels. This discovery inspired us to introduce a new method that is a combination of the two above.

The aim of our approach is to perform inference on the cluster level and at the same time provide a good balance between false positive and negative errors in the delineation of activation borders. We therefore propose a Gamma-Gaussian mixture model as a method to account for a distributions of T-values in SPMs (Woolrich et al., 2005) and set a threshold specific to the data at hand. A natural way to determine this threshold is to take the point which separates signal from noise. This point is the crossing between the Gaussian, the model corresponding to no activation, and the Gamma distribution, the model corresponding to positive activations, and provides a good trade-off between false positive and negative (voxel-wise) rates. Finally, once this threshold is established, topological inference via FDR correction over clusters (Chumbley and Friston, 2009) is used to correct for the number of tests performed while accounting for spatial dependencies across voxels, thereby explicitly controlling for Type I cluster rate. This heuristic approach combines advantages of the different methods mentioned above. Specifically it relies on a simple model of the SPM, allows adaptive thresholding, and accounts for multiple comparisons in the context of topological inference.

3.3.2 *Gamma-Gaussian mixture model*

Following Woolrich et al. (2005), the T-value distribution from a SPM covering all brain voxels is modelled using a Gamma-Gaussian mixture model, with the Gaussian distribution as a model for the null distribution (no activation) and Gamma distributions as models for the negative (deactivation) and positive (activation) distributions. Note that due to high degrees of freedom in a typical fMRI experiment, i.e. the number of time points greatly exceeds number of regressors, a normal distribution is good approximation of Student's T-distribution. In practice, three different models are fitted to the data, namely:

1. $P(x) = N(x|\mu, \sigma)$
2. $P(x) = \pi_N N(x|\mu, \sigma) + \pi_A \mathrm{Gamma}(x + \mu|k, \theta)$
3. $P(x) = \pi_D \mathrm{Gamma}(-(x + \mu)|k_D, \theta_D) + \pi_N N(x|\mu, \sigma) + \pi_A \mathrm{Gamma}(x + \mu|k_A, \theta_A)$

with x representing all the T-values, $p(x)$ the probability distribution, μ is the mean and σ the standard deviation of the Gaussian (N) component, k is the shape parameter and θ the scale parameter of the Gamma component(s), and π is the proportion/contribution of each component (N for Gaussian / noise, A and D for Gamma / activation-deactivation).

Model 1 is fitted using maximum likelihood estimator, and Models 2 and 3 are fitted using an expectation-maximization algorithm (Dempster et al., 1977). In all three models, the Gaussian component represents the noise. In Model 2, the Gamma component corresponds to the activations. In Model 3, Gamma corresponds to the activation and deactivation classes. Note that Gamma components are shifted by the estimated mean of the noise (Gaussian) component (the non-spatial model described in Woolrich et al., 2005 did not incorporate such shift). The Gaussian distribution is a natural choice to model noise, while the Gamma distributions have the advantage of being restricted to cover only values above (activation) or below (deactivation) the Gaussian mean (see Figure 3.3). This helps to force these components to fit the tails of the distribution. For each model, BIC is calculated and the model with the highest score is selected. Although only Model 2 can be preferred as some signal is expected, fitting all three models offers much more flexibility. In particular, compared to other approaches (e.g. Pendse et al., 2009), the explicit model selection via BIC allows the case when no signal is present (Model 1) to be determined, and avoids having to attribute subjectively model components to noise or (de)activations, i.e. Models 2 and 3. Similarly, in the case that deactivations are present, the mean of the noise component in Model 2 can be biased because the left tail is not well estimated and so is the positive Gamma component; having an explicit model for this case (Model 3) allows for deactivations to be present without interfering with the threshold. In the case that Models 2 or 3 are selected, each voxel is assigned a label (activation, deactivation and noise) corresponding to the component with the highest posterior probability. In these cases, the highest T-value among voxels belonging to the noise class is chosen as the new cluster forming threshold.

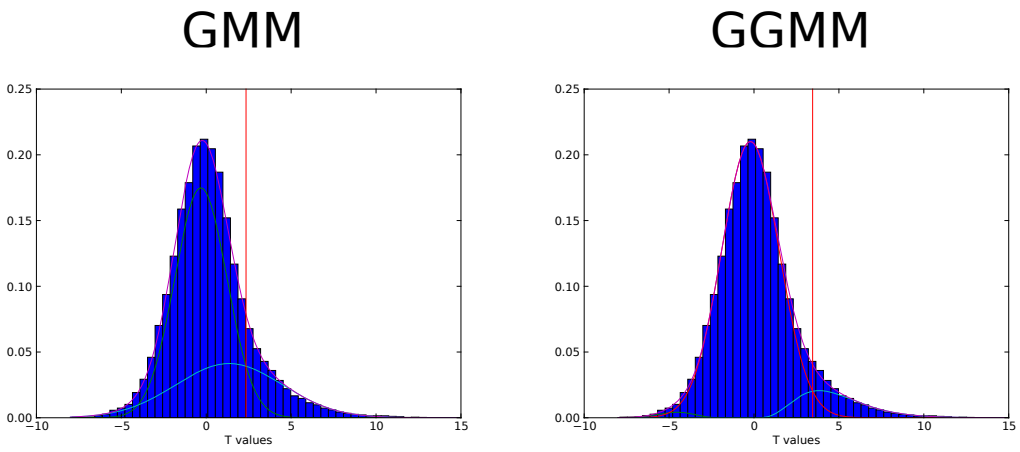

Figure 3.3: Fit of the Gaussian (GGM) and Gamma-Gaussian (GGMM) mixture models for an example dataset. In this example in the GGM model the deactivation component was fitted to the mode of the distribution representing noise. This shows how the GGM model can pose problems with respect to interpretability of the components. The GGMM model, however, provides additional constrains that restrict the activation and deactivation components to the tails of the distribution.

3.3.3 *Thresholding procedure*

Models 2 and 3 allow a probability of being active to be assigned to every voxel. This probability is used to find a threshold that corresponds to a point in which the probabilities of positive Gamma and Gaussian are equal, i.e. the crossing point between the two distributions. This equal probability threshold thus separates signal from noise. At this stage, topological FDR is used to control for false positive clusters (Chumbley and Friston, 2009). In the situation when Model 2 or 3 is selected in the first stage, thus providing evidence of true activation, but none of the clusters survive the topological FDR step, a heuristic threshold is applied to make sure that some activation is found. In this case, the cluster with the highest sum of T-values is labelled as active. We have found that this situation can arise in a few clinical cases, and this heuristic approach solves the issue. An overview of the method can be found in 3.4. A freely available implementation of the method is available at `https://github.com/chrisfilo/Adaptive-Thresholding` for both Nipype (python) and SPM8 (Matlab©).

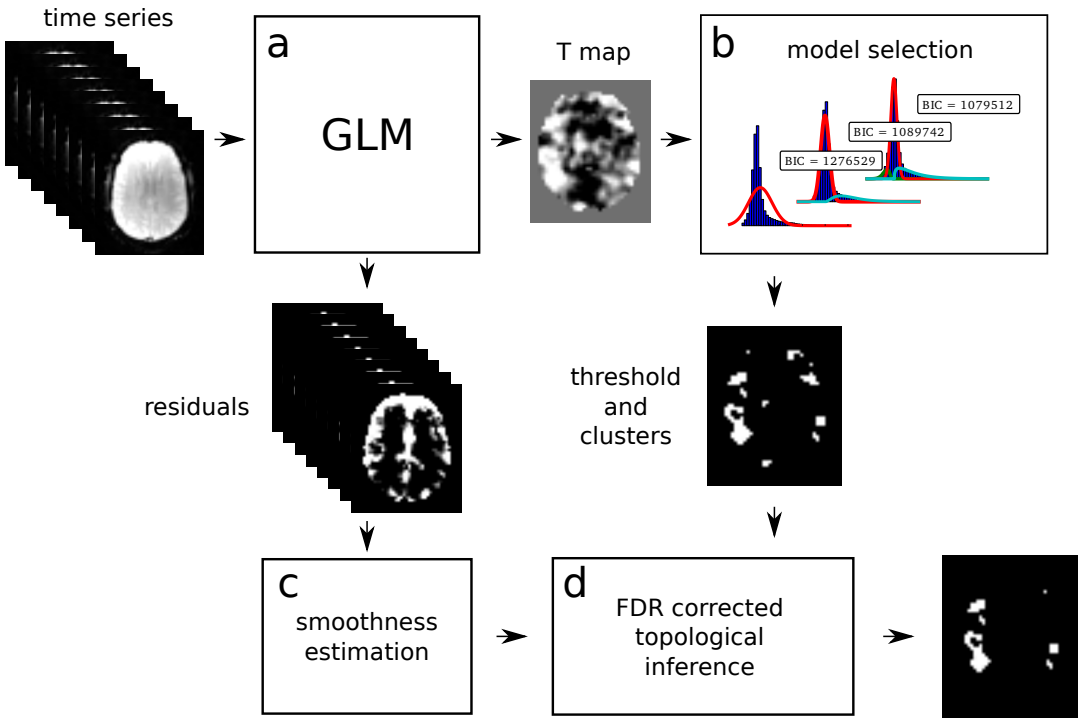

Figure 3.4: Overview of the topological FDR inference using our Gamma-Gaussian mixture model to set adaptively the cluster forming threshold. GLM produces T-maps and residuals (a). Three models are fitted to the voxels from the T-map (b). Models include a combination of deactivation (green), noise (red), and activation (cyan) components. Smoothness of the image is estimated from the residuals (c). Threshold estimated from the winning model (b) and smoothness of the image are used to perform topological inference on cluster extent (d).

3.3.4 *Simulations*

To compare the performance of Adaptive Thresholding to Topological FDR using fixed thresholds, a total of 2500 time series were simulated. Each simulated time series included eighty planes of 128×128 elements. Half of the planes included just normally distributed noise ($\mu = 0$, $\sigma^2 = 1$) and the second half included a pattern of activation added to the noise. The pattern consisted of six squares of different sizes (4×4, 8×8, 12×12, 16×16, 20×20, and 24×24). Because temporal aspects of the fMRI signal such as autocorrelation were not the focus of this research, the time series consisted of only two blocks, namely 40 planes of "rest" followed by 40 planes of "task". All of the planes were convolved with a Gaussian filter of FWHM of 6 mm. The height of the pattern, representing the strength of the signal, was also varied (0.04, 0.08, 0.16, 0.32, and 0.64) and for each of the 5 signal strengths, data (signal+noise) were simulated 500 times (for an example simulation see Figure 3.6).

Time series generated in this way were fitted with a GLM model with a single regressor, and no autoregression, high-pass filtering, or convolution with a hemodynamic response function. Because neither the simulated signal nor the fitted model included any temporal dependencies, the selected design (40 "rest" followed by 40 "task" planes) was no different from any other combination, e.g. 5 "rest" followed by 5 "task" blocks repeated 8 times. A single contrast was estimated and thresholded using topological FDR with 3 different cluster forming thresholds. Two fixed cluster forming thresholds were used across all 2500 SPMs, specifically a p-values of 0.05 with FWE correction (T-value of 4.47) and 0.001 uncorrected (T-value of 3.19). These thresholds were chosen as they correspond to defaults values used in the SPM software package (`http://www.fil.ion.ucl.ac.uk/spm/`) and we refer to them as **Fixed Threshold (FT) (0.05 FWE and 0.001)**. This contrasts with the cluster forming thresholds obtained with the Gamma-Gaussian mixture model which by nature change with the data. Note that for each map, all three Gamma-Gaussian models were always fitted and the model that best described the data according to our BIC was selected to set the cluster forming threshold. In these simulations, Model 2 was always the best model since there was always some signal plus noise, which also showed that the model selection worked. We refer to these thresholds as **Adaptive Threshold (AT)**. These simulations therefore allow the performance of AT and FT to be compared in terms of false positive and false negative cluster rates, spatial accuracy, and influence of global signal variation (see Figure 3.6).

3.3.4.1 *False positive and negative cluster rates*

A false positive cluster was defined as a supra-threshold group of connected voxels that did not overlap to any extent with the squares in the true activation pattern. By analogy, the false negative cluster rate was defined as the rate of true patterns that were not detected, i.e. missed. Comparison of AT with FT were performed in a pair-wise fashion for every simulated time series. First, false positive and negative cluster rates were calculated for all three thresholding methods. Second, the differences (trade-off) between false positive and negative rates were computed. Third, the difference between AT and the two default FT values (0.001 uncorrected and 005 FWE corrected) for the absolute value of the trade-offs were obtained. Finally, a

percentile bootstrap, resampled with replacement of the differences between thresholding methods, was used to estimate p-values and confidence intervals of the mean differences and multiple tests correction was applied using the Benjamin-Hochberg (B-H) method maintaining FDR at the 0.05 level (Benjamini and Hochberg, 1995). Computing the difference between false positive and negative rates allowed testing for the average improvement of AT over the two default FT values in terms of trade-off, i.e. values around 0 mean a good balance between the two types of error. However, if the method gives two very large errors it can still give a good trade-off. We thus also computed the total sum of type I and type II errors, ensuring that AT doesn't lead to overall larger errors.

3.3.4.2 *Spatial accuracy*

Spatial accuracy was defined as the difference between the overestimation and underestimation of cluster's borders, i.e. it reflects if the cluster borders were well delineated. For a given true cluster, the degree of underestimation was defined as the number of voxels that were falsely declared as not active, and the degree of overestimation was defined as the number of voxels that were falsely declared as active. Using these definitions, cluster borders can be simultaneously overestimated (voxels declared active that should not be) and underestimated (voxels declared non-active that should not be; see Figure 3.5). Note that only true positive clusters that were observed in all thresholding methods were used for this analysis to make the count fair between the three thresholds. In addition, each cluster size was analyzed separately. Comparisons between AT and FT were performed in a pair-wise manner using a percentile bootstrap on the Harrell-Davies (H-D) estimates of the median Harrell and Davies (1982) differences. Multiple test correction was applied using B-H method maintaining FDR at the 0.05 level.

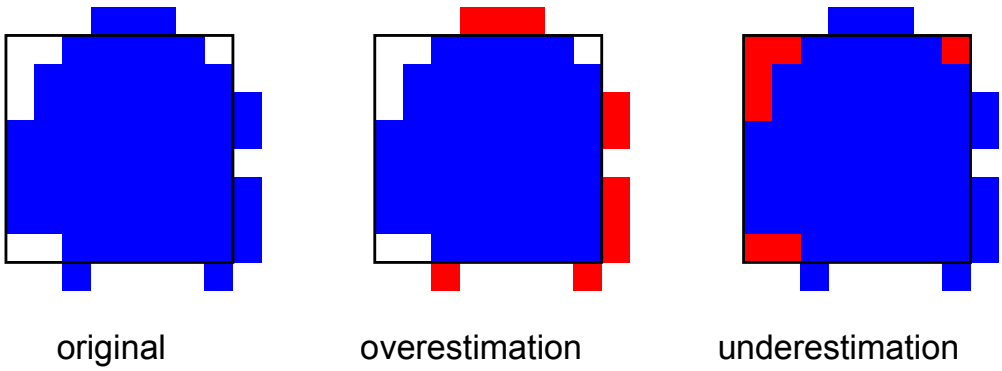

Figure 3.5: Overestimation and underestimation of the border explained. Left: original pattern (black square) and estimated cluster (blue). Middle: overestimation of the border marked in red. Right: underestimation of border marked in red.

3.3.4.3 *Influence of global effects*

One major confound that can influence thresholding results is a global (occurring in all voxels) signal change that is correlated with the stimuli. This has been commonly referred to in the literature as the "global effect" (Friston et al., 1990; Aguirre et al., 1998; Gavrilescu et al., 2002; Junghöfer et al., 2005; Murphy et al., 2009). This global effect results in a shift of all T-values by a constant. We simulated this effect by taking all the T-maps of signal height 0.08 and adding a random constant (normally distributed, $\mu = 0$, $\sigma^2 = 1$) to all values. T-maps created this way were thresholding using AT and the two default FT values. Here only simulations with low SNR were manipulated to investigate the noisiest scenario. Dice coefficients (Dice, 1945) were computed for every simulation between the thresholded shifted and unshifted maps. This allowed the reliability of thresholding methods to be investigated in the context of global effects. Comparison between AT and FT was performed using a percentile bootstrap of the mean of the pair-wise differences between Dice coefficients.

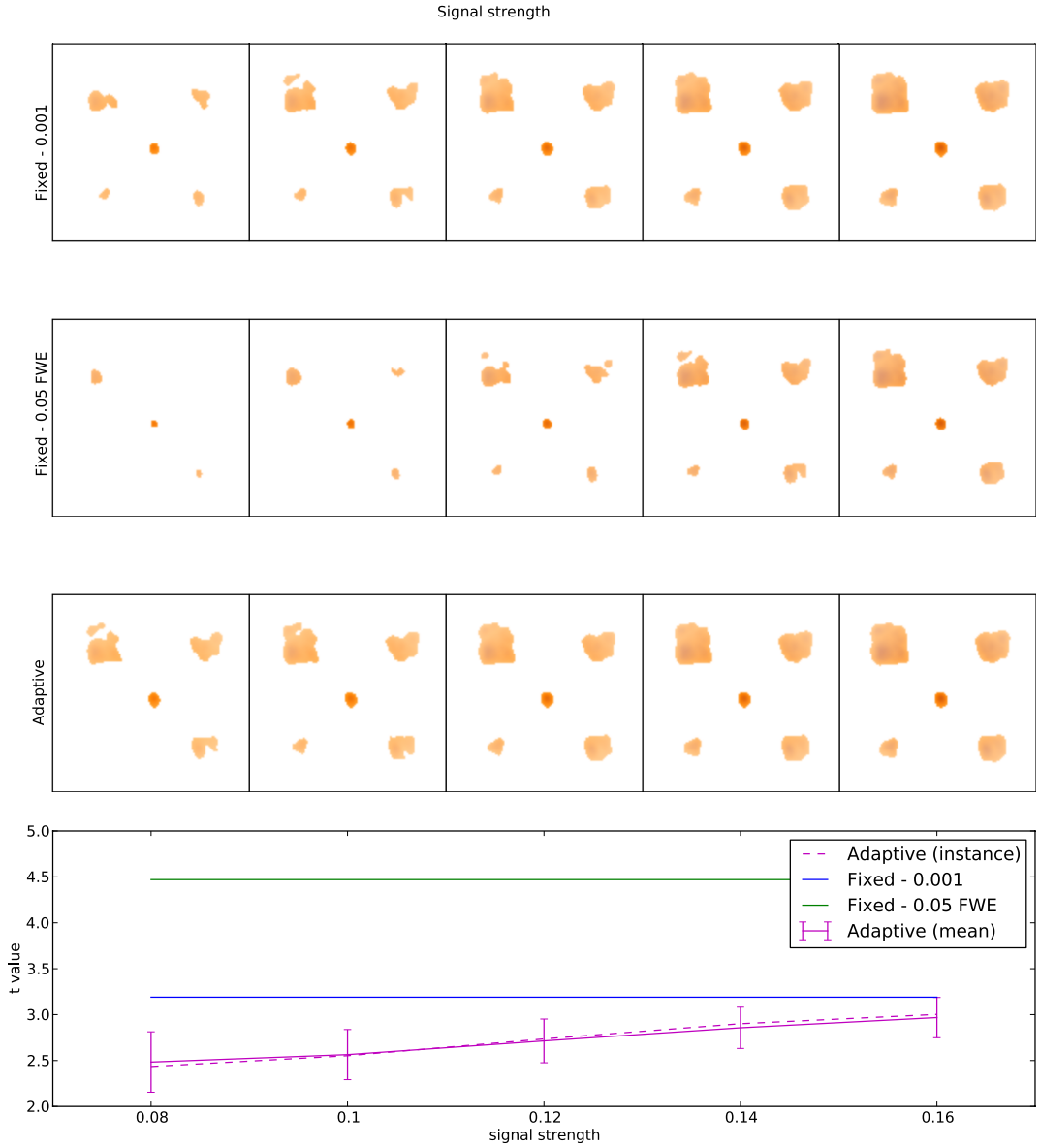

Figure 3.6: Example of the simulated data. Each square shows the effect of the topological FDR procedure with different cluster-forming threshold and SNR level. The bottom row shows how AT methods adjust the threshold according to the signal strength.

3.3.5 *Results*

3.3.5.1 *False positive and negative cluster rates*

In terms of sensitivity or false negative clusters, AT outperformed both default FT values. The difference was largest for lower SNR and a FT of p=0.05 FWE corrected. In the case of FT of 0.001 uncorrected, AT was more sensitive only for SNR values below 0.14 (Table 3.2 and Figure 3.7c). This increase in sensitivity for AT, especially at low SNR, also came with a higher number of false positive clusters than FT (see Figure 3.7a). However, this increase in false positive clusters was comparatively small to the gain in sensitivity such that the total number of errors was similar to FT; in fact even better than FT in most cases (see Table 3.3). Statistical analysis of the differences between false positive and negative clusters shows that AT has a better trade-off than both default FT values (Table II), with the biggest advantage for low SNR values. With high SNR, AT and FT (0.001 uncorrected) gives similar results (Figure 3.7d).

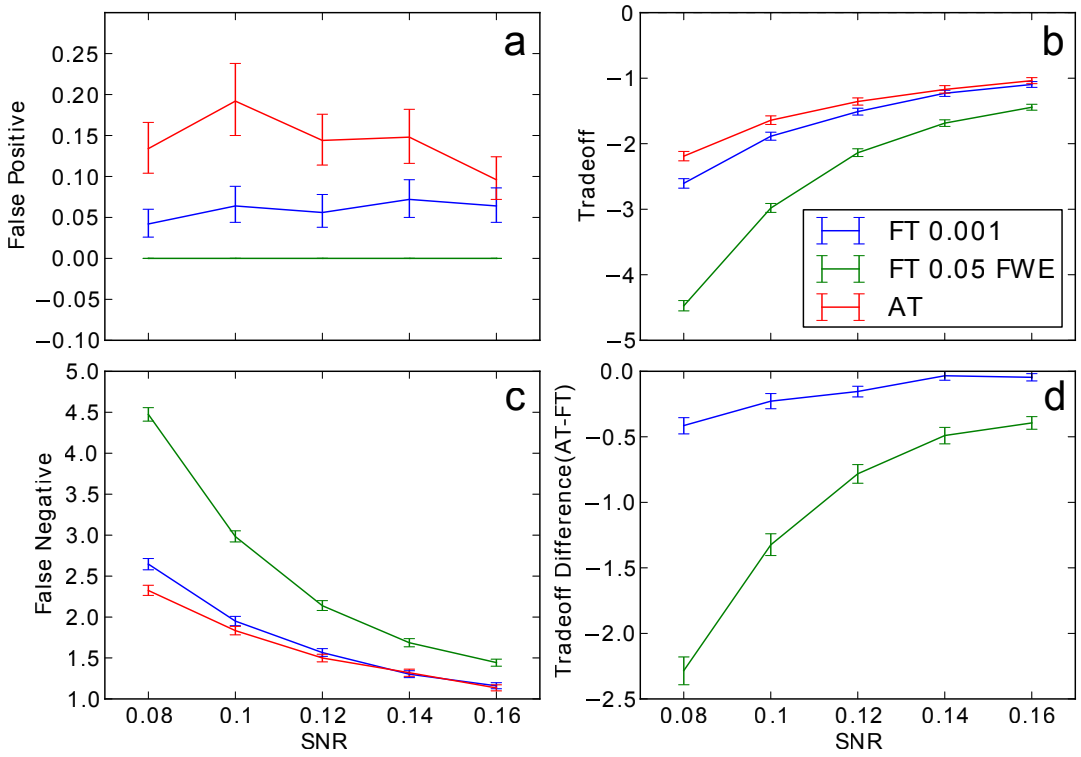

Figure 3.7: False positive and negative cluster rates. On the left are displayed the mean false positive and negative cluster rates. On the right are displayed the mean clusters trade-off and the difference between AT and the two default FT values for this trade-off. Whiskers represent 95% confidence intervals estimated using a percentile boostrap for each SNR independently (uncorrected for multiple comparisons).

3.3.5.2 *Spatial accuracy*

Due to the fact that the smallest cluster was found by all of the thresholding methods in only a handful of runs, it was excluded from further analyses; in other words there

Table 3.2: Statistical analysis the pair-wise difference (AT-FT) comparison of cluster error trade-off. Q-values correspond to p-values corrected for multiple comparisons using the Benjamin-Hochberg method for controlling FDR.

		SNR				
		0.08	0.1	0.12	0.14	0.16
AT - FT 0.001	high CI	-0.350	-0.168	-0.118	0.002	-0.018
	mean	-0.414	-0.228	-0.154	-0.034	-0.046
	low CI	-0.480	-0.288	-0.194	-0.070	-0.076
	q-vals	**<0.0001**	**<0.0001**	**<0.0001**	0.059	**0.003**
AT - FT 0.05 FWE	high CI	-2.182	-1.242	-0.712	-0.428	-0.346
	mean	-2.284	-1.324	-0.782	-0.490	-0.394
	low CI	-2.390	-1.404	-0.856	-0.550	-0.444
	q-vals	**<0.0001**	**<0.0001**	**<0.0001**	**<0.0001**	**<0.0001**

were not enough true positives to reliably estimate border accuracy. For the remaining cluster sizes, AT outperformed both default FT values in terms of underestimation of borders, i.e. it showed fewer false negative voxels (see Figure 3.8b), but at the same time it performed worst in terms of overestimation with more false positive voxels (see Figure 3.8a). However, the difference was such that AT had a better overall spatial accuracy, i.e. trade-off between over and underestimation (see 3.4 and Figure 3.8c and Figure 3.8d). AT provided a statistically significant improvement in terms of the border over/under estimation when compared to both of the two FT values. As in the cluster analysis the effect was stronger for lower SNR levels, although in case of the highest tested SNR, 0.16, FT 0.001 performed equally well as AT.

Table 3.3: Statistical analysis of the pair-wise difference (AT-FT) comparison of the total number of errors (false positive + false negative). Q-values correspond to p-values corrected for multiple comparisons using the Benjamin-Hochberg method for controlling FDR.

		SNR				
		0.08	0.1	0.12	0.14	0.16
AT - FT 0.001	high CI	-0.168	0.072	0.06	0.132	0.036
	mean	-0.23	0.012	0.022	0.094	0.006
	low CI	-0.292	-0.048	-0.018	0.058	-0.024
	q-vals	**<0.0001**	0.772	0.50222222	**<0.0001**	0.772
AT - FT 0.05 FWE	high CI	-1.91	-0.866	-0.424	-0.16	-0.162
	mean	-2.016	-0.956	-0.494	-0.218	-0.214
	low CI	-2.126	-1.044	-0.566	-0.276	-0.266
	q-vals	**<0.0001**	**<0.0001**	**<0.0001**	**<0.0001**	**<0.0001**

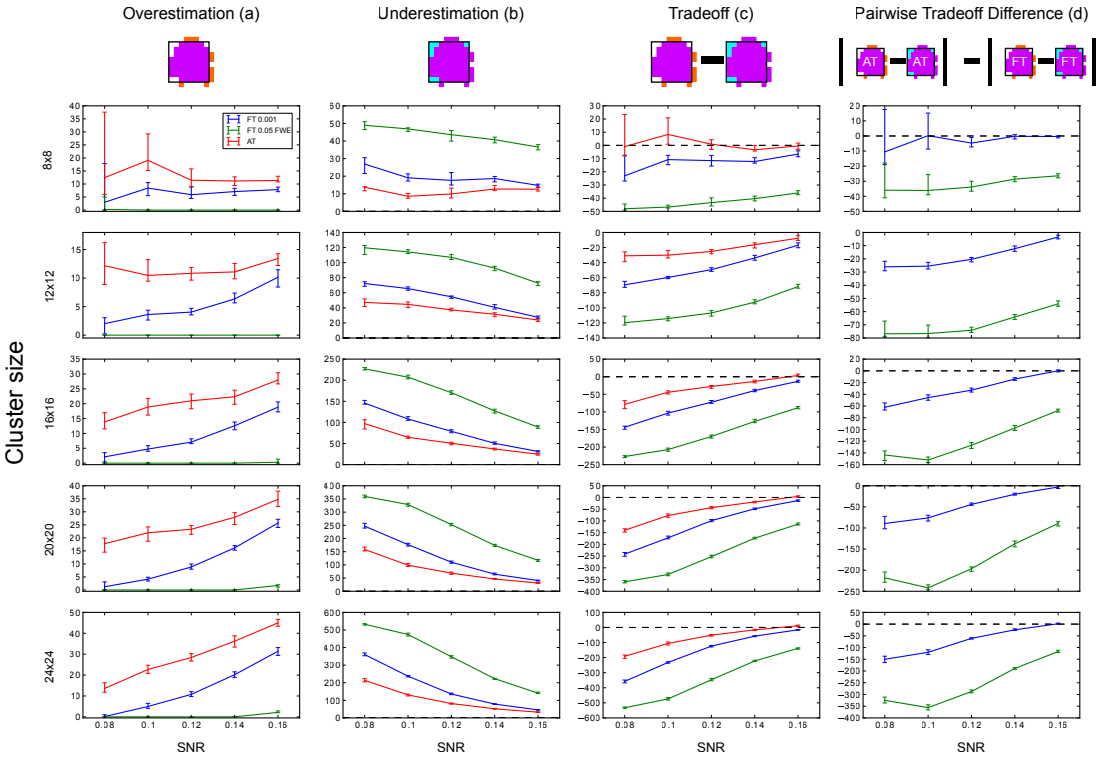

Figure 3.8: Illustration of over and underestimation and performances of the different thresholding methods. At the top is an illustration of an observed cluster (purple) over the true underlying signal (square outline). Estimated voxels outside of the true border are in orange and missed voxels inside the border are in cyan. Below, graphs represent four biases: H-D estimates of the median cluster extent overestimation, i.e. the number of false positive voxels for a particular cluster (a), H-D estimates of the median cluster extent underestimation, i.e. the number of false negative voxels for a particular cluster (b) the overestimation and underestimation trade-off, i.e. differences of the H-D estimates of the medians (c), and the pairwise comparison between AT and FT trade-offs (d). Each row corresponds to different cluster size and whiskers represent 95% confidence intervals. Due to the fact that the smallest cluster (4×4) was found by all of the thresholding methods only in a handful of runs it was excluded from this plot.

Table 3.4: Statistical analysis the pairwise difference (AT-FT) comparison of spatial accuracy tradeoff (see 3.8d). Q-Values correspond to p-values that were corrected for multiple comparisons using B-H method for controlling FDR.

		SNR									
		0.08		0.1		0.12		0.14		0.16	
		0.001	0.05 FWE	0.001	0.05 FWE	0.001	0.05 FWE	0.001	0.05 FWE	0.001	0.05 FWE
8x8	high CI	17.624	-17.949	15.924	-24.583	-0.985	-30.220	0.842	-26.872	0.084	-25.104
	H-D median	-10.574	-35.951	0.209	-36.159	-4.722	-33.886	-0.208	-28.597	-0.323	-26.397
	low CI	-19.000	-41.000	-8.616	-38.967	-7.310	-36.614	-1.961	-30.305	-1.244	-27.710
	q- value	0.720	**0.000**	0.985	**0.000**	**0.031**	**0.000**	0.824	**0.000**	0.179	**0.000**
12x12	high CI	-21.769	-67.618	-22.629	-70.496	-18.924	-71.547	-10.301	-62.223	-2.035	-51.868
	H-D median	-26.068	-76.810	-25.523	-76.621	-20.370	-74.153	-12.286	-64.090	-3.455	-54.108
	low CI	-29.245	-79.166	-27.671	-79.760	-22.375	-75.860	-14.309	-65.932	-4.912	-55.846
	q- value	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**
16x16	high CI	-54.657	-137.005	-40.853	-146.946	-29.412	-122.378	-11.873	-93.170	1.494	-65.270
	H-D median	-62.178	-143.510	-45.686	-152.128	-33.017	-126.985	-14.093	-97.528	0.032	-67.542
	low CI	-67.046	-153.732	-49.917	-156.135	-36.206	-132.445	-16.313	-101.854	-1.874	-70.395
	q- value	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	0.976	**0.000**
20x20	high CI	-72.921	-204.005	-69.943	-234.604	-40.932	-192.090	-17.569	-130.588	-0.399	-85.261
	H-D median	-89.445	-218.161	-76.181	-241.916	-44.158	-196.763	-20.186	-137.650	-3.039	-89.292
	low CI	-101.879	-228.995	-83.028	-247.105	-46.580	-202.319	-22.103	-144.764	-6.096	-95.166
	q- value	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.035**	**0.000**
24x24	high CI	-136.983	-311.630	-109.479	-344.577	-58.295	-280.437	-20.326	-185.564	3.943	-111.384
	H-D median	-150.097	-324.914	-120.186	-356.117	-60.786	-287.782	-24.314	-188.970	1.566	-117.165
	low CI	-163.942	-336.446	-129.655	-365.877	-64.013	-293.151	-27.205	-192.819	-0.169	-121.351
	q- value	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	0.094	**0.000**

3.3.5.3 *Comparison of AT to FT using lower cluster forming thresholds*

In this additional comparison we looked at two different fixed thresholds, specifically a low FT value was set to the mean threshold estimated by AT for low SNR cases and a high FT value was set to the mean threshold estimated by AT for high SNR cases. Pairwise trade-off difference showed that AT performs equally well as FT low and outperforms FT high for low SNR. This is reversed for high SNR. Therefore using the same fixed threshold (FT low or high) for any SNR situation leads to spatial inaccuracy of the border estimation which can be avoided by using AT (Figure 3.9).

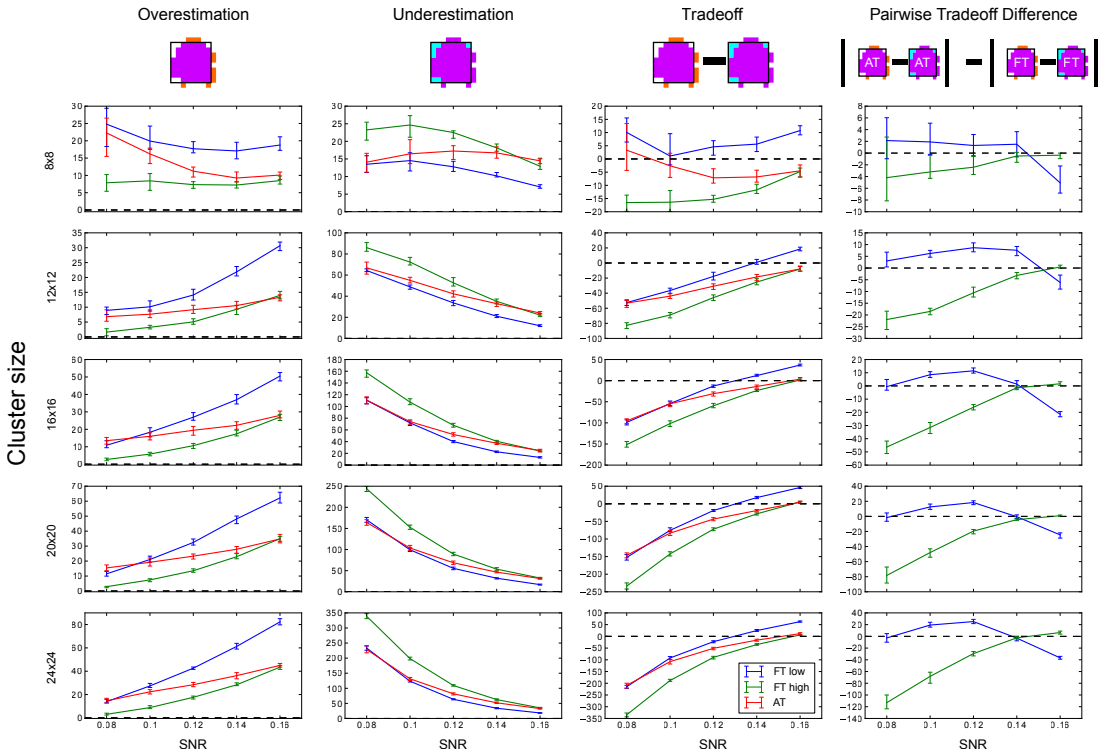

Figure 3.9: Illustration of over and underestimation, and performances of AT and FT with low cluster-forming thresholds. At the top is an illustration of an observed cluster (purple) over the true underlying signal (square outline). Estimated voxels outside of the true border are in orange and missed voxels inside the border are in cyan. Below, graphs represent four biases. On the left are displayed the H-D estimates of the median cluster extent overestimation (number of false positive voxels for a particular cluster). Next are displayed the H-D estimates of the median cluster extent underestimation (number of false negative voxels for a particular cluster). Next is displayed the overestimation and underestimation trade-off (differences of the H-D estimates of the medians). Finally on the right hand side is displayed the pairwise comparison between AT and FT trade-offs. Each row corresponds to different cluster size and whiskers represent 95% confidence intervals. Due to the fact that the smallest cluster (4×4) was found by all of the thresholding methods only in a handful of runs it was excluded from this plot.

3.3.5.4 *Influence of global effects*

Pair-wise difference between Dice coefficients for AT and FT show an overall higher immunity to global noise for AT than FT (mean difference: 0.32 for FT 0.001 uncor-

rected; p <0.0001 and 0.51 for FT 0.05 FWE; p < 0.0001). Global effects lead to a shift of the overall distribution such that the FT procedures created clusters of different sizes. By contrast, AT was able to recover from this confound by shifting the centre of the Gaussian in the mixture model, thus creating clusters of similar sizes. Looking at the correlation between the applied shift and the estimated mean of the Gaussian component (see Figure 3.10) showed that the Gamma-Gaussian mixture model accurately estimated this effect (r=0.99, p < 0.0001). Plotting Dice coefficient differences against the applied distribution shift (see Figure 3.10) showed that the increase in reliability came from this shift such that it varied proportionally with the absolute value of the applied shift (FT 0.001 uncorrected r=0.69; p<0.0001 and FT 0.05 FWE r=0.34;p<0.0001). This demonstrates how big an influence global noise can have on the thresholded maps. Due to flexibility in the assumptions of the noise distribution, in that the mean does not necessarily have to be zero, AT managed to accurately estimate the confounding shift. This lead to better recovery of the unshifted maps, which in real world would translate to better reliability for the same subject between two sessions.

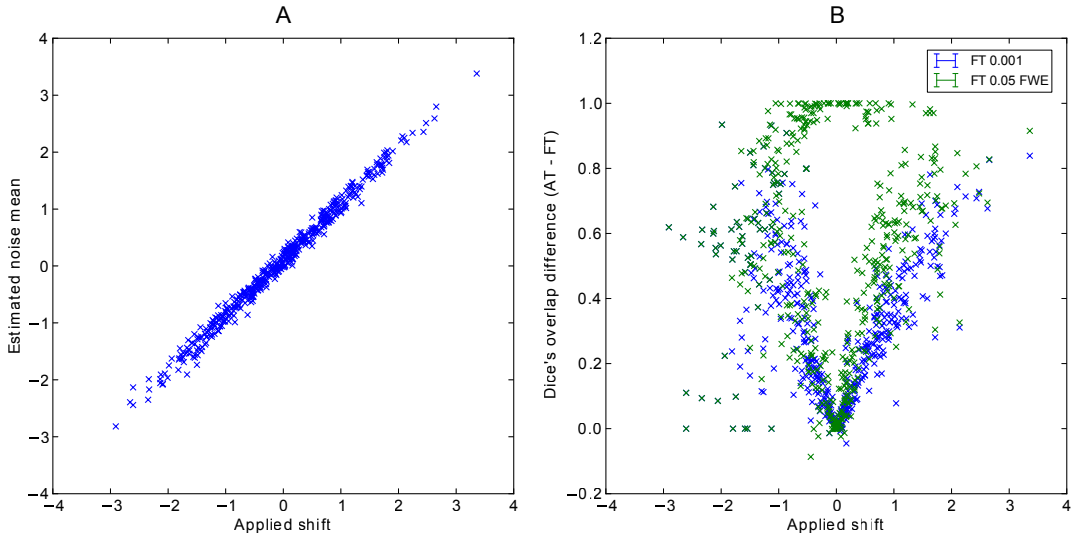

Figure 3.10: Estimated mean of the noise component versus the applied distribution shift (A) and the improvement of AT over FT with respect to the applied distribution shift (B).

3.4 DISCUSSION

Single subject fMRI analyses have different requirements than group studies mainly because the SNR is often lower, and one wants to reveal specific or expected areas and delineate their spatial extent. For these reasons, a fixed threshold strategy is rarely adopted and each subject's T-value map tends to be thresholded differently. Here, we propose a method that thresholds each subject's statistical map differently, but follows an objective criterion rather than a subjective decision. Indeed, we show that our adaptive thresholding method outperforms default fixed thresholds both in terms of trade-off between Type I and Type II cluster error rates and in terms of spatial

accuracy. This increase in spatial accuracy can also be inferred from the reliability results. While validity and reliability can be separated in various conditions, we can infer that, for fMRI, the most valid voxels are the ones detected reliably. Valid and reliable voxels usually correspond to voxels located at the core of a cluster while non-valid and non-reliable voxels are located at the cluster borders. Since AT leads to higher reliability than FT, we can infer that it also improves clusters delineation in real data sets.

In the analysis of our simulations we have looked at levels of false positive and false negative clusters and voxels at different levels of noise and true activation patch size. The combined cost of the two errors (false positive and false negative) was assessed by looking at the sum as well as the difference between them. Another way of looking at similar classification problems is to use a Receiver Operating Characteristic (ROC). This technique plots the ratio of true positives vs. true negatives across a range of thresholds. One can compare different classifiers by calculating the are under the ROC curve. It has been successfully used to evaluate classifiers in medicine, psychology, and biometrics. In the context of presurgical planning, however, it does not show the full picture. In this analysis we wanted to evaluate the performance of certain fixed threshold levels or formalized strategies to obtain thresholds in different noise situations. The goal of this research is to propose a certain thresholding technique (together with a set of parameters) not to evaluate a classifier across different threshold levels. Therefore we developed and used set of new “tradeoff" measures resembling the ROC in many ways but focusing at a single threshold level.

A major source of noise in fMRI time series relates to global effects. Because of the shift of the overall T-value distribution below or above zero, a fixed threshold strategy can lead to the under or overestimation of the true signal. By contrast, we show that AT can correct for “global effects” by shifting the mean of the Gaussian component in our Gamma-Gaussian mixture model. A similar approach has been used before to remove global effect biases in a session variability study by Smith et al. (2005), but not in context of thresholding statistical maps.

Mixture models have been used previously to threshold statistical maps. Most recently Pendse et al. (2009) have used a mixture of Gaussians to improve FDR control by estimating the empirical null. There are two major differences between this and our approach. Firstly, inference is performed on the cluster level as described by Chumbley and Friston (2009), and directly incorporates spatial dependencies between voxels. Secondly, when it comes to border delineation, we are interested in the balance between false positive and negative errors. Controlling for voxelwise FDR does not solve the problem of false negative errors, which as we argue above, are very important in the clinical context and for single subject analyses in general. The closest method to our approach is work presented by Woolrich et al. (2005). Their model also uses a Gamma-Gaussian mixture model, but incorporates spatial information through MRF instead of Gaussian Random fields. Such a model is harder to fit than the Gaussian Random fields approach due to the problems of finding the right spatial regularization coefficients. Also, both approaches do not assume o centrality for the noise component, whereas our model shifts the activation and deactivation Gamma distribution according to the estimated Gaussian (noise) mean, thereby providing immunity to global noise.

Nonetheless, we acknowledge that future improvements of the Adaptive Thresholding method would benefit from incorporating a fully Bayesian framework. This would no only involve using the Discreet Markov Random Field as a prior to achieve spatial regularization, but also incorporating the information about the expected activation as well as regions of interest in a formal way. The Bayesian approach would allow to combine the uncertainty of the smoothness estimation and the uncertainty of the intensity derived class parameters. Even though the presented solution works (as we have shown through simulations) a Bayesian method would be easier to extend at the same time providing a better formalization of the assumptions about the modelled data.

In our method we decided to choose a cluster forming threshold that would minimize the sum of voxelwise false positive and false negative errors. Modelling the T-values distributions using a mixture of gamma and Gaussian distributions allows such an optimization to be performed. Higher thresholds yield more false negative errors and lower thresholds yield more false positive errors. However, when it comes to the sum of all errors there is an optimal threshold which is equal to the crossing point between the Gaussian and Gamma distributions. Our simulations confirmed this theoretical relation (see Table 3.3). There have also been other attempts at creating adaptive thresholding methods. One of the most notable is Activity Mapping as Percentage of Local Excitement (AMPLE) (Voyvodic, 2006). In this technique, T-values are scaled by a local, within Region of Interest (ROI), maximum value just before thresholding. This results in reduced sensitivity to sample size and increased test-retest reliability (Voyvodic, 2012). However, this approach does not assume any formal model of noise and signal and does not incorporate spatial information, although this might not be necessary for small ROIs. It does, on the other hand, apply different thresholds for different parts of the brain. In principle it is very likely that characteristics of noise and signal are not stationary across the brain, but finding rois to fit models locally is not trivial. AMPLE uses rois that are atlas derived, manually drawn (Voyvodic, 2006; Voyvodic et al., 2009), or semi-automatically discovered from the same activation signal (Voyvodic, 2006). We aimed at keeping our method as automated as possible to reduce user input and subjectivity. Additionally using parcellations derived from activation signal to establish local parameters used for thresholding the same activation may introduce "double dipping" biases. Nonetheless, we can see a potential extension of the method in which mixture model could be fitted separately to different brain regions. In such approach parcellation and local thresholding should be done in an iterative way so one would inform the other until reaching convergence.

Because AT separates signal from noise, it is expected to reduce the false negative rate. Indeed, simulations show that AT has better Type II cluster error rates than FT, but this comes at the price of creating more false positive clusters. However, overall it achieves a better balance in terms of detection. One possible explanation for this is that AT tends to use lower cluster forming thresholds than the default FT values and thus a good balance could be achieved simply by using a lower fixed threshold. Additional analyses (see 3.9) using two such low fixed thresholds, one corresponding to the mean threshold estimated with AT at high SNR and the other with AT at low SNR, show that this was not the case and that AT always outperforms FT because it adjusts to the estimated strength of the signal, thereby providing a lower

threshold for weak signals and higher threshold for stronger signals. This results in fewer false negative clusters for weak signal cases and fewer false positives for strong signal cases. Despite the good balance obtained between false positive and negative clusters in our simulations, this method does not provide any guaranteed statistical properties (however, guarantees made by other methods are as good as their assumprions). It is more of a heuristic approach based on sound assumptions than an analytical solution. A possible extension of the method that could improve sensibility is to fix the cluster forming threshold to a certain point, e.g. 0.05, on the cumulative density function of the signal distribution rather than using a point of equal probability between signal and noise. This would control explicitly for the expected voxel-wise Type II error rate. However, because this approach would not include information about the characteristic of noise, it will not be as accurate in terms of spatial extend and reliability.

Finally, because AT provides a higher spatial accuracy and adapts to noise, it also leads to an increase in reliability. In the context of single subject fMRI analysis, and in particular for data used in clinical procedures such as presurgical planning, it is worth noting that spatial accuracy is essential. Of particular interest here, AT showed much lower underestimation than FT, which may be useful in clinical situations. Increased spatial reliability in healthy controls also means that one can be confident that the method will more often detect valid clusters as suggested by the reduced false negative rate in the simulations. Overall, AT therefore achieves a better balance than FT approaches, and provides a new tool for reliably and objectively thresholding multiple single-subject SPMs.

3.5 SUMMARY

In this chapter we have introduced a new way of thresholding statistical maps developed with presurgical mapping application in mind. We have shown through simulations improved spatial accuracy and smaller susceptibility to global effect artefacts. In the next chapter we will look at test-retest reliability of fMRI and show how adaptive thresholding improves it.

4

RELIABILITY OF COMMONLY USED MAPPING TASKS

4.1 INTRODUCTION

This work has been presented at OHBM 2012 (Gorgolewski et al., 2012b), and has been accepted for publication in NeuroImage (Gorgolewski et al., 2013).

In the previous chapter we have introduced a new thresholding method and shown through simulations that it outperforms its competitors. Simulations are, however, a simplified version of reality. By definition they include assumptions which can have significant influence on the results. This is the price one has to pay for being able to access the ground truth. There is, however, another approach to evaluating inference methods. Based on assumption that a good method should be resistant to noise we can claim that it should yield the same result when measuring the same phenomenon multiple times. This is also known as test-retest reliability. It is not only useful for evaluating data processing methods, but also data acquisition techniques. This is a very important issue in presurgical planning since there are many variants of the behavioural tasks and scanning parameters. Lastly fMRI reliability, even though studied for many years, is not well understood. There are many metrics for measuring reliability and theories on what factors can influence it. Therefore the aim of this chapter is threefold:

- to find out if AT provides more reliable results than FT,
- to investigate the reliability of behavioural task commonly used in fMRI presurgical mapping,
- to take a more general look at fMRI reliability by investigating the relation between different measures of reliability and how different confounding factors can contribute to them.

We begin by a description of the test-retest reliability dataset used to answer the aforementioned questions. The three sections that follow include analysis details and results specific for each of the three aspects of reliability we are interested in.

4.1.1 *Behavioural tasks*

As we have discussed in the previous chapters, fMRI can map the cortex in a non invasive way. For every brain area of interest a separate behavioural task has to be performed. The most commonly used tasks correspond to common post operative neural deficits such as: hemiplegia (motor cortex), receptive aphasia (Wernicke's area), expressive aphasia (Broca's area), and hemineglect (right parietal lobe). Tasks have to be designed in such way that allows delineating specific activation and provides reliable location and extent of the eloquent cortex. This information is the basis of the decision that surgeon will make concerning the risk of the procedure. In the following

study we have investigated five tasks in terms of the intersubject variability in order to assess how much a single scan can be trusted.

Focusing only on a few areas in the brain might seem crude but in the clinical context some faculties are more important than others. For example the ability to speak and move is more important to protect than subtle personality changes that subjects can experience in the case of frontal cortex damage.

4.1.1.1 *Word repetition task*

Initial attempts to localize speech related brain areas were based on lesion studies. Development of in-vivo functional scanning techniques such as PET and fMRI stimulated many new studies on the topic. Language skills can be divided into verbal (speech) and non verbal (i.e. writing/reading). The word repetition task consists of hearing and immediately repeating (overtly) a word and is aimed at mapping speech recognition areas (Wernicke). It does not involve higher level language skills (such as grammar) and therefore is well suited for avoiding post operative receptive, expressive, and conduction aphasia, but not anomia. The main concern about this task (apart from obvious motion artefacts caused by speaking) is that activation in primary auditory cortex often overlapping with activation in Wernicke's area related to speech comprehension (Binder et al., 1995). However, in the context of presurgical planning this is not a disadvantage since damaging auditory cortex should also be avoided.

4.1.1.2 *Verb generation task*

The two most common tasks that are supposed to elicit speech related brain activation are overt and covert speech generation. In the latter case subjects are asked to silently think of speaking instead of executing the motor action. This variation of the speech task was introduced to minimize the amount of motion related artefacts. Early PET study compared overt and covert speech tasks found different activation both in the language and motor areas (Bookheimer et al., 1995), suggesting that the overt task might not just be the overt task with extra motor activation on top of it. This was confirmed by fMRI studies (Huang et al., 2002) which found that overt tasks do not produce a simple superposition of covert and motor action activations. Differences were found both in the location and strength of the activation within the language related areas.

The motivation for using covert tasks —- reducing the amount of motion artefacts —- is also debatable. For example one study had to discard 5 out of 11 subjects due to motion artefacts (Phelps et al., 1997). In an attempt to quantify the amount of artefacts defined as activation found outside of the skull one study concluded that covert tasks produces stronger activation with fewer artefacts (Yetkin et al., 1995). On the other hand with proper training and head immobilization the influence of motion artefacts can be negligible (Huang et al., 2002).

To minimize the influence of movement one can incorporate sparse sampling techniques (Hall et al., 1999). In this approach gaps are inserted in the scanning session during which subject can perform the task. This was initially developed for auditory tasks in which scanner noise interferes with the stimuli. Because of the gaps the number of samples is smaller than for continues sampling, but at the same time the Time

of Repetition (TR) is longer which gives more time for T2* relaxation and improves SNR (Hall et al., 1999).

For speech tasks, sparse sampling has indeed shown improved motion related artefacts (Abrahams et al., 2003). The number of samples vs. SNR trade-off was also investigated for an auditory task. For certain combination of parameters, improvement from better SNR makes up for the smaller sample size (Schmidt et al., 2008).

Brain mapping in the context of speech has not only been used to decouple the language brain network, but also for presurgical planning. In this clinical application subjects are often confused, medicated and under the influence of the condition they are treated for such as a brain tumour. This makes scanning harder than in case of healthy controls. Getting feedback would allow making the decision as to whether there is a need to retrain the subject and rerun the scan. There have been attempts to filter out scanner noise from in scanner recordings to check if subjects were speaking during the task phase, but those were done in an offline post processing (Cusack et al., 2005). Doing this filtering online (in realtime) seems feasible since we know the timing of scanner noise with great precision, but this would involve installing extra hardware. Sparse sampling solves this issue —- scanning is ceased for the time when the subjects speak and the radiologist or scanner operator can listen in and check if the task is performed correctly.

Additionally it has been shown that overt speech task corresponds better to ECS (Petrovich et al., 2005) which is more relevant in the context of presurgical planning.

4.1.1.3 *Motor tasks*

One of the first presurgical mapping paradigms to be tested was on motor tasks due to a high signal strength, simple paradigm design. To obtain a fuller map of the motor cortex more than one motor task is recommended. The most commonly used are: finger, foot, and lips. This of course makes the scanning time longer which can be a problem in the case of agitated patients. In such cases, skipping the body area that is further from the tumour in its suspected cortical representation is recommended.

The way different movements can be contrasted between each other provides insights into different brain regions. The simplest case of one movement vs. rest (no movement) will elicit activation in primary and supplementary motor cortex. However, contrasting movement of one limb vs. its contralateral counterpart (i.e. left vs. right foot) should produce activation only in the primary motor area. In the case of three movements (and no rest), when contrasting between one body part vs. all the other primary motor areas will be accompanied with supplementary motor area exclusive only to this body part (Stippich, 2007). Therefore the most comprehensive sequence would include three movements and rest periods. This, however, is also the longest sequence.

Motor tasks carry significant risk of motion artefacts. Repetitive tongue movements with closed mouth, opposition of fingers D2-D5 to D1, and flexion and extension of toes without moving the ankle have been recommended in the past as eliciting the fewest movement artefacts (Stippich, 2007).

4.1.1.4 *Line bisection task*

The line bisection has been used in clinical practice to assess hemineglect. In this test patients are presented with a piece of paper with a printed horizontal line on it. The paper should be positioned directly in front of the patients who are asked to draw a vertical line in the middle of it. In case of spatial attention deficit such as hemineglect patients ignore one part of the visual field and draw the line far from the true middle.

There have been only a few fMRI studies using the line bisection task. Fink et al. were the first to attempt this task (Fink et al., 2000). They have used the so called landmark task; instead of asking the patient to bisect the line, as it is done in the clinical practice, they have been asked to tell if a line is pre-bisected correctly. This was contrasted with asking subjects to tell if a line was bisected at all or not. They have scanned 12 subjects (healthy male volunteers) and found activations in right superior posterior and right inferior parietal lobe, early visual processing areas bilaterally, the cerebellar vermis, and the left cerebellar hemisphere. Since they were expecting a lateralized response they have controlled for the hand used for giving the responses by doubling the number of tasks to cover all combinations. The same group has also assessed differences between vertical and horizontal landmark tasks (Fink et al., 2001). The locations of activations found were similar to the previous study. When contrasted vertical with horizontal lines, increased activation in early visual areas has been found. The authors concluded that difference in activation pattern between horizontal and vertical variants is mostly due to visual characteristics of the stimulus and not due to difference in the underlying task.

Flöel et al. (2005) have looked at the landmark task in subjects with atypical lateralization (language in the right hemisphere with attentions in the left hemisphere and both language and attention in the right hemisphere). Their control group confirmed the finding of Fink et al., namely activation in right superior and inferior parietal cortex. The atypical group with left-hemisphere dominance for spatial attention showed activation in regions homotopic to the control group. The group with both language and spatial attention in the right hemisphere did not differ in the presented activation pattern from the control group. The authors also briefly commented on the use of such tasks in presurgical planning, concluding that for patients with atypical hemispheric dominance activations outside of homotopic brain areas should be interpreted with caution.

The landmark task is, however, different from the line bisection task used in the clinical practice. Ciçek et al. (2009) have developed an fMRI paradigm in which subjects can move a notch on a horizontal line to properly bisect it. They have found that such tasks elicit activation in right Intra–Parietal Sulcus (IPS), anterior cingulated gyrus and right Lateral Peristriate Cortex (LPC). A conjunction analysis with the landmark task revealed left activations in IPS and LPC. However, the activation they found in the landmark task does not overlap with Fink et al. findings in all regions (Fink et al. additionally found activation in prefrontal and early visual areas). Çiçek et al. attributed this to differences in the visual presentation of the lines (corners vs. middle of the screen) and the control task.

4.2 TEST–RETEST RELIABILITY DATASET

4.2.1 *Participants and procedure*

A group of normal healthy volunteers without contraindications to MRI scanning were recruited using flyers distributed among University of Edinburgh staff in electronic and traditional form. To match the mean age of diagnosis of the glioma patients undergoing resection surgery (Ohgaki, 2009), all volunteers were over 50 years of age. Out of 11 volunteers, data from one participant were discarded due to problems with executing the tasks. Additionally one session of the word repetition task was discarder for one of the subjects. The remaining 10 subjects included four males and six females, of which three were left–handed and seven right–handed according to their own declaration, with median age at the time of first scan of 52.5 years (min = 50, max = 58 years). The study was approved by the local Research Ethics Committee.

4.2.2 *Behavioural tasks*

All the behavioural tasks were implemented using Presentation® Software (Neuro Behavioural Systems `http://www.neurobs.com/`). Stimuli synchronisation and presentation was provided by NordicNeuroLab hardware (`http://www.nordicneurolab.com/`). During the first scanning session, each subject was trained for each task with a few trials inside the scanner. Care was taken to make sure that volunteers understood and could properly perform the tasks.

4.2.2.1 *Overt word repetition*

Subjects had to repeat aloud words presented via headphones. The following instructions were used: "When you hear the word, repeat it immediately". A block design with 30 sec activation and 30 sec rest blocks was employed in conjunction with a sparse sampling data acquisition technique to present and record stimuli during the silent periods, with four trials used for training. After 2.5 sec of blank screen during which the fMRI data were acquired, subjects were presented with an auditory stimulus which consisted of a pre–recorded native British English speaker reading a noun chosen at random from a set of 36 nouns (759 msec sound tracks length, mean lexical frequency: 0.000087, min: 0.000005, max: 0.000392, std: 0.000098). This was followed by a question mark prompting the subject to repeat the word. Question marks disappeared after 1741 msec and the sequence was repeated 6 times (see Figure 4.1). The nouns used were randomised for every subject/session combination. A blank screen was also presented during rest periods. There were six activation/rest blocks for a total scan time of 7 min 40 sec. Subject responses were recorded using an MRI compatible microphone. During the scanning session, the radiography staff listened to check if the subject was executing the task correctly.

4.2.2.2 *Covert verb generation*

Subjects were asked to think of a verb complementing a noun presented to them visually. The following instructions were used: "When a word appears it will be a noun. Think of what you can do with it and then imagine saying 'With that I can . . . '

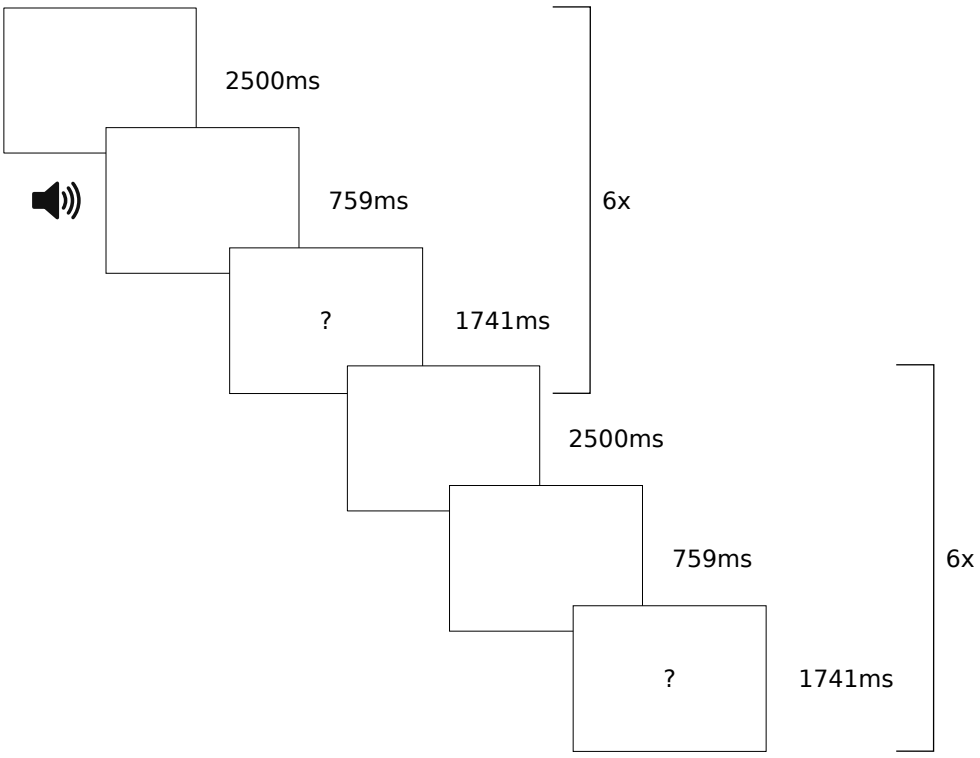

Figure 4.1: Overt word repetition paradigm design.

or 'That I can . . . ' ". A block design with 30 sec activation and 30 sec rest blocks was employed, with eight trials used for training. During the activation blocks, ten nouns were presented for 1 sec each followed by a fixation cross during which subject had to generate the response (see Figure 4.2). The nouns were chosen at random from a set of 70 nouns (mean lexical frequency: 0.000087, min: 0.000005, max: 0.000392, std: 0.000092). Rest blocks had an analogous structure but with each word replaced by scrambled visual patterns generated by scrambling the phase of the 'picture' of each word, i.e. the control patterns were matched in the amplitude spectrum. Seven activation/rest blocks were presented for a total scan time of 7 min 12.5 sec.

4.2.2.3 *Overt verb generation*

In this task, subjects were asked to say a verb complementing a noun presented to them visually (overt version of the previous task). The following instructions were used: "When a word appears it will be a noun. Think of what you can do with it and then say the corresponding verb: 'With that I can . . . ' or 'That I can . . . '". A block design with 30 sec activation/rest blocks was used in conjunction with a sparse sampling data acquisition technique to present and record stimuli during the silent periods. Eight trials were used for training. Four volumes were used for signal stabilization before stimulus presentation. After 2s of blank screen a word/scrambled image was presented for 0.5s (words were chosen from the same set as in the covert version of the task). This was followed by a question mark prompting the subject to

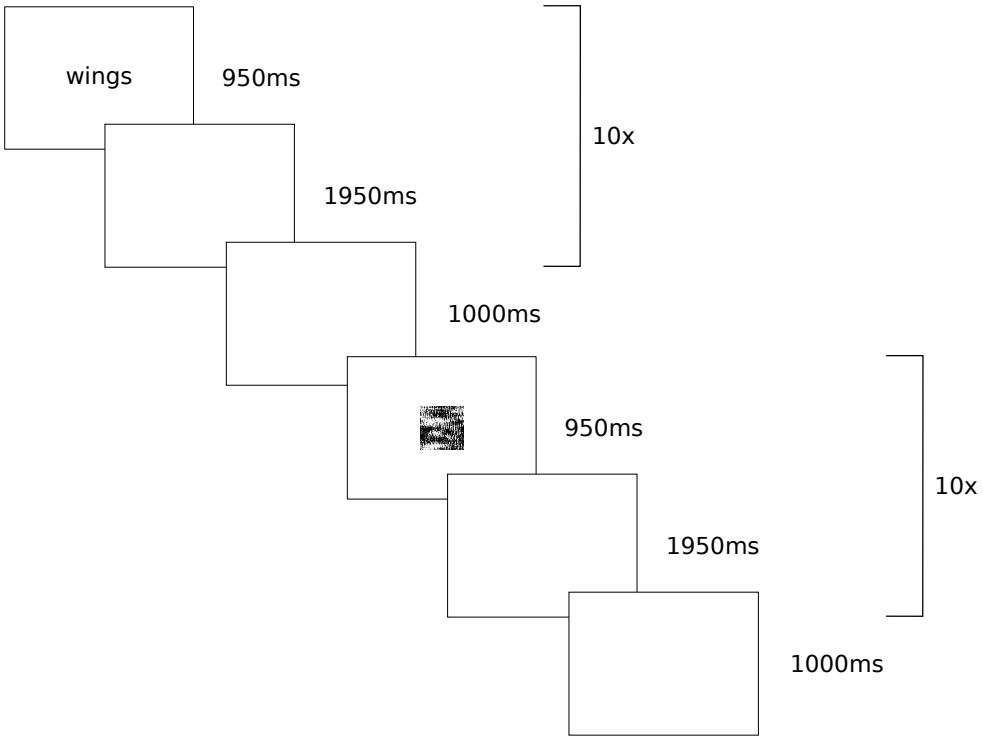

Figure 4.2: Covert verb generation paradigm design.

speak, which was presented for 2.5 seconds (see Figure 4.3). In each block 6 words followed by 6 scrambled images were presented. Both words and scrambled pictures were randomized between session and subjects. Seven activation/rest blocks were presented for a total scanning time of 7 min 20 sec. Subject responses were recorded using an MRI compatible microphone. During the scanning session radiography staff was listening in to check if the subject was executing the task correctly.

4.2.2.4 *Motor tasks*

Subjects had to move a body part corresponding to a picture. The following instructions were issued: "You have to tap your index finger when you see a picture of a finger, flex your foot when you see a picture of a foot, and purse your lips when you see a picture of lips". A block design with 15 sec activation periods and 15 sec rest periods was employed, with four trials used for training. In every block, subjects moved the index finger of their dominant hand, or flipped their dominant foot or pouched their mouth (see Figure 4.4). Movement was paced with a frequency of 0.4 Hz using a visual stimulus. There were five repetitions of each activation/rest block for a total scan time of 7 min 40 sec.

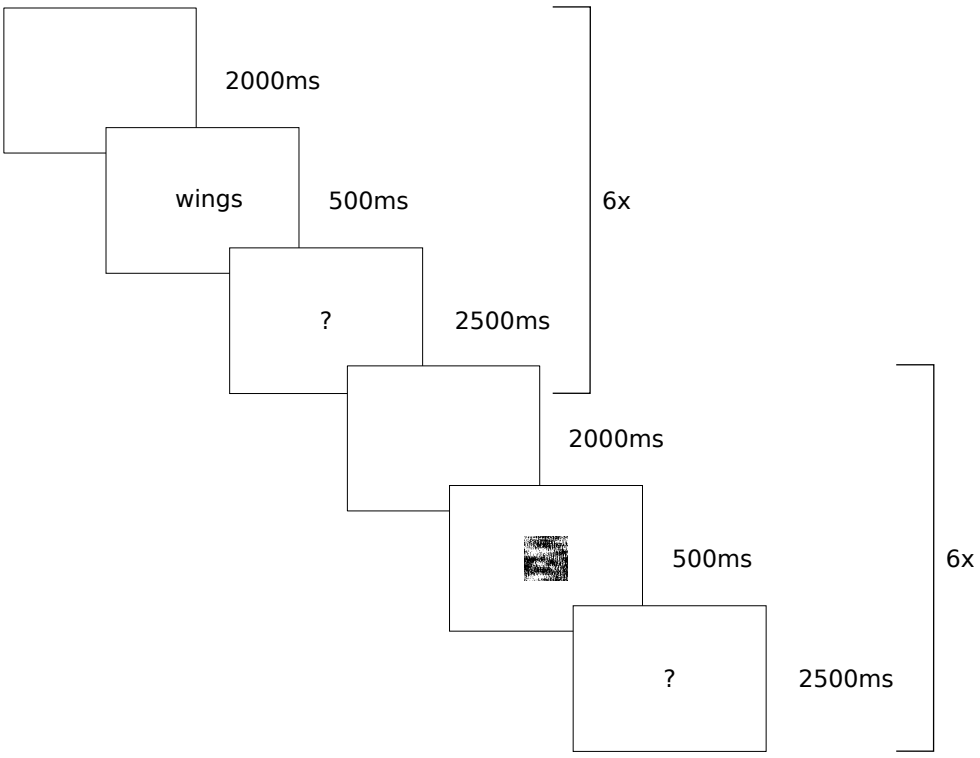

Figure 4.3: Overt verb generation paradigm design.

4.2.2.5 *Landmark*

Subjects performed two alternate tasks, namely tell if a horizontal line is crossed precisely in the middle (LANDMARK) and tell if a horizontal line is crossed at all (DETECTION). The following instructions were used: "Press the button with your left index finger if the line is bisected in the middle otherwise press the button with your right finger" or "Press the button with your left index finger if the line is crossed otherwise press the button with your right finger". A block design with 16.25 sec landmark/detection blocks was used, with ten trials used for training. Each task was preceded by an instruction screen which was presented for 8.25 sec with a rest period of 8 sec (see Figure 4.5). Each block consisted of 10 lines, four correct and six incorrect. Each line was presented for 525 msec and subjects had 1100 msec to respond before the next presentation. Lines were presented in the four corners of the screen. For the landmark task and incorrect trials, the crossing line was located at three different distances from the middle, specifically 12, 40, and 62 pixels from the true middle corresponding to 0.45, 1.5, and 2.325 degrees of visual angle. There were eight landmark/detection blocks for a total scan time of 9 min 55 sec. All trials were randomized and all responses were recorded.

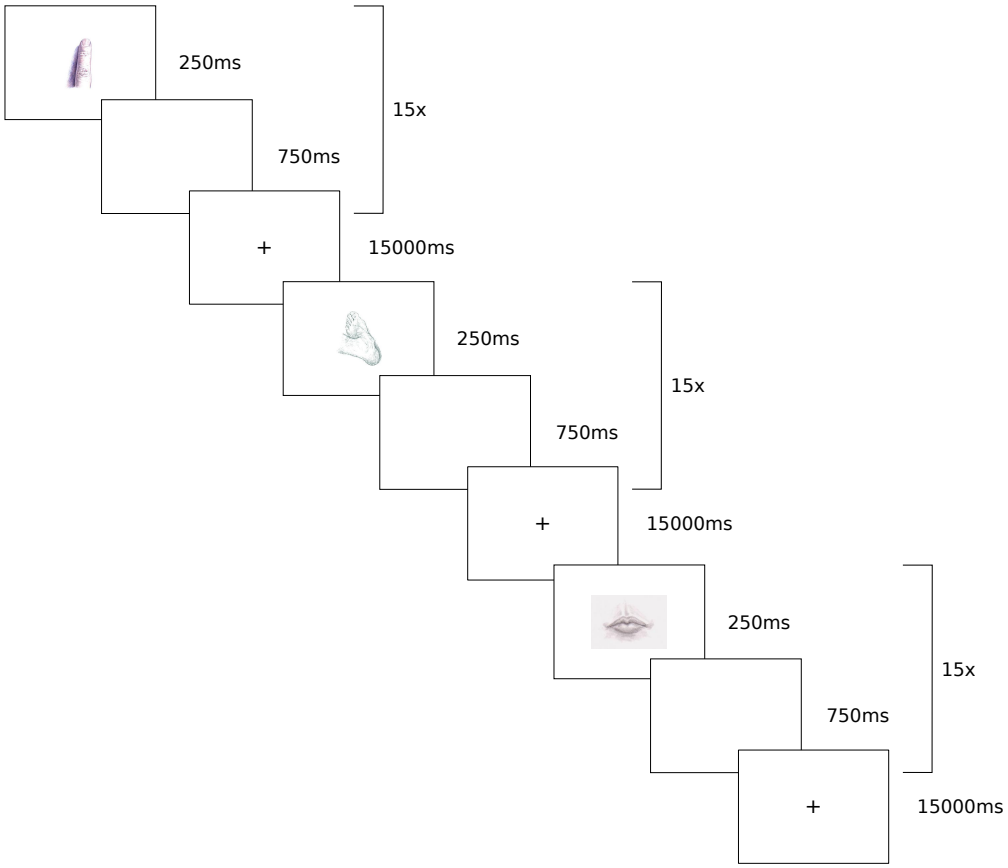

Figure 4.4: Motor paradigm design.

4.2.3 *Scanning sequence*

All scans were acquired on a GE Signa HDxt 1.5 T scanner at the Brain Research Imaging Centre (BRIC Edinburgh http://www.bric.ed.ac.uk/) located at Western General Hospital, Edinburgh. Each volunteer was scanned twice two (eight subjects) or three (two subjects) days apart (using the same sequence). All of the fMRI sequences shared the following parameters (unless otherwise stated): FOV=256 × 256mm, slice thickness 4mm, 30 slices per volume, interleaved slices order, voxel size 4 × 4 × 4mm, acquisition matrix 64 × 64, and TR=2.5s, flip angle=90°, TE=50ms. Scanning session consisted of the following acquisitions:

1. Overt word repetition task. 76 (4 + 6 blocks of 12). Sparse sampling (effective TR=5s, real TR=2.5s).
2. Covert verb generation task. 173 (4 + 7 blocks of 24 + 1) volumes.
3. Overt verb generation task. 88 (4 + 7 blocks of 12). Sparse sampling (effective TR=5s, real TR=2.5s).
4. Motor task. 184 (4 + 5 blocks of 36).

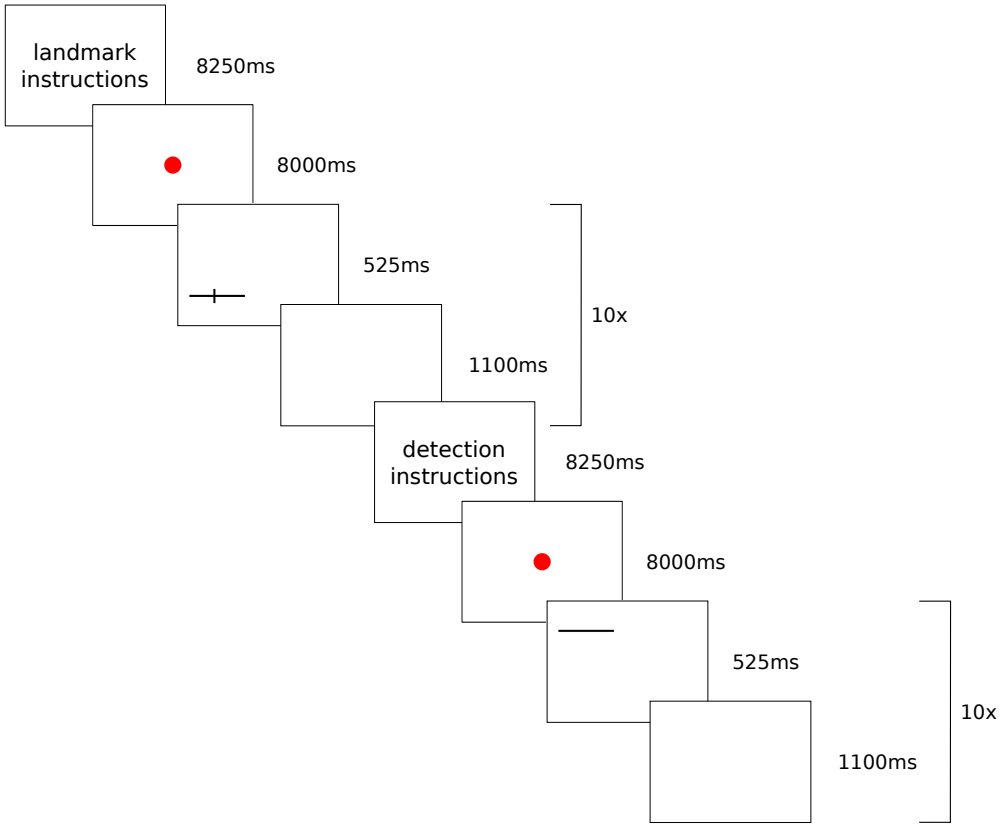

Figure 4.5: Landmark paradigm design.

5. Landmark task. 238 volumes.

6. T1 weighted coronal scan. FOV=256 × 256mm, slice thickness 1.3mm, 156 slices, voxel size 1 × 1 × 1.3mm, acquisition matrix 256 × 256.

The order of the verb generation tasks were counterbalanced across subjects such that half of the subjects did the task in the order: 1. Overt word repetition, 2. overt verb generation, 3. covert verb generation, 4. motor, 5. landmark; and the other half: 1. Overt word repetition, 2. covert verb generation, 3. overt verb generation, 4. motor, 5. landmark.

4.2.4 *Data analysis*

Data was processed using SPM (`http://www.fil.ion.ucl.ac.uk/spm/`) and FSL(`http://www.fmrib.ox.ac.uk/fsl/`) within the Nipype framework (`http://nipy.org/nipype/` - Gorgolewski et al., 2011b).

4.2.4.1 *Preprocessing*

For every subject T1 volumes from both sessions were coregistered, resliced and averaged. A DARTEL template was created using averages from all subjects (Ashburner,

2007). Additionally a brain mask was estimated from each average using BET (Smith, 2002).

The first 4 volumes, during which the scanner reaches steady state, of every EPI sequence were discarded and the remaining images were slice time corrected. Finger, foot, and lips sequences of left–handed subjects (3 subjects) were flipped along the Z–Y plane. For every subject, all slice time corrected volumes from all tasks and sessions were realigned and resliced to their mean to remove motion artefacts. The mean volume was coregistered to the T1–weighted volume between session average and the resulting affine transformation was applied to headers of realigned files. Each EPI volume was then normalized using the DARTEL template and corresponding flow field and smoothed with 8mm full width half maximum Gaussian kernel. Apart from the fact that smoothing improves SNR, it is necessary to maintain assumptions of the RFT which is being used for thresholding. The smoothed volumes supplemented with a previously estimated brain mask and realignment parameters were searched for artefacts using ArtifactDetection toolbox (`http://www.nitrc.org/projects/artifact_detect/`).

4.2.4.2 *First level analysis*

Each session was analyzed separately, with a GLM (Friston et al., 1994) being used to fit a design matrix consisting of an autoregressive filtering matrix (AR1) and task, realignement (6 parameters), high pass filter (128 Hz), and artefacts (one per artefact) regressors. Task regressors for overt verb generation, covert verb generation and overt word repetition were simple boxcart functions convolved with a canononical hemodynamic response function. For these tasks, a simple contrast including the single task regressor was used (= activation vs. baseline). In case of finger, foot and lips tasks, each body part was modelled with a separate boxcart regressor, while 3 simple contrasts per body part were computed as well as three contrasts between each body part against the two others. The design matrix for the landmark task included five event related regressors acquired from each subject/session experiment log: landmark stimuli with correct responses, landmark stimuli with incorrect responses, detection stimuli with any response (correct or incorrect), and detection and landmark stimuli with no response. This allowed four contrasts to be estimated: all landmark stimuli vs. all detection stimuli, only landmark stimuli with response vs. only detection stimuli with responses, only landmark stimuli with correct responses vs. only detection stimuli with any responses, and only landmark stimuli with correct responses vs. only landmark stimuli with incorrect responses. Only voxels within previously estimated brain mask were included in model fitting. For overview of preprocessing and first level analysis, see Figure 4.6.

4.2.4.3 *Second level (random effects) analysis*

For every subject and task, contrast volumes were averaged between the two sessions. These averages were then used in second level group analysis using the Holmes-Friston approach (Holmes and Friston, 1998), i.e. a one sample t–test on each contrast was run to estimate a group effect. The result of each t–test was thresholded using the topological FDR method (Chumbley and Friston, 2009) with a cluster extent probability threshold set to 0.05 after FDR correction. The cluster forming threshold was

set to the default SPM p-value of 0.001 uncorrected. Anatomical labelling of the activation areas was done using SPM Anatomy Toolbox (Eickhoff et al., 2006, 2007, 2005), Harvard-Oxford Cortical and Subcortical Atlases, and Talairach Atlas (Lancaster et al., 2007, 2000; Talairach and Tournoux, 1988).

4.2.4.4 *Reliability analysis*

VOLUME OVERLAP OF THREASHOLDED T MAPS

Single subject t maps were thresholded using adaptive thresholding (see Chapter 3) and topological FDR with fixed cluster forming threshold p=0.05 FWE corrected. Using the suprathreshold maps the between–session Dice overlaps was calculated. In the case where both maps were empty (no suprathreshold voxels), a Dice overlap of zero was assumed to penalize for lack of signal. In addition, to test if the tasks were reliable, the mean Dice overlap obtained for each subject and task was compared with the between-subject Dice overlap. The between-subject Dice overlap was obtained by computing the overlap between the thresholded map of every subject in Session 1 and the thresholded maps of all the other subjects in Session 1. The procedure was repeated for Session 2 and all Dice measures were averaged over sessions for each task. This allowed the testing of whether the overlap measured within-subjects was significantly greater than the overlap measured across all subjects, given that all subjects were in standard space. A percentile bootstrap test of the H-D median (Harrell and Davies, 1982) was used to estimate if the difference of within- and between-subject Dice overlap was statistically significant.

COMPARISON OF THESHOLDING METHODS

Dice coefficients and maximum distance differences obtained with the GGMM and the FT topological FDR approaches were compared for each contrast using a percentile bootstrap (bootstrap on the differences between measures). In each case, two robust measures of location (Harrell-Davis estimate of the median and 20% trimmed mean) were tested. The 95% confidence intervals and p-values are reported for each map separately. However, statistical significance was obtained using a FDR correction over all measures of location (q=5%) and all measures of dispersion (q=5%).

WITHIN VS. BETWEEN SUBJECTS

In addition to comparing thresholding methods, we also tested the internal validity of each task, that is we tested if dice coefficients and maximum distance differences were better within than between subject differences. A lack of difference means that the location and size of an activated brain region is not more reliable (reproducible) for a given subject than taking any other subject, making this task unsuitable in the clinical context. As for the comparison of thresholding methods, Harrell-Davis estimates of the median, 20% trimmed mean and percentile bootstraps (with independent resampling per group) were computed. Again, 95% confidence intervals and p-values are reported for each map separately but statistical significance was obtained using a FDR correction over all measures of location (q=5%) and all measures of dispersion (q=5%).

Results are reported for the full brain, and task specific ROI. These were constructed using probability maps available in the anatomy toolbox. For the mapping of the primary motor cortex, the whole left areas 4a and 4p were used. For Broca area, Brodmann areas 44 and 45 were used . For Wernicke area, area TE30 was used. For the auditory cortex, we used areas TE1, 1.1 and 1.2. Finally, for the landmark task, right Inferior Parietal Cotex and Superior Parietal Lobule were used. Masks were generated in the MNI space and resliced to DARTEL template dimensions.

4.2.4.5 *Intraclass correlation analysis*

For every task and contrast an intra class correlation coefficient (ICC) map was calculated(Caceres et al., 2009; Shrout and Fleiss, 1979). We used the ICC(3,1) variation, a two–way model (subjects vs. sessions) with no interaction and a consistency criteria; in other words allowing for a constant between–session effect such as learning. ICC(3,1) is an estimate of

$$\mathrm{ICC}(3,1) = \frac{\sigma_r^2}{\sigma_r^2 + \sigma_e^2} \tag{4.1}$$

where σ_r^2 is between–subjects (rows) variance and σ_e^2 is the between-sessions variance (variance of the residuals after removing the subject and session effect).

4.2.4.6 *Reduced dataset analysis*

MRI scanners in most hospitals are heavily used and therefore there is strong focus on keeping the scanning sequences short. Therefore using multiple runs or long scanning sessions is not feasible in clinical applications. Additionally many patients struggle with coping with the MRI scanner environment and following the behavioural paradigm of the fMRI experiment. This results in only a subset of data being useful for further analysis. To simulate this we have run the same processing pipeline on just the first 2, 3, and 4 first blocks of each task. The between session Dice overlap for those reduced datasets were computed to compare if there is a difference in performance of the two thresholding methods in question. Additionally for each session the Dice overlap between the result from each reduced dataset and the full dataset was calculated to test how quickly the final map is obtained.

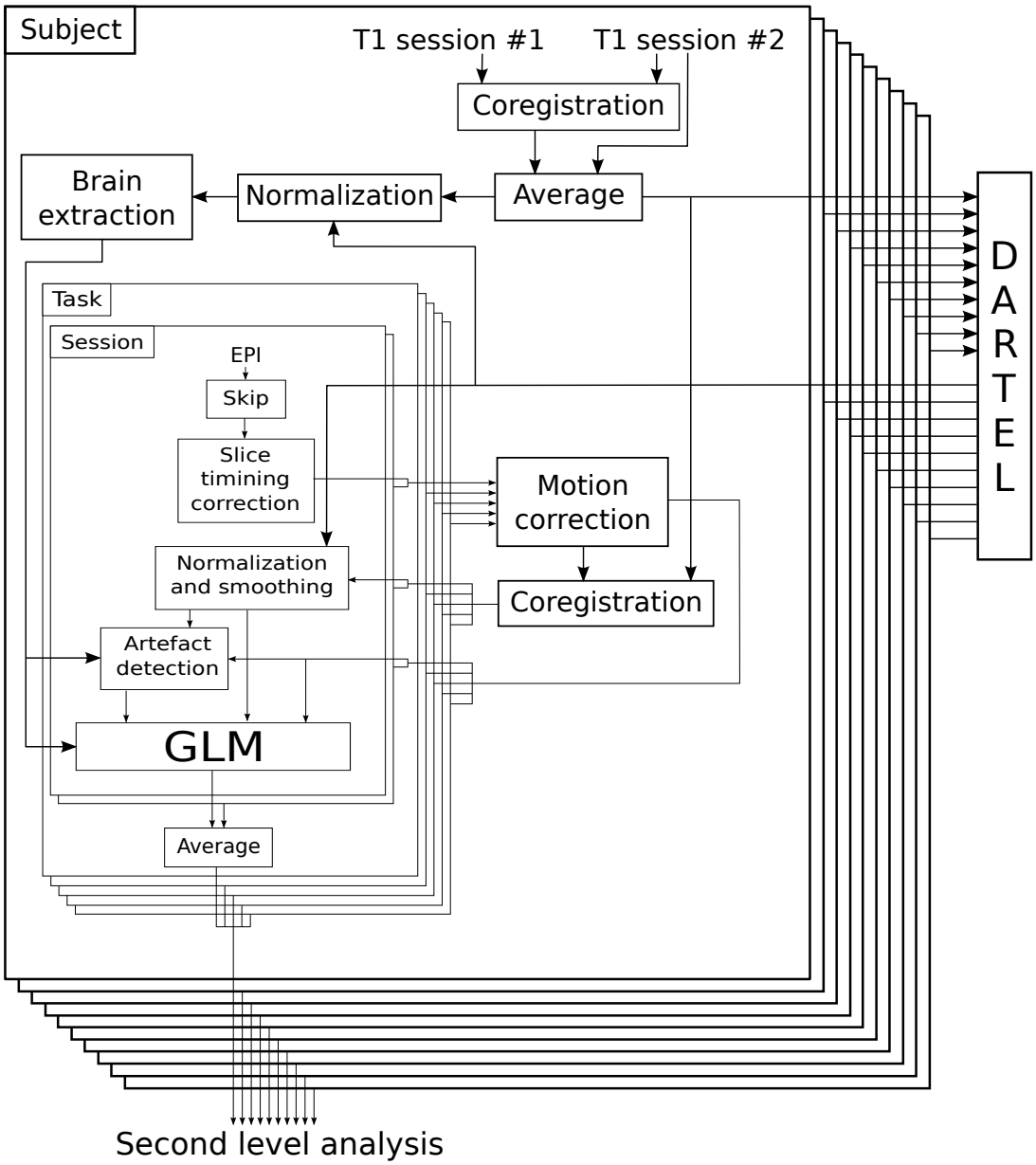

Figure 4.6: Overview of data processing.

4.2.5 *Group analysis*

4.2.5.1 *Language tasks*

The random effect analysis of overt word repetition revealed a strong activation over Superior Temporal Gyrus, mostly in the left and right Primary Auditory Cortex (areas TE 1.1, TE 1.2 and TE 3 - Morosan et al., 2001 and left Wernicke's Area (IPC-PF - Caspers et al., 2006). Additional activations were found in the Supplementary Motor Area (SMA), Brodmann Area (BA) 6, the Postcentral Gyrus (BA 3b) and the Cerebellum (see Figure 4.7). One of the subjects misunderstood the instructions and one of his sessions had to be excluded; his second session instead of average of two was used for the analysis. Apart from this single case, subjects replied to $98 \pm 3\%$ of the stimuli. In the case of covert verb generation, activations were observed in left Broca's area (BA 44 and 45), left Temporal Gyrus, left Inferior Parietal Lobule, SMA and left thalamus. For the overt version of the task, two subjects did not reply to any of the stimuli in both sessions and had to be discarded bringing the number of subjects in the group analysis to 8. Additionally one of the subjects replied only during one session and therefore results from his second session instead of an average of two were used. After excluding these cases subjects replied to $81 \pm 15\%$ of the stimuli. Similar results to the covert generation were obtained with additional activations of motor related areas (BA 4p and OP 4 — Eickhoff et al., 2010).

4.2.5.2 *Motor task*

Using simple contrasts (activations vs. rest) strong activations of the left precentral gyrus were observed respecting the known motor homunculus: (1) Foot contrast revealed activations near the top end of the contra-lateral precentral gyrus (extending to left SMA) and also showing ipsilateral cerebellar activation; (2) Finger contrast produced activation in the middle /lateral contra-lateral preentral gyrus and ipsilateral cerebellum; (3) Lips contrast produced bilateral activation in the inferior part of the precentral gyrus, but also SMA and cerebellum. Activations were also observed in the visual cortex over inferior occipital / fusiform gyri in response to the stimulus presentation.

Using the more complex contrasts (e.g. finger vs. others) produced similar results. Over motor regions, differences were observed for foot vs. others, additionally revealing activations of the ipsilateral precentral sulcus and for lips vs. others in which the SMA did not show significant activations anymore (see Figure 4.8 and Figure 4.9).

4.2.6 *Landmark task*

For "all stimuli" contrast activation was found in the right Middle and Inferior Temporal Gyrus (V5 and overlapping with Inferior Parietal Cortex (IPC)), SMA (bilaterally), right Superior Frontal Gyrus, left Postcentral Gyrus (BA 2), right Precentral Gyrus (BA 6), right Postcentral Gyrus (BA 2) overlapping with right Inferior Parietal Lobule (IPC), right Insula, right Supramarginal Gyrus (IPC), right Superior Parietal Lobule (SPL), and left Middle Occipital Gyrus.

"Only stimuli with responses" elicited activation mainly in the right Superior and Inferior Parietal Lobule (SPL), left Fusiform Gyrus, left Cerbellum, left Postcentral

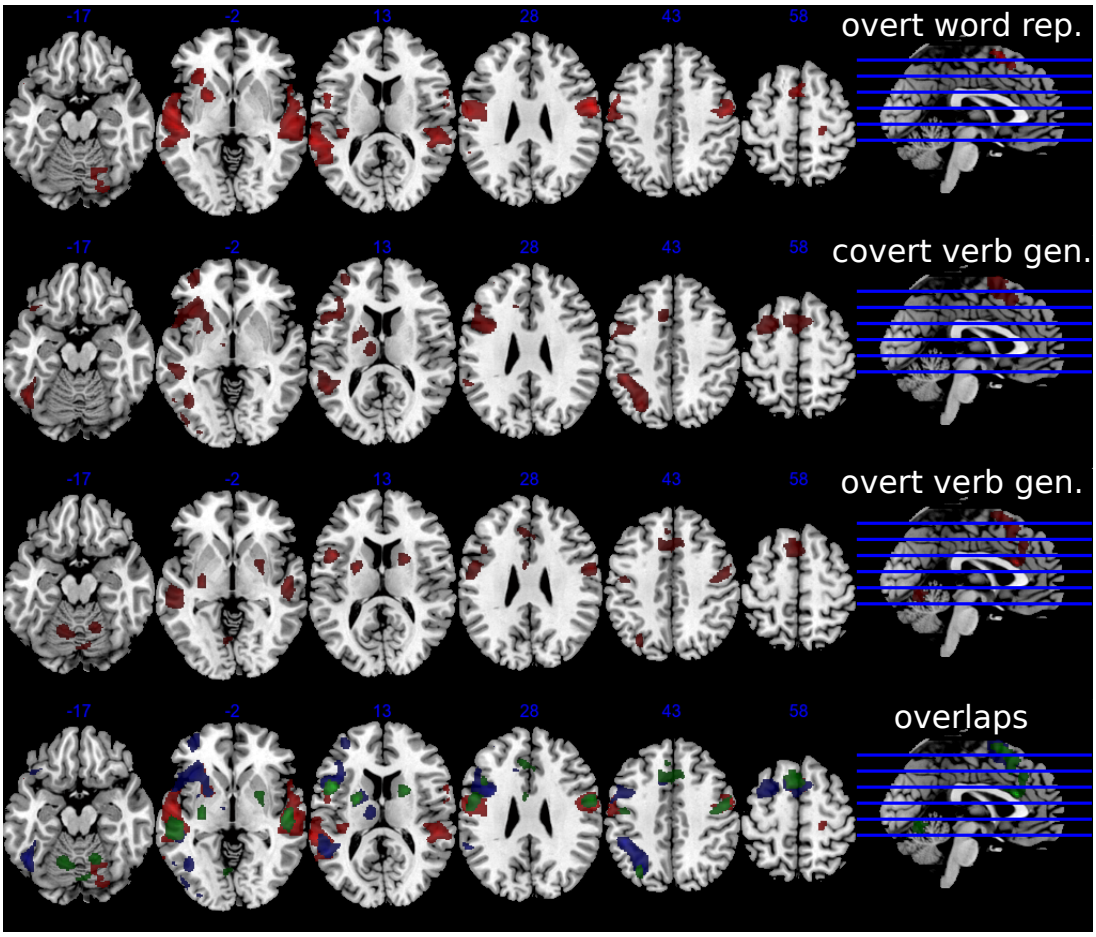

Figure 4.7: Results of the mixed effect analysis (thresholded using topological FDR at 0.05) for language tasks. From top to bottom: overt word repetition, covert verb generation, overt verb generation, overlap between overt word repetition (red), covert (blue) and overt (green) verb generation.

Gyrus (BA 2), right Inferior Temporal Gyrus, Precentral Gyrus (BA 6 - bilaterally), SMA (bilaterally), right Inferior Frontal Gyrus (BA 44 and 45), and left Calcarine Gyrus (BA 17).

"Only stimuli with correct responses" mostly replicated pattern of the previously described contrast with addition of Corpus Callosum and stronger activation in the left Visual Cortex.

"Correct vs. incorrect landmark responses" produced activation in the Occipital and Calcarine Gyri (bilaterally BA 17 and 18), left Putamen, left Fusiform and Lingual Gyri, left Cerebellum, left Inferior Frontal Gyrus, right Cuneus, right Thalamus (see Figure 4.10 and Figure 4.11).

Subjects responded to $87.25\% \pm 6.9888$ of the landmark stimuli and $78.5625\% \pm 4.7971$ of the detection stimuli. Out of all landmark stimuli, $73.0623\% \pm 9.562$ of responses were correct.

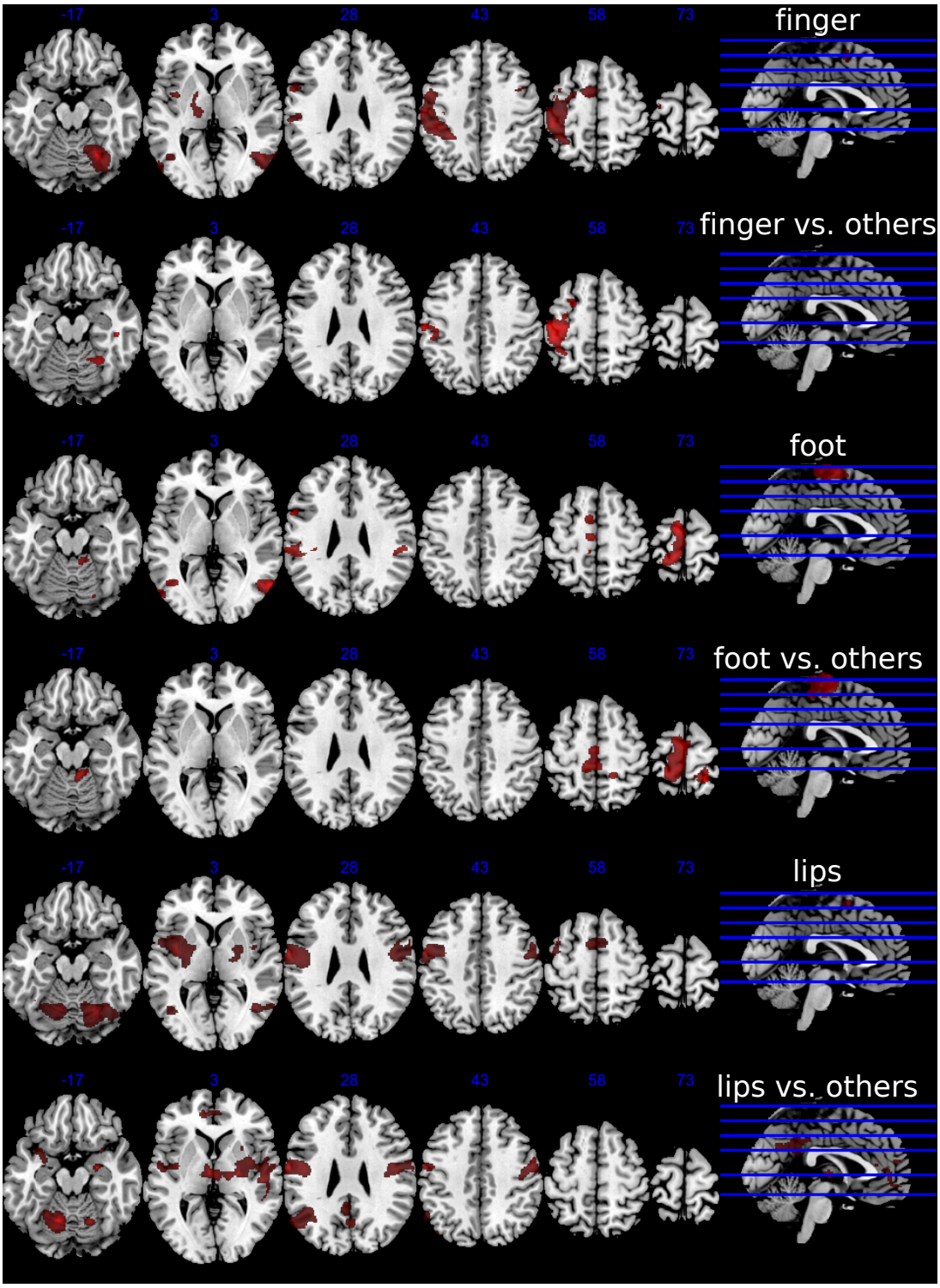

Figure 4.8: Results of the mixed effect analysis (thresholded using topological FDR at 0.05) for motor task. Contrasts from top to bottom: finger, finger vs. others, foot, foot vs. others, lips, lips vs. others.

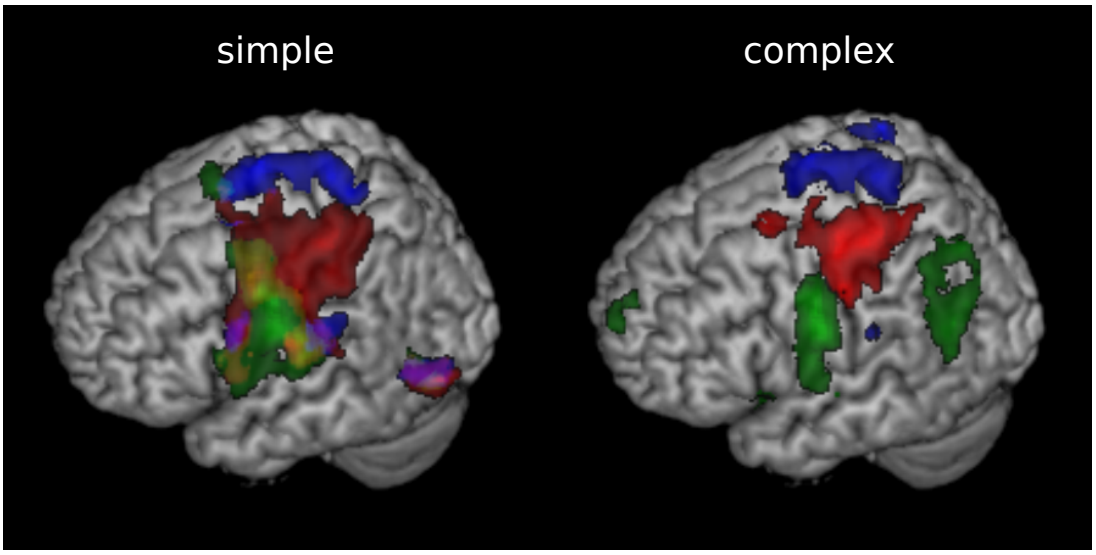

Figure 4.9: Motor task: 3D views. On the left: simple contrasts. On the right: complex (vs. others) contrasts. Finger (red), foot (blue), lips (green).

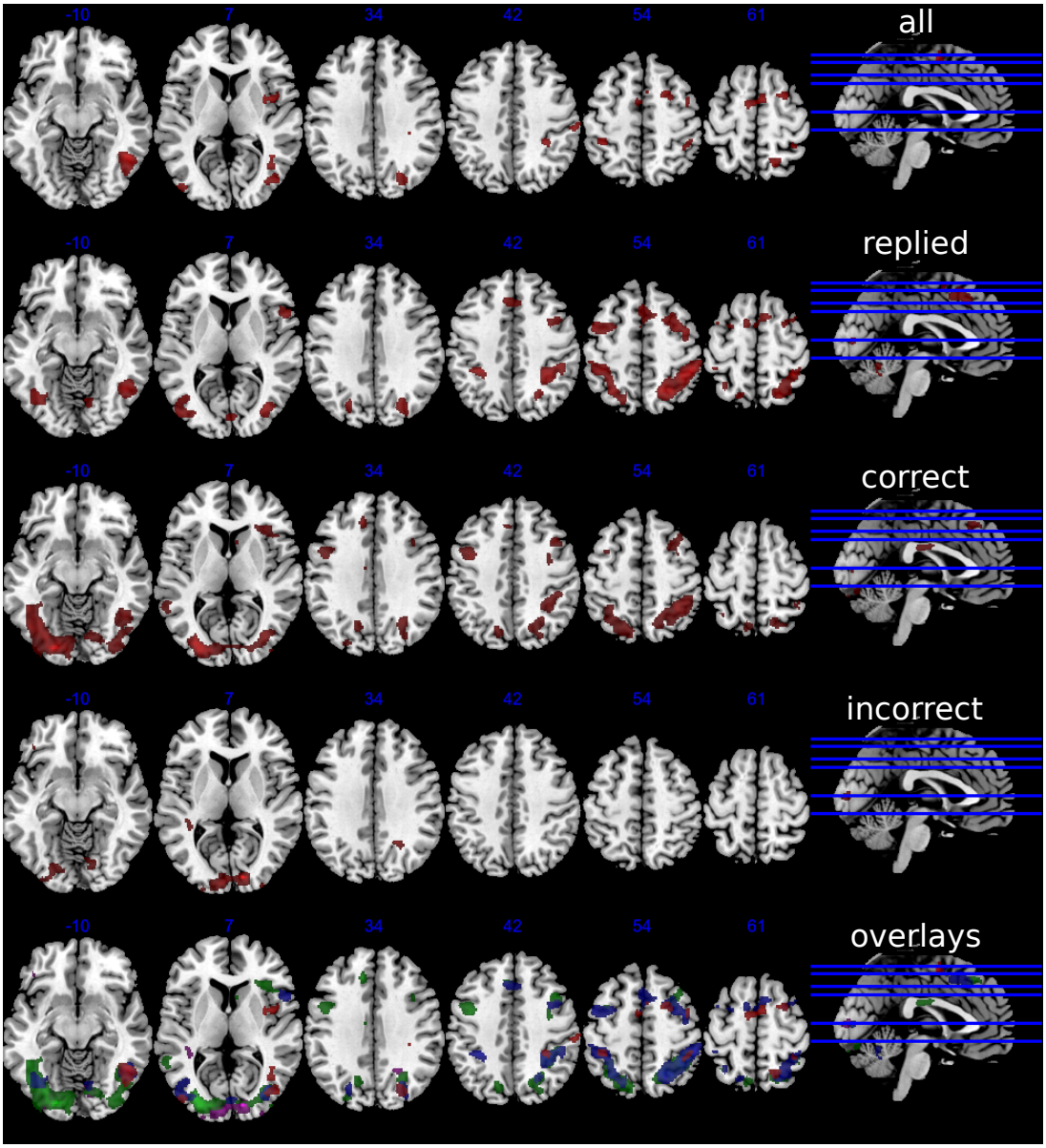

Figure 4.10: Results of the mixed effect analysis (thresholded using topological FDR at 0.05) for landmark task. From top to bottom: all stimuli, only stimuli with responses, only stimuli with correct responses, only stimuli with incorrect responses, overlap between the four above: all stimuli (red), only stimuli with responses (blue), only stimuli with correct responses (green), only stimuli with incorrect responses (purple).

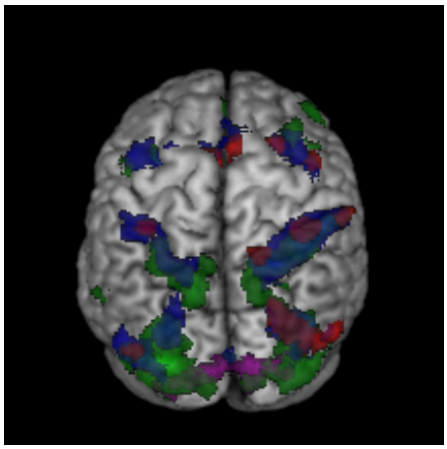

Figure 4.11: Landmark task: 3D views. Overlap between: all stimuli (red), only stimuli with responses (blue), only stimuli with correct responses (green), only stimuli with incorrect responses (purple).

4.3 RELIABILITY OF THRESHOLDING METHODS

Depending on the SNR not all statistical maps survive thresholding. In other words in some cases thresholding yields an empty map. Since the tasks used were straightforward and validated by the mixed effect analysis, we have assumed, for the time being, that every session should elicit some activation. Our working hypothesis was that AT will be able to find any activation within the region of interest for more sessions than FT. These results are summarised in Table 4.1. In all cases AT was at least as good as FT and in many cases found activation in more subjects. A McNemar test show that overall, Adaptive Thresholding outperforms Fixed Thresholding to detect activations (chi-squared(1) = 21.8065, p-value = 3.016×10^{-6}. Table 4.2).

Table 4.1: Percentage of subjects that produced any activation within the ROI. Contrasts where AT performed better than FT shown in bold.

Task	Contrast	Session 1		Session 2	
		Fixed Thr	Adaptive Thr	Fixed Thr	Adaptive Thr
Language Tasks	Overt word repetition	100.00%	100.00%	100.00%	100.00%
	Covert verb generation	100.00%	100.00%	100.00%	100.00%
	Overt verb generation	**75.00%**	**100.00%**	71.43%	71.43%
Motor Tasks	Finger	**90.00%**	**100.00%**	**80.00%**	**100.00%**
	Foot	100.00%	100.00%	**90.00%**	**100.00%**
	Lips	100.00%	100.00%	100.00%	100.00%
	Finger vs. Other	**90.00%**	**100.00%**	**80.00%**	**100.00%**
	Foot vs. Other	100.00%	100.00%	100.00%	100.00%
	Lips vs. Other	100.00%	100.00%	100.00%	100.00%
Landmark Task	Task All Greater Than Control All	**20.00%**	**30.00%**	**0.00%**	**50.00%**
	Task Answered Greater Than Control Answered	**60.00%**	**70.00%**	**60.00%**	**80.00%**
	Task Correct Greater Than Control Answered	**0.00%**	**50.00%**	**10.00%**	**60.00%**
	Task Correct Greater Than Task Incorrect	0.00%	0.00%	0.00%	0.00%

Table 4.2: Contingency table for activation within ROI for FT and FT thresholding methods. Sums are split between the first and second sessions. McNemar's $\chi^2 = 21.8065$, df = 1, p-value = 3.016×10^{-6}.

		Fixed Thr		
		Found	Not found	
Adaptive Thr	Found	89+86	12+17	101+103
	Not found	2+0	23+23	25+23
		91+86	35+40	126+126

To further investigate the performance of the two thresholding strategies we examined the thresholded maps. To make the comparison fair we have included only those subject/task/contrast combinations that yielded suprathreshold activation for both sessions and both thresholding methods. Based on mixed effect analysis and Table 4.1 we have excluded the three basic motor contrasts and all landmark contrasts except the "task answered vs. control answered" contrast. Multiple comparison correction was applied to the family of all 28 tests.

Investigation of shorter scanning times (re–analysis taking only 2, 3, or 4 first blocks of the scans) was performed by applying the same threshold to each subset. The comparison of thresholding methods used the same combinations of subject/task/contrast as for the full data set analysis.

4.3.1 *Language tasks*

Among all language tasks, AT and FT methods had similar reliability levels. The sign of the difference was always in favour of the adaptive technique but it did not reach significance for the three language tasks (see Table 4.3).

The analysis of the reduced datasets shows that adaptive thresholding recovers more signal with fewer datapoints (see 4.12). In all of the tasks, the advantage of AT over FT is higher for smaller datasets. For overt word repetition, the AT yielded a higher reliability than the FT with 2, 3 and 3 blocks of the data, while it also obtained 50% overlap with the final map using only half of the data. For overt verb generation, AT yielded to a higher reliability than the FT with 2, 3 and 4 blocks of the data, while it also obtained 50% overlap with the final map using 4 blocks of the data. For covert verb generation, the AT was not more reliable than the FT (except with 3 blocks of the data). However, it allowed 50% of the final map to be recovered using only 1 block of the data (vs. 3 blocks for the FT).

4.3.2 *Motor task*

In the motor task, the AT method was more reliable than FT method when contrasting hand and foot movements to other movements. Methods did not differ for the contrast lips vs. other movements (see Table 4.4).

In the motor tasks, as in the case of the language tasks, adaptive thresholding showed bigger improvement over fixed thresholding in smaller datasets (shorter se-

Table 4.3: Comparison of between session Dice overlaps between FT and AT for language tasks. All p values are two sided. Q values obtained by applying Benjamin–Hochberg False Discovery Correction to all p-values from Table 4.4 and Table 4.5.

	Overt Word Repetition (N=9)			
	H-D median		tmean	
	ROI	full	ROI	full
Fixed Thr	0.422022	0.354664	0.385309	0.348007
Adaptive Thr	0.585318	0.462406	0.575587	0.454410
CI	[0.009461 0.317416]	[-0.007999 0.253925]	[0.015558 0.305255]	[0.001280 0.228477]
p	0.024667	0.110667	0.011333	0.046667
q	0.088667	0.258222	0.063467	0.118788
	Covert Verb Generation (N=10)			
	H-D median		tmean	
	ROI	full	ROI	full
Fixed Thr	0.595107	0.527597	0.591907	0.522306
Adaptive Thr	0.639919	0.564095	0.637900	0.550031
CI	[-0.070474 0.112740]	[-0.077556 0.098734]	[-0.068632 0.146562]	[-0.074999 0.097227]
p	0.828000	0.572667	0.821333	0.663333
q	0.828000	0.697159	0.828000	0.742933
	Overt Verb Generation (N=5)			
	H-D median		tmean	
	ROI	full	ROI	full
Fixed Thr	0.391453	0.432157	0.398388	0.436665
Adaptive Thr	0.463781	0.434408	0.461588	0.438166
CI	[-0.075740 0.224551]	[-0.087042 0.133443]	[-0.085262 0.242001]	[-0.087574 0.142898]
p	0.301333	0.448000	0.334000	0.455333
q	0.519556	0.579515	0.519556	0.579515

Table 4.4: Comparison of between session Dice overlaps between FT and AT for motor task. All p values are two sided. Q values obtained by applying Benjamin-Hochberg FDR correction to all p-values from Table 4.3 and Table 4.5. Tests with q-values below 0.05 highlighted in bold.

	Finger vs. Other (N=7)			
	H-D median		tmean	
	ROI	full	ROI	full
Fixed Thr	**0.693010**	0.619350	**0.675953**	0.627243
Adaptive Thr	**0.802078**	0.647907	**0.792059**	0.626737
CI	**[0.025522**	[-0.118262	**[0.029282**	[-0.116193
	0.225890]	0.094456]	**0.219961]**	0.103365]
p	**0.000667**	0.292000	**0.000000**	0.428667
q	**0.004667**	0.519556	**0.000000**	0.579515
	Foot vs. Other (N=10)			
	H-D median		tmean	
	ROI	full	ROI	full
Fixed Thr	**0.495056**	0.502437	**0.501763**	0.485329
Adaptive Thr	**0.713597**	0.533353	**0.712552**	0.533903
CI	**[0.118172**	[-0.061228	**[0.115414**	[-0.055593
	0.331657]	0.147792]	**0.339045]**	0.138543]
p	**0.000000**	0.721333	**0.000000**	0.639333
q	**0.000000**	0.776821	**0.000000**	0.742933
	Lips vs. Other (N=10)			
	H-D median		tmean	
	ROI	full	ROI	full
Fixed Thr	0.805357	0.532662	0.794346	0.525366
Adaptive Thr	0.857566	0.464320	0.853359	0.464550
CI	[0.003405	[-0.102888	[0.001001	[-0.104175
	0.213174]	-0.021559]	0.215953]	-0.008844]
p	0.025333	0.014667	0.041333	0.029333
q	0.088667	0.068444	0.115733	0.091259

Table 4.5: Comparison of between session Dice overlaps between FT and AT for landmark task. All p values are two sided. Q values obtained by applying Benjamin–Hochberg False Discovery Correction to all p-values from this table. Tests with q-values below 0.05 highlighted in bold.

	Task Answered Greater Than Control Answered			
	H-D median		tmean	
	ROI	full	ROI	full
Fixed Thr	0.258408	0.240337	0.250415	0.269457
Adaptive Thr	0.428450	0.347364	0.444036	0.366116
CI	[-0.072476 0.616191]	[-0.125631 0.336689]	[-0.072476 0.616191]	[-0.090491 0.295687]
p	0.313333	0.402667	0.280667	0.318667
q	0.519556	0.579515	0.519556	0.519556
n	3	4	3	4

quences —- see Figure 4.13). AT also managed to recover the final map quicker from fewer datapoints than FT. For the finger vs. others contrast, AT yielded a higher reliability than the FT with 2 blocks, 3 blocks and 4 blocks of the data, while it also obtained 80% overlap with the final map using only 2 blocks of the data (vs. 70% with 4 blocks of the data with the FT). For the foot vs. others contrast, AT yielded to a higher reliability than the FT with 2 blocks, 3 blocks and 4 blocks of the data. It also obtained 70% overlap with the final map using only 2 blocks of the data (vs. 70% with 4 blocks of the data with the FT). For the lips vs others contrast, AT and FT did not differ due to higher variance between sessions. However, AT allowed 80% of the final map to be recovered with only 3 blocks of the data (vs. 4 blocks of the data for FT).

4.3.3 *Landmark task*

For the landmark task, only 1 contrast was tested (showing IPL activation in the RFX analysis) and in this case, the reliability of both methods was similar (see Table 4.5).

Due to the event related nature of the landmark paradigm reduced dataset analysis was not possible without changing the design matrix and contrasts, for example some subjects did not give any incorrect responses during first two blocks. Because this would make the comparison meaningless and due to the fact that landmark task did not show reasonable overall reproducibility (see next section) this analysis was not included.

4.3.4 *Thresholding method reliability and global effects*

Mapping of the parameter space (see Figure 4.14) showed that many combinations of thresholds can lead to high Dice overlap, and that the highest values were obtained when different thresholds between sessions were used. The reason behind this phe-

nomenon is that maximum T-values are often shifted between sessions as evidenced by looking at the joint distribution of T-values. Indeed the tail of the joint distribution is off–diagonal (see Figure 4.14b), meaning that voxels in the second scan session have higher or lower T-values than the same voxels in the first session. This effect is mostly observed when there is a shift of the overall distribution, i.e. in the context of a "global effect" (Friston et al., 1990) such as when temporal noise correlates with the stimuli sequence and affects the whole brain. Such a between session shift of T-values in a test–retest study has recently been reported by (Raemaekers et al., 2012). AT attempts to estimate and correct for this effect by allowing the Gaussian component to have non-zero mean and having the "activation" and "non-activation" components range fixed to that mean, leading to a choice of a pair of thresholds optimal in terms of Dice overlap (see Figure 4.14c).

Overall AT shows better sensitivity than FT, as it was able to find activation more often. We have also found using a test-retest dataset that it provides more reliable maps. Additionally we have shown that due to its adaptive nature it is able to recover signal with fewer samples than FT.

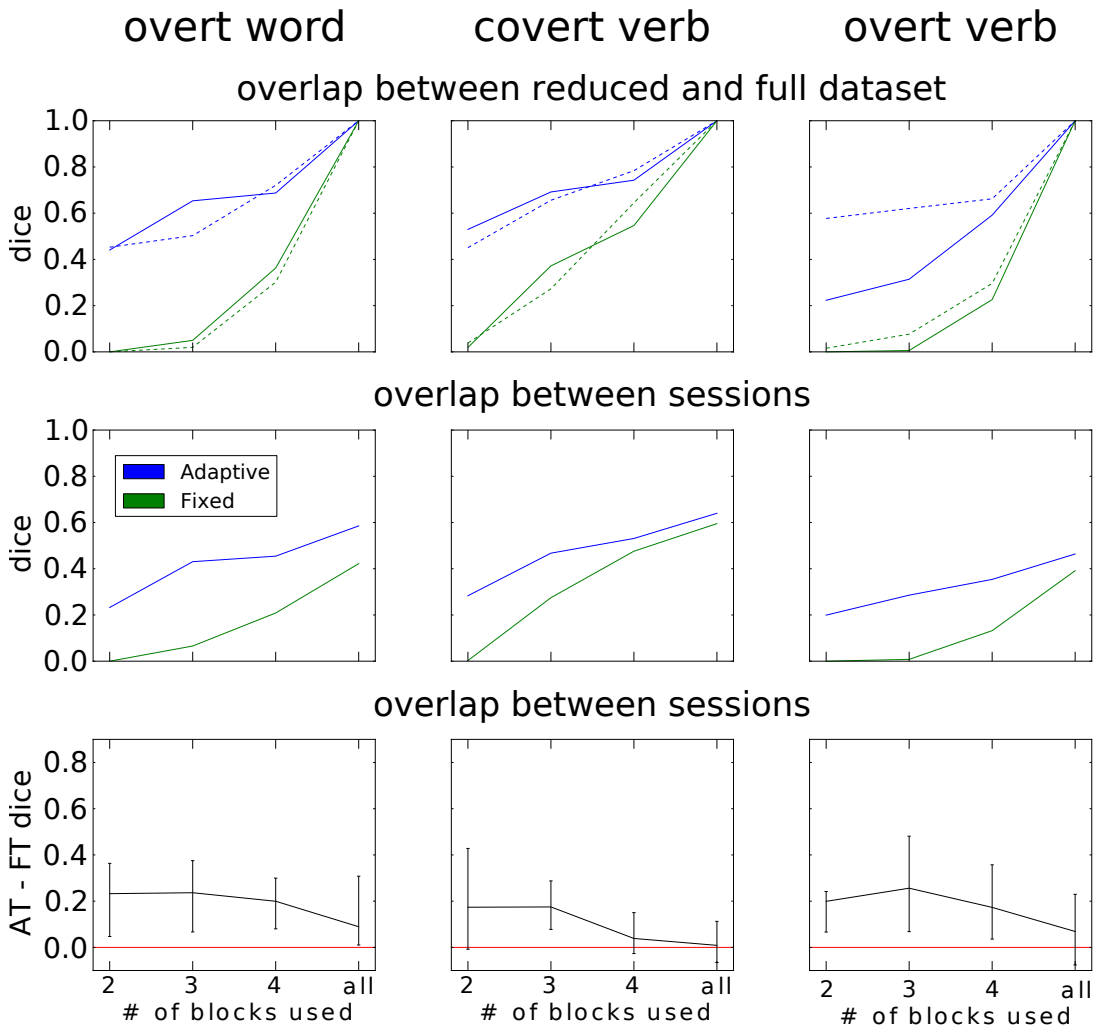

Figure 4.12: Results of applying AT and FT to data from just the first 2, 3, 4 or all blocks of the three evaluated language tasks. Top: H–D medians of Dice's coefficients between thresholded data of the reduced dataset and all data available for first (solid lines) and second (dashed lines) sessions, middle: H–D medians of between session overlaps, bottom: H–D medians of paired (within subjects) differences of Dice's coefficients, with confidence intervals bootstrapped at 0.05.

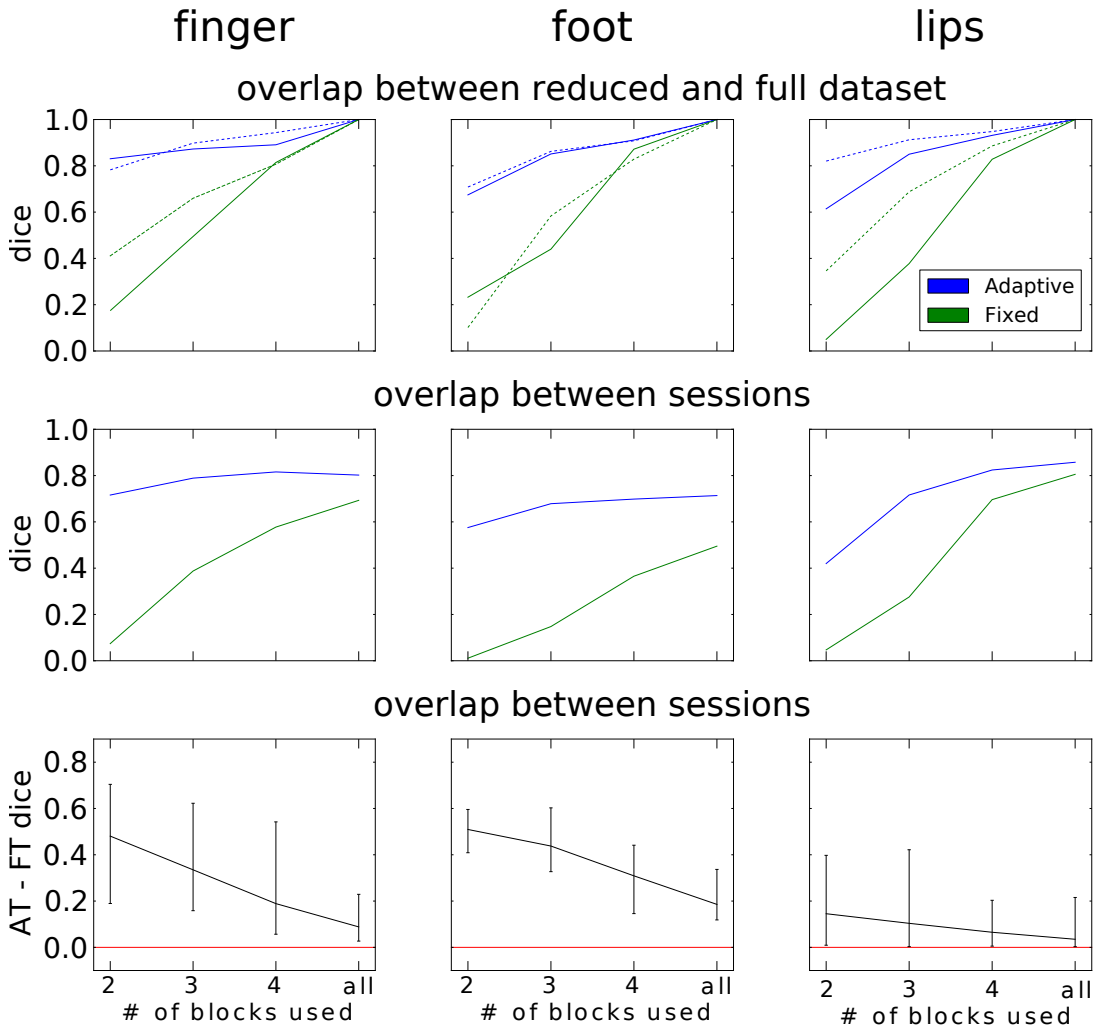

Figure 4.13: Results of applying AT and FT to data from just the first 2, 3, 4 or all blocks of three motor contrasts. Top: H-D medians of Dice's coefficients between thresholded data of the reduced dataset and all data available for first (solid lines) and second (dashed lines) sessions, middle: H-D medians of between session overlaps, bottom: H-D medians of paired (within subjects) differences of Dice's coefficients, with confidence intervals bootstrapped at 0.05.

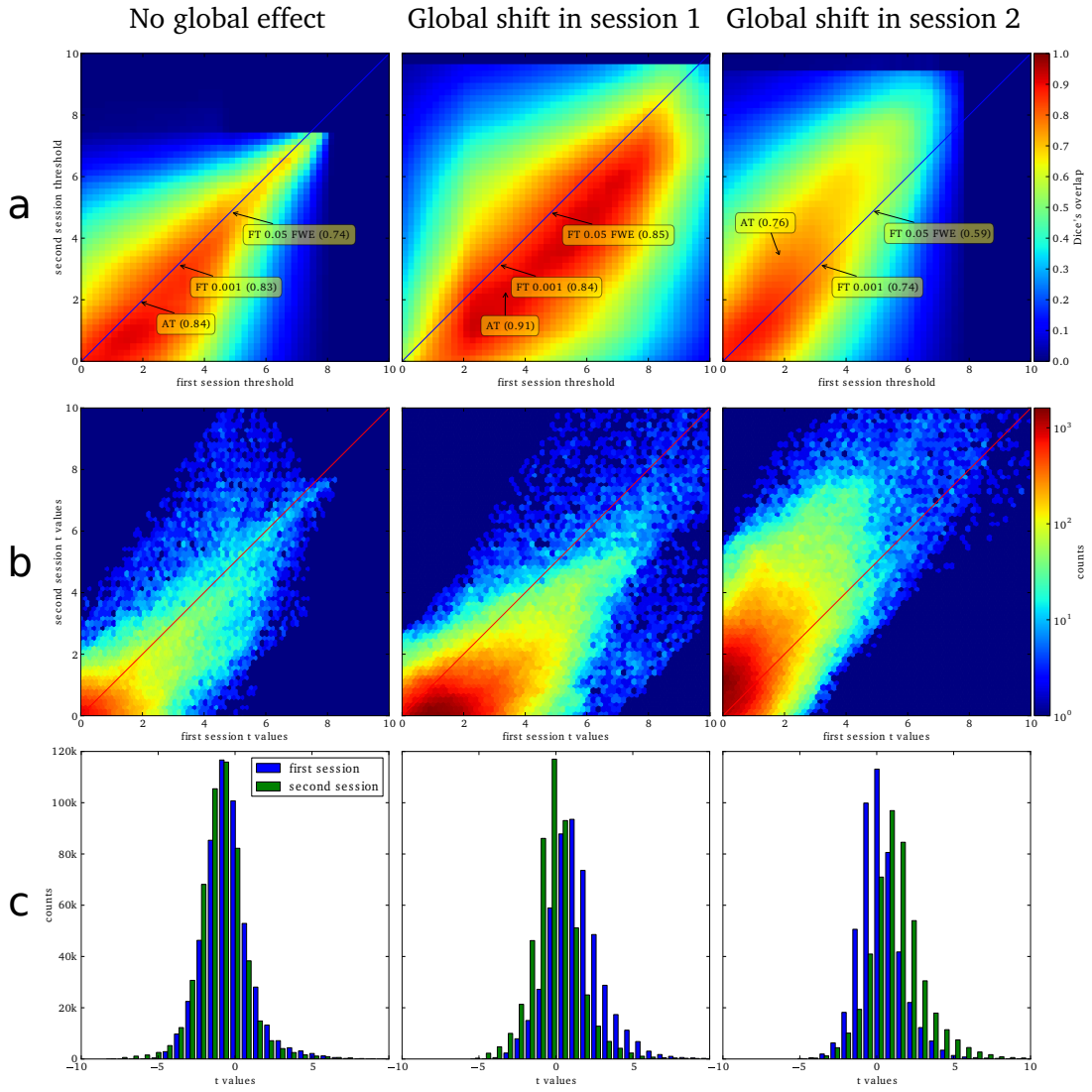

Figure 4.14: Analysis of the T-map reliability of three selected subjects. The top row (a) shows between session Dice coefficients for different pairs of cluster forming thresholds. The middle row (b) shows the upper right quadrant of the joint distribution of the unthresholded T-values, while the bottom row (c) shows distributions of T-values from the first and the second session. "No global effects" (example from finger contrast for subject 1) illustrates the case where choosing the same threshold for both sessions is the optimal course of action; the joint distribution confirms this by showing lack of a consistent between session value shift, while AT manages to infer this without having access to the joint distribution. "Global shift session 1" (example from lip movement contrast for subject 2) shows a shift of values between the sessions. This is clear not only from the joint distribution but from the two separate distributions. This allows AT to choose a lower threshold for the second session and optimize the Dice coefficient value. "Global shift session 2" (example from foot contrast for subject 3) indicate a shift in the opposite direction.

4.4 RELIABILITY OF TASKS

The goal of these comparisons was to test if the variance between two sessions was smaller than variance between subjects, i.e. evaluate the test-retest reliability of the tasks. Subjects who did not show activation in both sessions were again excluded. Based on the results from the previous section showing the superiority of the AT method, all analyses were performed using the adaptive strategy.

4.4.1 *Language tasks*

Overall, the three tasks show a higher reliability within than between subjects (see Table 4.6). For the overt word repetition task, activations restricted to Wernicke area were, however, not reliable whereas the full brain pattern of activity was. Covert and overt word repetition tasks were also reliable when considering the full brain pattern of activity, but only the covert verb generation task showed reliable results over Broca area.

Looking at the ICC values for ROI, there was a close similarity with the RFX analyses (Figure 4.15).

For overt word repetition, maximum T–values of the RFX analyses matched maximum ICC values over Wernicke area and maximum single subject maps. Thus, in Wernicke area, despite its lack of reliability within subjects, there was a signal increase consistent across subjects averaged over sessions, which also has similar amplitudes in both sessions and is strong enough to be seen in all subjects; at the max of RFX x=-54, y=-24, z= 1.5, ICC = 0.59 and 0.55% of subjects showed an activation. This pattern of results was different for the contra-lateral homologue. Maximum T-values of the RFX analyses did not correspond to maximum ICC and single subject maps. For instance, in the anterior STS, there was no consistency across subjects (RFX analysis not significant) despite reliable amplitude increase over sessions and within subjects.

For the covert verb generation task RFX analysis showed activation mainly in the left hemisphere despite the fact that there were high ICC values in both hemispheres. Stacked overlaps (frequency map of intersections of thresholded areas from both sessions for all subject included in the analysis) reflected the ICC findings with a slight incline towards lateralized activation; at the max of RFX x=-43.5, y=18, z=10.5, ICC = 0.77 and 40% of subjects showed an activation. For the overt verb generation task both RFX and stacked overlaps showed activation in posterior parts of the mask. To a certain extent, patterns of ICC in the left hemisphere did correspond with this activation, but this was not the case for the right hemisphere. Here low or even negative ICC [1] values translated into suprathreshold T values of RFX analysis and non zero stacked overlap; at the max of RFX x=-45, y=10.5, z=13.5 ICC = -0.61 and none of subjects showed an activation.

1 Theoretically, ICC is non negative since this is defined as a ratio of variances (which are positive values). However, estimates of ICC can be negative. This is explained by the fact that both session and subject effects are modeled and variances are estimated using a two-way ANOVA model. For the variance estimators not to be biased they have to allow for negative values (otherwise estimates of zero value would not average to zero). In general, this indicates low inter-session reliability. A more detailed explanation together with examples can be found in Taylor (2010).

Table 4.6: Comparison between within and between subject variance for language tasks. All p values are two sided. Q values obtained by applying B-H FDR to all p-values for each contrast separately. Tests with q-values below 0.05 highlighted in bold.

	Overt word repetition (N within =9, N between = 36)			
	H-D median		tmean	
	ROI	full	ROI	full
within	**0.585318**	**0.462406**	**0.575587**	**0.454410**
between	**0.340059**	**0.223096**	**0.309611**	**0.220995**
CI	**[0.069453**	**[0.118324**	**[0.052999**	**[0.133929**
	0.419928]	**0.347355]**	**0.394757]**	**0.327865]**
p	**0.010000**	**0.000000**	**0.015333**	**0.000000**
q	**0.013333**	**0.000000**	**0.015333**	**0.000000**
	Covert verb generation (N within =10, N between = 45)			
	H-D median		tmean	
	ROI	full	ROI	full
within	**0.639919**	**0.564095**	**0.637900**	**0.550031**
between	**0.321044**	**0.249421**	**0.341339**	**0.248647**
CI	**[0.052376**	**[0.069289**	**[0.048845**	**[0.068382**
	0.491096]	**0.421119]**	**0.457233]**	**0.416956]**
p	**0.017333**	**0.012667**	**0.016667**	**0.008667**
q	**0.017333**	**0.017333**	**0.017333**	**0.017333**
	Overt verb generation (N within = 7, N between = 21)			
	H-D median		tmean	
	ROI	full	ROI	full
within	0.309291	0.337577	0.302969	0.326379
between	0.055683	0.069443	0.053890	0.065974
CI	[-0.012147	[0.022319	[-0.004863	[0.032864
	0.508635]	0.465002]	0.502423]	0.446286]
p	0.069333	0.038000	0.055333	0.026000
q	0.069333	0.069333	0.069333	0.069333

4.4.2 *Overt vs. covert verb generation*

Between subject variability for covert verb generation was much lower than for the overt verb generation as observed with ICC and stacked overlap (see Figure 4.16).

Distributions of ICC over Broca's area for covert and overt variants of the verb generation task show a heavier tail (negative ICC) for the overt task, meaning that opposite activation patterns were observed between sessions. In comparison, only a few voxels showed negative ICC in the covert task making, it more reliable. Within subject (between sessions) overlaps of the overt and covert verb generation task were also compared using a paired bootstrap test. Adaptive thresholding was used to create suprathreshold maps and comparison was restricted to Broca's area and its right homologue (see Figure 4.17). The test yielded significant difference (p=0.001) in favour of the Covert verb generation task with 6 out of 7 subjects in the comparison showing a bigger between session overlap in the covert version of the task.

4.4.3 *Motor task*

All tested motor contrasts yielded significantly bigger within subject than between subject reliability (Table 4.7).

Comparison of RFX analysis, ICC and stacked overlaps maps (see Figure 4.18) show a good consistency: this paradigm leads to strong activations in each subject and over both sessions (finger: x=-46.5, y=-18, z=57, ICC =0.49 and 70% of subjects showed an activation, foot: max of RFX x=-10.5, y=-39, z=70.5 , ICC = 0.14 and 80% of subjects showed an activation, lips: max of RFX x=-57, y=-6, z=39, ICC =-0.15 and 80% of subjects showed an activation).

4.4.4 *Landmark task*

In the only contrast of the landmark task that showed any activation in the right parietal lobe (answered trials landmark vs. answered trials detection), there was no significant difference between within and between subject overlaps (Table 4.8). Both variances were high (small overlaps), and in the case of analysis restricted to an ROI, the average overlap between different subjects was bigger than within the same subjects; however, this difference was not statistically significant.

Peak activation of the RFX analysis corresponded with low ICC values (Figure 4.19). On the other side peak ICC values were observed for the same voxels that showed any non zero values on the stacked overlaps map (note on the stacked map the maximum overlap is N=3). Thus IPL activation is small and varies between sessions, although it is consistent across subjects.

4.4.5 *Evaluating tasks through their test-retest reliability*

We have developed a new technique for looking at the internal validity of a task. Comparing between and within subject Dice overlaps enables a choice between which tasks are suitable for surgical planning. We have used this technique to choose the covert verb generation task over its overt variant and discard the landmark task as

Table 4.7: Comparison between within and between subject variance for motor tasks. All p values are two sided. Q values obtained by applying Benjamin-Hochberg False Discovery Correction to all p-values for each contrast separately. Tests with q-values below 0.05 highlighted in bold.

	Finger vs. Other (N within =10, N between = 45)			
	H-D median		tmean	
	ROI	full	ROI	full
within	**0.761132**	**0.619930**	**0.756824**	**0.608450**
between	**0.562011**	**0.338060**	**0.567423**	**0.346512**
CI	**[0.030310 0.342250]**	**[0.071173 0.391105]**	**[0.041751 0.326450]**	**[0.070452 0.388408]**
p	**0.013333**	**0.004667**	**0.008000**	**0.004667**
q	**0.013333**	**0.009333**	**0.010667**	**0.009333**
	Foot vs. Other (N within =10, N between = 45)			
	H-D median		tmean	
	ROI	full	ROI	full
within	**0.713597**	**0.533353**	**0.712552**	**0.533903**
between	**0.557540**	**0.313623**	**0.544175**	**0.321310**
CI	**[0.076407 0.263564]**	**[0.086434 0.310117]**	**[0.087887 0.271345]**	**[0.086559 0.299603]**
p	**0.000000**	**0.000667**	**0.000000**	**0.000667**
q	**0.000000**	**0.000667**	**0.000000**	**0.000667**
	Lips vs. Other (N within =10, N between = 45)			
	H-D median		tmean	
	ROI	full	ROI	full
within	**0.857566**	**0.464320**	**0.853359**	**0.464550**
between	**0.657252**	**0.239766**	**0.646546**	**0.267673**
CI	**[0.109267 0.275757]**	**[0.082572 0.326871]**	**[0.116745 0.277179]**	**[0.067742 0.306335]**
p	**0.000000**	**0.002667**	**0.000000**	**0.004667**
q	**0.000000**	**0.003556**	**0.000000**	**0.004667**

Table 4.8: Comparison between within and between subject variance for landmark task. All p values are two sided. Q values obtained by applying Benjamin-Hochberg False Discovery Correction to all p- for each contrast separately. Tests with q-values below 0.05 highlighted in bold.

	Landmark (N within =10 N between = 45)			
	H-D median		tmean	
	ROI	full	ROI	full
within	0.065536	0.113074	0.081616	0.126288
between	0.197776	0.091920	0.184178	0.098469
CI	[-0.253168 0.124880]	[-0.117770 0.220737]	[-0.216622 0.104472]	[-0.100957 0.199243]
p	0.277333	0.791333	0.308000	0.745333
q	0.616000	0.791333	0.616000	0.791333

not being reproducible on single subject level. In addition the technique might provide valuable insights for neuroimaging in general as the comparison with RFX and ICC allows those areas showing strong and reliable signals (RFX, ICC, and stacked overlap) to be distinguished from those with weak but consistent signal (high t values on the group level, but low ICC and stacked overlap); such patterns might reflect differences inherent to the computations being performed and the necessity of such regions in the task involved.

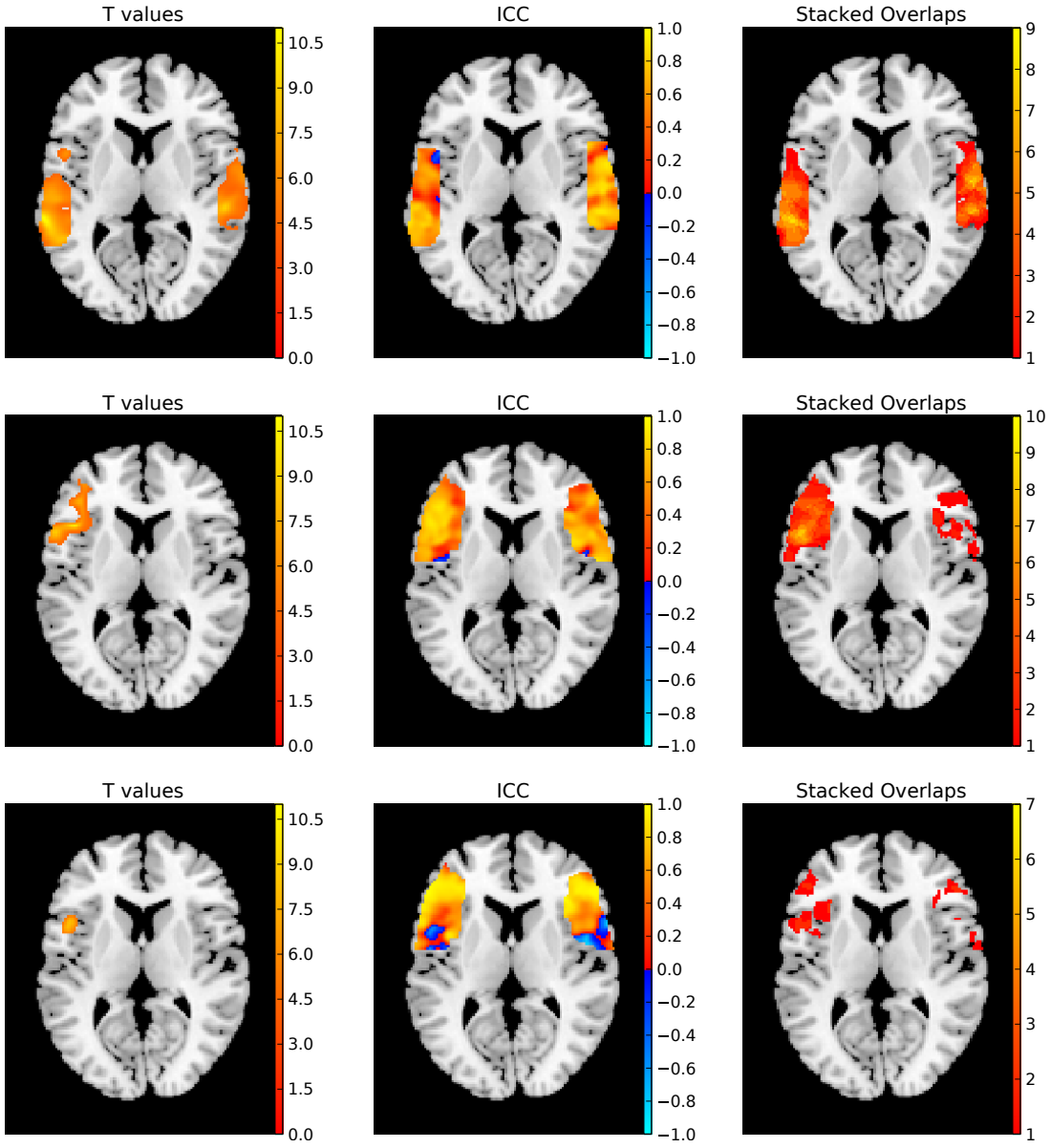

Figure 4.15: Comparison of mixed effect analysis, ICC, and stacked overlaps. All maps were restricted with their respective ROI masks. Mixed effects map was thresholded using topological FDR with 0.05 cluster-forming threshold and 0.05 cluster extent threshold. Stacked overlaps map was thresholded at 1. Tasks from the top: overt word repetition, covert verb generation, and overt verb generation.

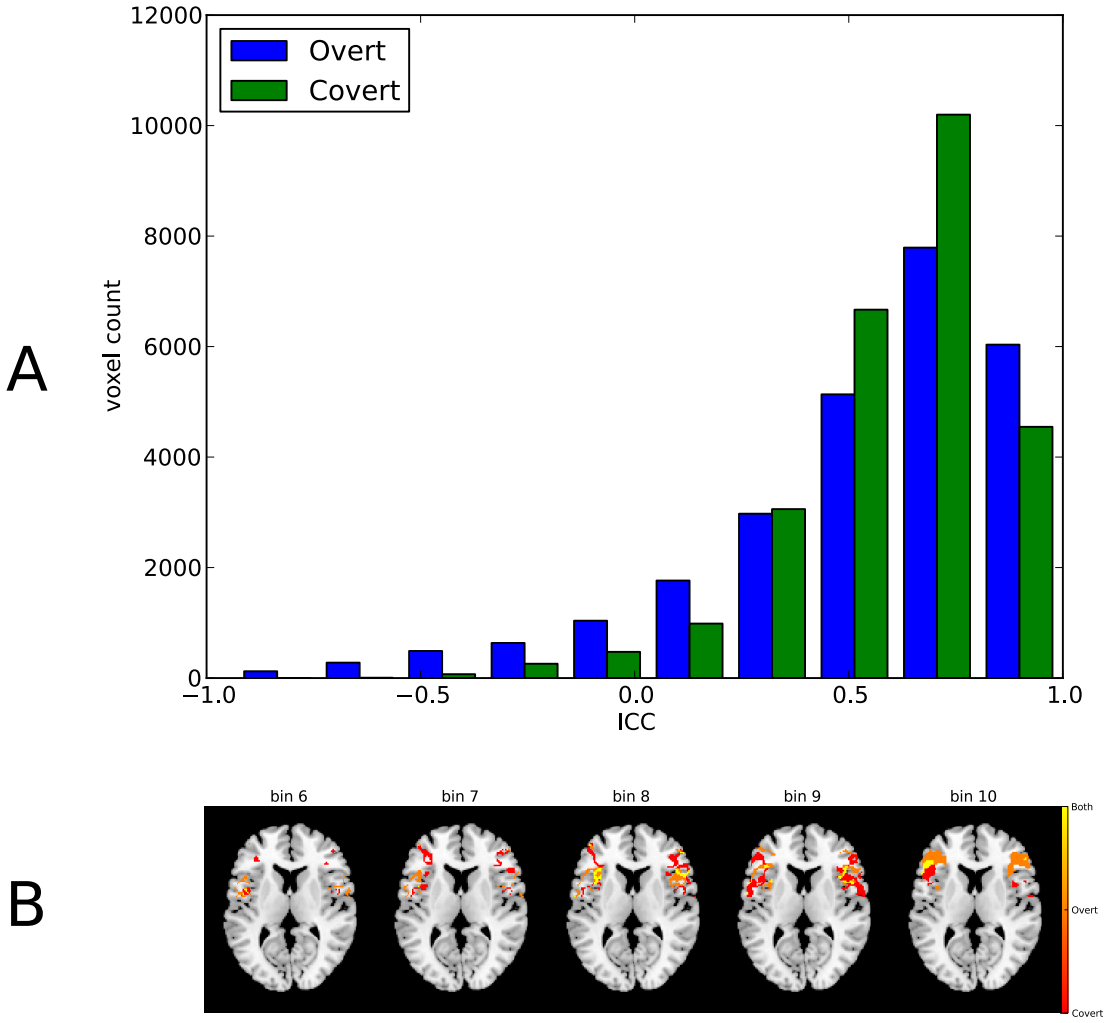

Figure 4.16: A) Distributions of covert (green) and overt (blue) ICC values within the ROI. (B) Map showing where the ICC distributions are different.

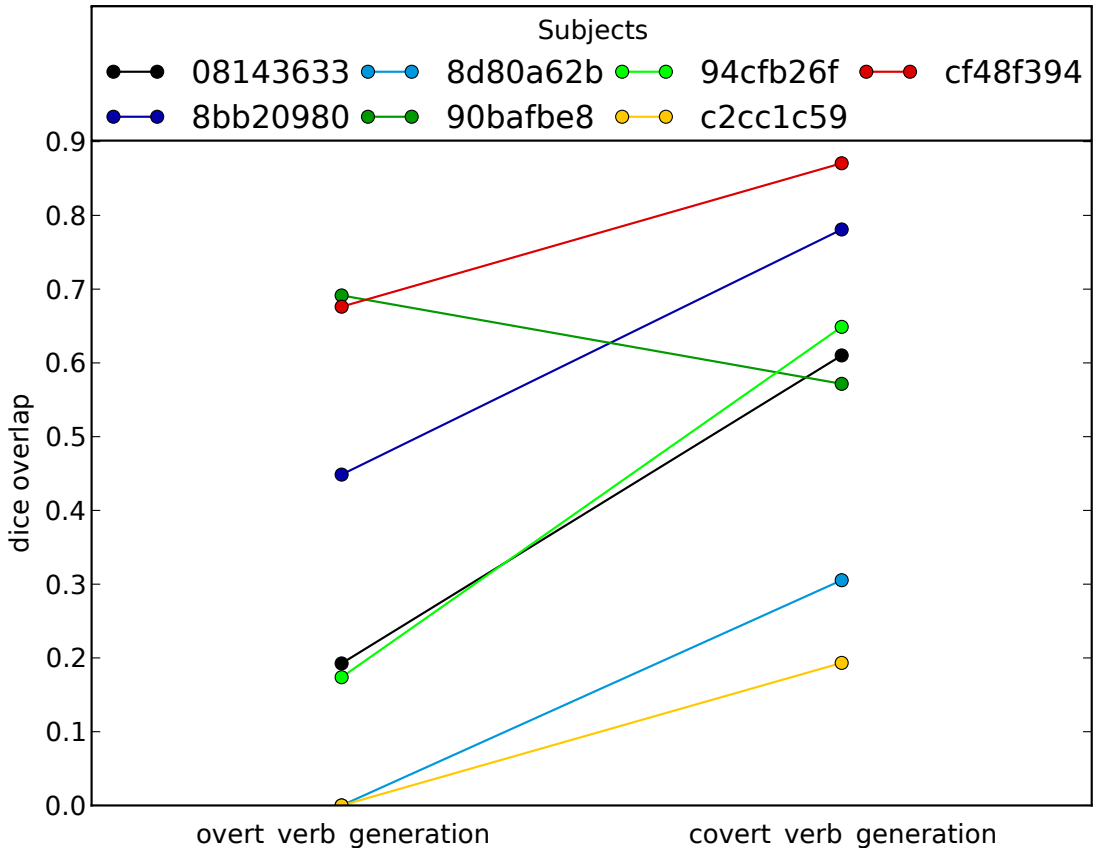

Figure 4.17: Dice coefficients (within ROI) for covert and overt verb generation tasks. Adaptive Thresholding on the left and Fixed Thresholding on the right.

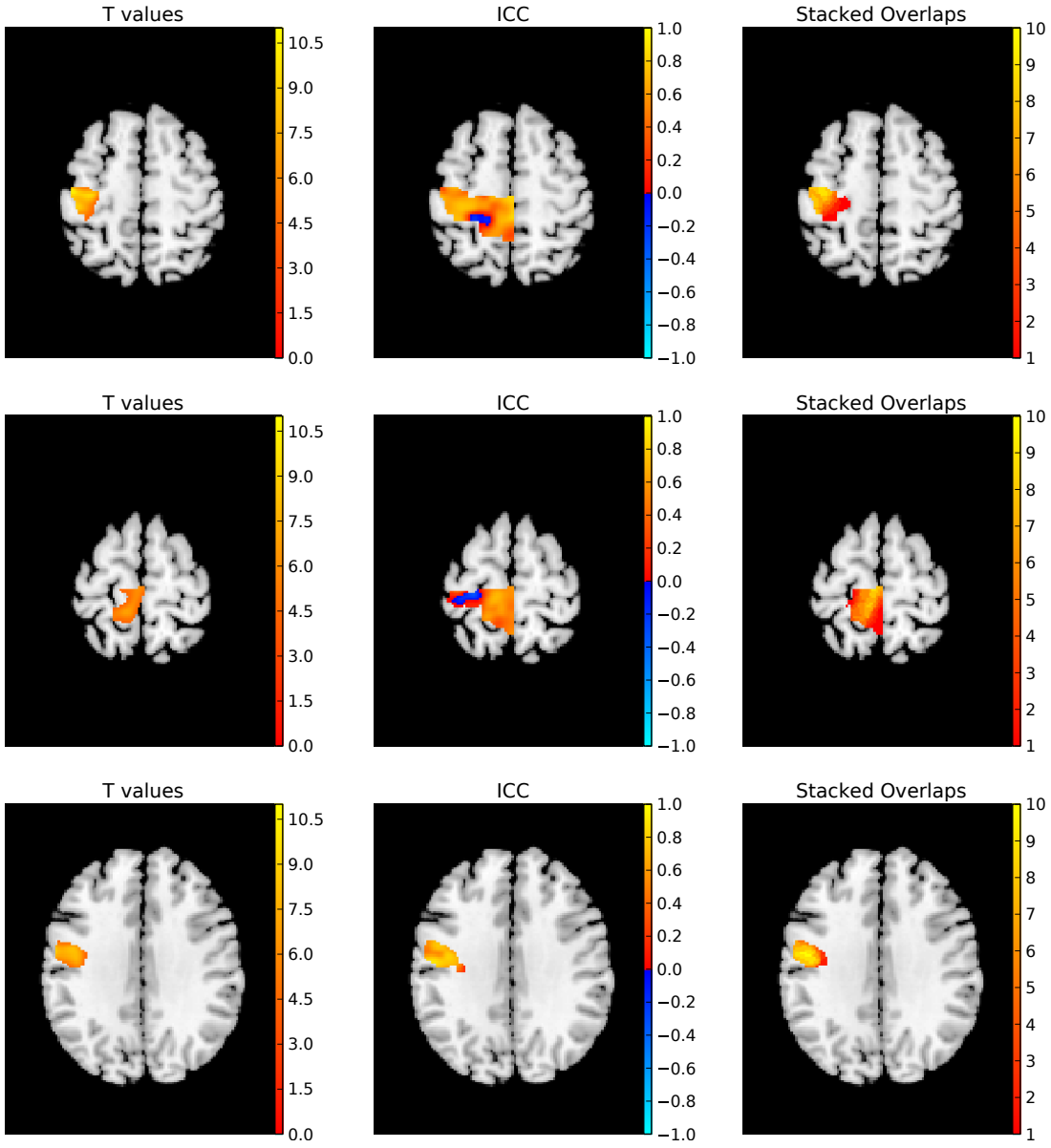

Figure 4.18: Comparison of mixed effect analysis, ICC, and stacked overlaps. All maps were restricted with their respective ROI masks. Mixed effects map was thresholded using topological FDR with 0.05 cluster-forming threshold and 0.05 cluster extent threshold. Stacked overlaps map was thresholded at 1. Tasks from the top: finger vs. others, foot vs. others, and lips vs. others.

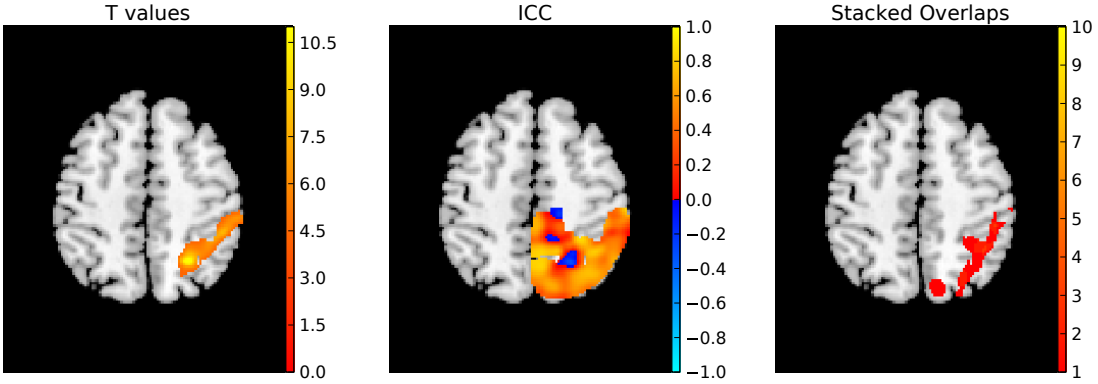

Figure 4.19: Comparison of mixed effect analysis, ICC, and stacked overlaps. All maps were restricted with the ROI mask. Mixed effects map was thresholded using topological FDR with 0.05 cluster-forming threshold and 0.05 cluster extent threshold. Stacked overlaps map was thresholded at 1. Landmark task – task answered greater than control answered contrast.

4.5 RELIABILITY METRICS AND CONFOUNDING FACTORS

4.5.1 *Introduction*

Having used test–retest reliability to compare different thresholding methods and behavioural paradigms we turn to a more general problem of relations between different ways to measure reliability and factors that can influence it. Despite the fact that fMRI has been used in thousands of studies, many of which have been independently replicated, there is as yet no consensus on how reliable fMRI measurements are (Bennett and Miller, 2010). At the same time it is widely accepted that fMRI can provide valuable insights into the human brain even when used on the single subject level. In other words, the result of analysing fMRI time-series is not random. However, it is also accepted that there is some variability in the results that cannot be accounted for by experimental variables. Understanding this variability of fMRI is crucial to delineating limits of fMRI as a research tool.

Despite the main theme of this thesis it is worth noting that single subject fMRI is not limited to presurgical mapping. It potentially can be used as a diagnostic tool (Raschle et al., 2012) and a way to plan and monitor rehabilitation (Dong et al., 2006). It is also being used to define individual functional regions of interest (ROIs) through functional localizer tasks (Duncan et al., 2009).

In comparison to group studies, the change of focus in single subject studies is reflected in a different approach to analysing data. The Holmes–Friston (Holmes and Friston, 1998) approach discards uncertainty of the first level analysis, the within-subject variance, by using each subject's contrast maps instead of t-maps. The uncertainty that influences the group level results comes from the between–subject variance. In contrast, a single subject examination relies on t-maps, instead of beta parameters maps, and thus depends on within–subject variance. This difference between which variance is relied upon has implications for what levels and metrics of reliability are suitable for group and single subject analyses. For group studies, it is reasonable to look at the within– and between–session variance of contrast maps as well as the similarity of thresholded and unthresholded group level t-maps. By contrast, for single subject studies, it is the within– and between–session variance of the BOLD signal and the similarity of t-maps that are relevant.

In the previous analyses we have primarily used volume overlap as a simple measure to quantify reliability. This method has the advantage of examining the final product of the neuroimaging analysis, the t-maps, and the same procedure applies to group or single subjects maps. However, overlap values heavily depend on the threshold applied to the t-maps, since the cluster overlap measures decrease with increasing threshold (Fernández et al., 2003; Duncan et al., 2009). Additionally, when used over the whole brain rather than for a specific cluster of interest, different thresholds can lead to different activation maps but a similar measure of overlap. Finally, this technique is sensitive to borderline cases; two very similar t-maps, one slightly above a threshold and another slightly below, would give a false impression of high variability (Smith et al., 2005). Nonetheless, thresholded maps are the typical end product of fMRI analyses and are used for ROI definitions. Furthermore, in the context of presurgical planning, which is the main focus of this work, where single subject thresholded maps are used, their variability is a major concern.

Nonetheless there are other levels at which one can look at reliability of fMRI. ICC is one of the other popular reliability measures. We have used it in previous analyses because it is widely acknowledged in the community. It was initially used in psychology to asses between raters variability (Shrout and Fleiss, 1979), but has been adapted to measure reliability (McGraw and Wong, 1996) by replacing judges/raters by repeated measurement sessions. Since this metric combines both between-subject and between-session variance, it is suitable for providing insights into random effect group analyses. However, the same value of ICC can come from both high σ_r^2(beetween subjects variance) and low σ_e^2(between sessions varaince) or low σ_r^2 and high σ_e^2, which makes the comparison between tasks harder. ICC is in fact more heavily influenced by between-subject variance than between-session variance (the variable of interest), making its usefulness as a quality estimator for group studies debatable. From the single subject point of view, between-subject variance is irrelevant and therefore it is more informative to consider only between-session variance. Furthermore, in contrast to volume overlap, it is not the variance of contrast maps (between–subject) that must be considered but the variance of t maps (contrast maps weighted by error). In the same way volume overlap is sensitive to the selected threshold, t value variability in ICC can be influenced by the design matrix used in GLM. This involves regressors, the hemodynamic response function (HRF) and contrasts definitions. For instance, Caceres et al. (2009) found that one can have highly correlated time–series but with a poor model fit leading to low reliability. They concluded that the wrong HRF model can lead to low reliability. However, inadequate regressors and contrast could also lead to similar results.

Apart from the issue of how to measure fMRI reliability, a further important question is what causes the lack of reliability in the first place and how this could be prevented. One of the suspected sources of variation in brain activation patterns is the possibility that different cognitive strategies and therefore different neuronal responses are produced by different subjects. These effects don't necessarily have to be task related. In a block design experiment, it would be enough that the subject consistently performs different mental tasks during the rest period to provide significantly variable results. The influence of this kind of variability is very hard to quantify because of the lack of access to the true neuronal activation patterns. It is, however, very likely that the type of task can reduce this "cognitive noise". For example, a simple finger tapping task involving primary motor cortex requires fewer possible cognitive strategies than the Iowa Gambling Task. Other possible sources of reduced reliability are easier to quantify. These include, but are not limited to, scanner noise, subject motion, and between–session coregistration errors. Even though these confounds have been recognized in the literature numerous times, to our knowledge, there is no analysis on how much they contribute to reliability metrics. To date, the only study examining such effect was performed by Raemaekers et al. (2007) who showed a positive correlation between "sensitivity" (average absolute t value) and between–session volume overlap.

In this section we will try to quantify how different measures of reliability relate to each other and how they are influenced by confounding factors.

4.5.2 *Reliability measurements and confounds*

To simplify the analysis we have not used the overt verb generation task in the following section. This decision was justified by the fact that it performed worse than its covert counterpart which we have shown in the previous section.

4.5.2.1 *Measuring reliability*

In addition to volume overlap of thresholded t maps described in the previous section the following additional measurements were performed:

BETWEEN–SESSION CORRELATION ON TIME–SERIES After the EPI sequences have been realigned, normalized, spatially smoothed, and detrended using second order polynomials, Pearson correlation coefficients between first and second session time-series were calculated for each voxel and then averaged. This allowed the determination of the similarity of the measurements before any statistical or HRF models had been fitted.

BETWEEN-SESSION VARIANCE OF UNTHRESHOLDED T MAPS t-maps were first corrected for global effects using estimates of the mean of the noise distribution estimated using AT (see Chapter 3). The mean of the squared between–session differences was calculated

$$t_{diff} = \frac{1}{n}\sum_{i}(t_{i1} - t_{i2})^2 \tag{4.2}$$

where n is the number of voxels, t_{i1} and t_{i2} are the ith voxel t values from the first or second session respectively. This measure is equivalent to the between-session component of Intraclass Correlation Coefficient (ICC), but adapted here for single subject analysis. The mean t_{diff} across subjects is inversely proportional to ICC, assuming constant between-subjects variance across sessions. The derivations for this relation depend on the assumptions made while calculating ICC (for k=2 sessions case):

1. ICC(1) assumes no session (learning) effects and is defined for one voxel as following

$$ICC(1) = \frac{\sigma_r^2}{\sigma_r^2 + \sigma_w^2} \tag{4.3}$$

where σ_r^2 is between subjects variance and σ_w^2 is defined in the following manner

$$\sigma_w^2 = \frac{\sum_{j=1}^{n}(t_{1j} - t_{2j})^2}{n} = \frac{\sum_{j=1}^{n} t_{diff}(j)}{n} \tag{4.4}$$

where n is the number of subjects, t_{1j} and t_{2j} are t values for subject j for first and second sessions respectively, $t_{diff}(j)$ is t_{diff} for subject j. Therefore:

$$((ICC(1) \propto \frac{1}{\sigma_w^2}) \cup (\sigma_w^2 \propto t_{diff})) \Rightarrow ICC(1) \propto \frac{1}{t_{diff}} \tag{4.5}$$

2. ICC(3,1) assumes session effects (learning) and is defined as

$$ICC(3,1) = \frac{\sigma_r^2}{\sigma_r^2 + \sigma_e^2} \tag{4.6}$$

where σ_e^2 is defined in the following manner

$$\sigma_e^2 = \frac{\sum_{j=1}^{n}((t_{1j} - \overline{t_1}) - (t_{2j} - \overline{t_2}))^2}{n-1} \tag{4.7}$$

where $\overline{t_1}$ and $\overline{t_2}$ are across subjects mean t values for first and second sessions respectively. Since:

$$\begin{aligned}((t_{1j} - \overline{t_1}) - (t_{2j} - \overline{t_2}))^2 &= ((t_{1j} - t_{2j}) - (\overline{t_2} - \overline{t_1}))^2 \\ &= (t_{1j} - t_{2j})^2 - 2(t_{1j} - t_{2j})(\overline{t_2} - \overline{t_1}) + (\overline{t_2} - \overline{t_1})^2 \\ &= (t_{1j} - t_{2j})^2 + (\overline{t_2} - \overline{t_1})((\overline{t_2} - \overline{t_1}) - 2(t_{1j} - t_{2j}))\end{aligned}$$

Therefore:

$$\begin{aligned}\sigma_e^2 &= \frac{\sum_{j=1}^{n}[(t_{1j} - t_{2j})^2 + (\overline{t_2} - \overline{t_1})((\overline{t_2} - \overline{t_1}) - 2(t_{1j} - t_{2j}))]}{n-1} \\ &= \frac{\sum_{j=1}^{n}[t_{diff}(j) + (\overline{t_2} - \overline{t_1})((\overline{t_2} - \overline{t_1}) - 2(t_{1j} - t_{2j}))]}{n-1}\end{aligned}$$

The relation between σ_e^2 and $t_{diff}(j)$ depends on the contributions of the session effects mainly through the $\overline{t_2} - \overline{t_1}$ part of the equation. However, for small or no session effects the relation still holds:

$$((ICC(3,11) \propto \frac{1}{\sigma_e^2}) \cup (\sigma_e^2 \propto t_{diff})) \Rightarrow ICC(3,1) \propto \frac{1}{t_{diff}} \tag{4.8}$$

4.5.2.2 *Measuring confounding factors*

For each of the above measurements, a multiple regression model with the task, scanner noise, subject motion (total displacement, stimuli/motion correlation, and interaction between task and stimuli/motion correlation) and coregistration error as regressors was fitted to the data (for the design matrix see Figure 4.20). The rows of the design matrix corresponded to subject/task combinations (i.e. each subject was present multiple times – once per each task).The relative importance bootstrap technique (Ulrike Grömping, 2006) with the Lindeman–Merenda–Gold metric (Lindeman et al., 1980) was used to assess the contribution of independent variables to the total explained variance. In short, this technique estimates relative importance by generating combinations of the given model and weighting contributions to the total R^2 by the order of adding variables. The estimates are boot–strapped 200 times to establish confidence intervals.

SCANNER NOISE

To estimate the noise due to scanner related fluctuations, the temporal Signal to Noise Ratio (tSNR) was measured

$$tSNR = \frac{1}{n}\sum_i \frac{\mu_i}{\sigma_i} \tag{4.9}$$

where n is the number of voxels, μ_i and σ_i are the mean and the standard deviation of the ith voxel across time. The average was taken across all voxels within the brain mask. Before calculating tSNR, the time-series were truncated by discarding the first four volumes, realigned to remove motion confounds and detrended using second order polynomials.

SUBJECT MOTION Two metrics were used to characterize motion: total displacement and stimulus by motion correlation. Total displacement (Wilke, 2012) measures in a single variable the overall motion using realignment parameters from every EPI volume. This measure has the advantage of capturing cortical voxel displacement due to both translation and rotation. Subject motion was characterised here by an average over this parameter from both sessions. Stimulus/motion correlation allowed the influence of motion on regressors of interest (and thus beta values) to be measured. For every design matrix (80 design matrices: 4 tasks $\times$ 10 subjects $\times$ 2 sessions), we measured the correlation between the regressors of interest and motion regressors using multiple regression models. The dependent variable of this model were the stimuli regressors (after HRF convolution) multiplied by the contrast vector, whilst the 6 motion parameters were used as independent variables, This way, for every design matrix, we were able to calculate R^2 —- the percentage of stimuli variance explained by motion. As for total displacement, values from the two sessions were averaged.

COREGISTRATION ERROR Inaccuracies of coregistering EPI volumes between two sessions was characterised by the correlation ratio (Roche et al., 1998) between mean EPI volumes from the two sessions. This metric measures functional dependencies between voxel intensities and has been previously used as a registration cost function. The correlation ratio was calculated on brain–masked volumes.

4.5.2.3 *Measuring relations between reliability metrics*

To investigate the relationships between reliability metrics, robust Spearman correlations with outlier removal (Rousselet and Pernet, 2012) were computed between each pair of measurements before and after fitting the multiple regression models accounting for confounds.

For each subject, the HD estimate of the median of t_{diff} and each time–series were also computed for three different ROIs: the area activated in both sessions (overlap), the area activated in one (either the first or second) of the sessions, and the area not activated in any of the sessions. Correlations were then computed to test whether the voxelwise reliability measures (t_{diff} and timeseries correlations) were significantly different between these regions.

4.5.3 *Non–Dice based reliability*

Low mean correlation values were observed on voxel time–series across the four tasks (range 0.07 to 0.17). Time–series correlations were not homogenous through the whole brain and higher values were observed within ROI (range 0.12 to 0.23)

Table 4.9: Reliability measurements obtained across the full brain and within ROI. Significant differences in within and between subjects Dice overlaps are marked in bold. For the motor task the average across three contrasts is shown in the brackets).

		Mean timeseries correlation	mean t_{diff}
Finger	Full brain	0.101 ± 0.040	1.98 ± 0.505
Foot			2.10 ± 0.524
Lips			2.44 ± 1.15
All			2.17 ± 0.80
Finger	Motor cortex	0.230 ± 0.067	2.03 ± 0.78
Foot			2.50 ± 0.53
Lips			1.85 ± 0.70
All			2.13 ± 0.73
Verb generation	Full brain	0.070 ± 0.084	3.58 ± 1.15
	BA44/45	0.120 ± 0.086	4.39 ± 2.28
Word repetition	Full brain	0.090 ± 0.031	2.83 ± 0.68
	Auditory cortices	0.255 ± 0.066	3.42 ± 1.24
Landmark	Full brain	0.135 ± 0.054	1.69 ± 0.28
	Right IPL	0.173 ± 0.063	1.97 ± 0.37

compared to the whole brain (Table 4.9). This indicates that for 'activated' regions, time-series were more similar than for non activated regions.

The opposite pattern of results was observed with t_{diff} (the between session variance of T values). We observed lower t_{diff} values for the whole brain (range 1.36 to 6.1) than within ROI, (range 1 to 8.4), but high t_{diff} values indicate lower reliability. However, as we show later, there was no clear relation between absolute t values ('activated' area) and t_{diff}. For an example map showing the reliability measures of one subject see Figure 4.21.

4.5.4 *Contribution of scanner noise, subject motion and coregistration errors to between-session variance.*

4.5.4.1 *Data modelling*

Since multiple regression is based on correlations, it requires a non-zero variance of the explaining factors. As shown by the correlogram between all of the confounding factors (Figure 4.22), the explaining factors have a reasonable spread of values (for example total displacement ranges from 0.2 to 1.4mm). We also looked at the contributions of the number of artefact volumes found by the ArtDetect algorithm used in preprocessing. These volumes are selected based on the signal intensity and motion signals and added as a confounding regressor (one per artefact) to the single subject design matrix. On average there were 1.75 artefacts in motor tasks, 0.27 in word repetition, 1 in verb generation, and 2.95 in line bisection. Despite the fact that the tasks

differed significantly in terms of these numbers ($F(5,53) = 4.121$ $p = 0.003$) adding them to the multiple regression model used to analyze reliability did not yield significant improvements in the model fit (similar adjusted R^2). Similarly, the model used here was the most parsimonious among a set of models where motion regressors were modelled either as a single parameters, split per task or both (see Table 4.10).

4.5.4.2 *Model results*

Fitting task, scanner noise, subject motion and coregistration error to the time–series correlation values led to a R^2 of 40.47% ($F(10,28) = 1.903$, p=0.08767; adjusted R^2 = 19.2%) with a large contribution of the task (17.53%) and subject motion 20%. When tested on t_{diff} (the between-session differences of t values — a component of the ICC measure), the model yielded a higher R^2 of 67% ($F(14,44) = 6.479$, $p = 8.171e-07$; adjusted R^2=57%) with again a large contribution of the task (40.32%) but also of motion (24%) and scanner noise (11.02%). Finally, when fitted to the Dice values, the model produced an R^2 value of 71% (F(14,44)=7.86, p=5.947e-08, adjusted R^2 62%) with again a major contribution of the task (42.68%) and motion (23%).

Overall task-induced variations is a major single contributor to reliability (18%, 31%, and 43% respectively). This could be explained by the high variability of the landmark task compared to others. If we sum up contributions from all motion related regressors (total displacement, stimuli/motion correlation, and interaction between task and stimuli/motion correlation) it also explains a large portion of the variance (20%, 24%, and 23% respectively). Interestingly, this was not the actual amount of motion that mattered the most (i.e. total displacement), but the correlation between the stimulus presentation (paradigms) and motion. Scanner noise and coregistration confounds add little to the equation, accounting only for 6%, 2% and 6% respectively (see Table 4.11 and Figure 4.23).

No matter how we measured reliability, out of the most commonly reported in previous reliability studies of confounds (scanner noise, subject motion, and coregistration error) only subject motion has a high contribution. To further verify this findings, we reran the reliability analysis on data acquired using the same pipeline but without motion correction (no realignment with runs, no motion parameter regressors and artefact detection in the design matrix). Turning off these corrections decreased the Dice overlap by 20% ($t(58) = 3.0795$, $p = 0.003166$), increased t_{diff} by 28% ($t(58) = -4.4787$, $p = 3.578e-05$) but did not influence time-series correlation significantly (8% decrease; $t(38) = 1.6644$, $p = 0.1043$). It is worth noting that for Dice and t_{diff}, turning off motion lead to changes equivalent to the amount of variance that can be explained by motion regressors, that is motion lead to a decreases in T value reliability and thus a decrease in map overlap.

4.5.5 *Relationships between reliability metrics*

No significant correlations were observed between time–series correlations, t value variance and Dice coefficients. Weak negative correlations were observed between time–series correlations and t_{diff}, however these weak effects disappeared once confounds were accounted for (see Figure 4.23). At the same time regressing out the confounds strengthen the relation between t_{diff} and Dice making it statistically signifi-

Table 4.10: Goodness of fit of different combinations of motion regressors. All models included task, scanner noise and corregistration error regressors. The best model (in bold) was chosen based on the adjusted R^2 (total variance explained accounting for the number of regressors) and Akaike information criterion (AIC) on the Dice measure. Despite the fact that the models that include total displacement and task interaction had slightly better AIC and adjusted R^2 results for time–series correlation and between–session variance, they also experienced colinearity problems due to high correlation between task and interaction regressors. Because of this co–linearity estimation of relative importance was nearly impossible and a simpler model was chosen.

	Time-series correlation	t_{diff}	Dice
Total displacement and stimulus by motion as unique regressors (2 regressors)	R^2 = 0.327 Adj. R^2 = 0.175 AIC = -104.388	R^2 = 0.626 Adj. R^2 = 0.557 AIC = 132.423	R^2 = 0.568 Adj. R^2 = 0.489 AIC = -43.701
Total displacement as 1 regressor across all tasks and split by tasks + stimulus by motion as 1 regressor across all tasks (6 regressors)	R^2 = 0.430 Adj. R^2 = 0.227 AIC = -104.894	R^2 = 0.662 Adj. R^2 = 0.554 AIC = 136.396	R^2 = 0.584 Adj. R^2 = 0.452 AIC = -35.912
Total displacement as 1 regressor across all tasks + stimulus by motion as 1 regressor across all tasks and split by tasks (6 regressors)	**R^2 = 0.404** **Adj. R^2 = 0.192** **AIC = -103.1326**	**R^2 = 0.673** **Adj. R^2 = 0.569** **AIC = 134.4486**	**R^2 = 0.7143** **Adj. R^2 = 0.623** **AIC = -58.02556**
Total displacement and stimulus by motion split by tasks (8 regressors)	R^2 = 0.5378577 Adj. R^2 = 0.297 AIC = -107.009	R^2 = 0.7136207 Adj. R^2 = 0.574 AIC = 136.686	R^2 = 0.7249795 Adj.d R^2 = 0.590 AIC = -50.260
Total displacement and stimulus by motion as 1 regressor across all tasks and split by tasks (10 regressors)	R^2 = 0.5378577 Adj. R^2 = 0.297 AIC = -107.009	R^2 = 0.7136207 Adj. R^2 = 0.574 AIC = 136.686	R^2 = 0.7249795 Adj. R^2 = 0.590 AIC = -50.260

Table 4.11: Relative contribution in percentage (with 95% confiance intervals) of task, scanner noise, subject motion and coregistration error to time-series correlation, between sesion variance and Dice overlap.

	Time-series correlation	t_{diff}	Dice
Task	17.54%	31%	42.68%
	[7.31 49.19]	[20.4 48.25]	[22.39 63.76]
Scanner noise	1.57%	11.84%	4.48%
	[0.57 22.66]	[1.67 25.95]	[0.5 16.23]
Subject motion	7.18%	0.48%	4.95%
(total displacement)	[0.63 23.44]	[0.03 4.46]	[1.14 15.56]
(stimuli/motion correlation)	4.5%	17.96%	3.84%
	[1.14 26.35]	[4.35 40.08]	[1.23 14.21]
(taskstimuli/motion correlation)	8.4%	5.35%	14.52%
	[1.84 36.4]	[1.60 17.17]	[4.2 25.81]
Coregistration error	1.28%	0.66%	0.96%
	[0.18 10.95]	[0.16 3.42]	[0.24 6.67]

cant (= 3.11 vs. 0.05 = 2.43). The direction of the relation (r=-0.44) makes conceptual sense (smaller differences in t values lead to higher overlaps). To investigate further a possible (non–monotonic) relationship between these variables, all voxels from each task/contrast were pooled together to create a series of scatter plots between t values and time–series correlations and t_{diff}.

Most voxels with high t values show increased time–series correlation for all tasks. The same is true for negative t values, which indicates that even though the negative t values are not usually of interest, they are stable between–sessions even on the time–series level (Figure 4.24a). It is also worth noting that there were many voxels with high correlation but low t value. These indicate a reliable signal not captured by the design matrix. When restricting the analyses to overlapping vs. non–overlapping activated areas, the highest time–series correlation values were observed in the overlapping (those that by definition will have high t values) rather than non–overlapping areas (Figure 4.25), confirming that high t values relates to reliable voxels (time–series).

No relationship was observed between mean t values and the variance (t_{diff}). Highest t_{diff} values (poorest reliability) were observed for t values close to zero, but one has to bear in mind that those values were also the most common (see Figure 4.24b). There were, however, differences in the observed patters between tasks. The distribution of t_{diff} across mean t values was almost uniform for verb generation. This was in contrast to the lips task for which the highest t_{diff} values were observed almost exclusively for the voxels with t values close to zero. The landmark task, on the other hand, showed a smaller spread in both t values and their between-session variance. When restricting the analyses to overlapping vs. non-overlapping activated areas, we noticed that mean t_{diff} in the overlapping area was no different than in the parts of the brain that were not active in either of the two session, but there was a significant

increase of t_{diff} for non overlapping active areas (see Figure 4.25). This relation can even be observed on the individual subjects maps (see Figure 4.21). In other words, t_{diff} is bigger in regions that were active in one of the sessions, but not in both of them. These are usually the borders of suprathreshold clusters.

We have shown that reliability is mostly influenced by the task and motion (especially correlated with the stimuli). Scanner noise and coregistration errors on the other hand, have little influence on reliability. We have also shown that relationships between reliability measures are not straightforward. This means that one cannot make decision about reliability of suprathreshold maps based on ICC alone.

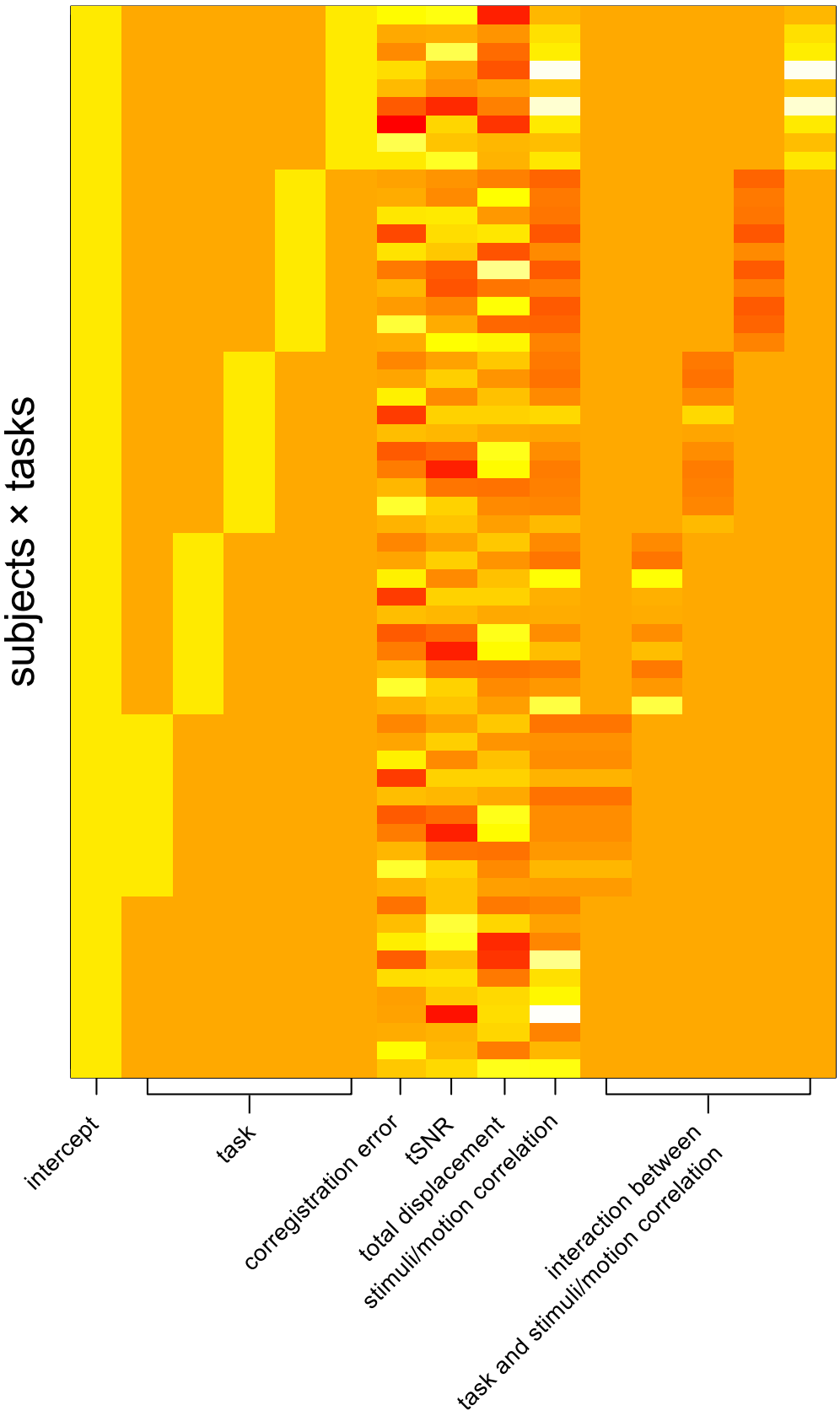

Figure 4.20: Design matrix of the multiple regression model used to establish contributions of confounding factors to reliability measures. Each row in this matrix corresponds to one subject and one task/contrast (therefore each subject entered the analysis multiple times - once per each task/contrast). Please note that since the relative importance technique has been used on top of this model the order of the regressors (columns) does not matter.

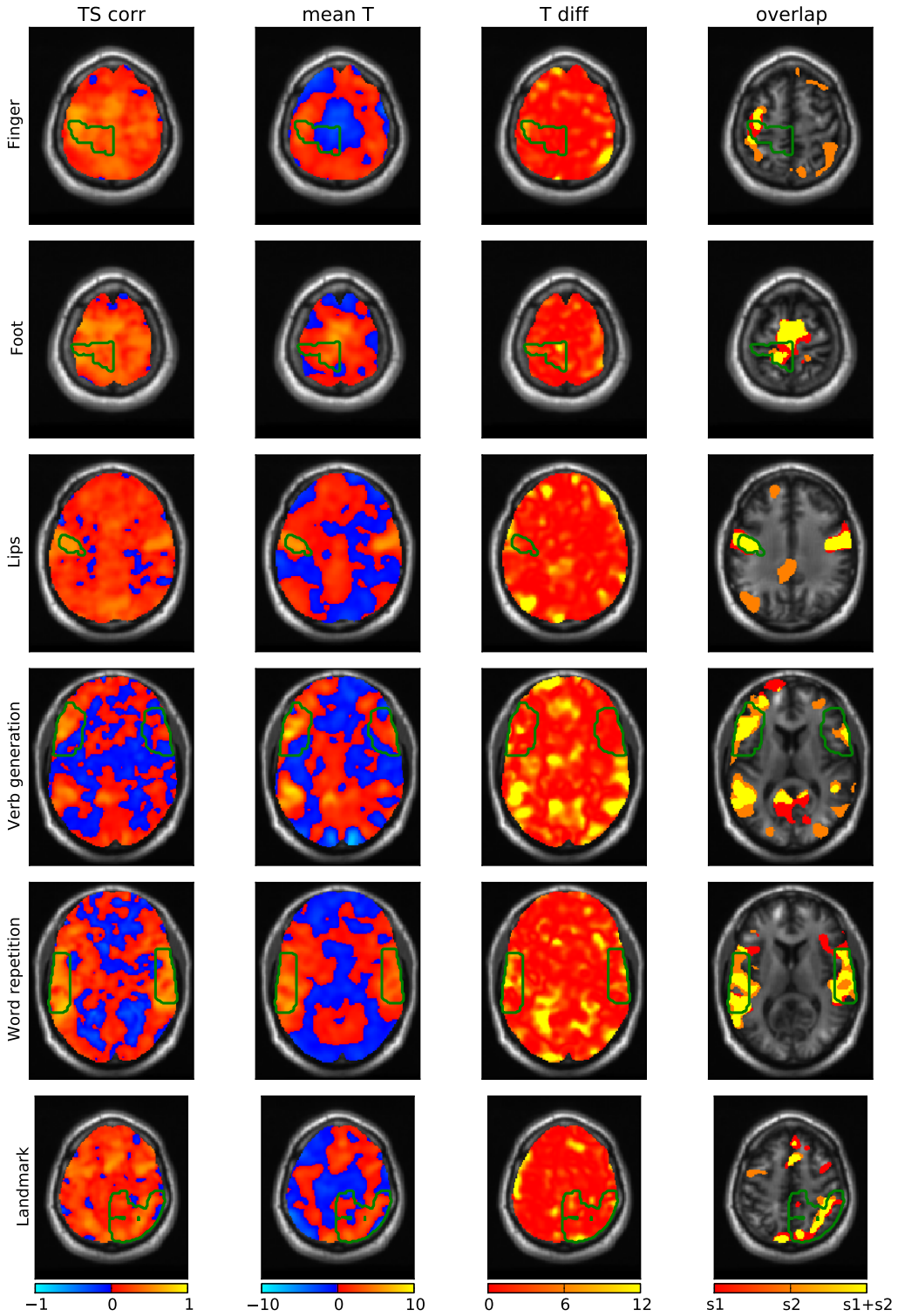

Figure 4.21: Brain statistical maps from a representative subject (subject 2). There is a spatial correspondence between time-series correlation (TS corr) and mean t maps, but there is no correspondence between t_{diff} and mean t. Additionally most heat points of the Tdiff maps overlap with non-overlapping active areas (orange and red colours in the overlap column).The anatomical ROIs are marked in green.

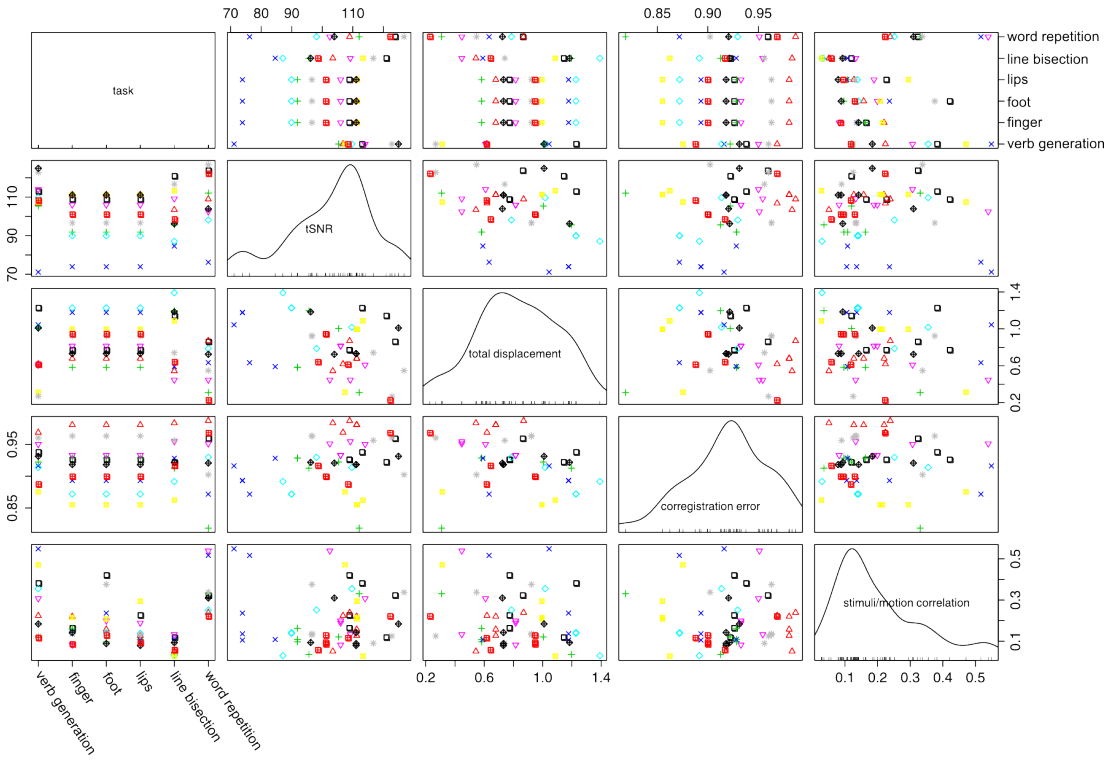

Figure 4.22: Distributions of modelled explanatory factors to reliability. Combinations of symbols and colours of points represent different subjects.

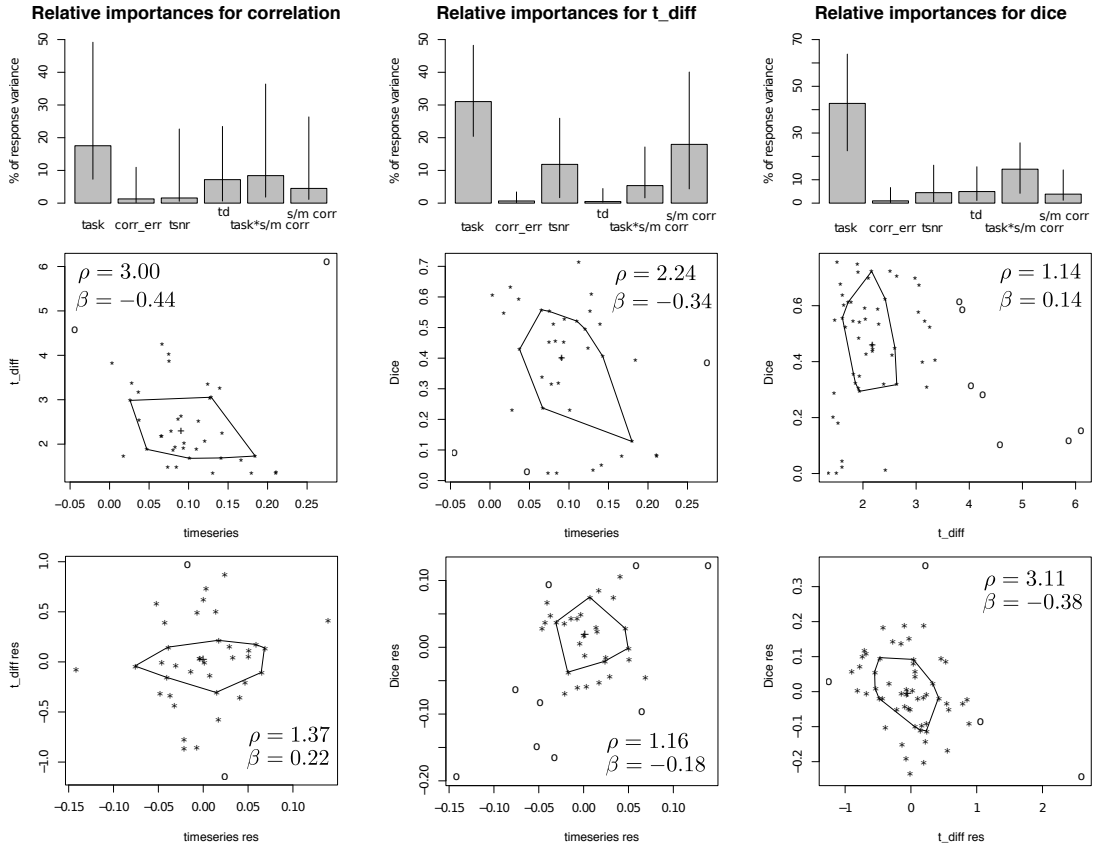

Figure 4.23: Relationships between reliability measures and confounds. Top row: percentage of timeseries correlation, tdiff, and Dice variance explained by task, scanner noise (tsnr), motion (total displacement (td), stimulus by motion correlation across tasks (s/m corr) and split by task (tasks/m corr)), and corregistration error (corr_err). Middle and bottom row: Scatter plots between reliability metrics (middle = raw metrics, bottom = metrics after removing confounds). The skip correlation was computed on data points after outlier removal (outliers are marked as circle —- polygons highlight the centre (75th decile) of the data cloud).

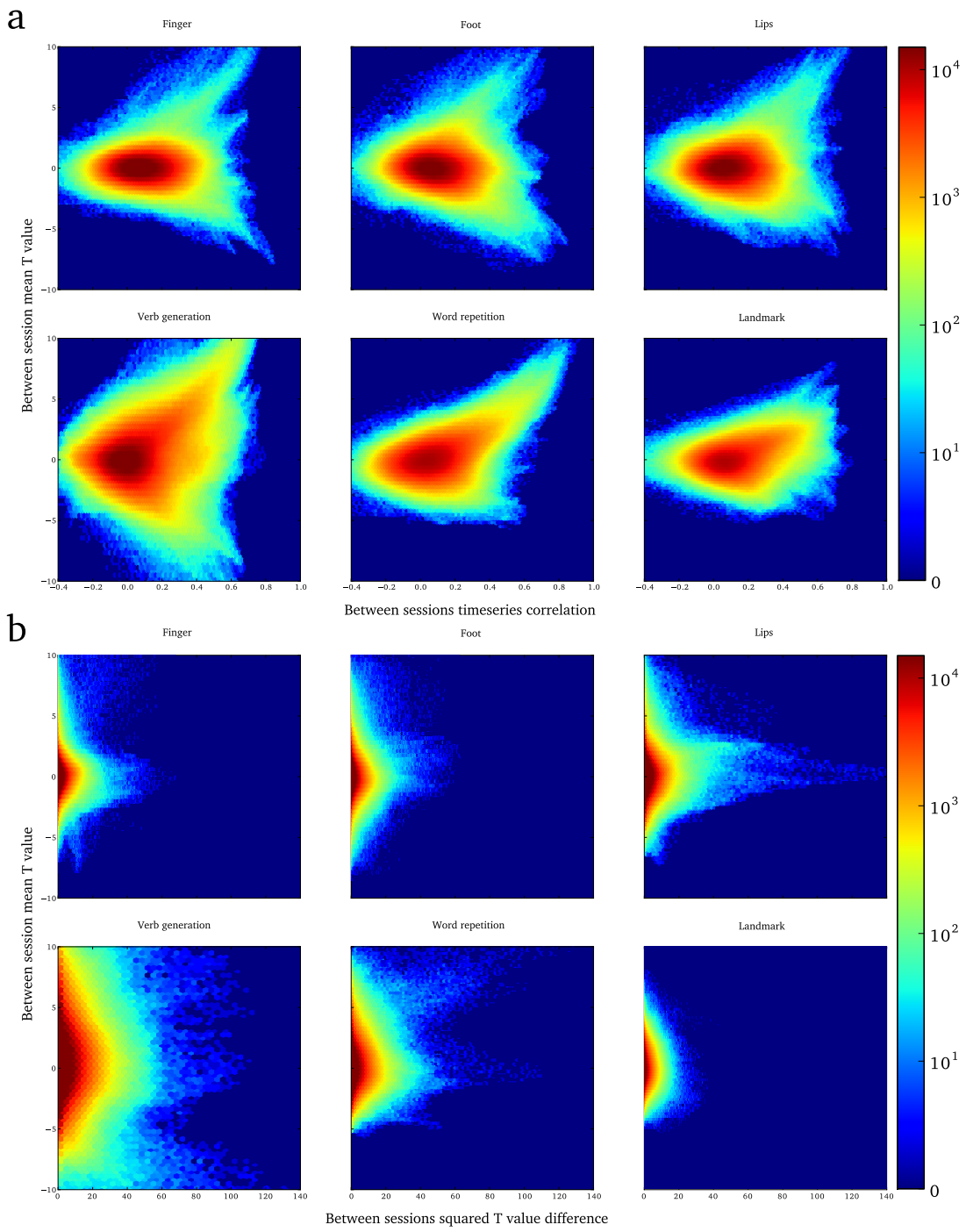

Figure 4.24: Joint distributions of mean t values and timeseries correlations (a) and mean t values and t_{diff} (b). Voxels from all subjects were pooled together.

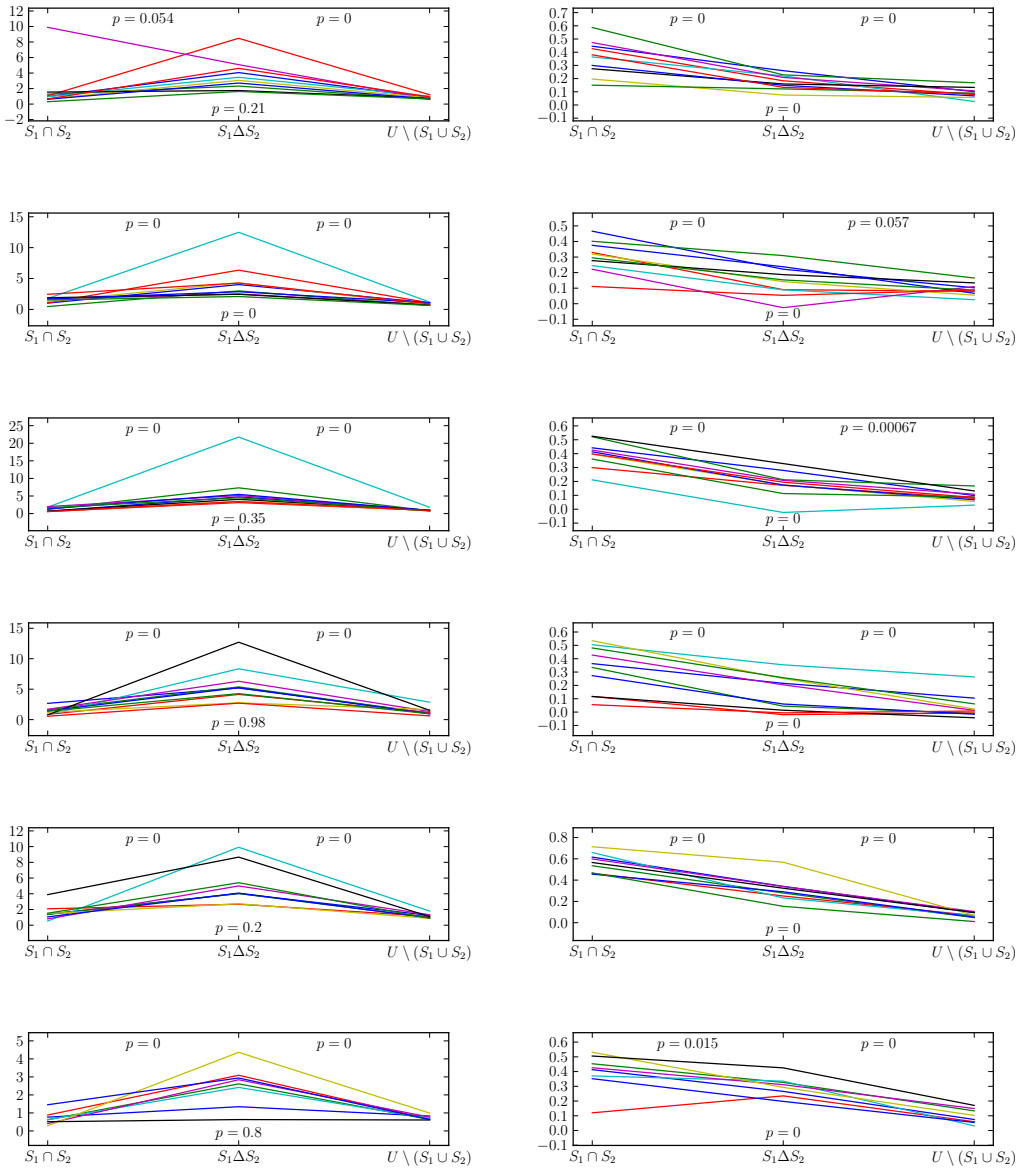

Figure 4.25: ROI analysis of voxel–wise reliability metrics. H-D median of t_{diff} and time-series correlation for three different ROIs: area activated in both sessions, area activated in either the first or second session, area not activated in any session. P values were estimated using pair–wise (within–subject) one sample bootstrap test. Each colour represents a different subject.

4.6 DISCUSSION

4.6.1 *Reliability of thresholding methods*

Using AT, we found activations in the expected area of the brain in more cases than using FT, and paired comparisons of activations maps show bigger between sessions overlap with AT than FT for most tasks/contrasts. Analysis of reduced datasets also shows that significant more reliable voxels in expected areas can be observed with AT when less data is being used. We can hypothesise that each task/contrast has an upper bound reproducibility (influenced by SNR, variability of underlying neuronal response etc.) that can be reached when using infinite scanning time. With such hypothetical dataset both thresholding methods should perform equally well. However, in practice, scanning sequences have finite length and subjects do not always manage to follow the paradigm throughout the whole experiment. AT managed to recover reproducible maps with fewer datapoints than FT. In other words AT was able to reach the aforementioned upper bound of reproducibility quicker than FT.

The main difference between AT and FT is the way the cluster-forming threshold is chosen. In FT it depends on the smoothness of the data and number of voxels in question (when defined as FWE corrected p-value). Despite that dependency, the actual threshold expressed in T values does not vary much between the sessions. In contrast AT tries to estimate the distribution of noise (using a Gaussian distribution) and signal (using a shifted Gamma distribution). This accounts for the SNR of the dataset and global signal changes that can cause distribution shift (non-zero mean of the noise Gaussian). These global signal changes are major sources of between-session variation (Raemaekers et al., 2012), and accounting for them in the thresholding process can be the cause of improved single subject test-retest reliability.

4.6.2 *Reliability of tasks*

Overt word repetition is a reliable localizer of the Wernicke's area Group analysis of the overt word repetition task confirmed the validity of our paradigm. As expected activation was observed in regions commonly associated with speech comprehension and production: auditory cortex, Wernicke's and Broca's areas. Additionally activation in the motor cortex and SMA was observed, most likely related to the motor component of the speaking activity. In the single subject analysis within subject (between sessions) reliability was statistically significantly higher than between session. This is a strong evidence for suitability of this task for presurgical cortical mapping.

Covert variant of the verb generation task is more reliable than its overt counterpart One of the goals of this study was to compare overt and covert verb generation tasks in terms of their single subject reliability. The paradigms were validated by group analysis which produced activations in Broca's area, auditory cortex, and SMA. Auditory cortex activation was stronger in the overt version of the task where activation in left visual cortex was stronger in the covert variant. However, only the covert verb generation showed a higher within subject than between subject reliability over Broca's area. In addition, comparing between session reliability between tasks in a paired manner (within subjects) we observed that 6 out of 7 subjects had a higher reliability with the covert task.

Having twice as many data points (due to continuous vs. sparse acquisition) outweighs the advantage of having higher SNR. One can also argue that motion artefacts could have caused this result, but we have tried to decrease their influence by preprocessing steps (motion correction, artefact detection, inclusion of motion and outlier regressors in the design matrix). Informal analysis of the estimated motion parameters did not show major differences between the two tasks. Our findings indicate that covert verb generation should be used in favour of the overt verb generation. However, the question if the underlying cognitive and neuronal mechanisms in these two tasks are comparable remains open. Finger, foot and lips areas can be reliably mapped in the motor cortex Not surprisingly motor tasks performed well showing expected activations in the motor cortex and reliable single subject results.

4.6.3 *Landmark task is not reliable enough for presurgical applications*

The group analysis revealed activations in locations previously reported in the literature. The only region active in the "all stimuli", "only stimuli with responses", "only stimuli with correct responses" was the right Lateral Occipital Cortex, Contrasting answered trials of Landmark vs. detection revealed activations of the SPL and IPL. Both results are in agreement with previous findings (Fink et al., 2000, 2001). However, within subject reliability was very poor (not significantly different from between subject) and in many subject/contrast, no activation was found at all. Additionally regions reported by group level analysis corresponded with low ICC values. The explanation for this discrepancy might lie in the fact that Holmes-Friston group analysis procedure discards error term (residuals) from the first level. ICC and single subject analysis will down weight noisy voxels. This task seems thus unsuitable for single-subject mapping.

4.6.4 *Reliability metrics and confounding factors*

Studies involving fMRI are complex and easily influenced by many factors. This is not only because the subject in question, the human brain, has intricate and not fully understood hemodynamics. The data acquisition and processing is a multilevel complicated process (Savoy, 2005). In this study, we investigated how different factors can contribute to between-session variance. We found that about 30-40 % of the observed single subject reliability (unthresholded or thresholded T-maps) can be explained by the task used and that among confounding factors, motion is the main problem accounting for about 20% of the variance.

4.6.5 *Choosing the right metric*

One important aspect of this study is the application of different methods of measuring reliability. Specifically, we assessed three different ways of measuring reliability, from the correlation of time-series, to t values and thresholded t maps. In addition, compared to many previous studies (Raemaekers et al., 2007; Caceres et al., 2009), we have not restricted our measurements to a predefined ROI or split analyses between different ROIs. This decision was motivated by the fact that reliability and activations

are not strictly related (see e.g. Caceres et al., 2009) and in some cases like Dice, ROI analysis introduces a selection bias. It is therefore misleading to assess a task only by the reliability within a predefined ROI. Our decision to use t_{diff} as the measure of between session variability of unthresholded maps was mostly driven by the ability to relate it to Dice overlap measure. First, we decided to use t-values instead of beta values because t-values are influenced by residual noise and thus reflect better acquisition (scanner) related variance. Second, t_{diff} captures the variance of t-values that translate directly into the extents of suprathreshold regions. Finally, t_{diff} can be related to ICC. As mentioned in the Introduction, this choice is of course only relevant if one is looking at single subject reliability, and $beta_{diff}$ could be more appropriate for group reliability.

Despite the fact that Dice overlap was previously being criticized as a reliability measure (Smith et al., 2005), we have still included it in our analysis. Thresholding as any form of dimensionality reduction can introduce biases and we agree that calculating overlaps of thresholded maps is a rough estimate of reliability. However, let us not forget that the thresholded maps are what the end result of an fMRI analysis is. Papers describing group studies are presenting and making claims about thresholded maps. The same applies to the single subject domain. Neurosurgeons plan and execute procedures based on thresholded maps. Functional localizers produce ROIs which are nothing less than thresholded statistical maps. We acknowledge problems with analyzing thresholded maps (that is why we have included two other reliability metrics) and at the same time we try to minimize their influence. Importantly, we have used the same method as Smith et al., 2005) for correcting for global effect (a t value distribution shift derived from a Gamma-Gaussian model).

Indeed, global effects in the context of single subject test-retest reliability has also been a topic of a recent work by Raemaekers et al., 2012). In their approach they fitted a line to session 1 vs. session 2 scatter plots. This allowed them to estimate between session variance as the variance orthogonal to this line. This is a variant of global effect correction used in our work. Their approach allowed the amount of shift applied to the t values to be in linear relation to them. In other words, in our model this line can be shifted from the centre of the data cloud, but keeps the 45 degrees angle. However, the approach we used (Smith et al., 2005) is more flexible as it allows applying the correction to one session without knowing anything about the other, i.e. the model is fitted using single session distribution, not the joint scatter plot. Additionally, when applied to the full brain, a linear fit to the joint distribution of values from two sessions would be driven by values close to zero and thus not capturing the shape of the tails which are the activated voxels (see Fernández et al., 2003).

Finally, we found that good time–series reliability is a necessary but not sufficient condition for good t map reliability. For example, one could observe a good correlation between time-series of two sessions, but a large difference in t values, a case that may correspond to a poor model fit, i.e. some regions may activate similarly in both sessions, in relation to the task, but not with the stimulus or block onsets described in the design matrix —- such regions can be captured by, e.g. independent components analyses. More intriguingly, we have observed a similar effect in the relationship between t values and thresholded maps. Small between-session differences in t values are necessary for a good suprathreshold overlap, but not sufficient, because a high

threshold can lead to low Dice overlap. For instance, the task that performed the worst (landmark) in terms of Dice, was the one showing the lowest t_{diff} values. This brings us to a paramount question, namely what makes a good task/analysis? The task should be reliable, but this is not the full answer, because it can be reliable in not measuring any meaningful activation. In other words we don't only want low between-session t value variance, but also high t values consistently across sessions. Dice overlap captures this property due to thresholding, since only high t values that survive thresholding contribute to the overlap. We showed here that using Dice, one can compare within- vs. between-subject overlap, and a reliable task can be defined as having a significantly higher degree of overlap within- than between-subjects.

4.6.6 *Explanatory factors*

The type of task was the main explanatory factor on our reliability metrics, which can explain the large variance observed across different studies (Bennett and Miller, 2010). Here, one can argue that the large effect observed depends essentially on the landmark detection task which failed to produce any suprathreshold clusters more often than the other tasks. This indeed can explain the effect over Dice overlap measurements, but not on t_{diff}. The observed between-session t value differences were actually lower for the landmark task than for the other tasks. It is therefore a case where one can observe differences between small t values not yielding any statistically significant activation. As already mentioned in the Introduction, this result also highlights the need to differentiate reliability of the fMRI signal (single subjects) from reliability of contrast maps (group studies) since a small BOLD signal but with a low between-subject variance gives significant group results.

The fact that the type of task can have such a big influence on reliability should perhaps not be surprising. First of all, the tasks in our study were not only different in terms of the behavioural paradigms (or in other words what the subject was meant to do during the scan), but also in terms of acquisition parameters. Word repetition used sparse sampling which in theory should improve SNR (Hall et al., 1999), although at the cost of the number of volumes acquired. Scanning time and therefore the number of volumes acquired ranged from seven up to almost ten minutes. All the tasks were executed in blocks, but the landmark task used event related regressors to restrict the response to correct answers only. All these factors can influence reliability on a purely data acquisition level. Further studies with systematic variation of these parameters, for example sparse/non-sparse, block/event related and number of volumes acquired, would be necessary to establish their exact contribution to reliability.

Apart from the data acquisition aspect of different tasks there is one more important reason explaining the observed influence of the task type on reliability. Different tasks involve different neuronal populations and can incorporate different cognitive strategies. For a given task the same observed behavioural response, such as generating a verb, can be achieved by different neuronal subsystems, hence eliciting different BOLD reaction. We hypothesise that this "cognitive freedom" is different for different tasks. For example, a simple finger tapping task is most likely to be executed in a similar fashion each time. In contrast, a more sophisticated task involving language generation or spatial attention could involve different neuronal subsystems each time.

This might be part of the explanation why in our study the landmark task did not perform well in terms of single subject reliability.

Scanner noise, and coregistration errors have previously been suggested to contribute to reliability (Bennett and Miller, 2010; Caceres et al., 2009; Fernández et al., 2003). Even though we have found such relationships, their magnitude was surprisingly small. Both of the confounding factors we investigated have been accounted and corrected for in the data processing pipeline. Scanner noise, for example, can be influenced by signal dropouts due to failing coils. Smoothing can mitigate this to a certain extent by improving tSNR, although this is achieved by the loss of spatial accuracy. Volumes with sudden signal dropout are also either removed or accounted for in the design matrix during the artefact detection step. As for the coregistration step, it is perhaps not surprising that modern algorithms managed to realign brain volumes of the same person scanned using the same sequence on the same scanner. Our results therefore suggest that thanks to advances in data processing methods, issues such as scanner noise and coregistration errors are not the most important contributing factors to between-session variance. This is, however, true only within normal working conditions. For example a serious scanner malfunction would inevitably result in poor reliability.

Subject motion on the other hand had non-negligible influence on reliability. It was the largest confounding factor (the 2nd largest explanatory variable) and for time-series correlation, even explained more than the task. Comparison of realigned vs. non realigned data confirmed those results by showing equivalent changes in t_{diff} and Dice. Only correlations on time-series were not significantly affected by turning off motion correction (-8%) despite a large portion of the variance explained by motion regressors on realigned data. In the present context this is difficult to explain. One possibility is that using Pearson correlation is not efficient enough to fully capture changes in reliability given the various limitations related to data range restriction, curvature, or heteroscedasticity (Wilcox, 2005). Importantly for planning fMRI experiment, we have found that motion correlated with the stimuli explains the lack of reliability much better than absolute motion. On time-series correlation (i.e. before model fitting) both total displacement and motion correlated with the stimuli mattered whilst for t_{diff} and Dice, only motion correlated with the stimuli mattered. This can be explained by the fact that t-values depends strongly on the signal correlated with the task (beta value) while the whole time-series correlation is also affected by the overall motion. This finding has implication towards behavioural task design and poses a question of theoretical upper limit on single subject reliability of motion related tasks. It is, however, important to also acknowledge the limitations of our modelling approach. We did not control explicitly the levels of confounding factors. In this study we have looked at the relation between reliability and measured (but not induced experimentally) confounds. What we are reporting is how much these factors explain the variability in reliability measures. This approach has some obvious limitations – for example if all of the subjects were expressing substantial motion but of an identical level there would be no variance within the confounding factor and it would yield no explanatory value. However, as we shown in Figure 4.22 we have a reasonable spread of combinations of values of confounds across subjects and tasks which allows concluding reasonably on the contribution of each factor.

4.7 SUMMARY

In this chapter we have shown that AT can detect and create more reproducible areas of activation then FT. This effect was even more pronounced in case of fewer data points (shorter scanning times).

Using AT, we determined a new approach to select tasks suitable for single subject analyses. Among the tasks chosen, motor tasks showed very satisfactory reliability. Most of language tasks (overt word repetition, covert verb generation) also produced reproducible activations. However, overt variant of the verb generation task did not perform as well as the covert counterpart. The advantage of more natural and controlled task and higher SNR due to longer TR was outweighed by higher number of data points and fewer motion related artefacts. Finally, the landmark task does not provide single subject activations that are reproducible enough to be used for cortical mapping.

To conclude, we have shown that task and motion (especially correlated with the stimuli) can have significant detrimental effect on reliability. Coregistration errors and scanner noise, however, contribute much less to observed reliability variance

5 CLINICAL PILOT STUDY

5.1 INTRODUCTION

This work has been done in collaboration with Prof. Iain Whittle

In the previous chapters we have established a methodology for performing presurgical planning using fMRI. This included a new adaptive method for automatic thresholding of statistical maps and selection of behavioural tasks suitable for providing reliable results. In the following chapter we present results from the application of these tools to a group of patients with brain tumours. Our goal was to test the methods and paradigms and check correspondence between ECS and fMRI mapping, in relation with pre-surgical behavioural deficits.

5.2 METHODS

5.2.1 *Patient population*

Eighteen patients diagnosed with a brain tumour were included in the study (8 females, 10 males). The median age was 42.5 years (min = 25, max = 75 years). Patients were recruited based on the tumour location (nearby motor cortex, Wernicke and Broca's areas) and their suitability for surgery. The most common tumour types were GBM and meningioma (see Figure 5.1). The study was approved by the local Research Ethics Committee.

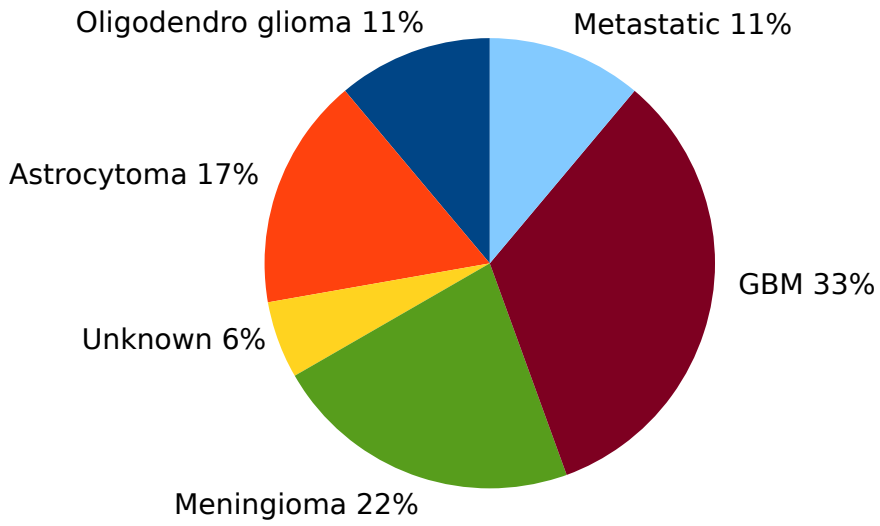

Figure 5.1: Distribution of tumour types found in patients included in the study.

5.2.2 *Scanning procedure*

Based on the findings from the previous chapter we selected three fMRI paradigms: motor (with the finger vs. others, foot vs. others, and lips vs. others contrasts), covert verb generation (for mapping Broca's area) and overt word repetition (for mapping the auditory cortex and Wernicke's area). For details of scanning procedures and data processing see Chapter 3 and Chapter 4. In all cases AT was used to create the suprathreshold maps.

5.2.3 *Behavioural tasks*

To assess behavioural deficits prior to the surgery a series of neuropsychological tests have been administered. Hand dexterity has been assessed using The Nine Hole Peg Test (9HPT —- see Grice et al., 2003; Mathiowetz et al., 1985. In this test the patient was asked to put nine wooden pegs into a square board with nine holes (all pegs and holes are the same, and order does not matter). After performing a test trial to get accustomed with the task, patients were timed on a second trial. Timing started with the patient grasping the first peg and finished when they inserted the last peg in the last remaining hole. Cut off time was 120 seconds. Patients were tested for both hands separately. According to the norms in Mathiowetz et al. (1985), the test was considered to be unsuccessful if the patient finished the task in a time longer then the mean + two standard deviations for the corresponding sex and group age.

To asses gait and mobility performance a 10 meter walk test was performed. In this simple test patients were asked to walk at a comfortable pace in a 10 meter stretch of corridor. The time to cover this distance was measured and the average speed was calculated. This speed was compared to the sex/age matched healthy controls norm (Bohannon, 1997). Again, the test was considered to have failed if the speed was lower than norm mean minus two standard deviations. Based on these two tests patients were divided into two subgroups: those with a motor deficit and those without. Failing at least one test was sufficient to be assigned to motor deficit group.

Additionally, six patients with speech and language related syndromes underwent two more tests, namely the Rey Auditory Verbal Learning test (RAVLT see Rey, 1941; Schmidt, 1996 and the Controlled Oral Word Association Test (COWAT —- see Loonstra et al., 2001). RAVLT involves presentation and memorizing several lists of words followed by a delay, and a recognition task with distractors. In COWAT, patients were asked to name as many words beginning on letters "F", "A", and "S" within 60 seconds for each letter. As in the previous examples, norms were used to classify the tests results (Geffen et al., 1990; Loonstra et al., 2001). A score lower than the sex/age matched average reduced by two standard deviations was considered a failure. Patients that failed at least one language test were labelled as having a language deficit.

5.2.4 *Surgery*

All patients were operated on by Prof. Ian Whittle. Prior to the procedure fMRI data were used for planning and assessing the risks. Based on the fMRI data six patients did not undergo ECS. Out of remaining twelve, nine had the stimulation points recorded; missing data for three patients was due to technical problems or lack of consent from the patient. After craniotomy, but before the resection, the site of ECS was recorded by taking a photograph of the brain with paper labels and reference ruler. Stimulation was then performed using a bipolar electrode with amplitude ranging between 2.5 and 4 mA, and frequency of 60Hz.

5.2.5 *Cortical mapping assessement*

For each case fMRI scans were visually inspected to asses the distance between the active area and the tumour. Each fMRI map (finger, foot, lips, Wernicke, Broca) was labelled as being close to the tumour if this distance was less than 10mm. Each case was also assessed based on the expected location of the activation. If the activation was not found in its typical location (M1 for finger/foot/lips, BA 44/45 for Broca, BA 22 for Wernicke) or the pattern was unusual (such as asymmetrical lips activation) the map was labelled as "unexpected". These changes were taken as signs of either changes of location due to mass effects or plasticity.

For cases with recorded ECS location, each stimulation point was positioned manually on the structural MRI map using the intraop photographs and post surgery notes. This allowed the assessment of the correspondence between fMRI and ECS by labelling each stimulation point as a true positive (fMRI activation and positive ECS response), true negative (no fMRI activation and lack of ECS response), false positive (fMRI activation, but no ECS response), or false negative (no fMRI activation, but an ECS response) —- Figure 5.2. Each fMRI map was assessed separately. Therefore, for example, a stimulation point eliciting finger response at a location predicted by the finger activation map is a true positive for that map, but a true negative for the foot map.

ECS was also used to classify the stimulation points in terms of proximity to the tumour. If a positive ECS stimulation was found within 10mm from the tumour based on post-operative notes and photographs, the case was labelled as being close to the tumour. Using the behaviouraly derived labels (deficit/no deficit; motor/language) and distance between the eloquent cortex and the tumour (derived from fMRI and ECS), eight contingency tables were generated, namely two (fMRI/ECS) for each task (finger/foot/lips/speech). McNemar tests were calculated for the contingency tables to establish the statistical significance of the relations.

5.3 RESULTS

5.3.1 *Behavioural test*

Three patients did not manage to complete the 9HPT. Among the remaining subjects, theaverage time to complete was 15.55 sec (std: 3.54) for the right hand and 16.46

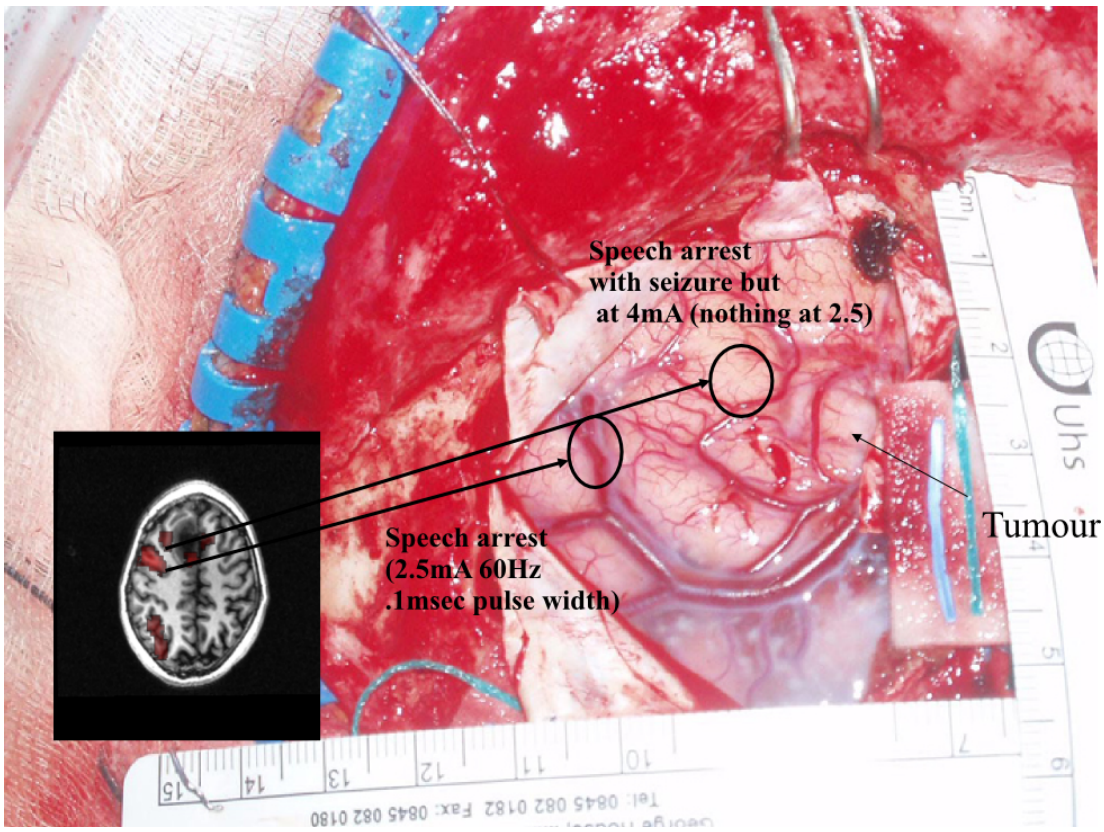

Figure 5.2: One of the surgical cases with ECS. Arrows shows correspondence between the stimulation points and fMRI.

sec (std: 7.26) for the left hand. The average pace in the 10 meter walk test was 1.25 m/sec (std: 0.36).

For the six patients that performed the language test, the average RAVLT score was 52.66 (std: 21.19), and average COWAT score was 27.83 (std: 10.24). The high spread of the measurements is due to differences in age and case severity.

5.3.2 *fMRI mapping success rate and cortical plasticity*

On average fMRI maps were successfully created in 92.96% of the cases (for task break down see Table 5.1). Unexpected activation pattern or location was found in 39.4% of the cases (for task break down see Table 5.1). An example of an abnormal activation pattern is shown in Figure 5.3.

Table 5.1: Rates of successfull mappings and mappings that provided unexpected location for different tasks.

	finger	foot	lips	Wernicke	Broca
success rate	100.00%	94.44%	94.44%	75.00%	88.89%
unexpected location	44.44%	27.78%	55.56%	25.00%	33.33%

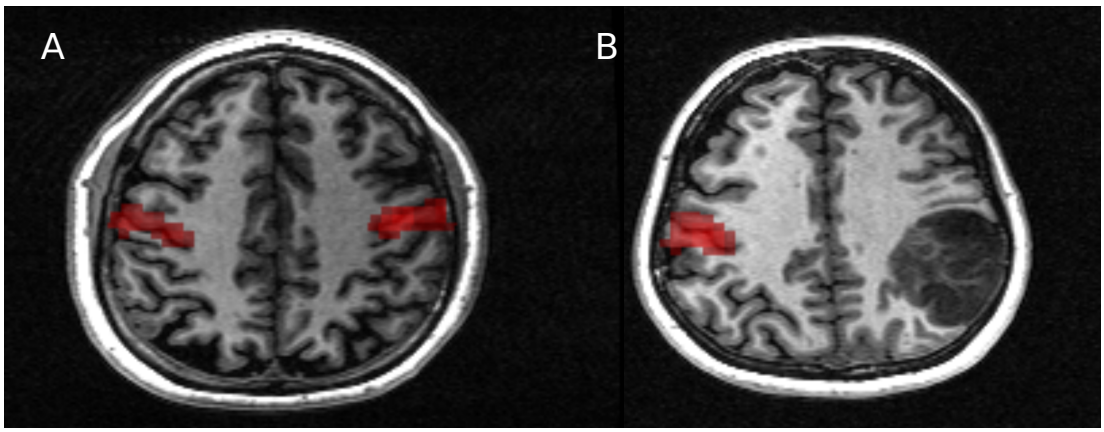

Figure 5.3: fMRI maps of lips representiation in the motor cortex. Normally lips movement is represented bilaterally on the same level (A). However, in presence of a tumour this pattern can change, by shifting the ipsitumoural representation of lips (B).

5.3.3 *Correspondence between ECS and fMRI*

We have found 100% correspondence between fMRI and ECS for both motor (see Table 5.2, Table 5.3, and Table 5.4) and speech (see Table 5.5). However, there were many more negative ECS stimulation points than positive ones (49 vs. 9). Nonetheless, no false positives or false negatives were observed.

Table 5.2: Correspondence between ECS stimulation and fMRI map for finger movement

		ECS	
		pos	neg
fMRI	pos	4	0
	neg	0	54

Table 5.3: Correspondence between ECS stimulation and fMRI map for foot movement

		ECS	
		pos	neg
fMRI	pos	1	0
	neg	0	57

5.3.4 *Distance between the eloquent cortex and the tumour vs. pre operative behavioural deficits*

No significant relationship between the preoperative deficits measured using the behavioural tests and proximity of the eloquent cortex to the tumour have been found. This was true for distances measured using both ECS and fMRI. All subjects were used to calculate the fMRI distance vs. motor deficits (see Table 5.6). The nine sub-

Table 5.4: Correspondence between ECS stimulation and fMRI map for lips movement

		ECS	
		pos	neg
fMRI	pos	2	0
	neg	0	56

Table 5.5: Correspondence between ECS stimulation and fMRI map for speech

		ECS	
		pos	neg
fMRI	pos	2	0
	neg	0	18

jects that had ECS location recorded were used to calculate ECS distance vs. motor deficits (see Table 5.7). The six subjects that performed the RAVLT and COWAT tests were used to calculate fMRI distance vs. language deficits (see Table 5.8). Finally, the four subjects that had both ECS locations recorded and performed the RAVLT and COWAT tests were used to calculate the ECS distance vs. language deficits (see Table 5.9).

Table 5.6: Correspondence between fMRI measured distance between the tumour and the eloquent cortex and motor deficits. chi-squared = 0.9, df = 1, p-value = 0.3428

		Motor deficit	
		yes	no
fMRI	close	4	7
	far	3	4

Table 5.7: Correspondence between ECS measured distance between the tumour and the eloquent cortex and motor deficits. chi-squared = 0, df = 1, p-value = 1

		Motor deficit	
		yes	no
ECS	close	1	3
	far	2	3

Table 5.8: Correspondence between fMRI measured distance between the tumour and the eloquent cortex and language deficits. chi-squared = 0.25, df = 1, p-value = 0.6171

		Language deficit	
		yes	no
fMRI	close	0	3
	far	1	2

Table 5.9: Correspondence between ECS measured distance between the tumour and the eloquent cortex and language deficits. chi-squared = 0, df = 1, p-value = 1

		Language deficit	
		yes	no
ECS	close	0	1
	far	1	2

5.4 DISCUSSION

5.4.1 *fMRI mapping for surgery*

fMRI mapping was performed for 18 tumour patients with very high success rate. 92% of the attempted maps were created successfully. fMRI mapping never failed completely; for any given subject at most one (and in most cases none) of the tasks failed. These rare failures were mainly caused by patients failing to understand the task instructions, the involved skill being severely impaired by progressing disease (i.e. troubles moving a finger during motor mapping), and seizure related artefacts. We also observed lower success rate for language related tasks than motor tasks. This is most likely due to the fact that those tasks are more complicated both in terms of instructions as well as the neuronal circuitry involved. It also cannot be ruled out that physiological changes related to the growing neoplasm have influenced fMRI signal and disrupted the mapping procedure.

Out all successfully created maps, 39% showed patterns uncommon for healthy controls. This is to be expected considering the effect of a slowly growing mass such as a tumour. Some tissue has been compressed and rewiring have most likely occurred to compensate and maintain functionality. There were differences between tasks in terms of frequency of these abnormal patterns (ranging from 25% to 55%). One has to remember, however, these abnormalities has been assessed by visual inspection and might be biased because of that. For example, consider the task for mapping the region responsible for moving lips (55% cases had unexpected pattern). Departure from its healthy symmetrical and bilateral pattern is easier to spot than for other changes (see Figure 5.3). On the other hand there is much more variance in language related areas (33% of unexpected patterns for Wernicke's and 33% for Brocka's areas) in healthy controls so the norm, and departure from it, is harder to define.

5.4.2 *fMRI vs ECS*

These data show that fMRI has a 100% correspondence with the ECS measurements. This suggests that fMRI may be a viable replacement for ECS, however, one has to keep in mind the limitations of this pilot study. Only 9 patients were included in this comparison since ECS was not performed or the stimulation locations were not recorded for the other subjects. Additionally there were many more negative stimulation points than positive ones (49 vs. 9). The sample was heterogeneous in terms of age and tumour type and location. To fully establish correspondence between fMRI and ECS using methodology presented in this thesis, a bigger study has to be performed.

Comparing ECS with fMRI has proven to be challenging due to many factors. Even though we have established an automated and adaptive procedure for thresholding fMRI maps, ECS suffers from thresholding issues as well. Depending on amplitude of the stimulation current used one may or may not elicit a response. Additionally ECS mapping is intrinsically limited by the procedure time and the size of the craniatomy. Additionally the methodology used in this pilot study for relating the stimulation

location to presurgical MRI scans can be inaccurate. In the future a neuronavigation system could be used to intraoperatively record the position of stimulation points (O'Shea et al., 2006).

5.4.3 *Behavioural correspondance*

No significant relation has been found between preoperative behavioural deficits and proximity of the eloquent cortex to the tumour. This could be caused by the small and heterogeneous sample. In terms of the ECS measured distance, one has to keep in mind that ECS was only performed just before the resection and the deficit might have been caused by subcortical functional areas. fMRI correspondence may be improved with delineation of the tumour, calculating the distance automatically and fitting a linear model. Behavioural deficit might also be more correlated to difference in activation pattern before and after the disease onset. However, such a dataset would be hard to come by. Additionally some aspects of the observed deficits might be caused by oedema compressing the eloquent cortex rather than the tumour itself or to damage to the white matter which could be assessed using Diffusion Tensor Imaging (DTI).

One has to ask, however, if ECS is the absolute gold standard that all other mapping methods should be aspiring to. Despite the already mentioned amplitude selection problem, the understanding of how ECS influences the cortex is limited (Borchers et al., 2012). Stimulation at the same location can lead to either inhibition of excitation of behaviour such as hearing voices vs. deafness. Additionally regional non-specificity has been reported – where different stimulation locations elicit the same response (Borchers et al., 2012). Such findings put the status of ECS as the gold standard under question.

One of the most promising results of this study is that in one third of the cases fMRI based information was enough for the surgeon to perform the procedure. Patients were not woken up during the procedure and ECS was not performed. This reduced the complexity and length of the procedure making it safer and cheaper. One has to keep in mind, however, that this finding is very preliminary. A randomized clinical trial comparing fMRI and ECS on a well balanced and homogenous population of patients needs to be performed.

Last but not least another big achievement of this study is to show that despite common practice of manual thresholding of statistical maps (O'Donnell et al., 2011; O'Shea et al., 2006), fMRI mapping can be performed in a fully automatic way with the help of our AT algorithm. Automation improves reproducibility, decreases influence of potential biases and creates a standard for using presurgical mapping as an out hospital service.

5.5 SUMMARY

In this chapter we have described a pilot study on 18 tumour patients and shown that using methods described in previous chapters we were able to create maps with excellent correspondence with ECS. Additionally, thanks to the use of presurgical

fMRI and our methods, ECS did not have to be performed in one third of the cases making the surgery shorter, safer and cheaper.

In the next chapter we will describe a neuroinformatics framework that facilitated this research and its deployment in the clinical setting.

6

NIPYPE - A NEUROIMAGING DATA PROCESSING FRAMEWORK

6.1 INTRODUCTION

This work has been undertaken in collaboration with Christopher D. Burns, Cindee Madison, Dav Clark, Yaroslav O. Halchenko, Michael L. Waskom, Satrajit S. Ghosh and other contributors to the Nipype project (see `http://github.com/nipy/nipype/contributors`). This work has been published in Frontiers of Neuroinformatics (Gorgolewski et al., 2011b) and has been presented on OHBM 2010 (Ghosh et al., 2010) and OHBM 2012 (Gorgolewski et al., 2012a).

In this chapter we will introduce and describe a novel data processing framework that was used to conduct the analyses described in the previous chapters. Even though this framework was to some extent developed to facilitate this research it addresses much broader questions. We will attempt to describe it in a general way and use the simulations, reliability and clinical studies as examples of how it can be used. Nonetheless its potential applications are beyond this particular project and we believe that it can improve the way data processing is performed in the neuroimaging community.

Over the past twenty years, advances in non-invasive in vivo neuroimaging have resulted in an explosion of studies investigating human cognition in health and disease. Current imaging studies acquire multi-modal image data (e.g., structural, diffusion, functional) and combine these with non-imaging behavioral data, patient and/or treatment history and demographic and genetic information. Several sophisticated software packages (e.g., AFNI, BrainVoyager, FSL, FreeSurfer, Nipy, R, SPM) are used to process and analyze such extensive data. In a typical analysis, algorithms from these packages, each with its own set of parameters, process the raw data. However, data collected for a single study can be diverse (highly multi-dimensional) and large, and algorithms suited for one dataset may not be optimal for another. This complicates analysis methods and makes data exploration and inference challenging, and comparative analysis of new algorithms difficult. Additionally the heterogeneous nature of the neuroimaging software ecosystem makes it difficult to translate methods developed for cognitive human neuroscience studies into solutions ready for clinical use.

6.1.1 *Current problems*

Here we outline issues that hinder replicable, efficient and optimal use of neuroimaging analysis approaches.

1. *No uniform access to neuroimaging analysis software and usage information.* For current multi-modal datasets, researchers typically resort to using different soft-

ware packages for different components of the analysis. However, these different software packages are accessed, and interfaced with, in different ways, such as: shell scripting (FSL, AFNI, Camino), MATLAB (SPM) and Python (Nipy). This has resulted in a heterogeneous set of software tools with no uniform way to use these tools or execute them. With the primary focus on algorithmic improvement, academic software development often lacks a rigorous software engineering framework that involves extensive testing, documentation, and integration/compatibility with other tools. This often necessitates extensive interactions with the authors of the software to understand their parameters, their quirks and their usage.

2. *No framework for comparative algorithm development and dissemination.* Except for some large software development efforts (e.g., SPM, FSL, FreeSurfer), most algorithm development happens in-house and stays within the walls of a lab, without extensive exposure or testing. Furthermore, testing the comparative efficacy of algorithms often requires significant effort (Klein et al., 2010). In general, developers create software for a single package (e.g., VBM8 for SPM), create a standalone cross-platform tool (e.g., Mricron) or simply do not distribute the software or code (e.g., normalization software used for registering architectonic atlases to MNI single subject template - Hömke, 2006).

3. *Methods developed for cognitive neuroscience research are hard to convert into clinical tools.* Many new neuroimaging algorithms have been developed to facilitate research on both healthy and diseased populations. However, the process of translating these advancements into methods that could improve clinical practice (for better diagnosis, procedure planning and evaluation) has been difficult. Due to heterogeneity of the available implementations it is troublesome to create self-contained, automatic and easy to use data workflows that would be suitable for clinical trails.

4. *Neuroimaging software packages do not address computational efficiency.* The primary focus of neuroimaging analysis algorithms is to solve problems (e.g., registration, statistical estimation, tractography). While some developers focus on algorithmic or numerical efficiency, most developers do not focus on efficiency in the context of running multiple algorithms on multiple subjects, a common scenario in neuroimaging analysis. Creating an analysis workflow for a particular study is an iterative process dependent on the quality of the data and participant population (e.g., neurotypical, presurgical, etc). Researchers usually experiment with different methods and their parameters to create a workflow suitable for their application, but no suitable framework currently exists to make this process efficient. Furthermore, very few of the available neuroimaging tools take advantage of the growing number of parallel hardware configurations (multicore, clusters, clouds and supercomputers).

5. *Method sections of journal articles are often inadequate for reproducing results.* Several journals (e.g., PNAS, Science, PLoS) require mandatory submission of data and scripts necessary to reproduce results of a study. However, most current method sections do not have sufficient details to enable a researcher knowledgeable in the domain to reproduce the analysis process. Furthermore, as discussed above,

typical neuroimaging analyses integrate several tools and current analysis software does not make it easy to reproduce all the analysis steps in the proper order. This leaves a significant burden on the user to satisfy these journal requirements as well as ensure that analysis details are preserved with the intent to reproduce.

6.1.2 *Current solutions*

There were several attempts to address those issues by creating a pipeline engine. Taverna (Oinn et al., 2006), VisTrails (Callahan et al., 2006) are general pipelining systems with excellent support for web services, but they do not address problems specific to neuroimaging. BrainVisa (Cointepas et al., 2001), MIPAV (McAuliffe et al., 2001), SPM include their own batch processing tools, but do not allow mixing components from other packages. Fiswidgets (Fissell et al., 2003), a promising initial approach, appears to have not been developed and does not support state of the art methods. A much more extensive and feature rich solution is the LONI Pipeline (Dinov et al., 2009, 2010; Rex et al., 2003). It provides an easy to use graphical interface for choosing processing steps or nodes from a predefined library and defining their dependencies and parameters. Thanks to advanced client-server architecture, it also has extensive support for parallel execution on an appropriately configured cluster (including data transfer, pausing execution, and combining local and remote software). Additionally, the LONI Pipeline saves information about executed steps (such as software origin, version and architecture) thereby providing provenance information (Mackenzie-Graham et al., 2008).

However, the LONI Pipeline does not come without limitations. Processing nodes are defined using eXtensible Markup Language (XML). This "one size fits all" method makes it easy to add new nodes as long as they are well-behaved command lines. However, many software packages do not meet this criterion. For example, SPM, written in MATLAB, does not provide a command line interface. Furthermore, for several command line programs, arguments are not easy to describe in the LONI XML schema (e.g., ANTS – Avants and Gee, 2004). Although it provides a helpful graphical interface, the LONI Pipeline environment does not provide an easy option to script a workflow or for rapidly exploring parametric variations within a workflow (e.g., VisTrails). Finally, due to restrictive licensing, it is not straightforward to modify and redistribute the modifications. For summary of comparison of other solutions see Table 6.1.

To address issues with existing workflow systems and the ones described earlier, we created Nipype (Neuroimaging in Python: Pipelines and Interfaces —- Gorgolewski et al., 2011b), an open source, community-developed, Python-based software package that easily interfaces with existing software for efficient analysis of neuroimaging data and rapid comparative development of algorithms. Nipype uses a flexible, efficient and general purpose programming language —- Python —- as its foundation. Processing modules and their inputs and outputs are described in an object-oriented manner providing the flexibility to interface with any type of software (not just well behaved command lines). The workflow execution engine has a plug-in architecture and supports both local execution on multi-core machines and remote execution on clusters. Nipype is distributed with a BSD license allowing anyone to

make changes and redistribute it. Development is done openly with collaborators from many different labs, allowing adaptation to the varied needs of the neuroimaging community.

Table 6.1: Feature comparison of selected pipeline frameworks. BrainVisa, MIPAV and SPM were not included due to their inability to combine software from different packages.

	Local multi-processing[1]	Grid engine	Scripting support	XNAT	Web Services[2]	Platforms	Graphical User Interface	Designed for neuroimaging
Taverna	Yes	PBS	Java, R	Yes	Yes	Mac, Unix, Windows	Yes	No
VisTrails	Yes	n/a	Python	Yes	Yes	Mac, Unix, Windows	Yes	No
Fiswidgets	No	n/a	Java	No	No	Mac, Unix, Windows	Yes	Yes
LONI	No	DRMAA	No	Yes	No	Mac, Unix, Windows	Yes	Yes
Nipype	Yes	SGE, PBS, Condor, IPython	Python	Yes	No	Mac, Unix	No	Yes

6.2 IMPLEMENTATION DETAILS

Nipype consists of three components (see Figure 6.1): 1) interfaces to external tools that provide a unified way for setting inputs, executing and retrieving outputs; 2) a workflow engine that allows creating analysis pipelines by connecting inputs and outputs of interfaces as a directed acyclic graph (DAG); and 3) plugins that execute workflows either locally or in a distributed processing environment (e.g., Torque[3], SGE/OGE). In the following sections, we describe key architectural components and features of this software.

6.2.1 *Interfaces*

Interfaces form the core of Nipype. The goal of Interfaces[4] is to provide a uniform mechanism for accessing analysis tools from neuroimaging software packages (e.g., FreeSurfer, FSL, SPM). Interfaces can be used directly as a Python object, incorporated into custom Python scripts or used interactively in a Python console. For example, there is a Realign Interface that exposes the SPM realignment routine, while the MCFLIRT Interface exposes the FSL realignment routine. In addition, one can also implement an algorithm in Python within Nipype and expose it as an Interface. Interfaces are flexible and can accommodate the heterogeneous software that needs to be supported, while providing unified and uniform access to these tools for the user. Since, there is no need for the underlying software to be changed (recompiled or adjusted to conform to a certain standard), developers can continue to create software using the computer language of their choice.

An Interface definition consists of: (a) input parameters, their types (e.g., file, floating point value, list of integers, etc.) and dependencies (e.g., does input 'a' require input 'b'); (b) outputs and their types, (c) how to execute the underlying software (e.g., run a MATLAB script, or call a command line program); and (d) a mapping which defines the outputs that are produced given a particular set of inputs. Using an object-oriented approach, we minimize redundancy in interface definition by creating a hierarchy of base Interface classes to encapsulate common functionality (e.g. Interfaces that call command line programs are derived from the CommandLine class, which provides methods to translate Interface inputs into command line parameters and for calling the command.)

We use Enthought Traits[5] to create a formal definition for Interface inputs and outputs, to define input constraints (e.g., type, dependency, whether mandatory) and to provide validation (e.g., file existence). This allows malformed or underspecified inputs to be detected prior to executing the underlying program. The input definition also allows specifying relations between inputs. Often, some input options should not be set together (mutual exclusion) while other inputs need to be set as a group (mutual inclusion).

Currently, Nipype (version 0.6) is distributed with a wide range of interfaces (see Table 6.2.). Adding new Interfaces is simply a matter of writing a Python class defi-

3 http://www.clusterresources.com/products/torque-resource-manager.php

4 Throughout the rest of the chapter we are going to use upper case for referring to classes (such as Interfaces, Workflows etc...) and lower case to refer to general concepts.

5 http://code.enthought.com/projects/traits/

nition. When a formal specification of inputs and outputs are provided by the underlying software, Nipype can support these programs automatically. For example, the Slicer command line execution modules come with an XML specification that allows Nipype to wrap them without creating individual interfaces.

Table 6.2: Supported software. List of software packages fully or partially supported by Nipype. For more details see http://nipy.org/nipype/interfaces

Name	URL
AFNI	www.afni.nimh.nih.gov/afni
BRAINS	www.psychiatry.uiowa.edu/mhcrc/IPLpages/BRAINS.htm
Camino	www.cs.ucl.ac.uk/research/medic/camino
Camino-TrackVis	www.nitrc.org/projects/camino-trackvis
ConnecomeViewerToolkit	www.connectomeviewer.org
dcm2nii	www.cabiatl.com/mricro/mricron/dcm2nii.html
Diffusion Toolkit	www.trackvis.org/dtk
FreeSurfer	www.freesurfer.net
FSL	www.fmrib.ox.ac.uk/fsl
Nipy	www.nipy.org/nipy
NiTime	www.nipy.org/nitime
Slicer	www.slicer.org
SPM	www.fil.ion.ucl.ac.uk/spm
SQLite	www.sqlite.org
PyXNAT	www.github.com/pyxnat,xnat.org

6.2.2 *Nodes, MapNodes, and Workflows*

Nipype provides a framework for connecting Interfaces to create a data analysis Workflow. In order for Interfaces to be used in a Workflow they need to be encapsulated in either Node or MapNode objects. Node and MapNode objects provide additional functionality to Interfaces, e.g. creating a hash of the input state, caching of results and the ability to iterate over inputs. Additionally, they execute the underlying interfaces in their own uniquely named directories (almost like a sandbox), thus providing a mechanism to isolate and track the outputs resulting from execution of the Interfaces. These mechanisms allow not only for provenance tracking, but aid in efficient pipeline execution.

The MapNode class is a sub-class of Node that implements a MapReduce-like architecture (Dean and Ghemawat, 2008). Encapsulating an Interface within a MapNode allows Interfaces that normally operate on a single input to execute the Interface on multiple inputs. When a MapNode executes, it creates a separate instance of the underlying Interface for every value of an input list and executes these instances independently. When all instances finish running, their results are collected into a list and exposed through the MapNode's outputs (see Figure 6.2D). This approach improves

granularity of the Workflow and provides easy support for Interfaces that can only process one input at a time. For example, the FSL 'bet' program can only run on a single input, but wrapping the BET Interface in a MapNode allows running 'bet' on multiple inputs.

A Workflow object captures the processing stages of a pipeline and the dependencies between these processes. Interfaces encapsulated into Node or MapNode objects can be connected together within a Workflow. By connecting outputs of some Nodes to the inputs of others, the user implicitly specifies dependencies. These are represented internally as a directed acyclic graph (DAG). The current semantics of Workflow do not allow conditionals and hence the graph needs to be acyclic. Workflows themselves can be a node of the Workflow graph (see Figure 6.1). This enables a hierarchical architecture and encourages Workflow reuse. The Workflow engine validates that all nodes have unique names, ensures that there are no cycles, and prevents connecting multiple outputs to a given input. For example in an fMRI processing Workflow, preprocessing, model fitting and visualization of results can be implemented as individual Workflows connected together in a main Workflow. This not only improves clarity of designed Workflows but also enables easy exchange of whole subsets. Common Workflows can be shared across different studies within and across laboratories thus reducing redundancy and increasing consistency.

While a neuroimaging processing pipeline could be implemented as a Bash, MATLAB or a Python script, Nipype explicitly implements a pipeline as a graph. This makes it easy to follow what steps are being executed and in what order. It also makes it easier to go back and change things by simply reconnecting different outputs and inputs or by inserting new Nodes/MapNodes. This alleviates the tedious component of scripting where one has to manually ensure that the inputs and outputs of different processing calls match and that operations do not overwrite each other's outputs.

A Workflow provides a detailed description of the processing steps and how data flows between Interfaces. Thus it is also a source of provenance information. We encourage users to provide Workflow definitions (as scripts or graphs) as supplementary material when submitting articles. This ensures that at least the data processing part of the published experiment is fully reproducible. Additionally, exchange of Workflows between researchers stimulates efficient use of methods and experimentation.

6.2.3 *Example - building a Workflow from scratch*

In this section, we illustrate to the creation and extend a typical fMRI processing Workflow, as applied to the healthy subjects (session 1) presented in Chapter 4 and to patients in Chapter 5. This fMRI Workflow can be divided into two sections: 1) preprocessing and 2) modeling. The first one deals with cleaning data from confounds and noise, and the second one fits a model to the cleaned data based on the experimental design. Here we will present how to set up only two steps: 1) skipping initial volumes (to remove artefacts related to scanner stabilization) and 2) slice timing correction (temporal resampling of the data to make all voxel from the same volumed

aligned to the same point in time). We use FSL and SPM Interfaces to define the processing Nodes.[6]

```
from nipype.pipeline.engine import Node, Workflow

skip = Node(interface=fsl.ExtractROI(), name="skip")
skip.inputs.t_min = 4
skip.inputs.t_max = -1

slice_timing = Node(interface=spm.SliceTiming(), name="slice_timing")
slice_timing.inputs.num_slices = 28
slice_timing.inputs.time_repetition = 2.5
```

We create a Workflow to include these two Nodes and define the data flow from the output of the realign Node (*realigned_files*) to the input of the smooth Node (*in_files*). This creates a simple preprocessing workflow.

```
preprocessing = Workflow(name="preproc_func")
preprocessing.connect(skip, "roi_file", slice_timing, "in_files")
```

The remaining preprocessing steps as well as the modeling Workflow is constructed in an analogous manner. Modeling is implemented by first defining Nodes for model design, model estimation and contrast estimation. We again use SPM Interfaces for this purpose. However, Nipype adds an extra abstraction Interface for model specification whose output can be used to create models in different packages (e.g., SPM, FSL and Nipy). The nodes of this Workflow are: SpecifyModel (Nipype model abstraction Interface), Level1Design (SPM design definition), ModelEstimate, and ContrastEstimate.

We create a master Workflow that connects the preprocessing and modeling Workflows, adds the ability to select data for processing (using DataGrabber Interface) and a DataSink Node to save the outputs of the entire Workflow. Nipype allows connecting Nodes between Workflows. We will use this feature to connect *realignment_parameters* and *smoothed_files* to modeling workflow.

The DataGrabber Interface allows the user to define flexible search patterns which can be parameterized by user defined inputs (such as subject ID, session etc.). This Interface can adapt to a wide range of directory organization and file naming conventions. In our case we will parameterize it with subject ID. In this way we can run the same Workflow for different subjects. We automate this by iterating over a list of subject IDs, by setting the iterables property of the DataGrabber Node for the input *subject_id*. The DataGrabber Node output is connected to the realign Node from preprocessing Workflow.

DataSink on the other side provides means for storing selected results in a specified location. It supports automatic creation of folders, simple substitutions and regular expressions to alter target filenames. In this case we store the statistical (T-maps) resulting from contrast estimation. A Workflow defined in this way (see **??**workflowfig:patient_workflow

6 Some essential input parameters such as slice order were omitted in this presentation to save space.

) is ready to run. This can be done by calling the run() method of the master Workflow.

If the run() method is called twice, the Workflow input hashing mechanism ensures that none of the Nodes are executed during the second run if the inputs remain the same. If, however, a highpass filter parameter of *specify_model* is changed, some of the Nodes (but not all) would have to rerun. Nipype automatically determines which Nodes require rerunning.

6.2.4 *Iterables — Parameter space exploration*

Nipype provides a flexible approach to prototyping and experimentation with different processing strategies, through the unified and uniform access to a variety of software packages (Interfaces) and creating data flows (Workflows). However, for various neuroimaging tasks, there is often a need to explore the impact of variations in parameter settings (e.g., how do different amounts of smoothing affect group statistics, what is the impact of spline interpolation over trilinear interpolation). To enable such parametric exploration, Nodes have an attribute called iterables.

When an iterable is set on a Node input, the Node and its subgraph are executed for each value of the iterable input (see Figure 6.2). Iterables can also be set on multiple inputs of a Node (e.g., somenode.iterables = [('input1', [1,2,3]), ('input2', ['a', 'b'])]). In such cases, every combination of these values is used as a parameter set (the prior example would result in the following parameter sets: (1, 'a'), (1, 'b'), (2, 'a'), etc.). This feature is especially useful to investigate interactions between parameters of intermediate stages with respect to the final results of a workflow. A common use-case of iterables is to execute the same Workflow for many subjects in an fMRI experiment and to simultaneously look at the impact of parameter variations on the results of the Workflow.

It is important to note that unlike MapNode, which creates copies of the underlying interface for every element of an input of type list, iterables operate on the subgraph of a node and create copies not only of the node but also of all the nodes dependent on it (see Figure 6.2).

6.2.5 *Parallel Distribution and Execution Plug-ins*

Nipype supports executing Workflows locally (in series or parallel) or on load-balanced grid-computing clusters (e.g., SGE, Torque or even via SSH) through an extensible plug-in interface. No change is needed to the Workflow to switch between these execution modes. One simply calls the Workflow's run function with a different plug-in and its arguments. Very often different components of a Workflow can be executed in parallel and even more so when the same Workflow is being repeated on multiple parameters (e.g., subjects). Adding support for additional cluster management systems does not require changes in Nipype, but simply writing a plug-in extension conforming to the plug-in API.

The Workflow engine sends an execution graph to the plug-in. Executing the Workflow in series is then simply a matter of performing a topological sort on the graph and running each node in the sorted order. However, Nipype also provides additional

plugins that use Python's multi-processing module, use IPython (includes SSH-based, SGE, LSF, PBS, among others) and provide native interfaces to SGE, PBS/Torque, and Condor clusters. For all of these, the graph structure defines the dependencies as well as which nodes can be executed in parallel at any given stage of execution.

One of the biggest advantages of Nipype's execution system is that parallel execution using local multiprocessing plug-in does not require any additional software (such as cluster managers like SGE) and therefore makes prototyping on a local multicore workstations easy. However, for bigger studies and complex Workflows, a high-performance computing cluster can provide substantial improvements in execution time. Since there is a clear separation between the definition of the Workflow and its execution, Workflows can be executed in parallel (locally or on a cluster) without any modification. Transitioning from developing a processing pipeline on a single subject on a local workstation to executing it on a bigger cohort on a cluster is therefore seamless.

Rerunning workflows has also been optimized. When a Node or MapNode is run, the framework will actually execute the underlying interface only if inputs have changed relative to prior execution. If not, it will simply return cached results.

6.2.6 *The Function Interface*

One of the Interfaces implemented in Nipype requires special attention: The Function Interface. Its constructor takes as arguments Python function pointer or code, list of inputs and list of outputs. This allows running any Python code as part of a Workflow. When combined with libraries such as Nibabel (neuroimaging data input and output), Numpy/Scipy (array representation and processing) and scikits-learn or PyMVPA (machine learning and data mining) the Function Interface provides means for rapid prototyping of complex data processing methods. In addition, by using the Function Interface users can avoid writing their own Interfaces which is especially useful for ad-hoc solutions (e.g., calling an external program that has not yet been wrapped as an Interface).

6.2.7 *Workflow Visualisation*

To be able to efficiently manage and debug Workflow one has to have access to a graphical representation. Using graphviz (Ellson et al., 2002), Nipype generates static graphs representing Nodes and connections between them. In the current version four types of graphs are supported: orig – does not expand inner Workflows, flat – expands inner workflows, exec – expands workflows and iterables, and hierarchical – expands workflows but maintains their hierarchy. Graphs can be saved in a variety of file formats including Scalable Vector Graphics (SVG) and Portable Network Graphics (PNG) (see Figure 6.3 and Figure 6.5).

6.2.8 *Configuration Options*

Certain options concerning verbosity of output and execution efficiency can be controlled through configuration files or variables. These include, among others, *hash_method*

and *remove_unecessary_outputs*. As explained before, rerunning a Workflow only recomputes those Nodes whose inputs have changed since the last run. This is achieved by recording a hash of the inputs. For files there are two ways of calculating the hash (controlled by the *hash_method* config option): timestamp —- based only on the size and modification time and content — based on the content of the file. The first one is faster, but does not deal with the situation when an identical copy overwrites the file. The second one can be slower especially for big files, but can tell that two files are identical even if they have different modification times. To allow efficient recomputation Nipype has to store outputs of all Nodes. This can generate a significant amount of data for typical neuroimaging studies. However, not all outputs of every Node are used as inputs to other Nodes or relevant to the final results. Users can decide to remove those outputs (and save some disk space) by setting the *remove_unecessary_outputs* to True. These and other configuration options provide a mechanism to streamline the use of Nipype for different applications.

6.2.9 *Deployment*

Nipype supports GNU/Linux and Mac OS X Operating System (OS). A recent Internet survey based study showed that GNU/Linux is the most popular platform in the neuroimaging community and together with Mac OS X is used by over 70% of neuroimagers. There are not theoretical reasons why Nipype should not work on Windows (Python is a cross-platform language), but since most of the supported software (for example FSL) requires a Unix based OS Nipype is not tested on this platform.

We currently provide three ways of deploying Nipype on a new machine: manual installation from sources (`http://nipy.org/nipype/`), PyPi repository (`http://pypi.python.org/pypi/nipype/`), and from package repositories on Debian-based systems. Manual installation involves downloading a source code archive and running a standard Python installation script (distutils). This way the user has to take care of installing all of the dependencies. Installing from PyPI repository lifts this constraint by providing dependency information and automatically installing required packages. Nipype is available from standard repositories on recent Debian and Ubuntu releases. Moreover, the NeuroDebian (http://neuro.debian.net - Halchenko and Hanke, 2012 repository provides the most recent releases of Nipype for Debian-based systems and a NeuroDebian Virtual Appliance making it easy to deploy Nipype and other imaging tools in a virtual environment on other operating systems, e.g. Windows. In addition to providing all core dependencies and automatic updates NeuroDebian also provides many of the software packages supported by Nipype (AFNI, FSL, Mricron, etc), making deployment of heterogeneous Nipype pipelines more straightforward.

6.2.10 *Development*

Nipype is trying to address the problem of interacting with the ever changing universe of neuroimaging software in a sustainable manner. Therefore the way its development is managed is a part of the solution. Nipype is distributed under the Berkley Software Distribution license which allows free copying, modification and distribu-

tion, and additionally meets all the requirements of open source definition (see Open Source Initiative - http://www.opensource.org/docs/osd) and Debian Free Software Guidelines (http://www.debian.org/social_contract#guidelines). Development is carried out openly through a distributed version control system (git via GitHub - http://github.com/nipy/nipype) in an online community. The current version of the source code together with its complete history is accessible to everyone. Discussions between developers and design decisions are undertaken on an open access mailing list. This setup encourages a broader community of developers to join the project (the project has a constantly growing user and developer base – see Figure growth) and allows sharing of the development resources (effort, money, information and time). Additionally, because the project is not tied to one source of funding or one particular lab it is able to develop freely and adjust to the needs of a wider range of users. This leads to the ability to cover more use cases and further facilitates rapid prototyping. 41 contributors submitted 4020 commits leading to a code base of 38942 lines of code (after exclusion of white spaces and comments; these number were retrieved at 21/08/12). Within the Basic COCOMO framework (Boehm, 1984) and assuming an average yearly salary of 55000 the project would have cost 512,036 USD.

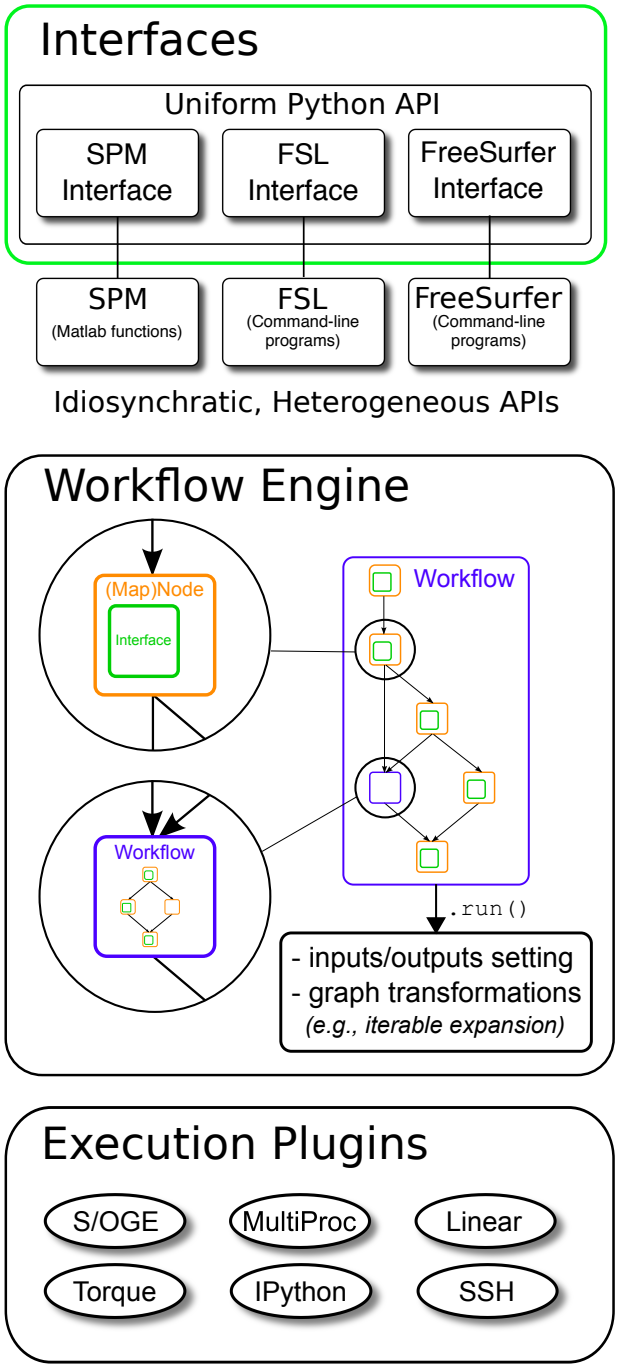

Figure 6.1: Architecture overview of the Nipype framework. Interfaces are wrapped with Nodes or MapNodes and connected together as a graph within a Workflow. Workflows themselves can act as a Node inside another Workflow, supporting a composite design pattern. The dependency graph is transformed before being executed by the engine component. Execution is performed by one of the plugins. Currently Nipype supports serial and parallel (both local multithreading and cluster) execution.

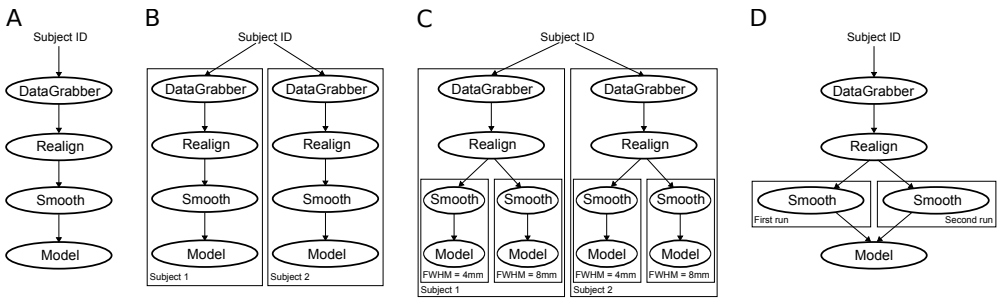

Figure 6.2: Workflow modification using iterables and MapNodes. If we take the processing pipeline A and set iterables parameter of DataGrabber to a list of two subjects, Nipype will effectively execute graph B. Identical processing will be applied to every subject from the list. Iterables can be used in a graph on many levels. For example, setting iterables on Smooth FWHM to a list of 4 and 8 mm will result in graph C. In contrast to iterables, MapNode branches within a node of the graph and also merges the results of the branches, effectively performing a MapReduce operation (D).

6.3 USAGE EXAMPLES

6.3.1 *A framework for comparative algorithm development and dissemination*

Uniform semantics for interfacing with a wide range of processing methods not only opens the possibility for richer Workflows, but also allows comparing algorithms that are designed to solve the same problem across and within such diverse Workflows. Typically, such an exhaustive comparison can be time-consuming, because of the need to deal with interfacing different software packages. Nipype simplifies this process by standardizing the access to the software. Additionally, the iterables mechanism allows users to easily extend such comparisons by providing a simple mechanism to test different parameter sets.

All simulations in Chapter 3 have been made using Nipype. Simulation generator Interface has been written for this purpose. It is an example of a pure Python interface that does not call any underlying software. It's purpose is to generate artificial timeseries of data (in NIfTI) format for given parameters (such as the true activation pattern, number of volumes, and noise). To generate multiple samples and explore the parameter space, iterables was used on these parameters (and a bogus "simulation id" parameters). Standard preprocessing and modeling workflows have been connected to this node and its outputs were fed to a set of different thresholding algorithms that were evaluated (see Figure 6.3). This was possible, because these algorithms have been previously "wrapped" in Nipype Interfaces and therefore were accessible in a unified way. This made adding and exchanging algorithms included in the comparison straightforward. The whole simulation was run on an Intel Core i7 machine with four independent processing units and was seamlessly parallelized using the "MultiCore" execution plugin. Debugging of the simulations was made less time-consuming by the caching mechanism included in Nipype. Thanks to internal tracking of values used for inputs only the nodes that had to rerun were rerun after correcting the simulation parameters.

Algorithm comparison is not the only way Nipype can be useful for a neuroimaging methods researcher. It is in the interest of every methods developer to make his or hers work most accessible. This usually means providing ready to use implementations. However, because the field is so diverse, software developers have to provide several packages (SPM toolbox, command line tool, C++ library etc.) to cover the whole user base. With Nipype, a developer can create one Interface and expose a new tool, written in any language, to a greater range of users, knowing it will work with the wide range of software currently supported by Nipype.

A good example of such scenario is ArtifactDetection toolbox (`http://www.nitrc.org/projects/artifact_detect/`). This piece of software uses EPI timeseries and realignment parameters to find timepoints (volumes) that are most likely artifacts and should be removed (by including them as confound regressors in the design matrix). The tool was initially implemented as a MATLAB script, compatible only with SPM and used locally within the lab. The current Nipype interface can work with SPM or FSL Workflows, thereby not limiting its users to SPM. The AT algorithm introduced in Chapter 3 is also distributed as a Nipype Interface.

6.3.2 *An environment for prototyping clinical neuroimaging workflows*

Nipype due to its abilities to combine heterogeneous pieces of software supplemented by homebrewed solutions can be used to create custom yet fully automatic workflows for clinical use. In Chapter 3 we introduced a data processing workflow used in a test-retest study. Its aim was to establish which methods and paradigms provide reliable single subject results. It was implemented using Nipype and self-contained reusable subworkflows (such as preprocessing, modelling, and group analysis). Thanks to this design, creating a clinical workflow to be used on a weekly basis on patients employed the shared code. The subworkflows shared between the two studies included preprocessing, modeling, and thresholding (see Figure 6.5). These tasks were integrating many heterogenous software solution (such as SPM, FSL and ArtifactDetect). However, due to the unified access provided by Nipype the integration was straighforward.

The clinical workflow was supplemented by automatic report creation. These reports included various quality-control information (see Figure 6.4) and the final thresholded maps overlaid on the structural scans. Additionally, for future use in a neuronavigation suite, the final maps were binarized and turned back to DICOM format. The whole workflow was deployed on a server based in the hospital where all the patients were scanned. Due to fully automatic nature of Nipype workflows data was analyzed straight after acquisition and did not require improvement from any personnel. This is an important feature that streamlines the process and cuts the costs of hiring qualified staff for manual data processing. The results were made accessible to authorized personnel (neurosurgeons and researchers on the study) through a secure network drive. Autogenerated reports were used to perform quality assessment to make sure that the results were not spoiled by artefacts or acquisition issues. Finally the thresholded maps were used by the neurosurgeon to evaluate the risks and plan the procedure (see Figure 6.6).

The final solution is fully automatic (partially due to the new thresholding algorithm) yet safe (thanks to quality-control). The development cost was minimized by reusing existing software solutions developed by the neuroimaging community and combined using Nipype.

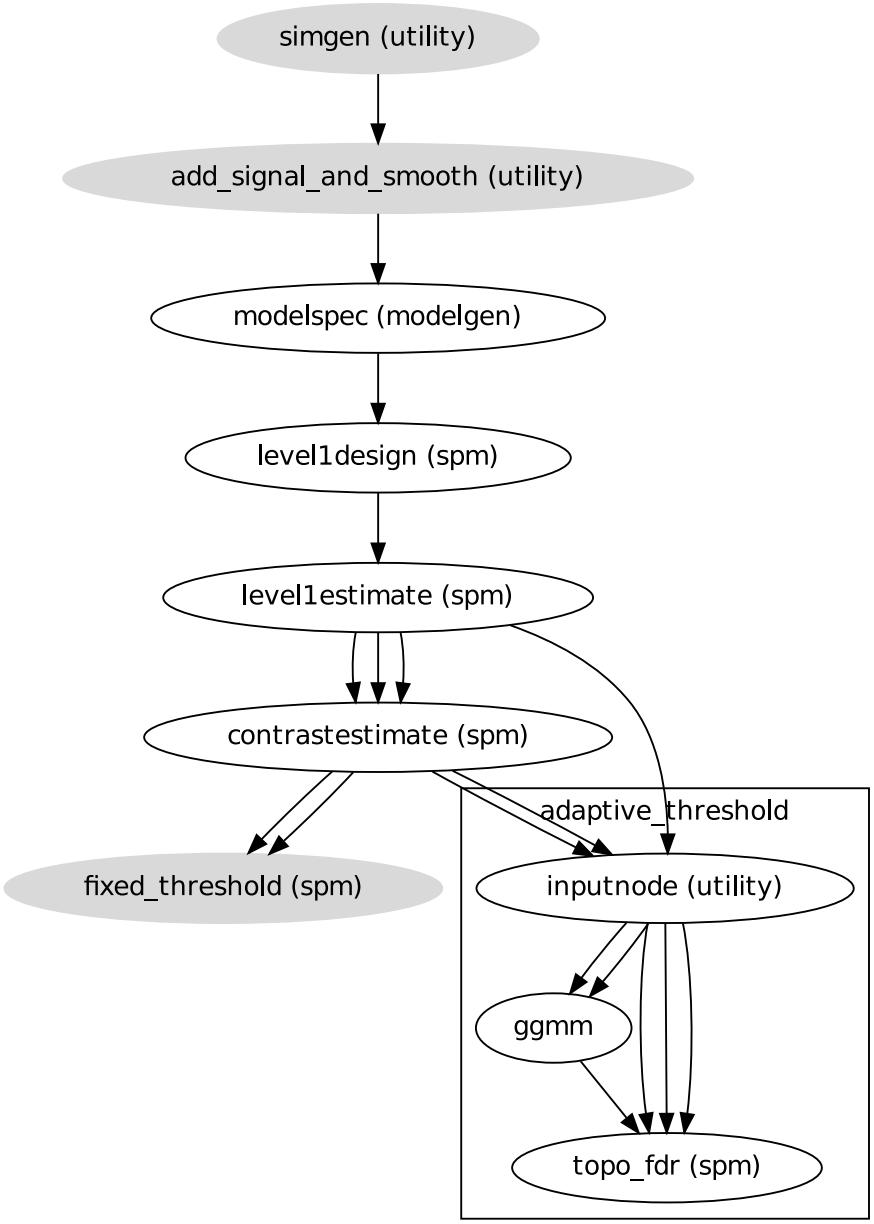

Figure 6.3: This pipeline was used to perform the second set of simulations described in Chapter 3. Rectangles represent subworkflow. Names in brackets correspond to the package the node s interface comes from. Nodes in gray are using iterables. In this example it allows iterating over different simulation ids (*simgen* node), SNR (*add_signal_and_smooth* node), and fixed threshold (*fixed_threshold* node).

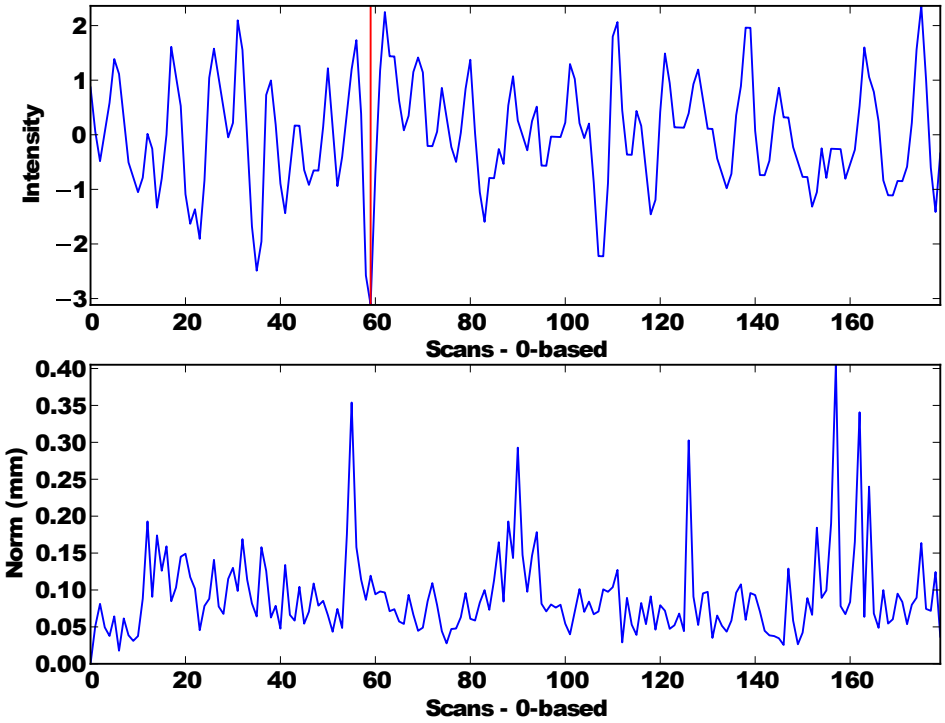

Figure 6.4: Part of the quality-control reports generated for clinical cases. This graph was generated using ArtifactDetect and was used to asses if the patient did not move too much in the scanner.

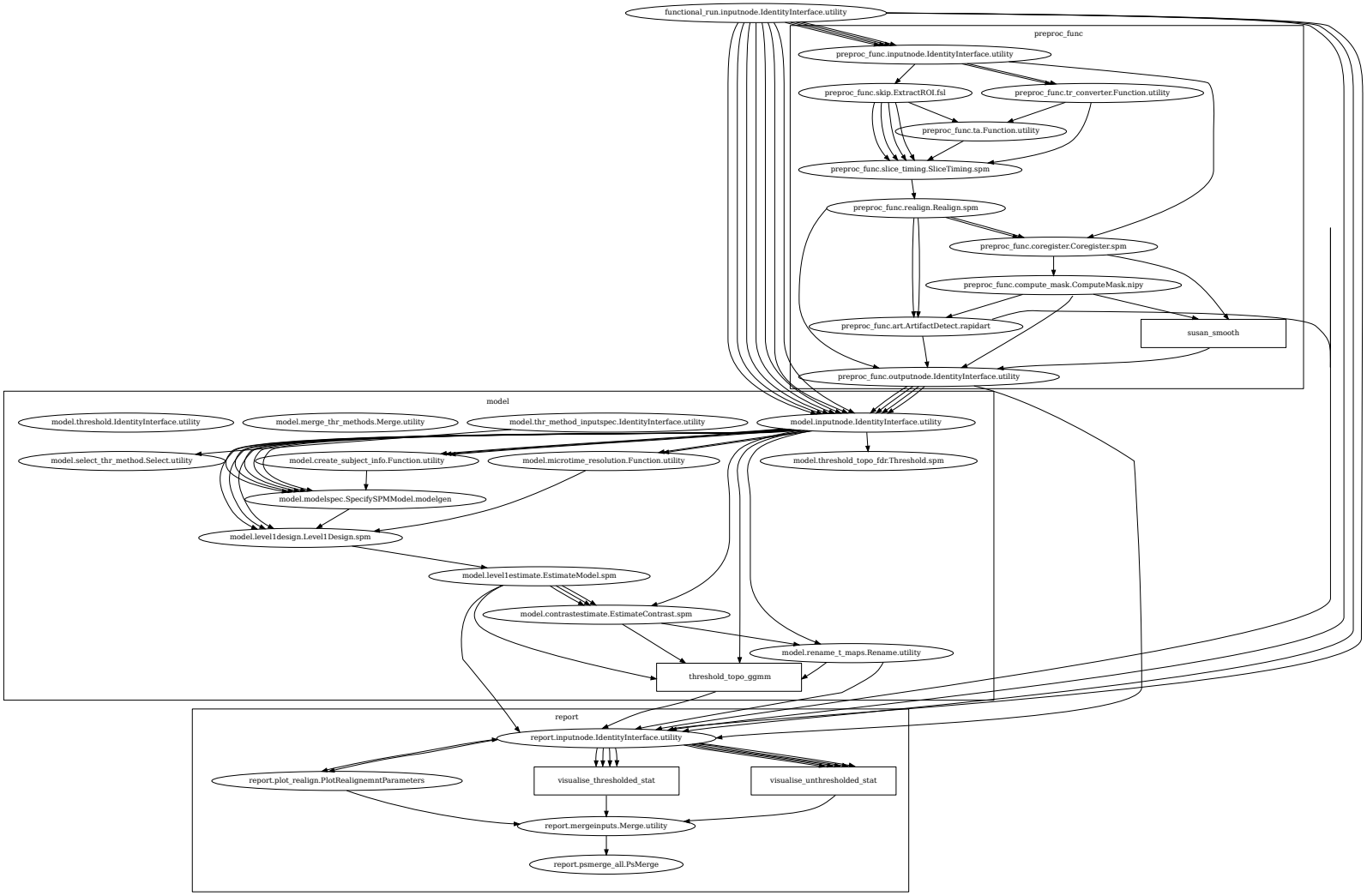

Figure 6.5: Graph showing the part of workflow used for processing data inChapter 4 and Chapter 5. Subworkflows are indicated by rectangular boxes. Nodes of some subsubworkflows are not shown for clarity reasons.

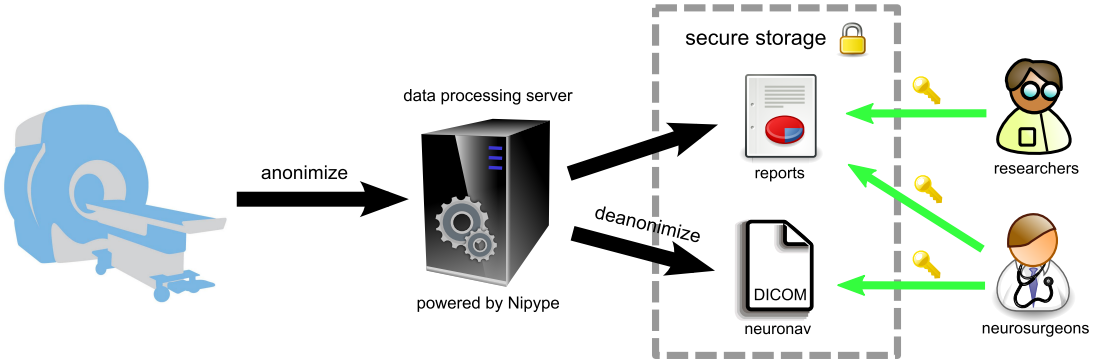

Figure 6.6: Schematic overview of the data flow and access set up for scanning patients. After anonymization, data is transferred to a processing server where Nipype is employed to create reports and DICOM files ready to be used in a neuronavigation suite. These are deanonymize to avoid potential patient mismatch. However, patient confidentiality is secured by keeping the deanonymized data in a secure network share which can be accessed only by researchers on the study and neurosurgeons.

6.3.3 *Computationally efficient execution of neuroimaging analysis*

A computationally efficient execution allows for multiple rapid-iterations to optimize a Workflow for a given application. Support for optimized local execution (running independent processes in parallel, rerunning only those steps that have been influenced by the changes in parameters or dependencies since the last run) and exploration of parameter space eases Workflow development. The Nipype package provides a seamless and flexible environment for executing Workflows in parallel on a variety of environments from local multi-core workstations to high-performance clusters. In the SPM workflow for single subject functional data analysis, only a few components can be parallelized. However, running this Workflow across several subjects provides room for embarrassingly parallel execution. Running this Workflow in distributed mode for 69 subjects on a compute cluster (40 cores distributed across 6 machines) took 1 hour and 40 minutes relative to the 32 minutes required to execute the analysis steps in series for a single subject on the same cluster. The difference from the expected runtime of 64 minutes (32 minutes for the first 40 subjects and another 32 minutes for the remaining 29 subjects) stems from disk I/O and other network and processing resource bottlenecks.

6.3.4 *Captures details of analysis required to reproduce results*

The graphs and code presented in the examples above capture all the necessary details to rerun the analysis. Any user, who has the same versions of the tools installed on their machine and access to the data and scripts, will be able to reproduce the results of the study. For example, running Nipype within the NeuroDebian framework can provide access to specific versions of the underlying tools. This provides an easy mechanism to be compliant with the submitting data and scripts/code mandates of journals such as PNAS and Science.

6.4 DISCUSSION

Current neuroimaging software offers users an incredible opportunity to analyze their data in different ways, with different underlying assumptions. However, this heterogeneous collection of specialized applications creates several problems: 1) No uniform access to neuroimaging analysis software and usage information; 2) No framework for comparative algorithm development and dissemination; 3) Methods developed in neuroscience are hard to translate in clinical reality; 4) Neuroimaging software packages do not address computational efficiency; and 5) Method sections of journal articles are often inadequate for reproducing results.

We addressed these issues by creating Nipype, an open-source, community-developed initiative under the umbrella of Nipy. Nipype solves these issues by providing uniform Interfaces to existing neuroimaging software and by facilitating interaction between these packages within Workflows. Nipype provides an environment that encourages interactive exploration of algorithms from different packages (e.g., SPM, FSL), eases the design of Workflows within and between packages, and reduces the learning curve necessary to use different packages. Nipype aims to address limitations of existing pipeline systems, thereby creating a collaborative platform for neuroimaging software development in Python, a high-level scientific computing language.

We use Python for several reasons. It has extensive scientific computing and visualization support through packages such as SciPy, NumPy, Matplotlib and Mayavi (Millman and Aivazis, 2011; Perez et al., 2011). The Nibabel package provides support for reading and writing common neuroimaging file formats (e.g., NIFTI, ANALYZE and DICOM). Being a high-level language, Python supports rapid prototyping, is easy to learn and adopt and is available across all major operating systems. At the same time Python allows the potential to seamlessly bind with C code (using Weave package) for improved efficiency of critical subroutines.

Python is also known as a good choice for the first programming language to learn (Zelle, 1999) and is chosen as the language for introductory programming at many schools and universities (`http://wiki.python.org/moin/SchoolsUsingPython`). Being a generic and free language, with various extensions available "out of the box", it has allowed many researchers to start implementing and sharing their ideas with minimal knowledge of Python, while learning more of the language and programming principles along the way. Later on, many such endeavors became popular community-driven FOSS projects, attracting users and contributors, and even outlasting the involvement of the original authors. Python has already been embraced by the neuroscientific community and is rapidly gaining popularity (Bednar, 2009; Goodman and Brette, 2009). The Connectome Viewer Toolkit (Gerhard et al., 2011), Dipy, NiBabel, Nipy, NiTime, PyMVPA (Hanke et al., 2009), PyXNAT (Schwartz et al., 2012) and Scikits-Learn are just a few examples of neuroimaging related software written in Python. Nipype, based on Python, thus has immediate access to this extensive community and its software, technological resources and support structure.

Nipype provides a formal and flexible framework to accommodate the diversity of imaging software. Within the neuroimaging community, not all software is limited to well behaved command line tools. Furthermore, a number of these tools do not have well defined inputs, outputs or usage help. Although, currently we use Enthought

Traits to define inputs and outputs of interfaces, such definitions could be easily translated into instances of XML schemas compatible with other pipeline frameworks. On the other hand, when a tool provides a formal XML description of their inputs and outputs (e.g., Slicer 3D, BRAINS), it is possible to take these definitions and automatically generate Nipype wrappers for those classes.

Nipype development welcomes input and contributions from the community. The source code is freely distributed under a Berkeley Software Distribution (BSD) license allowing anyone use of the software, and Nipype conforms to the Open Software Definition of the Open Source Initiative. The development process is fully transparent and encourages contributions from users from all around the world. The diverse and geographically distributed user and developer base makes Nipype a flexible project that takes into account needs of many scientists.

Improving openness, transparency, and reproducibility of research has been a goal of Nipype since its inception. A Workflow definition is, in principle, sufficient to reproduce the analysis. Since it was used to analyze the data, it is more detailed and accurate than a typical methods description in a paper, but also has the advantage of being reused and shared within and across laboratories. Accompanying a publication with a formal definition of the processing pipeline (such as a Nipype script) increases reproducibility and transparency of research. The Interfaces and Workflows of Nipype capture neuroimaging analysis knowledge and the evolution of methods. Although, at the execution level, Nipype already captures a variety of provenance information, this aspect can be improved by generating provenance reports defined by a standardized XML schema (MacKay, 2003).

Increased diversity of neuroimaging data processing software has made systematic comparison of performance and accuracy of underlying algorithms essential (for examples, see Klein et al., 2009, 2010. However, a platform for comparing algorithms, either by themselves or in the context of an analysis workflow, or determining optimal workflows in a given application context (e.g., Churchill et al., 2012), does not exist. Furthermore, in this context of changing hardware and software, traditional analysis approaches may not be suitable in all contexts (e.g., data from 32-channel coils which show a very different sensitivity profile, or data from children). Nipype can make such evaluations, design of optimal workflows and investigations easier (as demonstrated via the smoothing example above), resulting in more efficient data analysis for the community.

6.5 SUMMARY

In this chapter we presented Nipype, an extensible Python library and framework that provides interactive manipulation of neuroimaging data through uniform Interfaces and enables reproducible, distributed analysis using the Workflow system. It has been used to analyze all the data included in this dissertation which leads to improved performance and reproducibility. Nonetheless applications of Nipype go beyond the work presented in the previous chapters. It has encouraged the scientific exploration of different algorithms and associated parameters, eased the development of Workflows within and between packages and reduced the learning curve associated with understanding the algorithms, APIs and user interfaces of disparate packages. An open, community-driven development philosophy provides flexibility

required to address the diverse needs in neuroimaging analysis. Overall, Nipype represents an effort towards collaborative, open-source, reproducible and efficient neuroimaging software development and analysis for application in neuroscience and the clinic.

7 DISCUSSION

The aim of this work was to improve the way fMRI is used for planning surgical resections of brain tumours.

We focused first on the importance of the thresholding of statistical maps. On one hand, we showed that thresholding maps manually can lead to high error rates and is not reproducible between hospitals (Chapter 2). On the other hand, simulations showed that existing thresholding methods which were developed for group studies are not suitable for single subject clinical applications due to poor performance in terms of low SNR, no control for false negative errors, lack of concern with the delineation of the borders of activation, and poor reliability (Chapter 3). We have therefore introduced a new technique combining Gamma-Gaussian mixture models and topological FDR. We demonstrated through simulations that this new method outperforms other methods in terms of error trade-of: better balance between false positives and false negatives, better border delineation. Simulations, however, can be biased and limited. We have therefore conducted a test-retest reliability study to see if our new method could produce more reliable suprathreshold maps in real conditions, the results of which demonstrated that AT produced better results than FT.

A second important focus was to asses which paradigms were suitable for single subject use. This was performed by undertaking a test-retest reliability study comparing Dice overlap between sessions with Dice overlap between subjects. This approach lead to the selection of the covert verb generation, overt word repetition, and motor mapping (finger tapping, foot movement, and lips pouching) tasks as suitable for presurgical planning. Overt verb generation and landmark tasks were not reliable enough for such use. We also took advantage of the reliability dataset we collected to ask more general questions of how reliability can be measured and what factors influence it. We found little relation between the reliability of thresholded and unthresholded maps, and timeseries. This shows that if one is interested in single subject applications and will be using thresholded maps (which is the case for presurgical planing and functional ROI) looking at unthresholded maps is not enough. The decision how to create the suprathreshold map can have significant influence on the reliability. Additionally we have shown that scanner noise and co-registration error have little effect on reliability in contrast to patient motion. This is especially true when the motion was correlated with the stimuli. It therefore seems essential to have minimum motion during scanning and to model various regressors related to motion (6 motions parameters, their 1st derivative or "outlier" scans) to ensure good delineation of activated areas (assuming reliable areas are also the most valid).

Having established and tested the methodology of data processing and picked suitable behavioural paradigms, we performed a pilot study on 18 patients with brain tumours. The aim of this study was to check if our thresholding methods can perform equally well on the target brain tumour patient population as it did on the healthy controls in the reliability study. Additionally, we compared fMRI generated maps with ECS performed before the resection. Our method was able to successfully

create a useful map in 92% of the time and never failed completely for any patient. Additionally we have found 100% correspondence between ECS and fMRI generated maps, but we expect this number to fall when more samples are acquired in the future.

We concluded this dissertation by introducing and describing a data processing framework that has been developed to enable this researc - Nipyp. It enables rapid prototyping, provenance tracking, parameter exploration and efficient parallel execution. Its creation and maintenance also represents a new approach to scientific projects – a decentralized open source collaboration between many researchers. Nipype was paramount to performing research included in this dissertation, but at the same time addresses problems common to many projects involving neuroimaging and can therefore be used in wide range of applications.

7.1 OPEN QUESTIONS AND FUTURE DIRECTIONS

7.1.1 *Single subject statistical maps thresholding*

The thresholding method presented here belongs to a category of model based methods. This means that there is an underlying model of the data whose freedom is restricted by the set of parameters it can take. In case of AT this was the Gamma-Gaussian mixture model and the Gaussian field used in the topological FDR. The parameters of these models were fitted from the data, but if the model assumptions (such as all activation can be described by one Gaussian) were wrong it could provide inaccurate results. A way to avoid this issue is to use non-parametric methods that do not make many assumptions about the underlying data (Petersson et al., 1999). Such an approach would require resampling of the single subject data to establish the null distribution. This would require generating many permutations of the timeseries which can be difficult due to autocorrelations (Friman and Westin, 2005). There have been, however, some developments that might in the future make single subject resampling both feasible and efficient (Eklund et al., 2012). A non-parametric approach to thresholding statistical maps with presurgical applications has the potential to provide improved results.

Additionally the concept of all activation coming from the same distribution has to be questioned. A combined fMRI data driven parcellation with a regional mixture model can potentially provide more robust results by adapting to local differences in SNR, an approach similar to that described by Voyvodic (2012). How the parcellation could be driven by mixture modelling and vice versa remains an open question. One, however, has to be careful when using such an approach not to bias it through circularity, or in other words "double dipping" (see Kriegeskorte et al., 2009).

Our approach has a two-stage nature, the first being voxelwise and second clusterwise. There is, however, little information flowing between the two stages, since the crossing point from the mixture model is used as a cluster-forming threshold for topological FDR. This could be improved by applying a fully Bayesian approach. Such attempt would be very similar to Woolrich et al. (2005) with a different way of informing the model of spatial dependencies. This could be achieved by an approach similar to Kiebel et al. (1999), which is now the standard way of estimating smoothness in topological FDR. Despite mathematical neatness (a prominent feature Bayesian mod-

els) practical gains would have to be investigated experimentally. It could be the case that flow of uncertainty achieved through Bayesian reasoning would not provide significant improvements despite substantially higher computational cost, as it was the case with Holmes-Friston approach to group analysis (Mumford and Nichols, 2009).

7.1.2 *Using test-retest reliability to compare methods*

In Chapter 4 we have used a test-retest reliability study to compare two thresholding methods. In contrast to simulations such methodology is based on real data and allows methods to be compared without the need to access the ground truth. However, such an approach is susceptible to certain problems. Imagine a method that always gives the same answer no matter what the input is. Such a method would show perfect reliability, but at the same time being useless. This issue has been addressed by the NPAIRS framework (Strother et al., 2002), by comparing reliability with prediction power. However, the prediction means in this context inferring the design matrix from the data. This is useful for evaluating data processing methods up until the GLM stage. For any method beyond that stage (such as thresholding) prediction power is hard to define.

7.1.3 *Using fMRI in presurgical planning*

We have only covered some issues that arise in presurgical fMRI. There are remaining unsolved problems including artificial signal from draining veins which are especially visible in single subject fMRI which has low SNR (see Krings et al., 2001a) and the neoplasm altering HRF (Stippich, 2007). There are potential solutions to these problems that involve change of the scanning protocol. In the case of draining veins, a contrast enhanced scan can reveal their location and can be used to remove the spurious activation. Establishing the HRF alone without using any behavioural paradigm can be achieved by a breath-hold scan. Alternatively one can use model free approaches such as Finite Impulse Response to model HRF (Goutte et al., 2000). All of these approaches have to be investigated in the context of clinical applications.

Despite good correspondence between fMRI and ECS, fMRI is far from replacing ECS for all patient cases. The biggest issue with fMRI is that it does not show regions of the brain crucial to the execution of certain behaviour, but rather all those regions involved. This means that it might be potentially safe (as in postoperative deficit risk) to remove some of the tissue labelled as active through fMRI mapping. This can potentially lead to unnecessary partial resections where a more radical approach could have been done. There is, however, a subset of cases where fMRI can provide a significant improvement over ECS. When fMRI mapping shows that the eloquent cortex is far enough away from the tumour to perform a full resection, our results suggest that there is no need to perform ECS. This means that the patient does not need to be woken up, which makes the procedure shorter, safer and cheaper. fMRI has also the advantage of providing all of this information in advance before the procedure. This allows the surgeon to plan the procedure, consult a panel of specialists and the patient themselves. These advantages need to be quantified to address the real impact of fMRI on clinical practice. This leaves many questions such as: how can presurgical

fMRI mapping influence the surgeon's procedure plan? How accurately can fMRI asses the risks of surgery? Will patients appreciate a more informed decision about treatment? Most of these issues involve an interface between technology and clinical practice and investigating them will require expertise from fields of health and social sciences.

Additionally the question how to best present data to surgeons remains open. Following the false positives vs. false negatives analogy it might be better to present maps of areas safe to remove instead of regions to be avoid. Suprathreshold maps also do not convey any information about the uncertainty of the border of the activation. Some work has been done to bootstrap thresholded maps to create such confidence intervals (Kirson et al., 2008), but it has not yet been applied to surgical planning.

7.1.4 *The future of Nipype*

Nipype has filled a gap in the neuroimaging software ecosystem and has been appreciated by its users. It does not mean, however, that it is complete and cannot be improved. The way the lack of uniformity between software packages has been solved requires a laborious process of writing Interface code. This could be avoided in many cases if the community would agree on standards describing software behaviour. Such attempts have been made previously by the Slicer and LONI Pipeline, but they did not reach widespread popularity. Additionally provenance tracking should be standardized to allow improved meta-analysis and third party applications development. LONI Pipeline did introduce an XML Schema describing provenance (Mackenzie-Graham et al., 2008), but it was not used outside of their ecosystem. There is also an attempt to create a more general standard for reporting provenance by the World Wide Web Consortium (`http://www.w3.org/2011/prov`). One of these approaches needs to be picked and implemented in Nipype.

Last but not least despite the fact that a substantial development effort has been put into the workflow engine of Nipype it should be replaced in the future by an external solution. Execution of a set of tasks with dependencies and transfer of results is not a neuroimaging specific problem and there are dedicated solutions to perform it (for example see `http://www.taverna.org.uk/`). This would allow some development cost to be offloaded, while focusing on issues strictly specific to neuroimaging.

7.2 SUMMARY

We hope that advancements presented in this dissertation will lead to improved clinical practice, fewer awake craniatomies, smaller risk of postoperative deficits and more informed patients. Additionally some aspects of this work (such as reliability study and data processing workflow) have a broader impact and can influence the way fMRI is being used and interpreted.

BIBLIOGRAPHY

Abrahams, S., Goldstein, L. H., Simmons, A., Brammer, M. J., Williams, S. C. R., Giampietro, V. P., Andrew, C. M., and Leigh, P. N. (2003). Functional magnetic resonance imaging of verbal fluency and confrontation naming using compressed image acquisition to permit overt responses. *Human brain mapping*, 20(1):29–40. (Cited on page 54.)

Aguirre, G. K., Zarahn, E., and D'Esposito, M. (1998). The inferential impact of global signal covariates in functional neuroimaging analyses. *NeuroImage*, 8(3):302–6. (Cited on page 40.)

Anttila, A., Heikkilä, P., Nykyri, E., Kauppinen, T., Pukkala, E., Hernberg, S., and Hemminki, K. (1996). Risk of nervous system cancer among workers exposed to lead. *Journal of occupational and environmental medicine*, 38(2):131–6. (Cited on page 6.)

Ashburner, J. (2007). A fast diffeomorphic image registration algorithm. *NeuroImage*, 38(1):95–113. (Cited on page 61.)

Avants, B. and Gee, J. C. (2004). Geodesic estimation for large deformation anatomical shape averaging and interpolation. *NeuroImage*, 23 Suppl 1:S139–50. (Cited on page 126.)

Beckmann, C., Woolrich, M., and Smith, S. (2003). Gaussian/Gamma mixture modelling of ICA/GLM spatial maps. In *Ninth Int. Conf. on Functional Mapping of the Human Brain*. (Cited on page 31.)

Bednar, J. a. (2009). Topographica: Building and Analyzing Map-Level Simulations from Python, C/C++, MATLAB, NEST, or NEURON Components. *Frontiers in neuroinformatics*, 3(March):8. (Cited on page 145.)

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 57(1):289–300. (Cited on pages 29 and 39.)

Bennett, C. M. and Miller, M. B. (2010). How reliable are the results from functional magnetic resonance imaging? *Annals of the New York Academy of Sciences*, 1191(1):133–55. (Cited on pages 92, 111, and 112.)

Berger, M. S., Kincaid, J., Ojemann, G. A., and Lettich, E. (1989). Brain mapping techniques to maximize resection, safety, and seizure control in children with brain tumors. *Neurosurgery*, 25(5):786–92. (Cited on page 13.)

Bindal, R. K., Sawaya, R., Leavens, M. E., and Lee, J. J. (1993). Surgical treatment of multiple brain metastases. *Journal of neurosurgery*, 79(2):210–6. (Cited on page 12.)

Binder, J., Rao, S., Hammeke, T., and Frost, J. (1995). Lateralized human brain language systems demonstrated by task subtraction functional magnetic resonance imaging. *Archives of Neurology*, 52:593–601. (Cited on page 53.)

Blowers, L., Preston-Martin, S., and Mack, W. J. (1997). Dietary and other lifestyle factors of women with brain gliomas in Los Angeles County (California, USA). *Cancer causes & control*, 8(1):5–12. (Cited on page 6.)

Boehm, B. (1984). Software engineering economics. *Software Engineering, IEEE Transactions on*. (Cited on page 136.)

Bohannon, R. W. (1997). Comfortable and maximum walking speed of adults aged 20-79 years: reference values and determinants. *Age and ageing*, 26(1):15–9. (Cited on page 115.)

Bookheimer, S. Y., Zeffiro, T. a., Blaxton, T., Gaillard, W., and Theodore, W. (1995). Regional cerebral blood flow during object naming and word reading. *Human Brain Mapping*, 3(2):93–106. (Cited on page 53.)

Borchers, S., Himmelbach, M., Logothetis, N., and Karnath, H.-O. (2012). Direct electrical stimulation of human cortex - the gold standard for mapping brain functions? *Nature reviews. Neuroscience*, 13(1):63–70. (Cited on page 122.)

Brüstle, O., Ohgaki, H., Schmitt, H. P., Walter, G. F., Ostertag, H., and Kleihues, P. (1992). Primitive neuroectodermal tumors after prophylactic central nervous system irradiation in children. Association with an activated K-ras gene. *Cancer*, 69(9):2385–92. (Cited on page 7.)

Caceres, A., Hall, D. L., Zelaya, F. O., Williams, S. C. R., and Mehta, M. A. (2009). Measuring fMRI reliability with the intra-class correlation coefficient. *NeuroImage*, 45(3):758–68. (Cited on pages 64, 93, 109, 110, and 112.)

Callahan, S. P., Freire, J., Santos, E., Scheidegger, C. E., and Vo, T. S. H. T. (2006). VisTrails : Visualization meets Data Management. In *Proceedings of the 2006 ACM SIGMOD international conference on Management of data*, pages 745–747. (Cited on page 126.)

Caspers, S., Geyer, S., Schleicher, A., Mohlberg, H., Amunts, K., and Zilles, K. (2006). The human inferior parietal cortex: cytoarchitectonic parcellation and interindividual variability. *NeuroImage*, 33(2):430–48. (Cited on page 66.)

Chen, H., Ward, M. H., Tucker, K. L., Graubard, B. I., McComb, R. D., Potischman, N. A., Weisenburger, D. D., and Heineman, E. F. (2002). Diet and risk of adult glioma in eastern Nebraska, United States. *Cancer causes & control*, 13(7):647–55. (Cited on page 6.)

Chumbley, J. R. and Friston, K. J. (2009). False discovery rate revisited: FDR and topological inference using Gaussian random fields. *NeuroImage*, 44(1):62–70. (Cited on pages 23, 29, 31, 32, 35, 37, 49, and 62.)

Churchill, N. W., Oder, A., Abdi, H., Tam, F., Lee, W., Thomas, C., Ween, J. E., Graham, S. J., and Strother, S. C. (2012). Optimizing preprocessing and analysis pipelines

for single-subject fMRI. I. Standard temporal motion and physiological noise correction methods. *Human brain mapping*, 33(3):609–27. (Cited on page 146.)

Çiçek, M., Deouell, L. Y., and Knight, R. T. (2009). Brain activity during landmark and line bisection tasks. *Frontiers in human neuroscience*, 3:7. (Cited on page 55.)

Cocco, P., Dosemeci, M., and Heineman, E. F. (1998). Brain cancer and occupational exposure to lead. *Journal of occupational and environmental medicine*, 40(11):937–42. (Cited on page 6.)

Cointepas, Y., Mangin, J., Garnero, L., and Poline, J. (2001). BrainVISA: Software platform for visualization and analysis of multi-modality brain data. *NeuroImage*, (6):2001–2001. (Cited on page 126.)

Cordier, S., Mandereau, L., Preston-Martin, S., Little, J., Lubin, F., Mueller, B., Holly, E., Filippini, G., Peris-Bonet, R., McCredie, M., Choi, N. W., and Arsla, A. (2001). Parental occupations and childhood brain tumors: results of an international case-control study. *Cancer causes & control*, 12(9):865–74. (Cited on page 6.)

Cusack, R., Cumming, N., Bor, D., Norris, D., and Lyzenga, J. (2005). Automated post-hoc noise cancellation tool for audio recordings acquired in an MRI scanner. *Human brain mapping*, 24(4):299–304. (Cited on page 54.)

Dean, J. and Ghemawat, S. (2008). MapReduce: simplified data processing on large clusters. *Communications of the ACM*, 51(1):107. (Cited on page 130.)

Dempster, A., Laird, N., and Rubin, D. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 39(1):1–38. (Cited on page 36.)

Desmond, J. E., Sum, J. M., Wagner, a. D., Demb, J. B., Shear, P. K., Glover, G. H., Gabrieli, J. D., and Morrell, M. J. (1995). Functional MRI measurement of language lateralization in Wada-tested patients. *Brain*, 118(6):1411–1419. (Cited on page 19.)

Dice, L. (1945). Measures of the amount of ecologic association between species. *Ecology*, 26(3):297–302. (Cited on page 40.)

Dinov, I., Lozev, K., Petrosyan, P., Liu, Z., Eggert, P., Pierce, J., Zamanyan, A., Chakrapani, S., Van Horn, J., Parker, D. S., Magsipoc, R., Leung, K., Gutman, B., Woods, R., and Toga, A. (2010). Neuroimaging study designs, computational analyses and data provenance using the LONI pipeline. *PloS one*, 5(9). (Cited on page 126.)

Dinov, I. D., Van Horn, J. D., Lozev, K. M., Magsipoc, R., Petrosyan, P., Liu, Z., Mackenzie-Graham, A., Eggert, P., Parker, D. S., and Toga, A. W. (2009). Efficient, Distributed and Interactive Neuroimaging Data Analysis Using the LONI Pipeline. *Frontiers in neuroinformatics*, 3(July):22. (Cited on page 126.)

Dong, Y., Dobkin, B. H., Cen, S. Y., Wu, A. D., and Winstein, C. J. (2006). Motor cortex activation during treatment may predict therapeutic gains in paretic hand function after stroke. *Stroke; a journal of cerebral circulation*, 37(6):1552–5. (Cited on page 92.)

Duncan, K. J., Pattamadilok, C., Knierim, I., and Devlin, J. T. (2009). Consistency and variability in functional localisers. *NeuroImage*, 46(4):1018–26. (Cited on page 92.)

Dunn, O. (1961). Multiple Comparisons Among Means. *Journal of the American Statistical Association*, 56(293):52–64. (Cited on page 29.)

Eickhoff, S. B., Heim, S., Zilles, K., and Amunts, K. (2006). Testing anatomically specified hypotheses in functional imaging using cytoarchitectonic maps. *NeuroImage*, 32(2):570–82. (Cited on page 63.)

Eickhoff, S. B., Jbabdi, S., Caspers, S., Laird, A. R., Fox, P. T., Zilles, K., and Behrens, T. E. J. (2010). Anatomical and functional connectivity of cytoarchitectonic areas within the human parietal operculum. *The Journal of neuroscience*, 30(18):6409–21. (Cited on page 66.)

Eickhoff, S. B., Paus, T., Caspers, S., Grosbras, M.-H., Evans, A. C., Zilles, K., and Amunts, K. (2007). Assignment of functional activations to probabilistic cytoarchitectonic areas revisited. *NeuroImage*, 36(3):511–21. (Cited on page 63.)

Eickhoff, S. B., Stephan, K. E., Mohlberg, H., Grefkes, C., Fink, G. R., Amunts, K., and Zilles, K. (2005). A new SPM toolbox for combining probabilistic cytoarchitectonic maps and functional imaging data. *NeuroImage*, 25(4):1325–35. (Cited on page 63.)

Eklund, A., Andersson, M., Josephson, C., Johannesson, M., and Knutsson, H. (2012). Does parametric fMRI analysis with SPM yield valid results?-An empirical study of 1484 rest datasets. *NeuroImage*. (Cited on page 149.)

Ellson, J., Gansner, E., Koutsofios, L., North, S., and Woodhull, G. (2002). Graphviz—open source graph drawing tools. In *Graph Drawing*, pages 594–597. Springer. (Cited on page 134.)

Everitt, B. S. and Bullmore, E. T. (1999). Mixture model mapping of brain activation in functional magnetic resonance images. *Human Brain Mapping*, 7(1):1–14. (Cited on page 30.)

Fernández, G., Specht, K., Weis, S., Tendolkar, I., Reuber, M., Fell, J., Klaver, P., Ruhlmann, J., Reul, J., and Elger, C. E. (2003). Intrasubject reproducibility of presurgical language lateralization and mapping using fMRI. *Neurology*, 60(6):969–75. (Cited on pages 92, 110, and 112.)

Fink, G., Marshall, J., Shah, N., Weiss, P., and PW (2000). Line bisection judgments implicate right parietal cortex and cerebellum as assessed by fMRI. *Neurology*, 54(6):1324–1331. (Cited on pages 55 and 109.)

Fink, G., Marshall, J., Weiss, P., and Zilles, K. (2001). The neural basis of vertical and horizontal line bisection judgments: an fMRI study of normal volunteers. *NeuroImage*, 14(1 Pt 2):S59–67. (Cited on pages 55 and 109.)

Fissell, K., Tseytlin, E., Cunningham, D., Iyer, K., Carter, C. S., Schneider, W., and Cohen, J. D. (2003). Fiswidgets: a graphical computing environment for neuroimaging analysis. *Neuroinformatics*, 1(1):111–125. (Cited on page 126.)

FitzGerald, D. B., Cosgrove, G. R., Ronner, S., Jiang, H., Buchbinder, B. R., Belliveau, J. W., Rosen, B. R., and Benson, R. R. (1997). Location of language in the cortex: a comparison between functional MR imaging and electrocortical stimulation. *American Journal of Neuroradiology*, 18(8):1529–39. (Cited on page 19.)

Flöel, A., Jansen, A., Deppe, M., Kanowski, M., Konrad, C., Sommer, J., and Knecht, S. (2005). Atypical hemispheric dominance for attention: functional MRI topography. *Journal of Cerebral Blood Flow & Metabolism*, 25(9):1197–208. (Cited on page 55.)

Fox, P. T. and Raichle, M. E. (1986). Focal physiological uncoupling of cerebral blood flow and oxidative metabolism during somatosensory stimulation in human subjects. *Proceedings of the National Academy of Sciences of the United States of America*, 83(4):1140–4. (Cited on page 17.)

Friman, O. and Westin, C.-F. (2005). Resampling fMRI time series. *NeuroImage*, 25(3):859–67. (Cited on page 149.)

Friston, K., Frith, C., Liddle, P., Dolan, R., Lammertsma, A., and Frackowiak, R. (1990). The relationship between global and local changes in PET scans. *Journal of Cerebral Blood Flow & Metabolism*, 10(4):458–466. (Cited on pages 40 and 78.)

Friston, K. J. (2007). *Statistical Parametric Mapping: The Analysis of Functional Brain Images*. Elsevier/Academic Press. (Cited on page 19.)

Friston, K. J., Holmes, A. P., Worsley, K. J., Poline, J.-P., Frith, C. D., and Frackowiak, R. S. J. (1994). Statistical parametric maps in functional imaging: A general linear approach. *Human Brain Mapping*, 2(4):189–210. (Cited on pages 19 and 62.)

Friston, K. J., Worsley, K. J., Frackowiak, R. S. J., Mazziotta, J. C., and Evans, A. C. (1993). Assessing the significance of focal activations using their spatial extent. *Human Brain Mapping*, 1(3):210–220. (Cited on page 25.)

Gavrilescu, M., Shaw, M. E., Stuart, G. W., Eckersley, P., Svalbe, I. D., and Egan, G. F. (2002). Simulation of the Effects of Global Normalization Procedures in Functional MRI. *NeuroImage*, 17(2):532–542. (Cited on page 40.)

Geffen, G., Moar, K. J., O'hanlon, A. P., Clark, C. R., and Geffen, L. B. (1990). Performance measures of 16– to 86-year-old males and females on the auditory verbal learning test. *Clinical Neuropsychologist*, 4(1):45–63. (Cited on page 115.)

Genovese, C. R., Lazar, N. A., and Nichols, T. (2002). Thresholding of Statistical Maps in Functional Neuroimaging Using the False Discovery Rate. *NeuroImage*, 15(4):870–878. (Cited on pages 29 and 31.)

Gerhard, S., Daducci, A., Lemkaddem, A., Meuli, R., Thiran, J.-P., and Hagmann, P. (2011). The Connectome Viewer Toolkit: An Open Source Framework to Manage, Analyze, and Visualize Connectomes. *Frontiers in Neuroinformatics*, 5(June):1–15. (Cited on page 145.)

Ghosh, S., Burns, C., Clark, D., Gorgolewski, K., Halchenko, Y., Madison, C., Tungaraza, R., and Millman, K. J. (2010). Nipype: Opensource platform for unified and replicable interaction with existing neuroimaging tools. In *16th Annual Meeting of the Organization for Human Brain Mapping*. (Cited on page 124.)

Goodman, D. F. M. and Brette, R. (2009). The brian simulator. *Frontiers in neuroscience*, 3(2):192–7. (Cited on page 145.)

Gorgolewski, K., Bastin, M., Rigolo, L., Soleiman, H. A., Pernet, C., Storkey, A., and Golby, A. J. (2011a). Pitfalls of Thresholding Statistical Maps in Presurgical fMRI Mapping. In *Proceedings 19th Scientific Meeting, International Society for Magnetic Resonance in Medicine*, Montreal, Canada. (Cited on pages 21 and 22.)

Gorgolewski, K., Burns, C. D., Madison, C., Clark, D., Halchenko, Y. O., Waskom, M. L., and Ghosh, S. S. (2011b). Nipype: A Flexible, Lightweight and Extensible Neuroimaging Data Processing Framework in Python. *Frontiers in Neuroinformatics*, 5(August):13. (Cited on pages 61, 124, and 126.)

Gorgolewski, K., Halchenko, Y., Notter, M., Varoquaux, G., Waskom, M., Ziegler, E., and Ghosh, S. (2012a). Nipype 2012: more packages, reusable workflows and reproducible science. In *18th Annual Meeting of the Organization for Human Brain Mapping*, Beijing, China. (Cited on page 124.)

Gorgolewski, K., Storkey, A., Bastin, M., and Pernet, C. (2011c). Using a Combination of a Mixture Model and Topological FDR in the Context of Presurgical Planning. In *17th Annual Meeting of the Organization for Human Brain Mapping*, Quebec, Canada. (Cited on page 27.)

Gorgolewski, K., Storkey, A., Bastin, M., and Pernet, C. (2012b). Reliability of single subject fMRI in the context of presurgical planning. In *18th Annual Meeting of the Organization for Human Brain Mapping*, Beijing, China. (Cited on page 52.)

Gorgolewski, K. J., Storkey, A. J., Bastin, M. E., and Pernet, C. (2012c). Adaptive thresholding for reliable topological inference in single subject fMRI analysis. In *ICML Workshop on Statistics, Machine Learning and Neuroscience*, Edinburgh, UK. (Cited on page 27.)

Gorgolewski, K. J., Storkey, A. J., Bastin, M. E., and Pernet, C. R. (2012d). Adaptive thresholding for reliable topological inference in single subject fMRI analysis. *Frontiers in Human Neuroscience*, 6(August):1–14. (Cited on page 27.)

Gorgolewski, K. J., Storkey, A. J., Bastin, M. E., Whittle, I., and Pernet, C. (2013). Single subject fMRI test-retest reliability metrics and confounding factors. *NeuroImage*, 69:231–243. (Cited on page 52.)

Goutte, C., Nielsen, F. a., and Hansen, L. K. (2000). Modeling the haemodynamic response in fMRI using smooth FIR filters. *IEEE transactions on medical imaging*, 19(12):1188–201. (Cited on page 150.)

Grice, K. O., Vogel, K. A., Le, V., Mitchell, A., Muniz, S., and Vollmer, M. A. (2003). Adult norms for a commercially available Nine Hole Peg Test for finger dexterity. *The American journal of occupational therapy. : official publication of the American Occupational Therapy Association*, 57(5):570–3. (Cited on page 115.)

Gurney, J., Smith, M., and Bunin, G. (1999). Chapter III: CNS and miscellaneous intracranial and intraspinal neoplasms. In Ries, L., Smith, M., Gurney, J., Linet, M., Tamra, T., JL Young, and Bunin, G., editors, *Cancer Incidence and Survival among Children and Adolescents: United States SEER Program 1975–1995*. National Cancer Institute, SEER Program, Bethesda, MD. (Cited on pages 3 and 4.)

Hå berg, A., Kvistad, K. A., Unsgå rd, G., and Haraldseth, O. (2004). Preoperative Blood Oxygen Level-dependent Functional Magnetic Resonance Imaging in Patients with Primary Brain Tumors: Clinical Application and Outcome. *Neurosurgery*, 54(4):902–915. (Cited on pages 20 and 23.)

Halchenko, Y. O. and Hanke, M. (2012). Open is Not Enough. Let's Take the Next Step: An Integrated, Community-Driven Computing Platform for Neuroscience. *Frontiers in Neuroinformatics*, 6(22). (Cited on page 135.)

Hall, D. A., Haggard, M. P., Akeroyd, M. A., Palmer, A. R., Summerfield, A. Q., Elliott, M. R., Gurney, E. M., and Bowtell, R. W. (1999). "Sparse" temporal sampling in auditory fMRI. *Human brain mapping*, 7(3):213–23. (Cited on pages 53, 54, and 111.)

Hall, W. a., Liu, H., and Truwit, C. L. (2005). Functional magnetic resonance imaging-guided resection of low-grade gliomas. *Surgical neurology*, 64(1):20–7; discussion 27. (Cited on page 20.)

Hanke, M., Halchenko, Y. O., Sederberg, P. B., Hanson, S. J., Haxby, J. V., and Pollmann, S. (2009). PyMVPA: A python toolbox for multivariate pattern analysis of fMRI data. *Neuroinformatics*, 7(1):37–53. (Cited on page 145.)

Harrell, F. and Davies, C. (1982). A new distribution-free quantile estimator. *Biometrika*, 69(3):635–640. (Cited on pages 39 and 63.)

Hartvig, N. V. v. and Jensen, J. L. (2000). Spatial mixture modeling of fMRI data. *Human Brain Mapping*, 11(4):233–248. (Cited on page 30.)

Hirsch, J., Ruge, M. I., Kim, K., Correa, D. D., Victor, J. D., Relkin, N. R., Labar, D. R., Krol, G., Bilsky, M. H., Souweidane, M. M., DeAngelis, L. M., and Gutin, P. H. (2000). An Integrated Functional Magnetic Resonance Imaging Procedure for Preoperative Mapping of Cortical Areas Associated with Tactile, Motor, Language, and Visual Functions. *Neurosurgery*, 47(3):711–722. (Cited on page 20.)

Holmes, A. and Friston, K. J. (1998). Generalisability, random effects & population inference. *NeuroImage*, 7:S754. (Cited on pages 62 and 92.)

Hömke, L. (2006). A multigrid method for anisotropic PDEs in elastic image registration. *Numerical Linear Algebra with Applications*, 13(2-3):215–229. (Cited on page 125.)

Huang, J., Carr, T., and Cao, Y. (2002). Comparing cortical activations for silent and overt speech using event-related fMRI. *Human Brain Mapping*, 15(1):39–53. (Cited on page 53.)

Huettel, S. A., Song, A. W., and McCarthy, G. (2008). *Functional Magnetic Resonance Imaging, Second Edition*. Sinauer Associates, Sunderland, MA, USA, second edition. (Cited on pages 8 and 15.)

Inskip, P. D., Linet, M. S., and Heineman, E. F. (1995). Etiology of brain tumors in adults. *Epidemiologic reviews*, 17(2):382–414. (Cited on page 7.)

Jack, C., Thompson, R., Butts, R., Sharbrough, F., Kelly, P., Hanson, D., Riederer, S., Ehman, R., Hangiandreou, N., and Cascino, G. (1994). Sensory motor cortex: correlation of presurgical mapping with functional MR imaging and invasive cortical mapping. *Radiology*, 190(1):85. (Cited on page 19.)

Junghöfer, M., Schupp, H. T., Stark, R., and Vaitl, D. (2005). Neuroimaging of emotion: empirical effects of proportional global signal scaling in fMRI data analysis. *NeuroImage*, 25(2):520–6. (Cited on page 40.)

Kiebel, S., Poline, J., Friston, K., Holmes, A., and KJ (1999). Robust smoothness estimation in statistical parametric maps using standardized residuals from the general linear model. *NeuroImage*, 10(6):756–66. (Cited on pages 29 and 149.)

Kirson, D., Huk, A. C., and Cormack, L. K. (2008). Quantifying spatial uncertainty of visual area boundaries in neuroimaging data. *Journal of Vision*, 8(10):1–15. (Cited on page 151.)

Klein, A., Andersson, J., Ardekani, B. a., Ashburner, J., Avants, B., Chiang, M.-C., Christensen, G. E., Collins, D. L., Gee, J., Hellier, P., Song, J. H., Jenkinson, M., Lepage, C., Rueckert, D., Thompson, P., Vercauteren, T., Woods, R. P., Mann, J. J., and Parsey, R. V. (2009). Evaluation of 14 nonlinear deformation algorithms applied to human brain MRI registration. *NeuroImage*, 46(3):786–802. (Cited on page 146.)

Klein, A., Ghosh, S. S., Avants, B., Yeo, B. T. T., Fischl, B., Ardekani, B., Gee, J. C., Mann, J. J., and Parsey, R. V. (2010). Evaluation of volume-based and surface-based brain image registration methods. *NeuroImage*, 51(1):214–220. (Cited on pages 125 and 146.)

Kriegeskorte, N., Simmons, W. K., Bellgowan, P. S. F., and Baker, C. I. (2009). Circular analysis in systems neuroscience: the dangers of double dipping. *Nature neuroscience*, 12(5):535–40. (Cited on page 149.)

Krings, T., Foltys, H., Reinges, M. H., Kemeny, S., Rohde, V., Spetzger, U., Gilsbach, J. M., and Thron, A. (2001a). Navigated transcranial magnetic stimulation for presurgical planning–correlation with functional MRI. *Minimally invasive neurosurgery : MIN*, 44(4):234–9. (Cited on pages 13 and 150.)

Krings, T., Reinges, M. H. T., Erberich, S., Kemeny, S., Rohde, V., Spetzger, U., Korinth, M., Willmes, K., Gilsbach, J. M., and Thron, A. (2001b). Functional MRI for presurgical planning: problems, artefacts, and solution strategies. *Journal of neurology, neurosurgery, and psychiatry*, 70(6):749–60. (Cited on page 20.)

Krishnan, R., Raabe, A., Hattingen, E., Szelényi, A., Yahya, H., Hermann, E., Zimmermann, M., and Seifert, V. (2004). Functional Magnetic Resonance Imaging-integrated Neuronavigation: Correlation between Lesion-to-Motor Cortex Distance and Outcome. *Neurosurgery*, 55(4):904–915. (Cited on page 20.)

Lancaster, J., Tordesillas-Gutiérrez, D., and M (2007). Bias between MNI and Talairach coordinates analyzed using the ICBM-152 brain template. *Human brain*, 28(11):1194–1205. (Cited on page 63.)

Lancaster, J., Woldorff, M., and Parsons, L. (2000). Automated Talairach atlas labels for functional brain mapping. *Brain Mapping*, 10(3):120–131. (Cited on page 63.)

Lee, C. C., Ward, H. a., Sharbrough, F. W., Meyer, F. B., Marsh, W. R., Raffel, C., So, E. L., Cascino, G. D., Shin, C., Xu, Y., Riederer, S. J., and Jack, C. R. (1999). Assessment of functional MR imaging in neurosurgical planning. *American Journal of Neuroradiology*, 20(8):1511–9. (Cited on page 20.)

Lehéricy, S., Duffau, H., Cornu, P., Capelle, L., Pidoux, B., Carpentier, a., Auliac, S., Clemenceau, S., Sichez, J. P., Bitar, a., Valery, C. a., Van Effenterre, R., Faillot, T., Srour, a., Fohanno, D., Philippon, J., Le Bihan, D., and Marsault, C. (2000). Correspondence between functional magnetic resonance imaging somatotopy and individual brain anatomy of the central region: comparison with intraoperative stimulation in patients with brain tumors. *Journal of neurosurgery*, 92(4):589–98. (Cited on page 20.)

Lindeman, R. H., Merenda, P. F., and Gold, R. Z. (1980). *Introduction to bivariate and multivariate analysis*. Scott, Foresman, and Company, Glenview, IL. (Cited on page 95.)

Logothetis, N., Pauls, J., and Augath, M. (2001). Neurophysiological investigation of the basis of the fMRI signal. *Nature*, 412(6843):150–157. (Cited on page 17.)

Loonstra, a. S., Tarlow, a. R., and Sellers, a. H. (2001). COWAT metanorms across age, education, and gender. *Applied neuropsychology*, 8(3):161–6. (Cited on page 115.)

MacKay, D. (2003). *Information theory, inference, and learning algorithms*. Cambridge Univ Press, Cambridge. (Cited on page 146.)

Mackenzie-Graham, A. J., Van Horn, J. D., Woods, R. P., Crawford, K. L., and Toga, A. W. (2008). Provenance in neuroimaging. *NeuroImage*, 42(1):178–95. (Cited on pages 126 and 151.)

Majos, A., Tybor, K., Stefańczyk, L., and Góraj, B. (2005). Cortical mapping by functional magnetic resonance imaging in patients with brain tumors. *European radiology*, 15(6):1148–58. (Cited on page 20.)

Mathiowetz, V., Weber, K., Kashman, N., and Volland, G. (1985). Adult Norms for the nine hole peg test of finger dexterity. *The Occupational Therapy Journal of Research*, 5(1):24–27. (Cited on page 115.)

McAuliffe, M., Lalonde, F., McGarry, D., Gandler, W., Csaky, K., and Trus, B. (2001). Medical Image Processing, Analysis and Visualization in clinical research. *Proceedings 14th IEEE Symposium on Computer-Based Medical Systems.*, pages 381–386. (Cited on page 126.)

McGraw, K. O. and Wong, S. P. (1996). Forming inferences about some intraclass correlation coefficients. *Psychological Methods*, 1(1):30–46. (Cited on page 93.)

McKean-Cowdin, R., Preston-Martin, S., Pogoda, J. M., Holly, E. A., Mueller, B. A., and Davis, R. L. (1998). Parental occupation and childhood brain tumors: astroglial and primitive neuroectodermal tumors. *Journal of occupational and environmental medicine*, 40(4):332–40. (Cited on page 6.)

Millman, K. and Aivazis, M. (2011). Python for Scientists and Engineers. *Computing in Science & Engineering*, 13(2):9–12. (Cited on page 145.)

Morosan, P., Rademacher, J., Schleicher, a., Amunts, K., Schormann, T., and Zilles, K. (2001). Human primary auditory cortex: cytoarchitectonic subdivisions and mapping into a spatial reference system. *NeuroImage*, 13(4):684–701. (Cited on page 66.)

Mumford, J. a. and Nichols, T. (2009). Simple group fMRI modeling and inference. *NeuroImage*, 47(4):1469–75. (Cited on page 150.)

Murphy, K., Birn, R. M., Handwerker, D. A., Jones, T. B., and Bandettini, P. A. (2009). The impact of global signal regression on resting state correlations: are anti-correlated networks introduced? *NeuroImage*, 44(3):893–905. (Cited on page 40.)

Nichols, T. E. (2012). Multiple testing corrections, nonparametric methods, and random field theory. *NeuroImage*, 62(2):811–5. (Cited on page 22.)

O'Donnell, L., Rigolo, L., Norton, I., Westin, C., and Golby, A. (2011). fMRI-DTI modeling via landmark distance atlases for prediction and detection of fiber tracts. *NeuroImage*, 60(1):456–470. (Cited on pages 27 and 122.)

Ogawa, S., Lee, T. M., Kay, a. R., and Tank, D. W. (1990). Brain magnetic resonance imaging with contrast dependent on blood oxygenation. *Proceedings of the National Academy of Sciences of the United States of America*, 87(24):9868–72. (Cited on pages 16 and 19.)

Ogawa, S. and Sung, Y.-W. (2007). Functional magnetic resonance imaging. *Scholarpedia*, 2(10):3105. (Cited on page 19.)

Ohgaki, H. (2009). Epidemiology of brain tumors. *Methods in molecular biology (Clifton, N.J.)*, 472:323–42. (Cited on pages 5 and 56.)

Ohgaki, H. and Kleihues, P. (2005). Epidemiology and etiology of gliomas. *Acta neuropathologica*, 109(1):93–108. (Cited on pages 5 and 7.)

Oinn, T., Greenwood, M., Addis, M., Alpdemir, M. N., Ferris, J., Glover, K., Goble, C., Goderis, A., Hull, D., Marvin, D., Li, P., Lord, P., Pocock, M. R., Senger, M., Stevens, R., Wipat, A., and Wroe, C. (2006). Taverna: lessons in creating a workflow environment for the life sciences. *Concurrency and Computation: Practice and Experience*, 18(10):1067–1100. (Cited on page 126.)

O'Shea, J. P., Whalen, S., Branco, D. M., Petrovich, N. M., Knierim, K. E., and Golby, A. J. (2006). Integrated image- and function-guided surgery in eloquent cortex: a technique report. *The international journal of medical robotics and computer assisted surgery*, 2(1):75–83. (Cited on page 122.)

Patchell, R. A., Tibbs, P. A., Walsh, J. W., Dempsey, R. J., Maruyama, Y., Kryscio, R. J., Markesbery, W. R., Macdonald, J. S., and Young, B. (1990). A randomized trial of surgery in the treatment of single metastases to the brain. *The New England journal of medicine*, 322(8):494–500. (Cited on page 12.)

Pauling, L. and Coryell, C. D. (1936). The Magnetic Properties and Structure of Hemoglobin, Oxyhemoglobin and Carbonmonoxyhemoglobin. *Proceedings of the National Academy of Sciences*, 22(4):210–216. (Cited on page 16.)

Pellerin, L. and Magistretti, P. J. (1994). Glutamate uptake into astrocytes stimulates aerobic glycolysis: a mechanism coupling neuronal activity to glucose utilization. *Proceedings of the National Academy of Sciences of the United States of America*, 91(22):10625–9. (Cited on page 17.)

Pendse, G., Borsook, D., and Becerra, L. (2009). Enhanced false discovery rate using Gaussian mixture models for thresholding fMRI statistical maps. *NeuroImage*, 47(1):231–261. (Cited on pages 30, 31, 36, and 49.)

Penfield, W. and Boldrey, E. (1937). Somatic motor and sensory representation in the cerebral cortex of man as studied by electrical stimulation. *Brain*, 60(4):389–443. (Cited on page 13.)

Perez, F., Granger, B. E., and Hunter, J. D. (2011). Python: An Ecosystem for Scientific Computing. *Computing in Science & Engineering*, 13(2):13–21. (Cited on page 145.)

Petersson, K. M., Nichols, T. E., Poline, J. B., and Holmes, a. P. (1999). Statistical limitations in functional neuroimaging. II. Signal detection and statistical inference. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, 354(1387):1261–81. (Cited on pages 30 and 149.)

Petrovich, N., Holodny, A. I., Tabar, V., Correa, D. D., Hirsch, J., Gutin, P. H., and Brennan, C. W. (2005). Discordance between functional magnetic resonance imaging during silent speech tasks and intraoperative speech arrest. *Journal of neurosurgery*, 103(2):267–74. (Cited on page 54.)

Phelps, E. a., Hyder, F., Blamire, a. M., and Shulman, R. G. (1997). FMRI of the prefrontal cortex during overt verbal fluency. *Neuroreport*, 8(2):561–5. (Cited on page 53.)

Poldrack, R. A., Mumford, J. A., and Nichols, T. E. (2011). *Handbook of Functional MRI Data Analysis*. Cambridge University Press. (Cited on page 19.)

Raemaekers, M., du Plessis, S., Ramsey, N., Wuesten, J., and Vink, M. (2012). Test retest variability underlying fMRI measurements. *NeuroImage*, 60(1):717–27. (Cited on pages 78, 108, and 110.)

Raemaekers, M., Vink, M., Zandbelt, B., van Wezel, R. J. a., Kahn, R. S., and Ramsey, N. F. (2007). Test-retest reliability of fMRI activation during prosaccades and antisaccades. *NeuroImage*, 36(3):532–42. (Cited on pages 93 and 109.)

Raschle, N. M., Zuk, J., and Gaab, N. (2012). Functional characteristics of developmental dyslexia in left-hemispheric posterior brain regions predate reading onset. *Proceedings of the National Academy of Sciences of the United States of America*, 109(6):2156–61. (Cited on page 92.)

Rex, D. E., Ma, J. Q., and Toga, A. W. (2003). The LONI Pipeline Processing Environment. *NeuroImage*, 19(3):1033–1048. (Cited on page 126.)

Rey, A. (1941). L'examen psychologique dans les cas d'encéphalopathie traumatique. *Archives de psychologie*, 28:215–285. (Cited on page 115.)

Roche, A., Malandain, G., Pennec, X., and Ayache, N. (1998). The correlation ratio as a new similarity measure for multimodal image registration. *Lecture Notes in Computer Science*, 1496:1115. (Cited on page 96.)

Rodvall, Y., Ahlbom, A., Pershagen, G., Nylander, M., and Spännare, B. (1998). Dental radiography after age 25 years, amalgam fillings and tumours of the central nervous system. *Oral Oncology*, 34(4):265–269. (Cited on page 7.)

Rola, R., Raber, J., Rizk, A., Otsuka, S., VandenBerg, S. R., Morhardt, D. R., and Fike, J. R. (2004). Radiation-induced impairment of hippocampal neurogenesis is associated with cognitive deficits in young mice. *Experimental neurology*, 188(2):316–30. (Cited on page 11.)

Ron, E., Modan, B., Boice, J. D., Alfandary, E., Stovall, M., Chetrit, A., and Katz, L. (1988). Tumors of the brain and nervous system after radiotherapy in childhood. *The New England journal of medicine*, 319(16):1033–9. (Cited on page 7.)

Rousselet, G. A. and Pernet, C. R. (2012). Improving standards in brain-behavior correlation analyses. *Frontiers in Human Neuroscience*, 6(May). (Cited on page 96.)

Sartorius, C. J. and Wright, G. (1997). Intraoperative brain mapping in a community setting - Technical considerations. *Surgical neurology*, 47(4). (Cited on page 21.)

Savoy, R. L. (2005). Experimental design in brain activation MRI: cautionary tales. *Brain research bulletin*, 67(5):361–7. (Cited on page 109.)

Schlosser, M. J., McCarthy, G., Fulbright, R. K., Gore, J. C., and Awad, I. A. (1997). Cerebral vascular malformations adjacent to sensorimotor and visual cortex. *Stroke*, 28(6):1130–1137. (Cited on page 20.)

Schmidt, C. F., Zaehle, T., Meyer, M., Geiser, E., Boesiger, P., and Jancke, L. (2008). Silent and continuous fMRI scanning differentially modulate activation in an auditory language comprehension task. *Human brain mapping*, 29(1):46–56. (Cited on page 54.)

Schmidt, M. (1996). Rey auditory verbal learning test: A handbook. *Los Angeles, CA: Western Psychological Services*. (Cited on page 115.)

Schwartz, Y., Barbot, A., Thyreau, B., Frouin, V., Varoquaux, G., Siram, A., Marcus, D. S., and Poline, J.-B. (2012). PyXNAT: XNAT in Python. *Frontiers in Neuroinformatics*, 6(May):1–11. (Cited on page 145.)

Shrout, P. E. and Fleiss, J. L. (1979). Intraclass correlations: uses in assessing rater reliability. *Psychological bulletin*, 86(2):420–8. (Cited on pages 64 and 93.)

Smith, S. M. (2002). Fast robust automated brain extraction. *Human brain mapping*, 17(3):143–55. (Cited on page 62.)

Smith, S. M., Beckmann, C. F., Ramnani, N., Woolrich, M. W., Bannister, P. R., Jenkinson, M., Matthews, P. M., and McGonigle, D. J. (2005). Variability in fMRI: A re-examination of inter-session differences. *Human Brain Mapping*, 24(3):248–257. (Cited on pages 17, 49, 92, and 110.)

Stippich, C., editor (2007). *Clinical functional MRI: presurgical functional neuroimaging*. Springer, Berlin. (Cited on pages 54 and 150.)

Stippich, C., Blatow, M., and Krakow, K. (2007). Presurgical Functional MRI in Patients with Brain Tumours. In Stippich, C., editor, *Clinical Functional MRI*, chapter 4, pages 87–134. Springer, Berlin. (Cited on page 27.)

Strother, S. C., Anderson, J., Hansen, L. K., Kjems, U., Kustra, R., Sidtis, J., Frutiger, S., Muley, S., LaConte, S., and Rottenberg, D. (2002). The quantitative evaluation of functional neuroimaging experiments: the NPAIRS data analysis framework. *NeuroImage*, 15(4):747–71. (Cited on page 150.)

Talairach, J. and Tournoux, P. (1988). *Co-planar stereotaxic atlas of the human brain: 3-dimensional proportional system: an approach to cerebral imaging*. Thieme. (Cited on page 63.)

Tarnaris, A., O'Brien, C., and Redfern, R. M. (2006). Ganglioglioma with anaplastic recurrence of the neuronal element following radiotherapy. *Clinical neurology and neurosurgery*, 108(8):761–7. (Cited on page 10.)

Taylor, P. (2010). An introduction to intraclass correlation that resolves some common confusions. (Cited on page 82.)

Turkheimer, F. E., Aston, J. A., and Cunningham, V. J. (2004). On the logic of hypothesis testing in functional imaging. *European Journal of Nuclear Medicine and Molecular Imaging*, 31(5):725–732. (Cited on page 30.)

Ulrike Grömping (2006). Relative importance for linear regression in R: the package relaimpo. *Journal of Statistical Software*, 17(1). (Cited on page 95.)

Voyvodic, J. T. (2006). Activation mapping as a percentage of local excitation: fMRI stability within scans, between scans and across field strengths. *Magnetic resonance imaging*, 24(9):1249–61. (Cited on page 50.)

Voyvodic, J. T. (2012). Reproducibility of single-subject fMRI language mapping with AMPLE normalization. *Journal of magnetic resonance imaging*, 36(3):569–80. (Cited on pages 50 and 149.)

Voyvodic, J. T., Petrella, J. R., and Friedman, A. H. (2009). fMRI Activation Mapping as a Percentage of Local Excitation : Consistent Presurgical Motor Maps Without Threshold Adjustment. *Journal of Magnetic Resonance Imaging*, 759:751–759. (Cited on page 50.)

Wilcox, R. (2005). *Introduction to robust estimation and hypothesis testing*. Academic Press, Burlington, MA, USA, 2nd edition. (Cited on page 112.)

Wilke, M. (2012). An alternative approach towards assessing and accounting for individual motion in fMRI timeseries. *NeuroImage*, 59(3):2062–72. (Cited on page 96.)

Wong, O. and Harris, F. (2000). Cancer mortality study of employees at lead battery plants and lead smelters, 1947-1995. *American journal of industrial medicine*, 38(3):255–70. (Cited on page 6.)

Woodward, D. E., Cook, J., Tracqui, P., Cruywagen, G. C., Murray, J. D., and Alvord, E. C. (1996). A mathematical model of glioma growth: the effect of extent of surgical resection. *Cell Proliferation*, 29(6):269–288. (Cited on page 12.)

Woolrich, M., Behrens, T., Beckmann, C., and Smith, S. (2005). Mixture models with adaptive spatial regularization for segmentation with an application to FMRI data. *IEEE Transactions on Medical Imaging*, 24(1):1–11. (Cited on pages 23, 30, 31, 35, 36, 49, and 149.)

Worsley, K. J., Evans, a. C., Marrett, S., and Neelin, P. (1992). A three-dimensional statistical analysis for CBF activation studies in human brain. *Journal of Cerebral Blood Flow & Metabolism*, 12(6):900–18. (Cited on page 29.)

Yetkin, F. Z., Hammeke, T. a., Swanson, S. J., Morris, G. L., Mueller, W. M., McAuliffe, T. L., and Haughton, V. M. (1995). A comparison of functional MR activation patterns during silent and audible language tasks. *American Journal of Neuroradiology*, 16(5):1087–92. (Cited on page 53.)

Zelle, J. (1999). Python as a first language. In *Proceedings of 13th Annual Midwest Computer Conference*, volume 2. (Cited on page 145.)

COLOPHON

This document was typeset using the typographical look-and-feel `classicthesis` developed by André Miede. The style was inspired by Robert Bringhurst's seminal book on typography "*The Elements of Typographic Style*". `classicthesis` is available for both LaTeX and LyX:

http://code.google.com/p/classicthesis/

Happy users of `classicthesis` usually send a real postcard to the author, a collection of postcards received so far is featured here:

http://postcards.miede.de/

Final Version as of March 26, 2013 (`classicthesis` version 4.0).