Thesis Introduction

Worldwide, radiology continues to evolve. Not only do imaging techniques advance and become more sophisticated, but factors affecting human health change with every decade. The continued advancement of medical images (their acquisition and interpretation) puts a strain on medical specialists, even before individual patient needs are considered. Factors that influence the ability of the reader to deliver patient needs depend on not only the image, but also the readers' level of experience and expertise. Medical image acquisition, accuracy and interpretation have a hugely important role to play in patient safety.

In neurology, referred patients are most frequently sent for computed tomography (CT) and magnetic resonance (MR) imaging of the brain to shed light on the origin and impact of disease. Whilst most observer performance studies focus on screening and detection of a few abnormalities in a non-diseased population, the diseases of 'older-age' are often neglected and treated reactively i.e. when a multitude of signs and symptoms appear, not necessarily as a preventive measure. Owing to the difficulty of measuring performance and the nature of expert interpretation when the technology itself is changing; neuroradiology has not been considered extensively from an observer performance perspective and studies concerning visual search in this area are very thin on the ground.

Stroke is the focus of inquiry here for many reasons, but predominantly because urgent imaging of patients with quick feedback of image findings can reduce disability and save lives. If a further 10% of acute stroke patients received thrombolytic therapy within 3 hours of onset, over 1,000 people would regain independence per annum rather than rapidly deteriorate (DoH, 2006). Once treatment is administered and followed up effectively, patients can benefit from a further 5-10 years of life (Indredavik, 1999; DoH, 2007). The benefits of further treatment within this population are known. What isn't known is whether experts make errors of judgement within this clinical area, even if access to healthcare is increased.

Thesis and Chapter Arrangement

The aims and objectives of the research have been explored in-depth in this chapter and the overriding theme within this thesis is to gain an in-depth appreciation of how radiologists perform their duties within the field of neuroradiology and specifically, in stroke detection. In order to achieve these aims the following chapters have been arranged;

- **Chapter 1** provides an introduction to the thesis and goes onto provide an in-depth exploration and review of the background literature. This chapter justifies the present research, considers the specific aims and objectives, and sets the stage for the experimental studies that follow.

- **Chapter 2** provides an overview study design and implementation, including an overarching framework for the quantitative analysis that follows in experimental chapters 3, 4, 5, 6, 7, and 8.

- **Chapter 3** explores the feasibility study, which formed the basis of studies 1, 2, 3, 4 and 5.

- **Chapter 4** assesses observer performance in stroke interpretation and the influence of experience in multidimensional Computed Tomography imaging (study 1).

- **Chapter 5** assesses observer performance in stroke interpretation and the influence of experience in multidimensional Magnetic Resonance Imaging (study 2).

- **Chapter 6** constitutes a direct comparison of CT and MRI multidimensional imaging to assess whether modality preference enhances observer performance (study 3).

- **Chapter 7** is a follow-on study (4), which aims to assess how radiology consultant performance compares with a matched sample of Neurologists i.e. when level of clinical knowledge and experience is controlled, what is the impact on visual search and observer performance between disciplines?

- **Chapter 8** provides an insight into how inexperienced through to experienced readers assess patient clinical histories prior to image review; an exploration of how clinical information is appraised through eye movement analysis (study 5).

- **Chapter 9** explores the key findings, limitations, recommendations and implications that have arisen from the present work and discusses the overall contribution of knowledge for academic research, and clinical/ patient populations.

Literature Review

1.1 The Cultural Context

Images of the human musculoskeletal system have been accessible since Roëntgen's x-ray discovery in 1895, when he first visualised the bones in his wife's hand. Since then, cathode rays have been used with great success in the medical field to diagnose bone fractures, skeletal abnormalities and signs of disease. As imaging techniques advanced, a surge of research began in the 1940's, investigating the visual acuity of the eye (DeVries, 1943), the detection of pulmonary tuberculosis, and the development of multimodal imaging techniques to ascertain which techniques were superior at the time (Birkelo, 1947). Around this period of technological advancement, psychology and physics began to merge to examine perception of images, contrast rendition, resolution and how best to optimise the image output by altering radiation dose (Kundel, 2006).

Following on from this period of intense work, inter-rater reliability and diagnostic errors in the interpretation of medical images came to the fore of perception research; it was uncovered that interpretation errors not only existed between observers, but within observer judgement also, which was not originally anticipated by either radiologists or researchers. Statistical methods were developed to quantify observer performance and the "Rose-De Vries" Model was formulated to investigate psychophysiological aspects of vision. Signal-to-Noise Ratio (SNR) properties, including sensitivity (the ability to detect an abnormality in non-diseased populations e.g. screening) and specificity (the ability to rule out an abnormality in diseased populations) came to the fore of observer and image comparisons (De Vries, 1943; Rose, 1948).

Cognitive and perceptual psychological investigations teamed up with the physical sciences over many years as imaging methods advanced to include Computed Tomography (CT), Positron Emission Tomography (PET) and Magnetic Resonance Imaging (MRI) in the 1970's, allowing the investigation and reporting of medical errors, observer interpretation and decision-making. During this period of research activity in the 1980's, psychophysics emerged as a distinct discipline (Kundel, 2006). Over a decade later and medical image perception became its own sub discipline that fused multiple methods of scientific enquiry from differing research perspectives; cognitive, perceptual, psychological, physical, medical, radiological, computer and engineering sciences, all in pursuit of a common goal; to encourage interdisciplinary working when investigating radiological performance.

1.2 The Clinical Context: The National Health Service and Our Nation's Health

As Medical Image Perception research emerged and evolved in the early 1940's, so did the National Health Service in the UK. Prompted by the demand of the Second World War to treat injured soldiers, the health service became an imperative requirement, responding to national need for a structured and tax-funded provider of healthcare. The NHS was established on the 5th July 1948 and during its 62-year evolution to-date it has witnessed many operational shifts, particularly following the industrial revolution and advancements in medical knowledge, treatment and technology (Rivett, 1998).

Since the NHS began, new diseases and conditions have challenged our health, impacting upon treatment arms and healthcare practice. With an ageing population, neurodegenerative (e.g. Alzheimers' Disease and Dementia) and cardiovascular diseases (e.g. Myocardial Infarction and stroke) have now become more prevalent, affecting older adult well-being and shared resources. Due to lifestyle change in recent decades, many of these conditions were not present when the health service began operating and thus the service provider is continually in arrears when responding to increasing demand for neurology and surgical services. It is also essential for medical services to rule out the presence of disease when symptoms or manifestations indicative of malignant disease are present i.e. in benign tumours. Medical research too is always trying to offer the best treatments with optimum efficacy and minimum side-effects.

1.3 Neurology, Neurological Conditions and Service Delivery Today

Clinical Neurology is the medical specialty concerned with the diagnosis, treatment and, when appropriate, the continuing assessment and care of patients with diseases of the central and peripheral nervous system. In some cases, neurology is also responsible for muscular function. (ABN, 1996). Neurological conditions occur as a direct result of disease or damage to the nervous system, originating from an organic or functional cause i.e. they arise from within an individual (e.g. predisposition or heredity), as a result of an external influence (e.g. car accident), or a combination of the two i.e. lifestyle factors interacting with an innate predisposition. This group of disorders and their related symptoms can be the result of a 'one-off' clinical event (e.g. stroke), or a permanent, intermittent and/ or degenerative disease such as Dementia, Multiple Sclerosis and/ or Huntingdon's disease. They are not confined to a specific age or socioeconomic group of individuals, although

lifestyle factors and socioeconomic status have been strongly linked with vascular complaints (ABN, 1996; DoH, 2005).

The symptoms associated with these conditions affect people in many debilitating ways and can be as complex, individual and varied as neurological abilities. There are commonalities with symptoms affecting motor, sensory, cognitive and/ or communication abilities and individuals can be affected by one symptom or a combination simultaneously (DoH, 2005). By the year 2005, the Department of Health reported that approximately 10 million people across the UK had suffered or were suffering from a neurological condition, accounting for 20% of acute hospital admissions, with 350,000 people requiring daily living assistance and over 850,000 people providing care for those affected (NHS, 2005).

The Action on Neurology (NHS, 2005) paper reports that between one in six-eight consultations in primary care and one in five emergency admissions are neurological in origin. In 1999/ 2000 there were 195,700 outpatient referrals, and 134,300 more in 2003/4. As patient numbers increase but resources remain constant, a dissonance emerges between staff capability, access to treatment and available resources when meeting demand. Furthermore, neurological conditions can be complicated and time consuming to diagnose, requiring a combination of tests to uncover the root cause of a myriad of interconnected symptoms and functional problems. Disorders are diagnosed on the basis of previous medical history, family history, blood and cerebrospinal fluid (CSF) examinations as well as Electroencephalography (EEG), Electromyography (EMG), nerve conduction studies and multiple imaging investigations.

Whilst some neurological problems can be treated, rehabilitated or prevented, others are intermittent, progressive (e.g. Muscular Sclerosis) and/ or terminal (e.g. Motor Neuron Disease or stroke), requiring high quality palliative care to ensure the patient remains free from pain and discomfort in the final stages of life. Factors that influence time elapsed from first point of contact with the healthcare system, through to diagnosis and intervention include; the complexity of neurological complaints, staffing levels and resources, patient involvement, physician expertise, communication and team work, and the organisation of healthcare services, including imaging investigations.

1.4 Clinical Radiology: Why Imaging is Essential in Neurological Disorders

Imaging plays a key role in the patient treatment pathway. Without radiology, physicians would be treating patients 'blind' without key diagnostic information regarding their internal functioning. In the UK, Radiology is an information-rich specialty, which is almost completely digital producing varying 2D, 3D and 4D images with images being inspected on very high resolution computer monitors (rather than displays on illuminated light boxes) in dedicated reading environments. Digital imaging procedures available today include ultrasound, unenhanced or enhanced Computed Tomography (CT), Positron Emission Tomography (PET), Magnetic Resonance Imaging (MRI) and Diffusion-Weighted MRI (DWI).

Radiographic images are produced to examine the human musculoskeletal system as well as organ and metabolic function. Both CT and MRI modalities are frequently used in the prevention, identification, diagnosis and treatment of people who have suffered a neurological deficit. Anatomical changes and characteristics of neurological problems can be highly variable between individuals, posing different accuracy and interpretation problems for the observer.

1.4.1 Computed Tomography and Positron Emission Tomography

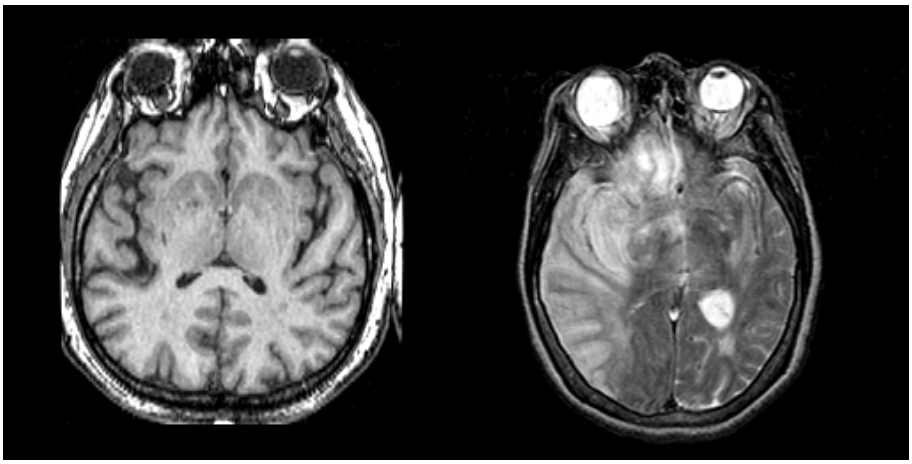
Computed Tomography is a form of tomography (sectional imaging) in which a computer controls the motion of the X-ray source over the body. Detectors process the data, and produce the required image via computer processing. CT is mostly applied to head, chest, cardiac and abdominal imaging and Positron Emission Tomography (PET) is frequently used to supplement CT imaging, particularly in the realm of oncology (Sureshbabu, 2005), and as seen in figure 1.1 below. PET detects metabolic changes associated with disease progression e.g. tumour development over time and can also be used to assess drug treatment efficacy (Herzog, 2007).

CT has widespread accessibility throughout western healthcare system and allows accurate visualisation of a number of conditions, predominantly after a few hours onset. However, the procedure exposes patients to radiation and many experience adverse reactions to contrast dye and subtle findings can be difficult to identify. CT provides quick, key and accessible information, facilitating a triage of treatments.

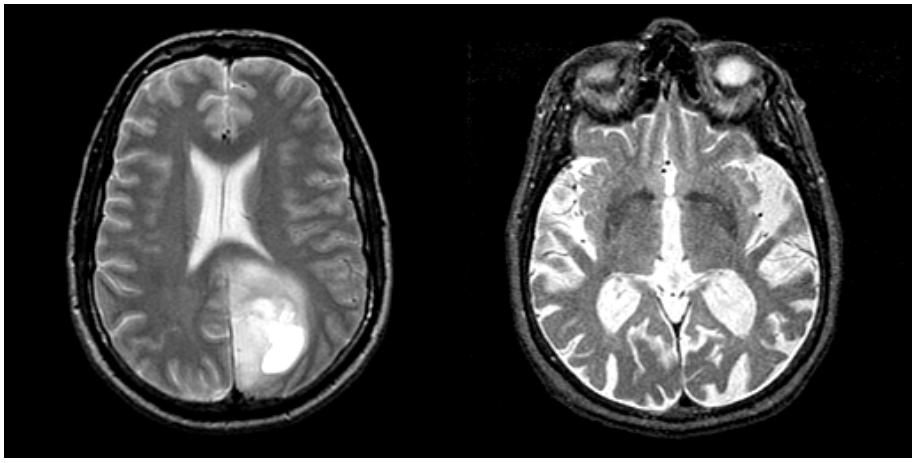
Figure 1.1 A whole-body scan (left with reversed grey scale for better display) reveals a metastasis in the abdomen (dotted line). The right-hand side shows transverse CT and PET images through the image slices. The PET/ CT overlay allows the exact localisation of the metastasis.

1.4.2 Magnetic Resonance Imaging

MR imaging systems use nuclear magnetic resonance, specifically proton movement, to produce images of water-based tissues (e.g. organs, muscle mass, adipose tissue, ligaments, and abnormalities including tumour growth and structural deficits) highlighting anatomical structure and function. MRI informs intravenous or catheterisation access and when the weighting of the image is inverted, MRI allows excellent visualisation of musculoskeletal anatomy also. The technique permits good visualisation of a large number of neurological abnormalities e.g. vascular disease and degenerative diseases such as Alzheimer's. The following figures highlight the differences in disease presentation;



Figures 1.2 & 1.3 demonstrate a healthy brain and a haemorrhagic stroke.



Figures 1.4 & 1.5 demonstrate the presence of a Glioma (tumour) and the generalised atrophy characteristic of Alzheimers' Disease.

1.5 What Clinical Condition is of Importance Here?

In the present thesis, stroke will be the focus of enquiry owing to its worldwide prevalence, the need for timely and accurate diagnosis and the pivotal role of radiology in detection and differentiation of stroke types to allow subsequent treatment. The importance of stroke detection will be described in the following sections;

1.5.1 Cardiovascular Disease and stroke Prevalence

Cardiovascular disease is a collection of problems relating to the heart, blood vessels and circulation, which encompasses myocardial infarction and stroke. Whilst heart failure is dealt with by cardiologists, stroke is treated within the neurology department. Worldwide, cardiovascular disease

(CVD) is a major health concern. By 2006, cardiovascular disease was reported to have claimed more lives than AIDS, TB and Malaria combined (DoH, 2006). In the UK alone, recent figures state that cardiovascular disease currently kills more people than all cancers combined, with stroke affecting 150,000 people per annum (DoH, 2006). When considered separately from other CVDs, stroke is the third leading cause of death in the US and UK, and the largest single cause of severe disability. Severe strokes leave 20% of individuals institutionalized and permanently disabled in a further 15-30%. 50-70% of stroke sufferers recover and can lead relatively normal lives, dependent upon the stroke severity and area of the brain affected (Asplund, Stegmayr & Peltonen, 1998), although depression and affiliated mental health problems can still cause suffering, even in the absence of severe disability (DoH, 2007).

In the US, stroke affects 700,000 people per annum and in 2007 the nationwide cost of stroke had risen to $62.7 billion (American Heart Association, 2007). In the UK, stroke costs the economy £7 billion per annum (DoH, 2007), which includes the financial loss from those out of work due to disability. Therefore, timely and accurate diagnosis is imperative when providing an efficient and efficacious radiology service and managing patient care and rehabilitation. Due to the adverse nature of patient outcomes and affiliated direct and indirect costs associated with stroke, a well run service is particularly important to provide quick identification, feedback to physician and treatment, enabling a reduction in the potential impact stroke can have on an individual (Gonzalez, 1999; Mullins, 2002; Kloska 2004; DoH, 2007).

1.5.2 What is a stroke? Definitions and Classifications

A stroke is a cerebrovascular incident or ‘attack’ on the brain which affects neurological, cognitive and physical functioning. There are two types of stroke: ischaemic and haemorrhagic. An ischaemic stroke occurs when an arterial vessel becomes blocked or compromised, by the build-up of atheroma or clot, preventing the adjoining region of the brain from receiving adequate blood supply and oxygen. A haemorrhagic stroke is caused by a burst blood vessel, which leaks blood into the surrounding cortex, causing damage and disturbing brain function (DoH, 2008).

A stroke can affect an individual in many different ways dependent upon the location and size of infarct. The most common symptoms are speech, vision, sensation, mobility and/or cognitive impairments. Unlike some neurological complaints, cerebrovascular disease is largely preventable and not necessarily a disease of ‘old age’. The main risk factors or precursors of stroke stem from lifestyle factors such as smoking, alcohol consumption and hypertension as a result of, or

exacerbated by, a lack of exercise, too much salt and saturated fat within our diets. A genetic or ethnic predisposition also enhance CVD likelihood in some people, for instance, African or Caribbean people are more likely to suffer a stroke than Caucasians. Also, whilst males are more likely to have a stroke than women, women are more likely to perish if they have one (WHO, 2004).

1.5.3 Small Vessel Disease and Changes

Small vessel disease (SVD) is common among older adults and can be a precursor to a more serious vascular incident. SVD occurs as a result of atherosclerosis i.e. a hardening and build up of plaque on the arterial wall, which can sometimes reduce cognitive function. Although the presence of arterial plaques do not immediately result in the same disabling symptoms characteristic of an ischaemic or haemorrhagic stroke, they need to be monitored and risk factors such as smoking, diabetes and hypertension need to be addressed or managed, to prevent a more serious vascular attack.

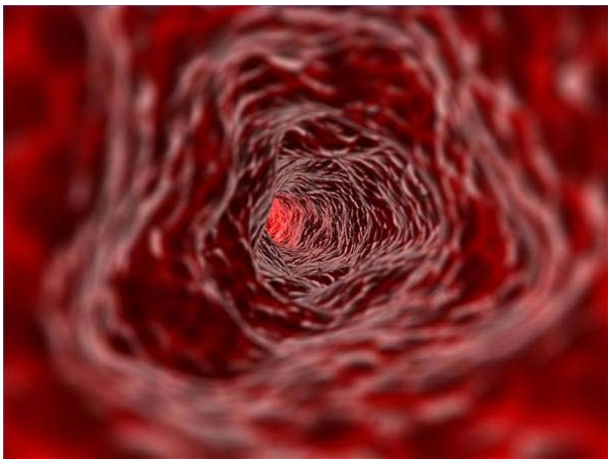
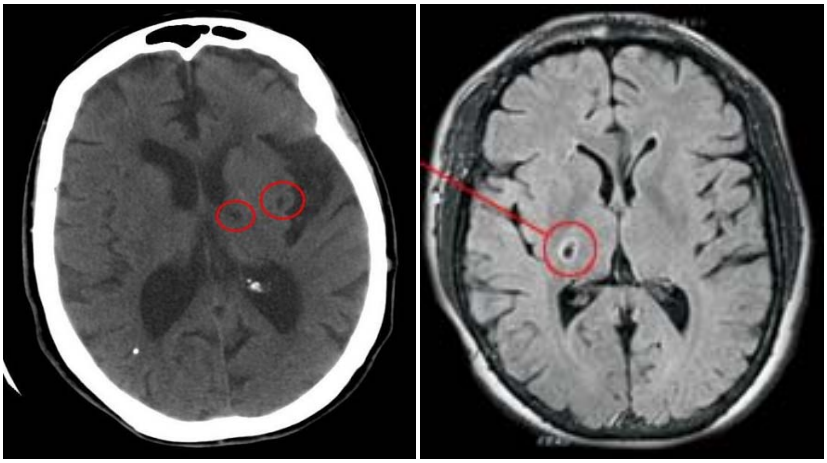


Figure 1.6 representing the inside of an arterial wall which has been affected by small vessel disease.

1.5.4 Focal Abnormalities or Lacunar Infarctions

Focal abnormalities or Lacunar Infarctions are classified as deep cerebral infarcts which are normally located in the basal ganglia. As defined by Osborn *et al.* (2005), Lacunar infarcts can literally be defined as a 'hole' in the white matter and range in size from microscopic to 15mm when an axial image is acquired. Focal abnormalities are very discrete and difficult to identify although they tend to cluster in multiple sites around the pons, caudate nuclei, internal capsule, putamen and thalamus. Typically they account for 15-20% of all strokes and can be most easily visualised with Magnetic Resonance Imaging (MRI) or Diffusion Weighted Imaging (DWI), if available. Unlike MRI, DWI provides information regarding the mobility of molecules within or between cell membranes, alternatively of the viscosity of water molecules in cerebral tissue (Chan et al., 2001).

Figure 1.7 represents the vessel change when a focal abnormality has occurred.



Figures 1.8 & 1.9 highlight focal abnormalities within CT (left) and MR (Right) images.

1.5.5 Acute stroke

Acute strokes are the second most common cause of death in the world and are classified as "an interrupted blood flow to the brain resulting in cerebral ischemia/ infarction with variable

neurological deficit" (Osborn *et al.,* 2005, *pp*I-4-77). Such interrupted blood flow is demonstrated in figure 1.10 below. The size of the infarct depends upon the degree of atherosclerosis (hardening and narrowing of the arteries) or whether a thrombus (blood clot) has travelled/formed and resulted in arterial blockage. Morphologically, the infarct appears as a 'wedge' shape within one or more grey matter vascular territories; alternatively, it may appear around the cerebral or territory peripheries.

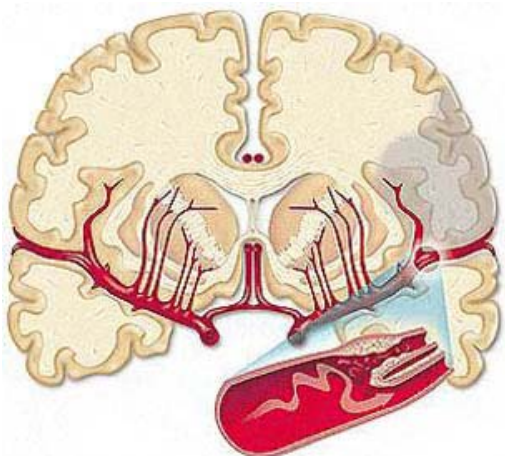
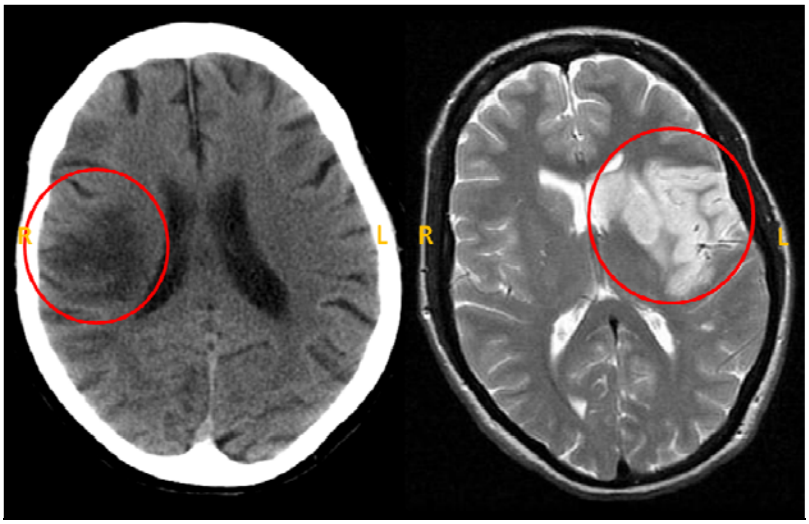


Figure 1.10 Represents how blood flow may become blocked or compromised by a thrombus or the build up of atheroma, resulting in a 'wedge' area of white and grey matter being affected.

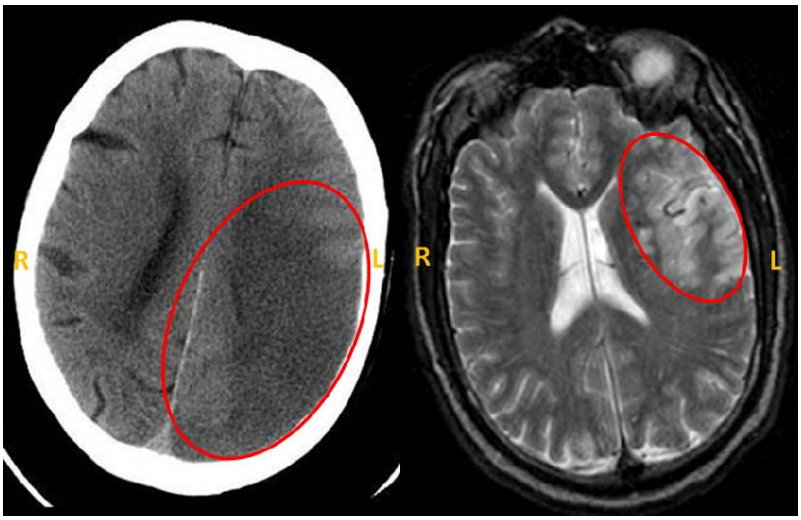


Figures 1.11 & 1.12 Demonstrate acute stroke presentation in CT and MRI.

The CT image (figure 1.11) demonstrates low attenuation (gradual loss in image intensity) in the right hemisphere of the patient. The MR image (figure 1.12) demonstrates a similar infarct and enhancement in the left hemisphere of the patient. When reading and reporting CT and MRI scans it is important to remember that the left side of the image represents the right side of the individual, and vice versa due to the acquisition process and hence the 'R' and 'L' labels on the images to remind the reader.

1.5.6 Subacute stroke

The subacute stroke type is characterised by infarcts that follow typical vascular pathways or territories through both white and grey matter. The period of time elapsed since first onset of ischemic 'insult' determines the size and extent of spread, typically classified between 2-14 days following initial arterial occlusion.

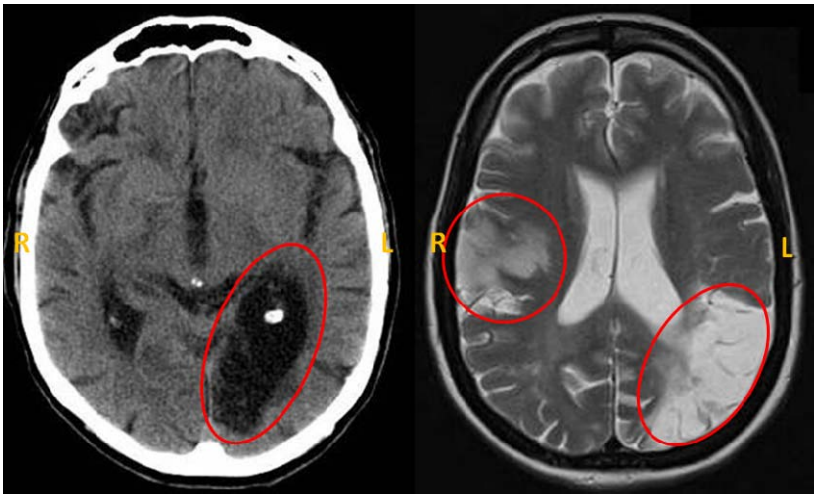


Figures 1.13 & 1.14 Represents the presentation of a subacute stroke in CT and MRI.

The above CT (figure 1.13) and MR image (figure 1.14) represent typical subacute left middle cerebral artery (MCA) infarctions. Both images highlight a low mass effect or 'spread' of damage to surrounding cortical regions, remembering that the left side of the image represents the right side of the individual, and vice versa due to the acquisition process.

1.5.7 Chronic stroke

Chronic strokes have the most severe effect on patient outcome of stroke types discussed in this chapter. A chronic stroke may follow a similar 'wedge-shaped' pattern, yet the cerebral volume loss is more marked and size can be more highly variable in a chronic infarct. Chronic strokes are likely to occur within any area served by a major blood supply such as both cerebral hemispheres (although unlikely in both simultaneously), cerebellum and brainstem. Due to the extensive pathological hallmarks, chronic strokes may be easily seen with either CT or MR imaging modalities (Osborn *et al.,* 2005).



Figures 1.15 & 1.16 Demonstrate chronic stroke presentation.

The above images demonstrate a distinct area of cerebral loss: the MR image (figure 1.16) also demonstrates a second area of ischemia that is 'older' (dark red circle) than the new area in the right hemisphere (bright red circle). In addition to the aforementioned stroke types, transient ischaemic attacks (TIA's) are also a form of 'mini' stroke. However, TIA's although they require monitoring via MRI, frequently resolve within 24 hours and due to their prognostic and treatment complexity, and have not been considered in the present thesis.

1.5.8 Stroke Treatment: What is the Current Standpoint?

The consultant neurologist manages patient treatment within the neurology department; however, radiologists are an integral part of this process. The current treatment of choice for ischaemic stroke is thrombolytic therapy, which must be administered intravenously within 3 hours of stroke onset and only when a radiologist is present to rule out a haemorrhagic stroke based on image findings (usually CT). Thrombolysis can reopen arterial occlusions, alleviating the pressure from a clot (Adams *et al.*, 1996). From recognising stroke symptoms, all involved must act quickly to increase the degree and likelihood of recovery, including non-medical professionals involved along the patient pathway who can speed-up access to urgent medical care by calling for an ambulance immediately. Individuals who have suffered a haemorrhagic stroke are not eligible for thrombolytic therapy, although quick entry into the acute care system will ensure their needs are monitored and met, even if palliative care is the most appropriate option.

Patient intervention, whether it be drug treatment, angiography (to highlight additional areas of blood vessel weakness appropriate for 'stenting') and/ or in rare cases, surgery, can only be administered following the detection (including localisation) and specification of the stroke type (i.e.

ischaemic or haemorrhagic), if present. The type of treatment is also dependent upon the size and 'age' of stroke i.e. how much brain function is affected and time since occurrence, as discussed above. The terms 'acute', 'subacute' and 'chronic' stroke types (as previously discussed) refer to the age and degree of spread or impact, a stroke has on functioning and prognosis.

1.5.8 What is the Imaging Protocol for stroke?

In suspected stroke cases, CT is the first port of call for emergency patients and is the modality of choice for ruling out intracranial haemorrhage (Mullins *et al*., 2002). However CT is not sensitive enough to detect small signs of the disease (DoH, 2008) such as small vessel disease or early parenchymal hypodensity. Due to enhanced detail, but also the increased expense of MRI, this modality is usually reserved for a second opinion following a CT scan and/ or identifying small vessel changes, which can be indicative of TIA. Although diffusion weighted imaging is reported to be superior to conventional magnetic resonance imaging, access is frequently limited due to availability and financial resources.

Despite, CT and MRI being the primary choice for identification of stroke, between the years of 1996 and 2003, there was a 31% increase in demand for CT examinations and a 68% increase in MRI services, demonstrating increased patient and service turnover in a relatively short period of time (BSNR, 2003). The British Society of Neuroradiologists (2003) considers the ratio of staff to demand 'a serious crisis'. Five years later and despite the Department of health issuing a ten-point plan promoting awareness and service delivery changes, aiming for the best possible care for all patients with TIA and stroke (DoH, 2007), they also admit that efficient and timely stroke treatment remains 'impractical' (DoH, 2008).

1.5.9 Which Modality is Considered Superior and Why?

CT and MRI are frequently used in combination, facilitating clinical decision-making in suspected stroke cases. When CT is compared with MR, the latter has much better soft tissue contrast resolution, without the ionising radiation. Both CT/ MR scanners generate multiple 2D images and 3D reconstructions but MR (although more expensive) has a variation of scanning properties, which can be altered and enhanced to detect different features, quicker, and in any plane e.g. axial, sagittal and/ or coronal, as depicted in figure 1.17 below (Fraunhofer, 2007). The advantage of MR is that there is no radiation and it is considered non-invasive, despite many patients experiencing claustrophobia during scan time.

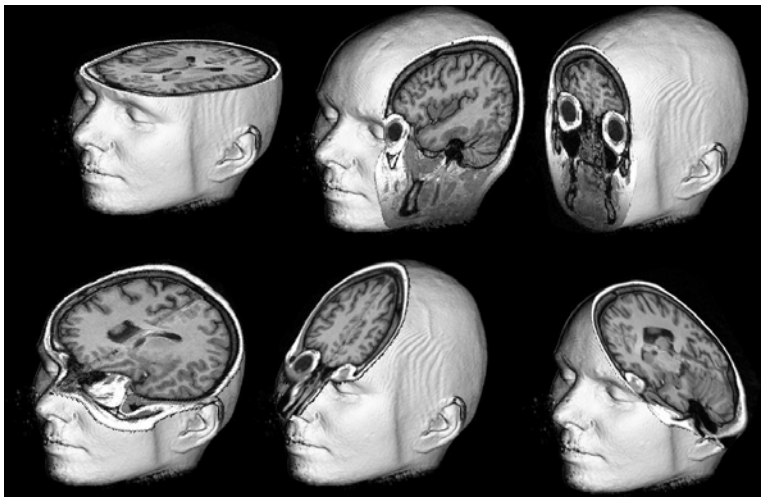


Figure 1.17 Represents how single images can be 'built up' to highlight axial, sagittal and transverse planes embedded within a 3D structure.

Radiologists have paid much attention to the comparison of sensitivity and specificity rates between image modalities. In a 1986 study by Haughton *et al.*, diagnostic accuracy in neuroradiology between CT and MR modalities was examined. Whilst MR imaging was considered more sensitive for spinal region abnormalities (88% MR sensitivity compared with 85% for CT), CT was considered more sensitive overall with 91% sensitivity rating and 82% for MRI. At that time, specificity was also perceived to be higher for CT (90%) than MR (81%) when 100 cases were viewed by two neuroradiologists and one research Fellow.

When stroke was the only condition being considered, mixed results were yielded. Gonzalez in 1999 compared rates of 18% sensitivity and 100% specificity for MR imaging and 45% sensitivity and 100% specificity for CT when 14 stroke patients were imaged within 6 hours of stroke-like symptoms. In a much larger review of 563 CT and 498 MR patients, Mullins *et al.,* (2002) uncovered that MR sensitivity had been underrated in the Gonzalez study by 40%, yet specificity remained constant at 100%. These studies and others have demonstrated a similar sensitivity of less than 50% for both conventional modalities, with neither modality proving superior to the other from an observer performance perspective (Mohr, 1995; Lansberg, 2000; Wintermark, 2007).

Recently, studies have reviewed the efficacy of diffusion-weighted MR imaging and found it be superior to conventional CT and MR, owing to superior lesion contrast and its ability to capture the diffusion of water molecules indicative of early physiological shifts (Gonzalez *et al.*, 1999; Lansberg, 2000). Gonzalez *et al.* reported 100% sensitivity and 86% specificity for DWI, yet with a larger sample size, Mullins *et al.,* (2002) reported a reverse trend of 97% sensitivity and 100% specificity.

Mullins also demonstrated that accuracy depended largely upon *when* the infarction had occurred and time elapsed until hospital admission. Following 12 hours of infarction onset, DWI accuracy was considered equivalent to CT. Perfusion CT, which generates 3D images of cerebral blood flow (Hoeffner., *et al,* 2004), has been shown to enhance conventional CT from 45% sensitivity to 76.3% and uncover lacunar infarctions previously missed by unenhanced CT readers, but CT still remains less sensitive than DWI (Kloska *et al.*, 2004). Despite modality assessment research, there is a paucity of research from a human observer perspective examining visual search itself in this clinical field.

1.6 Medical Image Assessment

"To strike a target, it is necessary to see it" (Gunderman, 2002. *pp*1239).

Medical imaging systems are the tools used to acquire an image, but the image alone is redundant without good sight and a wealth of prior medical knowledge. Before even considering the experience and the decision-making process of a reader, image assessment starts with the human visual system. The following section discusses the human visual system, how images are processed in the brain, and perception, prior to examining image interpretation, decision-making and performance between different observer types.

1.6.1 Anatomical Features of the Eye

Our eyes provide us with a wealth of visual information to interact safely and efficiently within our environment. Sight is made possible by the human visual system and the complex connections that exist between the eye and brain. The anatomy of the eye includes the cornea, iris, pupil, aqueous humour, lens, vitreous humour, retina, optic disc, fovea and the optic nerve and sheath, as demonstrated in Figure 1.18.

The eyes are held in the skull by extraocular muscles that control three types of actions; vergence, saccadic and pursuit movements. Vergence eye movements ensure that both retinas are focussed on a single object, despite how close or far away the object may be. These movements ensure the co-operation of each eye to achieve unified vision. Saccadic movements are the jerky eye movements, between fixations of objects that allow rapid scanning of a visual scene. Pursuit movements on the other hand, ensure a moving object can be 'pursued' within the visual field. A process known as accommodation allows objects to be fixated upon and processed by the brain, irrespective of distance from the observer (Carlson, 2001).

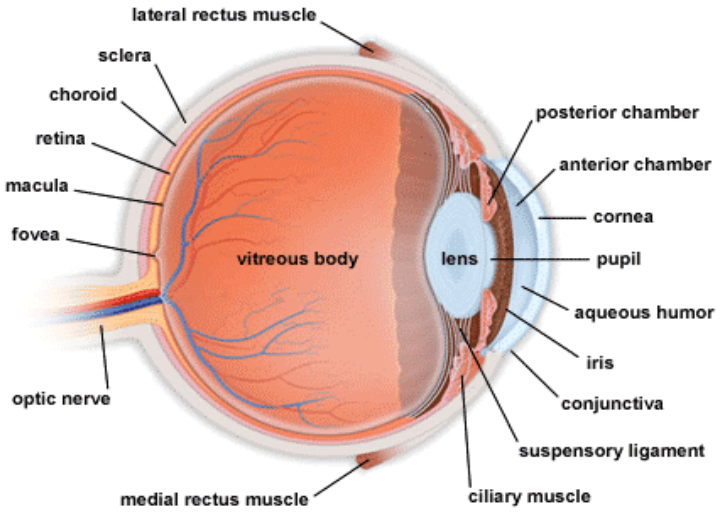


Figure 1.18 Demonstrating normal eye anatomy.

1.6.2 How Vision is Processed in the Brain.

An external image is 'transduced' into visual perception when light falls upon the retina and layers of photoreceptor cells (known as rods and cons) convert the light energy into neuronal signals (Duchowski, 2007).

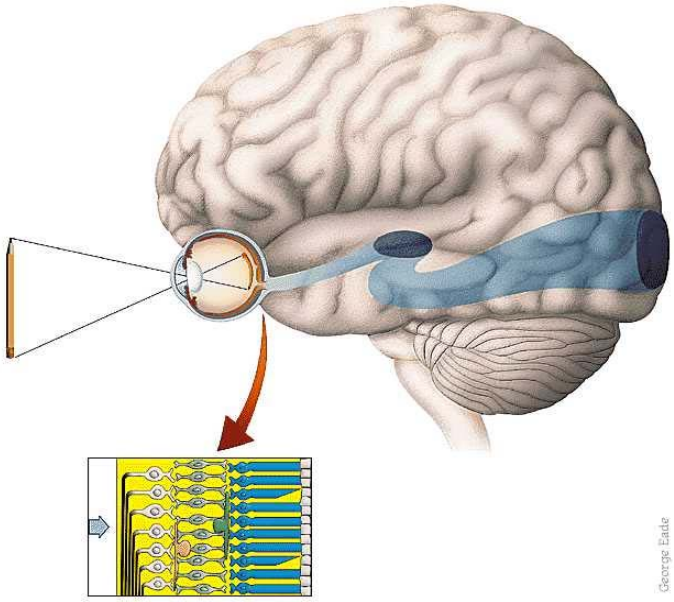


Figure 1.19 Representing the human visual neuronal pathway, consisting of photoreceptor, bipolar and ganglion cells.

It is the iris that controls the amount of light that reaches the retina. The Photoreceptor neurons that become activated by light transmit information to bipolar cells, which connect with ganglion cells by

synaptic information processing and communicate directly with the optic nerves. Optic nerves then carry this visual information to the primary visual cortex, located towards the rear of the brain (Carlson, 2001). A representation of this pathway is demonstrated above (figure 1.19).

1.6.3 Visual Perception and the Visual Field.

People can only glean the most detailed information from a very small region at any one time, although 20/20 vision is possible in the central visual field when visual abnormalities or disease are absent (Spector, 1990; Schwartz, 2010). It is estimated that the central area of visual interest which emanates from the fovea is only about 1-2 degrees in size and therefore, we are very selective about where we spend our time and energy when appraising each 'visual scene' (Tobii Technology Whitepaper, 2010). Our brains amalgamate external information available through neural code and construct the entire visual field that, ultimately we perceive and comprehend. This external-internal interaction is cumulatively known as 'visual perception' (Carlson, 2001, Schwartz, 2010).

Experimentally, researchers are able to use the fovea as a reference point from which to make inferences regarding visual perception, interest, attention and cognitive processing – people tend to gaze at interesting features for much longer than those that do not require much effort to see or 'understand' and do not require our fixed attention. Vision becomes 'perception' when the entire human visual system collaborates; from eye movements i.e. microsaccades, saccades, fixations, gaze duration, smooth pursuits etc through to neuronal and cognitive processing of visual stimuli, is combined and translated into meaningful interpretations (Carlson, 2001, Schwartz, 2010).

1.7 Image Interpretation: What is the Gold Standard?

It is well established that experts in any field hold a vast amount of theoretical knowledge and practical experience, which is informed by a vast number of hours on the job, objective measures of performance and feedback from others to critically evaluate one's own performance to a recognised standard (Ericsson *et al*., 2006). The question; 'what makes an expert radiologist?' frequently appears in research literature of this kind to determine how radiologists interpret medical images (Norman et al., 1992; Nodine & Krupinski, 1998).. One of the interesting features of a radiology professional is their exceptional aptitude for accurately (and consistently) detecting the location and diagnosing abnormalities, or excluding them, under resource and time pressures (Manning, Donovan & Crawford, 2006). The complexity of their job and relationship with technology has been of primary

concern in a number of studies aiming to uncover the psychological basis of how they perform their duties. It is still unknown whether radiologists are a product of their training and experience or whether they have a certain set of innate traits (e.g. visual search and memory) that predispose them to being superior medical 'problem-solvers' in the image realm (Norman *et al.,* 1992; Nodine & Krupinski, 1998).

1.7.1 Expertise Acquisition: How do you Become Expert at Interpreting Brain Images?

Radiology training programmes are based across NHS Trusts in the United Kingdom, sometimes in registered Radiological Academies (i.e. Norwich Radiology Academy, Peninsula Radiology Academy and Yorkshire and Humber Deanery). Trainees are expected to have a thorough understanding of general medicine and surgery prior to course registration (www.rcr.ac.uk, accessed February, 2011). They undertake a series of compulsory and elective modules centred on imaging of anatomy, structure and disease states as well as interventional radiology and therapy. In total, five years are devoted to training. Years 1-3 cover core training in science, radiation safety, interpretation and procedural skills. Years 4 and 5 are devoted to special interest clinical skills whilst undergoing supervised practice (RCR, 2007).

To become a registered neuroradiologist, trainees must devote 12 months to head and neck imaging, normally in the form of a master's degree in medical interpretation. Trainees should also become proficient in choosing the correct interventional technique for the type of problem under investigation, ensuring they are aware how contrast media and radiopharmaceuticals affect the human body and the clinical adverse events that might occur if used inappropriately. Of most relevance to the current research, neuroradiologists must have a sound understanding of normal anatomical features, disease manifestation and progression in head and neck imaging (RCR, 2007).

Multidisciplinary team participation ensures that trainees become competent at planning investigations, gathering experience of care planning and expected outcomes of disease, error detection in diagnosis and complications of treatment. The course is examined by regular direct observation through supervision by a radiology tutor or an external observer, regular formal reviews of skill acquisition and a final assessment of professional competence (RCR, 2004). Once trained, Neuroradiologists split their time between image assessments, interventional procedures, report writing and attending clinical meetings (BSNR, 2003).

1.7.2 Error in Radiology: What is the Potential for Error?

Medical image acquisition, accuracy and interpretation have an important role to play in contributing to patient safety, yet diagnostic errors in the interpretation of medical images have been reported since the 1940s and abnormalities have been either missed, or over-read, with various rates across numerous experimental studies. Since the 1960's, research in visual attention, perception and cognition has dominated 2D chest (e.g. Thomas & Lansdown,1963; Kundel, 1978; Gale *et al.*, 1979; Krupinski, 1993; Manning, 2004) and breast screening (Mello-Thoms, 2002; Scott *et al.*, 2008) in many countries, to identify the presence or absence of malignant or benign nodules.

Irrespective of image type, if an abnormal mass or structure is overlooked, there may be dire consequences for the patient (missed diagnosis and/ or suspension of treatment) and the medical team involved (potential disciplinary action or litigation). A number of variables impact upon decision-making: perceptual (the accuracy of image inspection), cognitive (medical knowledge, experience and decision-making) and ergonomic factors (seating, lighting and the environment used for interpretation) e.g. (Manning, Gale & Krupinski, 2005).

Specialists must examine the image and make prompt decisions, usually within seconds, about which aspects, if any, require further attention or action. Experts may fall short and miss critical abnormalities when they cease to recognise the limitations of medical image acquisition and/ or become satisfied with their primary visual search. Eye-tracking research has demonstrated that lesions in chest radiographs are looked at but not reported when satisfaction of search has occurred (Berbaum, 1990; Samuel *et al.*, 1995). In addition, experts may become so engrossed in their respective specialism that they discount irregularities in the scan that do not fit with their prior experience, inadvertently alienating themselves from perceiving and engaging with new possibilities and learning new imagery techniques or conditions (Donovan, 2006).

The NHS itself has been described as an 'error-prone' environment, frequently operating without adequate resources where a whistle-blowing individual approach is more likely adopted rather than a blame–free culture. In a blame-free culture, everyone works together and acknowledges human error, the need for regular top-up training and competence feedback. Errors within a blame–free culture are said to be more readily discussed and reasons behind the mistake teased out to ensure everyone learns from the scenario – a systems approach is taken to errors, rather than an individual being held accountable (Fitzgerald, 2001). Risk factors are dually monitored and errors can be prevented when communication and reporting is high between individuals, departments and NHS Trusts (DoH, 2000).

1.7.3 The Image

It is important to remember that medical images, as well as the readers and the environments in which they work, are not free from limitations. Apart from observer limits, there are a number of factors that influence image clarity and ability to depict anatomy accurately. Detailed anatomical features can become hidden within overlapping body structures on 2D images and unknown, or known, artefacts can distort or blur an image. Quantum limits also apply to image data, affecting image density and faulty technology has been said to account for between 10-30% of malpractice claims (Brenner *et al.,* 1998). It is well known that technical limitations of a chosen modality can affect accuracy e.g. spatial resolution, contrast, and noise also play a part in technical and reader efficacy (Metz, 2006).

The quality of the information and its presentation are continually being improved through computer science and engineering developments, however, as human observers, we are not particularly aware of spatial resolution, contrast, or noise variables, we mainly perceive, consolidate and interpret the meaning of the image dependent upon the information available to us at the time (Manning, Gale & Krupinski, 2005), and therefore image quality itself is not part of the present thesis.

1.7.4 The Digital Era and New Image Interpretation challenges

Radiology has moved away from film or hard-copy images and into the digital era, largely prompted by technological advances. Where radiology is not digitised; steps are being implemented to bring all National Health Service Trusts up to the same technological standard. Sakas (2002) stated technological advancement is essential to be more useful and efficient. Computer visualisation systems can now collate single images and 'build up' image representations to gather more information than ever before. Modern imaging methods can now produce hundreds of images, viewed in multiple sequences, offering free scrolling between sagittal, axial, and coronal orientations, even diagonal planes for the reader.

Each stride in technology impacts heavily upon the amount of human-computer interaction required from radiologists and resultant clinical performance (Phillips *et al.,* 2008). From an interpretation stand point, it is important that radiologists can make use of technology quickly and effortlessly whilst ensuring quality patient care and safety are not compromised (Krupinski, 2000). As multi-slice digital imagery, which involves incredibly large data sets, is now commonplace in the clinical setting, accompanied by an increasingly flexible viewing software, research must endeavor to

examine how multidimensional and multimodal imaging affects accuracy and efficiency in radiology. As further practical changes are implemented within the health service, further questions arise: Is clinical technology advancing at a speed that medical professionals might find overwhelming? What impact will further alterations have on the radiology workforce? How will the shift from 2D to 3D affect performance and clinical errors? How do incorporating larger data sets into the clinical setting affect practice, and how can researchers measure and predict the consequences?

Previous research has found mixed results with regard to the 2D-3D image problem. Krupinski (2000) reported that more errors tend to represent false negatives (incorrectly ruling out abnormalities in an image) rather than false positives (incorrectly marking a normal area as abnormal) with search, recognition, and decision errors occurring when assessing anatomical images. Only a handful of studies have explored factors affecting the interpretation of multidimensional imaging, yet these techniques will only become more important and widespread in the future (Sakas, 2002; Phillips *et al.*, 2008). This change in practice calls for observer performance studies to thoroughly encompass these complex and multimodal image type questions, yet only recently have experimental studies been designed to track eye movements across a number of medical images in radiology (Phillips *et al.,* 2005 & 2008).

1.7.5 Visual Search, Attention and Cognition

The visual scan paths and eye movement recordings represent the areas of an image that an observer finds most interesting. Prolonged gaze in certain areas can indicate conscious and/ or analytical thought processes. From these assumptions it is inferred that abnormalities are dwelt upon for long periods of time in specified areas or zones by observers. Eye-tracking provides an objective, experimental insight into how radiographers and radiologists appraise a complex medical image and how errors might occur.

Observers may ‘see’ a lesion but not recognise it as being abnormal, and thus it is not processed cognitively. Alternatively, an observer may miss a large amount of information because their attention was prioritised elsewhere, or they become so satisfied with the information that they had interpreted, that they no longer continue to investigate other possibilities within the scan, missing secondary or tertiary abnormalities; this type of behaviour is referred to as ‘Satisfaction of Search’ (Berbaum, 1990).

One of the well established ways of assessing observer performance across differing images when searching for abnormalities is also through eye-movement research methods (Kundel, 1978;

Mello-Thoms, 2002). Visual search studies in medical image perception allow conclusions to be drawn about the observers search behaviour and underlying cognitive processing and also gain an objective insight into human-computer interaction.

1.7.6 Observer Performance, Experience and Visual Search

Factors that influence the ability of the observer to meet the patients' requirements are not only dependent upon image quality and modality, but are also dependent upon the observers level of experience. Radiology studies using eye tracking technology have demonstrated novice and expert differences in location and duration of gaze, to inform skill acquisition and expertise. Experts examine medical images in 'longer, sweeping eye movements' but they also focus in on a smaller number of areas or 'zones' for a more detailed inspection than novice observers. Interestingly, as visual coverage alters through experience, so does speed and accuracy: experts make better decisions in much less time (Manning *et al.*, 2005). Visual search studies in radiology have shown consistent significant differences between novice and expert readers irrespective of reading task (Donovan, 2006).

It appears that experts operate a system of deduction, ruling out certain areas very quickly to economise effort and enabling attention to be redirected to more interesting clinical features. It appears that novice readers are unable to 'sweep over' an image and rule out certain zones with speed and confidence, avoiding a diagnostic 'faux pas'. Experts may well employ a strategic visual coverage pattern based upon prior knowledge of anatomical locations where abnormalities may cluster, and this ability, if it indeed exists, requires more attention.

Whilst it is expected expert readers are the optimum performers, studies demonstrate subtle differences in visual search between the transitions from novice to trainee and through to expert reader and distinctions between the groups may be less clear. Studies of radiographers highlighted no significant difference between pre-training radiographers and novice readers (Manning, 2006), but once training was underway or completed, a medically analytical mindset (coupled with experience of viewing multiple images), differentiated these radiographers from lay people (Nodine and Krupinski, 1998). Whether optimum visual search strategies are acquired as part of a radiology training programme or image interpretation is an innate ability in some individuals, akin to individual differences in face perception as demonstrated through functional MRI studies (Kanwisher, McDermott & Chun, 1997) cannot be 100% confirmed but evidence can be gathered which attempts to support or refute this view point. Conclusions from a study conducted by Nodine

and Krupinski (1998) into a non clinical visual search task, namely 'Where's Waldo' image, uncovered that whilst radiographers did not differ significantly from lay people in their overall task performance, what differentiated them from lay people was their 'structured' visual search patterns.

The differences in visual scanning between novice and experts may not only be influenced by direct experience and repeated exposure to clinical images. It may be that experts have developed complex 'cognitive maps' about anatomical features. Research suggests that experts focus on specific areas of the scan, can generate hypotheses and make complex decisions based upon mental schema of salient image features (Garlatti & Sharples, 1998), drawing upon problem-solving skills and synthesizing biomedical knowledge to support clinical reasoning (Rogers, 1995). If they do develop certain 'schema', can this phenomenon be measured and empirically investigated? In addition, experts may hold a bank of acquired information surrounding disease manifestations and progression. For instance, different types of arterial occlusion and haemorrhage pervade different vascular routes within the brain due to known cerebrovascular pathways. Experts should have preconceptions, not only about anatomy, but where another stroke might develop, what abilities might become impaired as a result and how to identify current and future problems from image data.

1.7.7 Eye Movements and Decision-Making

Nodine and Kundel (1990) reported the minimal dwell time for abnormality detection to occur in clinical studies is 0.9 seconds. In a series of studies by Manning *et al.* (2004), dwell time appeared to have implications for confidence in decision-making processes; missed lesions were actually dwelt on for an average time of 3.1s, indicating that they were recognised but not interpreted cognitively, therefore, perception maybe the key to fundamental errors. Dwell time appeared to have links with observer expertise; length of dwell time appeared to be positively correlated with inexperience on nodule detection tasks.

Manning *et al.,* (2005) found that correct negative decisions are made quickly and are positively associated with expertise with half of these decisions being made within one second of visual attention. Incorrect negative decisions individual image features that were gazed at for much longer (i.e. >3 seconds) indicating an extended interpretation of, and/ or cognitive dissonance surrounding the feature. Experienced radiologists ruled the presence of a possible nodule out within 2 seconds of foveal fixation. Whereas all naive decisions made after 3 seconds were incorrect; once over the 4.75 fixation time point, no negative decisions (either correct or incorrect) were attributed

by radiologists. Manning (2006) also confirmed that statistically, not much separated pre-training radiographers from novice readers in a nodule detection task, but fixations per film reduced as training continued.

1.8 Experience and Expertise in Radiology

1.8.1 Expertise and Consultant Performance

If medical training and years of consultancy are controlled for, what differences will exist (if any) in the visual search and accuracy patterns between two expert groups from differing specialties? As previously discussed in chapter 1, treating the patient without key information from a head scan, confirming early parenchymal signs of anatomical change can have catastrophic consequences. For instance, a patient who is given thrombolytic therapy for an intracranial haemorrhage, rather than ischaemic stroke, will bleed further internally and it is highly likely the patient will die as a result of this inappropriate intervention (Schriger *et al.,* 1998; Grotta, 1999; DoH, 2007).

Schriger *et al.,* (1998) compared the ability of 38 emergency physicians, 29 (largely community) neurologists and 36 general radiologists to rule out the presence of haemorrhage and rule in the administration of thrombolytic therapy following the presentation of 15 scans. All observers were tested with 5 initial scans and placed in either a standard or advanced category for the following ten scans to match reader ability to case difficulty. Observers were asked to rate the scans on whether thrombolytics could be administered to the patient by answering 'yes' or 'no'. If they answered 'no', participants had to state whether this decision was due to signs of haemorrhage or acute infarction.

The results demonstrated that radiology and neurology performance did not differ, with both groups scoring an average of 83% correct responses. Emergency medicine readers performed much worse with an average score of 67%. 40% of neurologists and 52% of radiologists scored 100% sensitivity, and overall study sensitivity was 82%. Specificity rates were not discussed, yet the authors alluded to some physicians having trouble differentiating between a haemorrhage and calcifications, and recent infarctions from existing abnormalities; both distinctions are important when thrombolytic drugs must be administered safely. The authors concluded that whilst some readers (predominantly those who admitted to reading CT scans on a regular basis) were capable of making an independent decision within the clinical workplace, the majority of readers were not, and went on to state that board certification alone did not guarantee competence in this area. The following

year the study was replicated at another centre with 70 scans and 16 readers from similar specialties with varying degrees of experience. Substantial variability was found both between readers and between specialties when Cohen's Kappa was used to compare inter-rater reliability (Grotta, 1999).

The Schriger study compares a population of predominantly community neurologists and general radiologists, who may or may not have had a special interest in neuroradiology. Between groups, their performance was not different and in subsequent studies interobserver agreement was fair to moderate (Grotta *et al.*, 1999). The present study explores not only stroke detection between neurologists and radiologists working in an acute care trust, but also the visual search patterns between the groups when examining the images which, to the authors' knowledge, has not been studied.

In addition, all prior studies discussed thus far have considered the image either in isolation from the clinical information, with it or have not specified either way. Images are rarely interpreted in isolation from the clinical information and performance by all parties might alter drastically dependent upon whether clinical information accompanies the images or not. The following section explores what impact, if any; clinical information has on image assessment and resultant performance.

1.9 Consultant Communication

Within the medical context, communication along the patient pathway occurs in a number of ways; from direct communication between nurses and clinicians on ward rounds, to staff within the radiology department and at multidisciplinary case conferences, to written communications e.g. medical history, reports and letters between healthcare practitioners from every specialty. Particular to the assessment of stroke; neurologists are required to provide a clinical history to accompany the medical assessment requested and radiologists are required to provide a definitive medical report of image findings. The reading and writing of reports is an integral aspect of patient care, which contributes towards the decision-making process surrounding the patient treatment protocol but does access to patient information prior to medical image assessment affect the decision-making process?

1.9.1 Does Patient Information alter Radiology Image Assessment?

The availability of clinical information is reported to impact upon resultant diagnostic accuracy, although the research to-date has been inconclusive. On first consideration, most would assume that clinical information enhances performance. Intuitively, some might argue it makes sense to know what is being looked for before you search for it. However, whilst some studies report increasing true positive ratings following presentation of clinical information (Schrieber, 1963) and agree with this assumption, others report no significant increase in performance (Good *et al*., 1990).

The effect of clinical information on diagnosis has been considered across a number of clinical domains; from fracture detection (Berbaum *et al.*, 1988) to ECG interpretation (Hatala, Norman & Brooks, 1999) with all studies aiming to assess whether prior availability of information results in a biased interpretation of image findings. For example, if a reader is aware of a specific condition prevalent within the population and the patient history matches the disease symptoms, is the reader more likely to provide a false positive account even if the image is normal? In 1981, Doubilet and Herman found a significant increase in accuracy but also false positive decisions of four radiologists in a radiograph detection task when information was present. In 1992, Norman, Brooks, Coblentz and Babcock investigated feature detection of bronchilitis and uncovered that both experts and novice readers were influenced by clinical information, although novices made more false-positive reports than experts on this chest radiograph detection task. In 1997, Tudor *et al.* measured the performance of five consultant radiologists on a plain radiograph detection task and although findings did not reach statistical significance, consultants made more false-positive diagnoses when information was present.

In a study particular to reporting on CT images where four readers examined 89 cases for tumour or vascular disease, accuracy values were 94.4% where information was withheld and 97.7% where information was given (McNeil, 1983). In a study of multiple anatomical regions; from pelvis through to head CT examination 17 years later, clinical information was found to increase the number of false positive diagnoses but reduced the number of false negatives in a sample of three radiology consultants. Clinical information appeared to strengthen consultant perceptions of insignificant findings if the accompanying information alluded to a positive diagnosis and worsen accuracy if it is was inaccurate. The authors reported that reliance upon information may be more likely in multidimensional imaging, which can be more challenging than interpretation of plain chest radiographs, due to the increase in images, whereby information can provide essential clues for abnormality localisation (Leslie *et al.*, 2000).

In studies pertaining solely to the detection of stroke and where clinical information was considered, Mullins *et al*. (2002) [a], uncovered that whilst there was no improvement of sensitivity in diffusion-weighted imaging, there was a significant improvement in sensitivity when readers were made aware of a likelihood of stroke in unenhanced CT cases. In this study the false positive ratio remained constant and did not increase with the accompanying clinical information, unlike the specificity ratings by Leslie *et al.* (2000).

Where neurology and radiology consultant performance was compared in a focal abnormality detection task of two ambiguous CT scans, which were randomised amongst 7 other scans (up to 9 slices of the brain were accessible). It was assumed in the study that neurologists would be more likely to over interpret or overlook findings than radiology readers, who were considered to be more objective. The results of the study found that neurologists were more likely to miss the focal abnormality than the radiologists, who were more sensitive, but that clinical information did not, in this instance, bias either reader group in either direction (Bonke *et al.*, 1989).

Upon consideration of the aforementioned studies, the impact of clinical information might be more complex than originally thought; information might indeed influence overall performance, might increase the number of true positives and/ or false positives, or it might not. In addition and to the authors' knowledge, there are no reported eye movement studies of reading and information processing itself with a clinical population.

1.10 Research Justification

Neurological problems are pervasive within our society, causing death, debilitation and suffering. CVD's are a subset of neurological deficits, and stroke is the third leading cause of death and disability in the US and UK. The effects of stroke, if spotted quickly can be reversed or diminished if haemorrhage is ruled out within 3 hours of onset by neurological and radiological intervention; "time is brain" (DoH, 2007). Therefore, it is important that departments collaborate quickly, effectively and stroke is ruled out or detected and diagnosed/ classified accurately to allow appropriate treatment to commence.

Every NHS Trust in the country encounters and treats stroke patients, but not all Trusts diagnose and treat stroke patients as quickly and efficiently as they could, owing to a number of contributory factors. Detection and classification errors have been widely reported in the chest and breast literature, however, errors of reporting have not been considered extensively within the neuroradiology research literature. Not only has observer performance been neglected, focussing

mainly on the image quality in the specialty, but there is a paucity of research relating to visual search in neuroradiology. As a result, there is little research into the search characteristics that define expert performance and how search differs along the expertise acquisition continuum.

Visual search in neuroradiology has been neglected for a number of reasons; mainly because imaging of stroke patients is reactive rather than preventative i.e. stroke patients are a diseased population, rather than a non-diseased population. Breast screening has received much attention as cancers, if sensitivity is high, can be treated in a similar fashion, with a high success rate and prevent an early death. Stroke is often seen as a disease of 'old-age' and is treated reactively i.e. when an infarct has occurred and symptoms appear. Stroke is less easy to prevent and treat as varied lifestyle factors play a significant role in CVD and brain intervention, unless stroke is treated with intravenous thrombolysis, is complicated and risky. In addition, visual search is much more difficult to monitor on brain imaging systems, rather than 2D images of the chest and breast, owing to the number of images and sequences available to the radiologist and the lack of available software to pair eye movement recordings to multiple image slices when radiologists free-scroll between image slices in multidimensional cases.

In addition to multidimensionality in stroke imaging, radiologists also compare multimodal cases i.e. some patients are imaged with CT and others with MRI dependent upon scanner accessibility, resources and patient presentation (i.e. emergency or outpatient), yet the two modalities have differing advantages and disadvantages. In the present study, it is intended to cross compare performance within and between modalities.

As well as the development of expertise, performance and visual search in neuroradiology in CT and MRI, only one or two research papers have compared how allied consultants compare in the same reading task i.e. radiologists compare with neurologists. In this instance, a baseline of experience and neurological knowledge is controlled for, but only radiologists have received the image interpretation training and intensive caseload experience.

And finally, irrespective of experience or specialty, images are rarely viewed in isolation from other influences in the clinical setting, namely preceding clinical information regarding patient status. Whilst this area has been explored in the research literature across multiple reading tasks, mixed findings have been reported regarding reader bias. The aforementioned literature review raises the following key research questions;

- What constitutes the gold-standard in image interpretation in stroke detection? If experts make errors, how and why do they occur?

- When using reporting and eye movements as an observable measure of performance and decision-making, how does visual search differ dependent upon level of experience and the decisions made?

- Does visual search differ between and within performance when multimodal methods are considered? And if experience level is controlled, how do a pair of matched consultants compare on a reading task?

- Does prior clinical information enhance, worsen or not affect observer performance and what information within the clinical report draws the observer's attention?

1.11 Ph.D. Aims and Objectives

- To determine the extent of error in stroke detection among radiologists, if any.

- To determine what characterises expert visual search in stroke detection and how performance and visual search differs along the expertise acquisition continuum.

- To determine how visual search differs across a number of medical images of the same patient, rather than a 2D visual search reading task.

- To determine not only how performance and visual search differs between groups in a single modality, but also differences within and between group performance in multimodal imaging i.e. CT and MRI.

- To determine how performance and visual search differ among a matched sample of radiologists and neurologists when level of experience is controlled.

- To determine whether the presence of clinical information biases image interpretation i.e. does the presence or absence of information alter accuracy, dependent upon reader group, stroke type or modality?

- Finally, to determine which aspects of the clinical information generate the most interest by different reader groups and why.

N.B. Please refer to page 2 for a list of chapters and studies therein.

Methodology

As visual search forms the basis of all studies within this thesis, the primary section of this chapter will explore eye movements – a brief recap of visual perception, followed by how eye movements were recorded in the subsequent experiments (section 3.1). Secondly, the overall design of the experimental studies is covered, including sample size estimation, participant selection and recruitment, study protocol, procedure and ethical considerations (section 3.2). The third section details the patient case selection and stroke type parameters, including dependent and independent study variables and counterbalancing of order effects within studies (section 3.3). Finally, factors are considered that contributed to the analysis of the observer performance data ahead of the experimental study chapters' i.e., Receiver Operating Characteristics and formulae applied to the eye movement data analysis (section 3.4).

2.1 Visual Perception and Eye Movements

Visual perception involves the co-operation of the entire human visual system; from eye movements i.e. microsaccades, saccades, fixations, gaze duration, smooth pursuits etc through to neuronal and cognitive processing of visual stimuli, which can then be combined and translated into meaningful interpretations of the external environment by an individual, as previously discussed in the literature review.

As people focus upon, and can only glean the most information from, a very small area within our environment, it is possible to (experimentally) use the fovea as a reference point from which to make inferences regarding visual interest, attention and cognitive processing – people tend to gaze at interesting features for much longer than those features that do not require much effort to see or 'understand' and do not require our fixed attention. It is estimated that the central area of visual interest which emanates from the fovea is only about 1-2 degrees in size and therefore, we are very selective about where we spend our time and energy when appraising each 'visual scene' (Tobii Technology Whitepaper, 2010). This proportion of the visual field can be expressed as a 'visual angle', which originates from the fovea. The area of the fovea and proportion of external information appraised is calculated using simple trigonometry depending on the viewing distance of the observer to the target such as a computer monitor.

2.1.1 Tracking and Recording Eye Movements: The Tobii Eye-Tracking System

The Tobii Gaze-Tracker is a remote system that connects to a Personal Computer (PC) and allows the recording of eye movements when observing images on a selected computer monitor, without interfering with the user environment. The gaze-tracker system uses infrared technology to monitor gaze, scan paths and dwell time information (i.e. saccades, fixation localisation and duration) back to the software programme, which then records this information from the eye-tracker. The infra-red beam highlights the pupil and the corneal reflection, which is then used to calculate the visual angle and central fixation point, expressed by 'x' and 'y' co-ordinates upon the computer monitor. As the gaze-tracker is a remote system it allows participants to move more freely (within predefined parameters), unlike head mounted eye tracking systems. The gaze-tracker allows computer-based experiments to be designed and conducted by simultaneously recording an observer's gaze from which experimental outputs can be collated and analysed using the Tobii software application Clear View. Additional outputs from experimental data include gaze plots (visual representations of the sequence and timing of visual attention), hot spots (a colour based indication of foveal fixations, visual coverage and gaze duration), and video replay of real-time participant eye movements, which are then overlaid upon original stimuli for research interpretation.

Figure 2.1 The Tobii x50 Gaze Tracker System.

To ensure the system operates effectively, an observer must be positioned 60 centimetres from the computer monitor and the observer must be looking directly at the centre of the screen. The position of the Tobii remote system should be at a 45 degree angle to capture the participants' eyes on the infrared beam. Once both eyes have been captured, a calibration exercise is performed using a five point scale, represented by five large 'pulsating' circles that move over the centre and circumference of the computer monitor. The observer must follow these points with their eyes to generate correction factors which are then applied to all subsequent eye movement data, as seen below in Figure 2.2.

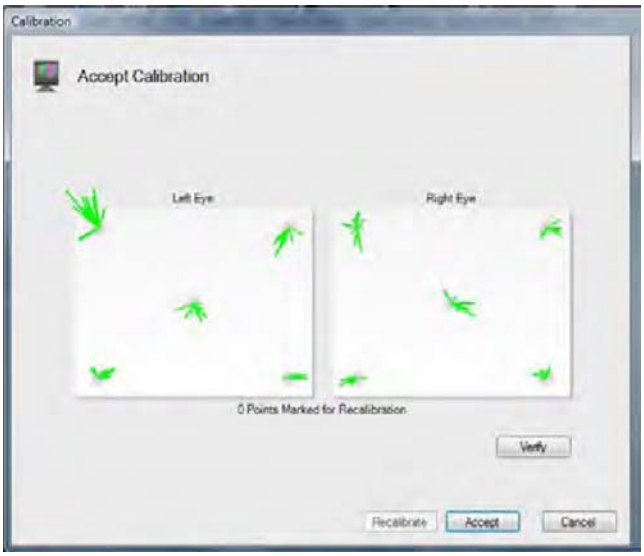


Figure 2.2 Screen shot of the calibration output (Tobii Technology).

Once this task is completed, the experiment may take place with validity (within the parameters of the system) ensured. Images were consistent for size and viewed on an LG Flatron L2000, black LCD monitor (Display area: horizontal; 16.8”, vertical; 12.0”, diagonal; 20.1”).

Below is a summary of the technical characteristics of the Tobii x50 system (table 2.1), demonstrating functionality, accuracy and any limitations of the system including drift, maximum gaze angles and the head-movement compensation error. On occasion, ‘drift’ can occur and the gaze-tracker output does not accurately resemble the eye movements of the reader. Both the reader and the reader’s environment can influence the degree of drift. In the following studies, experiments were conducted in the same locations to avoid any external impact upon the Tobii results and regular calibrations were performed at the beginning of each study and halfway through each experimental condition. Where considerable drift was present from moisture on the eye, or the natural physiology of participants’ eyes, it was decided these images would be excluded from the final analysis to ensure results remained robust.

Table 2.1 Technical characteristics of the Tobii x50 system (Tobii Technology AB, 2004).

Characteristic	Tobii x50
Physical design	Stand alone
Data output	Time stamp, Gaze position relative to stimuli for each eye (X and Y) Position in camera field of view of each eye (X and Y) Distance from camera of each eye Pupil size of each eye Validity code of each eye
Accuracy	0.5-0.7 degrees
Spatial resolution	0.35 degrees
Drift	< 1 degree
Freedom of head movement (W x H x D)	30 x 16 x 20 cm at 60 cm from tracker
Camera field of view	21 x 16 x 20 cm at 60 cm from tracker
Binocular tracking	Yes
Head-movement compensation error	< 1 degree compensation error for head translations in the three dimensions and rotations across the entire head movement space
Top head-motion speed	~ 10 cm/s
Time to tracking recovery	< 100 ms
Frame rate	50 Hz
Latency	35 ms
Max gaze angles	/ - 35 degrees
TFT Display	None
Connectors	Firewire, USB, Power
Infant AddOn option	Yes
Weight (including desk stand, but without case)	~ 3 kg

2.2 Design

In addition to research team discussions, three expert Radiologists were consulted over the design and methodology of the ongoing research; 1) The Norwich Radiology Academy Director, 2) Consultant Neuroradiologist at Norfolk and Norwich University Hospitals Trust, and 3) Consultant Neuroradiologist, Queens Medical Centre, Nottinghamshire NHS Trust, who provided advice regarding the purpose, suitability, design and methodology of the research project. As this project focussed upon the identification of stroke in medical images across different levels of expertise, three participants groups were decided upon; novice readers, trainee readers and expert or consultant readers throughout the series of studies.

Study 1 was conducted as a feasibility study, designed to assess whether the project would be possible with a larger case and sample size, and would produce meaningful results. Eight patient cases and their clinical histories and outcomes were selected from the Harvard Medical School bank of clinical cases to represent a spread of control, acute, subacute and chronic cases. These cases were edited and an experimental study was designed in much the same way as studies 2 and 4 – further details will follow. A selection of participants were identified and recruited namely; four novice readers were recruited from Loughborough University. Four naive participants were recruited from Loughborough University. One second year Specialist Registrar (trainee) was recruited from the Norwich Radiology Academy and three experienced Consultant Radiologists were recruited from the Norfolk and Norwich University Hospitals NHS Trust. The study formed the basis of studies 1, 2, 3, 4, and 5 as all participants were asked to comment on the design of the study. All participants provided positive feedback regarding the study design yet, the radiologists made recommendations regarding the appearance of the reporting sheets – the brain 'atlas' where the infarction locations were to be plotted was to become an abstract representation, rather than a detailed depiction of the middle slice with ventricular details etc. The results and conclusions of this study will be explored in more detail in the subsequent chapter.

2.2.1 Sample Size Estimation

The size of the effect to be detected in studies 1, 2 and 3 were sought on the basis of the effect-size observed in the pilot study. Effect-size was calculated in terms of Cohen's d. The formal Cohen's d calculation was used (Maxwell, 1990. pp 573) to estimate the sample size required. Sample size calculations were based on feasibility study results, previously examined with 8 consenting participants (novices=4, radiology trainee and experts=4).

The mean performance of participants when clinical information was received was subtracted from the mean performance of participants when clinical information was not received. On the basis of the absolute values of these figures, the mean difference for each expertise group (expert, novice) was derived. The mean difference for each expertise group was divided by the standard deviation of the differences for that expertise group in order to obtain Cohen's d (novices d=.50; experts d=.87). Pearson's Correlations were also calculated for the strength of relationship between scores at each level of the provision of information factor, separately for each expertise group (novices r = 0.58; experts r = 0.33). These figures were used to derive estimates of the required sample sizes based on designing a study with power of .80 at an alpha level of .05 by consulting the appropriate look-up table (Maxwell, 1990. *pp* 570, table 13.10).

To achieve a study with 80% power, given an alpha-level of .05, it was estimated that 18 experts would be required in each of two conditions in studies 2 and 3 relating to the provision of clinical information, whilst 34 novices in each of two conditions relating to the provision of clinical information. Whilst this number is achievable for the novice group, unfortunately, there would not be enough trainee/ experts to accommodate the power analysis from one NHS site. However, Obuchowski (2000) reported that with an image collection of 120 per modality, 10 observers per group would be acceptable.

It was therefore aimed to recruit 10 observers per group and if this was not achieved, the observed power of each study would be calculated to highlight any caveats associated with the study's findings (if the study was found to be underpowered). It is important to note that the majority of literature within this research area is published with very small sample sizes e.g. three or four expert radiologists alone and meaningful results are still achieved. Therefore, the aim to recruit 10 participants per condition exceeds that which has been achieved in many other published studies and is recommended by Obuchowski (2000) when 120 images were incorporated into each respective study.

2.2.2 Participants

This series of studies required a selection of participants with a comprehensive range of experience of reading medical images, therefore, three groups of participants i.e. novice, trainee and expert readers were required;

1) Novice participants were individuals from a non-medical background with little or no prior knowledge or experience of radiological image interpretation,

2) Trainee participants were registered on a radiology course and were required to specify whether they were in year one or two of core training, or postgraduate trainees specialising in head and neck Radiology. First and second year radiology trainees were required to state whether they had completed their neuroradiology module or not.

3) Expert radiologists were defined as those who had completed their core training, were registered practitioners with either a practising special interest in neuroradiology or had completed the full neuroradiology specialty training via the masters in head and neck radiology route. Although some radiologists worked in other clinical fields, all those with a special interest in neuroradiology were included in the study.

There were no restrictions on age or gender for this study and prospective participants were identified in one of three ways;

1) Novice participants were identified through Loughborough University staff and research students.

2) Trainee participants were identified through the Norwich Radiology Academy trainee lists.

3) Expert participants were identified by working with the Consultant's medical secretaries employed by the Radiology Academy or the Norfolk and Norwich University Hospitals NHS Trust.

Participants were primarily informed about the study by the author, although on occasion, participants may have been introduced to the study by the Norwich Radiology Academy Operational Director, if a trainee, and/ or the Consultants' medical secretaries at the main hospital site. Upon identification, participants were formally recruited and given a written participant information sheet and consent form to consider. Participants were also given ample opportunity to consider the research aims with the author and/or ask any questions prior to informed consent being granted. Two copies of consent forms were obtained - one for research purposes and one for Trust records. In total, 36 participants were recruited; ten novice readers, ten Specialist Radiology Registrars, eight Consultant Radiologists and eight Consultant Neurologists. The study was performed between Loughborough University, Norwich Radiology Academy and the Radiology Department, NNUH. As previous studies have considered the impact of room temperature and lighting conditions on radiology performance, recently McCarthy & Brennan, 2003, Brennan, 2007 and McEntee, 2009, room temperature and lighting conditions were controlled for at all image reading locations.

2.2.3 Case Selection and Group Allocation: Studies 2, 3, 4 and 5

Two-hundred and forty clinical images were selected and made anonymous from a bank of predetermined clinical cases at Norfolk and Norwich University Hospital Picture Archiving and Communications System (PACS) with the assistance of a resident Specialist Registrar in Radiology. Of these images, 120 were acquired using a conventional CT scanner and 120 acquired using a conventional MRI scanner. All axial image slices were extracted from forty-eight predetermined clinical cases.

Unfortunately, although accurate data could be found regarding stroke epidemiology (Feignin *et al.,* 2003), accurate data on lesion size per stroke classification could not be found, and therefore, the spread of cases was decided on the likelihood of the observer detecting each stroke

type and a representation of patient cases that enter the radiology department within a typical month at the radiology department. In addition, as the case population is a diseased population i.e. the collection of symptoms the patient presents with is likely to indicate an underlying neurological deficit, a quarter of case types were normal controls, some of which were more challenging than others i.e. they also had incidental clinical findings such as small vessel changes, which are often a precursor of an infarct. The cases were selected by the author and Specialist Registrar in Radiology on the basis of case spread and the radiology report information, which issued the final diagnosis and stroke classification. Both CT and MR cases represented by a spread of six normal controls, eight acute cases, six subacute cases and four chronic cases, were matched as closely as possible for each modality to ensure cross-comparability between studies.

Following case selection, five axial slices per case were extracted to represent between 75% and 100% of the abnormality throughout the full image 'stack'. Control cases had equivalent slices extracted to ensure consistency between experimental cases. All images were independently rated by two Consultant Neuroradiologists; one based at Norfolk and Norwich University Hospitals NHS Trust and the other at Addenbrookes Hospital, Cambridge. Where agreement could not be reached the images were viewed by a more senior Consultant at Addenbrookes Hospital, Cambridge for consensus. Following image selection and abnormality classification, a computer-based, eye-tracking study was subsequently developed to assess diagnostic accuracy and interpretation in stroke CT and MR imagery.

The independent variables of modality (CT, MRI), case severity (acute, subacute, chronic or normal aging control) and influence of clinical information (clinical information given or withheld) were assessed in a within and between participant design in each separate study (1 and 2) and the results were then combined to form studies 3 (CT versus MRI) and 4 (Consultant comparisons). To test for within and between participant differences, each participant in studies 1, 2, 3 and 4 received half patient cases with clinical information and the other half without. Six experimental conditions were outlined to counterbalance the order of presentation of the independent variables (i.e. the medical images, case severity and clinical information received) to control for order effects. A table providing an explanation of the conditions and the counter-balancing of order effects can be located in the following section 2.2.4. The dependent variable considered in the present series of studies was inspection strategy, reported confidence and diagnostic accuracy. All information regarding the influence of clinical information was extracted and analysed separately to form study 5 (chapter 8).

2.2.4 Counter balancing of order effects: Studies 2, 3, 4, 5 and 6

Table 2.2 Counter balancing of order effects

Group A	Info?	Group B	Info?	Group C	Info?	Group D	Info?	Group E	Info?	Group F	Info?
Acute.1	Yes	Subacute 2	Yes	Normal 3	Yes	Chronic 4	Yes	Subacute 5	Yes	Acute 7	Yes
Subacute 1	Yes	Chronic 2	Yes	Acute 3	Yes	Normal 4	Yes	Acute 5	Yes	Normal 6	Yes
Normal 1	Yes	Acute 2	Yes	Chronic 3	Yes	Subacute 4	Yes	Acute 6	Yes	Subacute 6	Yes
Chronic 1	Yes	Normal 2	Yes	Subacute 3	Yes	Acute 4	Yes	Normal 5	Yes	Acute 8	Yes
Subacute 2	X	Normal 3	X	Chronic 4	X	Subacute 5	X	Acute 7	X	Acute.1	X
Chronic 2	X	Acute 3	X	Normal 4	X	Acute 5	X	Normal 6	X	Subacute 1	X
Acute 2	X	Chronic 3	X	Subacute 4	X	Acute 6	X	Subacute 6	X	Normal 1	X
Normal 2	X	Subacute 3	X	Acute 4	X	Normal 5	X	Acute 8	X	Chronic 1	X
Normal 3	Yes	Chronic 4	Yes	Subacute 5	Yes	Acute 7	Yes	Acute.1	Yes	Subacute 2	Yes
Acute 3	Yes	Normal 4	Yes	Acute 5	Yes	Normal 6	Yes	Subacute 1	Yes	Chronic 2	Yes
Chronic 3	Yes	Subacute 4	Yes	Acute 6	Yes	Subacute 6	Yes	Normal 1	Yes	Acute 2	Yes
Subacute 3	Yes	Acute 4	Yes	Normal 5	Yes	Acute 8	Yes	Chronic 1	Yes	Normal 2	Yes
Chronic 4	X	Subacute 5	X	Acute 7	X	Acute.1	X	Subacute 2	X	Normal 3	X
Normal 4	X	Acute 5	X	Normal 6	X	Subacute 1	X	Chronic 2	X	Acute 3	X
Subacute 4	X	Acute 6	X	Subacute 6	X	Normal 1	X	Acute 2	X	Chronic 3	X
Acute 4	X	Normal 5	X	Acute 8	X	Chronic 1	X	Normal 2	X	Subacute 3	X
Subacute 5	Yes	Acute 7	Yes	Acute.1	Yes	Subacute 2	Yes	Normal 3	Yes	Chronic 4	Yes
Acute 5	Yes	Normal 6	Yes	Subacute 1	Yes	Chronic 2	Yes	Acute 3	Yes	Normal 4	Yes
Acute 6	Yes	Subacute 6	Yes	Normal 1	Yes	Acute 2	Yes	Chronic 3	Yes	Subacute 4	Yes
Normal 5	Yes	Acute 8	Yes	Chronic 1	Yes	Normal 2	Yes	Subacute 3	Yes	Acute 4	Yes
Acute 7	X	Acute.1	X	Subacute 2	X	Normal 3	X	Chronic 4	X	Subacute 5	X
Normal 6	X	Subacute 1	X	Chronic 2	X	Acute 3	X	Normal 4	X	Acute 5	X
Subacute 6	X	Normal 1	X	Acute 2	X	Chronic 3	X	Subacute 4	X	Acute 6	X
Acute 8	X	Chronic 1	X	Normal 2	X	Subacute 3	X	Acute 4	X	Normal 5	X

2.2.5 Procedure

Visual search behaviour of each participant was monitored using a Tobii X50 remote eye tracker, mounted below the computer monitor, which permits unobtrusive recording of saccadic eye movements. Participant eye movements were first calibrated on a 5-point scale (as previously discussed in section 2.1). A short presentation was viewed by all, per modality, to give a basic training on the clinical features of stroke, as presented by CT and MR imagery, together with a short lesion identification training exercise. Each participant was then instructed to gaze at a fixation point in the centre of the screen (to regulate the initial gaze point between participants) prior to image viewing. Patient axial slices were presented in the same order (within their counterbalancing condition), and therefore, participants could only scroll down through the 'stack' from the top of the cranium down to the basal spine region. To clarify, although participants navigated down through five images, they only viewed one image at one time thus all five images were not on the same screen at the same time.

Participants were not confined to a time limit by slice or case but were asked to appraise each case thoroughly and reach a decision regarding the presence or absence of an abnormality. Participants rated each case on a four-point Likert scale, namely whether a primary abnormality (i.e. stroke) was; 1) definitely present, 2) probably present, 3) probably absent, or 4) definitely absent. If an abnormality was considered present, participants were required to confirm the location of the infarct on the observer reporting sheet. If a participant considered there to be more than one abnormality present i.e. a secondary infarct, they were required to circle the second location and indicate which was the primary abnormality (i.e. the largest or most recent) with a '1' and which was the secondary (i.e. either an older infarct or a smaller more insignificant infarct) with a '2' on the brain atlas task. Radiology trainees and Consultant readers were encouraged to mark on the brain atlas if they were aware of the presence of small vessel changes and/ or lacunar infarctions, which are incidental clinical findings often associated with normal aging or a precursor to a stroke, and would provide additional information regarding the cortical tissue examined in visual search processing. Recalibration was performed after the appraisal of 12 cases in each modality to minimise the impact or control for drift. A visual representation of the main study protocols can be seen below in figures 2.3 and 2.4.

The maximum time taken for this study was estimated at 1 hour per modality. After examination of all 48 cases, all participants were informed of the specific aims and objectives of the

study, were able to receive feedback regarding their individual eye movements within and between modalities, and were thanked for their time. Where participants indicated their interest, research papers were distributed via electronic mail.

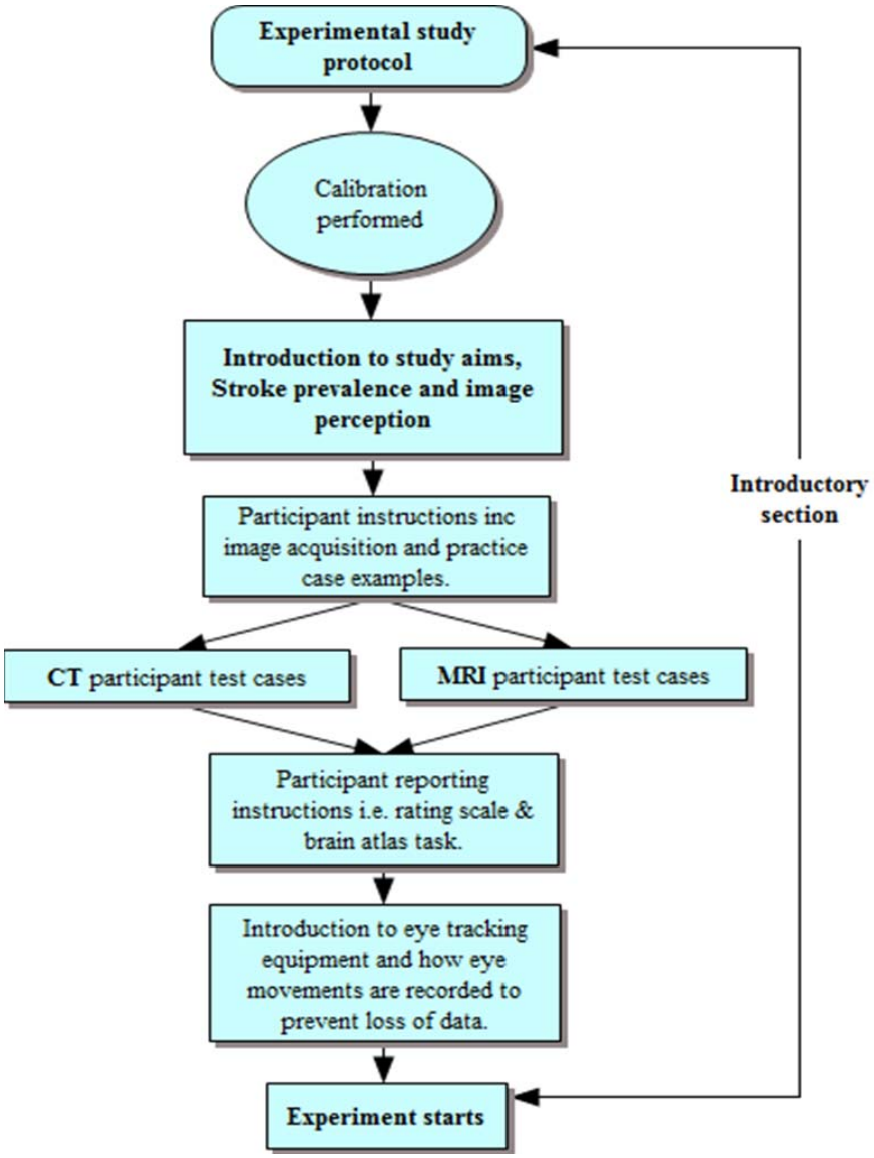


Figure 2.3 Represents the order of information given to participants prior to the start of each experimental condition.

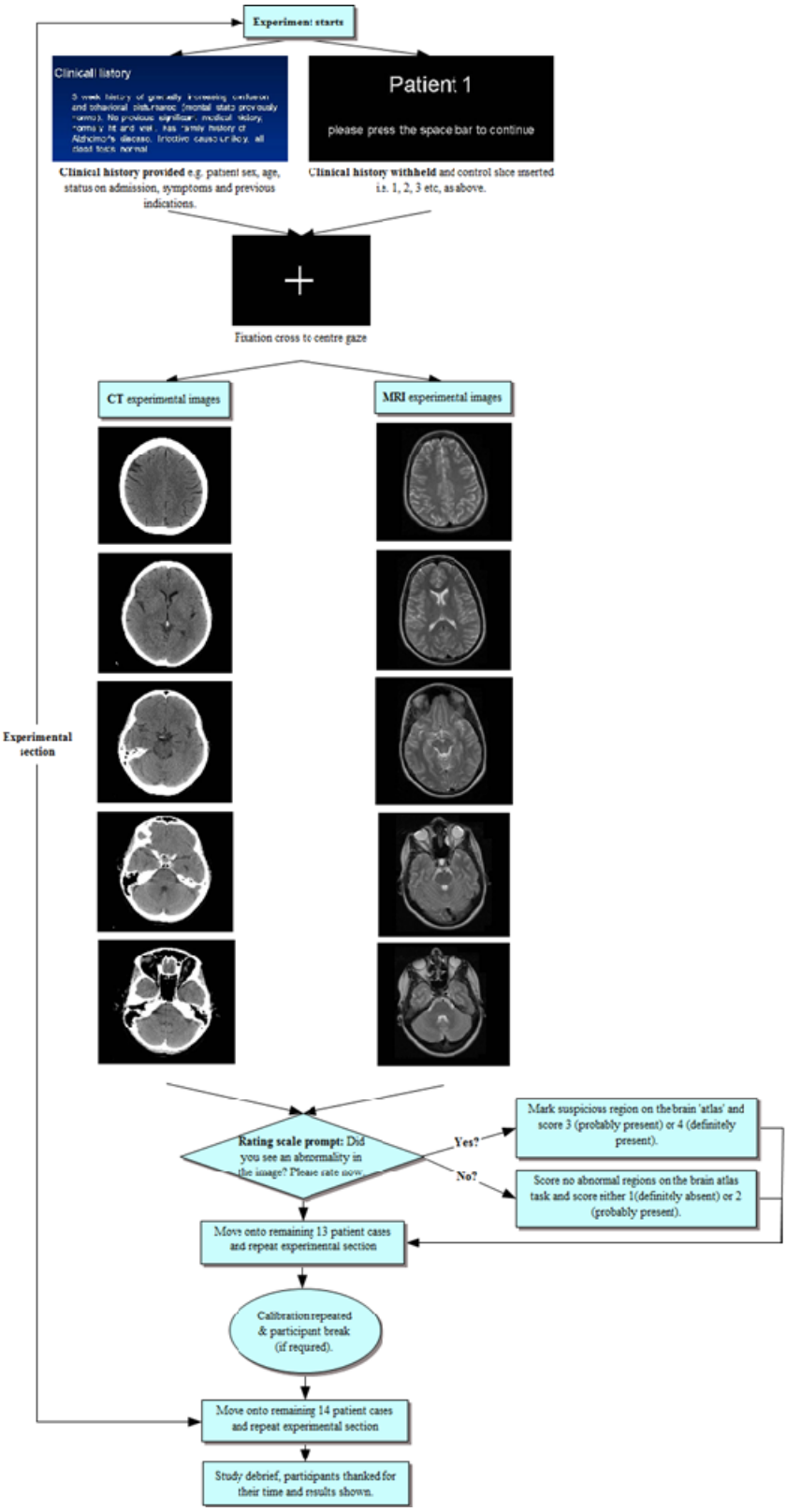


Figure 2.4 represents the order of information given to participants during the eye movement studies.

It is important to note here that all clinical information results were extracted and discussed separately in study 6, which explores the effect of clinical information on image appraisal in detail.

2.2.6 Ethical Considerations

Study invitation letters, consent forms and participant information sheets were prepared to ensure informed consent was granted by all participants. Participants were fully informed of the study aims and objectives. They were also informed that they were not obliged to participate within the study and could withdraw from the study, at any time and without providing a reason. NHS staff and students were informed their treatment within the Trust would not be affected if they did not participate or withdrew from the study.

All protocols and procedures were approved prior to commencing investigation, ensuring adherence to Loughborough University and National Research Ethics (NRes) committee guidelines. Professional indemnity and sponsorship agreements were sought from the University to proceed with Research Governance and Research Ethics procedures in line with NHS Trust requirements. All studies were approved by the Loughborough University Ethics Committee prior to submission to the NHS governing bodies.

Ethical approval was sought on the 8th August 2008 and East Norfolk and Waveney Research Governance Committee provisionally approved the series of studies on the 5th January 2009 (Ref: 2008RAD04S, 133-09-08). Full approval to proceed was officially granted by Norfolk Research Ethics Committee on 22nd September 2009. Following the 13-month ethical approval process, data collection with clinical participants commenced on the 30th September 2009 and ran over a four month period. During this time, the author was resident full-time between the Norwich Radiology Academy and the radiology department, Norfolk and Norwich University Hospitals NHS Trust.

All resultant data held was in accordance with the Data Protection Act (1998); all personal identifiers were removed from data before data entry and analysis and replaced by a detailed coding scheme. Personal identifiers were also removed from any disseminated reports or publications. Information regarding the research was stored on University laptop computers based in the Applied Vision Research Centre, Loughborough University. Computers and personal data files were password

protected. All research data was stored at the University in accordance with the Data Protection ACT (1998).

2.3 Stroke Case Parameters and Scoring Criteria

The following section provides a detailed explanation of the process of defining what classifies a stroke as acute, subacute or chronic and defines the process of converting a brain image into a scoring criterion, which then translates into an area of interest (AOI) being created and attributed to the case. This section also highlights how participant rated each case as true positive, false positive, false negative and true negative, or a combination of the four categories i.e. one true positive and one false positive markings were rated and attributed.

Primarily, it was decided that normal cases would include cases that were free from an infarct but could have small vessel changes present. Acute strokes were those that had occurred and the image acquired within 7 days. Subacute strokes were those that had occurred and been imaged between 7 and 30 days and chronic strokes were those that had taken place and an image acquired following 30 days. All cases were extracted from PACS on the basis of these parameters and the description within the radiology report and clinical case file.

2.3.1 Establishing a 'Gold Standard' and Assessing Inter-Rater Agreement

Following case selection by the Specialist Registrar, each experimental case was independently rated by the lead Neuroradiologist at Norfolk and Norwich University Hospitals NHS Trust and a senior Neuroradiologist at Addenbrookes Hospital, Cambridge. Following completion of case ratings, a Cohen's Kappa statistical test was performed on the confidence ratings for primary abnormalities on all cases to assess the degree of inter-reader agreement. An overall K value of .67 was produced for both modalities combined which shows a good strength of agreement between readers.

Unfortunately, the Kappa scores could not take into consideration the location scores of each reader, only the degree of confidence per case as demonstrated by ratings on a continuous scale. When reader locations were considered, inter-rater reliability appeared stronger. The readers only disagreed on one primary abnormality in CT case CRW and one normal MRI case NPM out of forty

eight cases. Where other disagreements occurred, they were regarding the presence of focal abnormalities and small vessel changes, which can appear with normal aging and may not contribute to the patients' symptoms or significantly affect normal functioning but must be monitored as they can be a precursor to stroke. Both abnormalities are considered fairly subjective as they are very small in nature. Where agreement could not be met regarding cases NPM and CRW, and subtle clinical features, a third and final opinion was sought from the head of department at Addenbrookes Hospital, Cambridge. Following case consideration, abnormal regions were highlighted to ensure ease of identification by the research team. A scoring criterion was subsequently developed following case consensus.

2.3.2 Brain image TO scoring criteria

The following figures 2.5 and 2.6 provide an overview regarding how the scoring criterion for each case was assigned and the degree of location 'tolerance' was considered within the scoring criteria prior to scoring participant ratings.

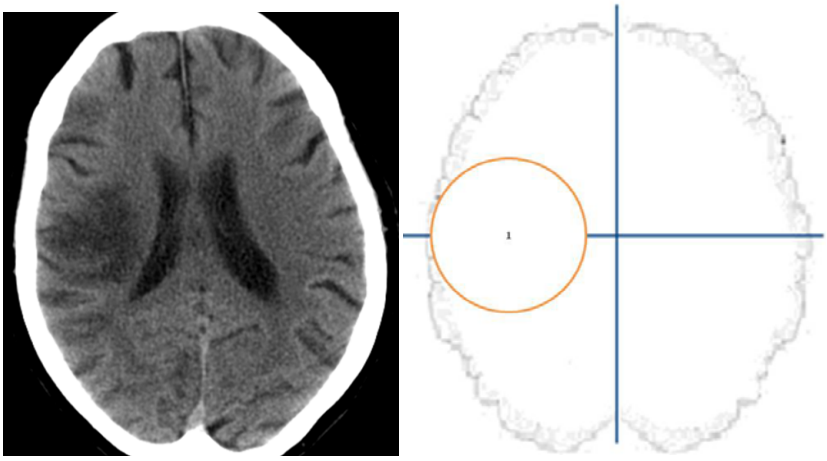


Figure 2.5 The original acute CT image (left) and the resultant study scoring criterion for the case on the right.

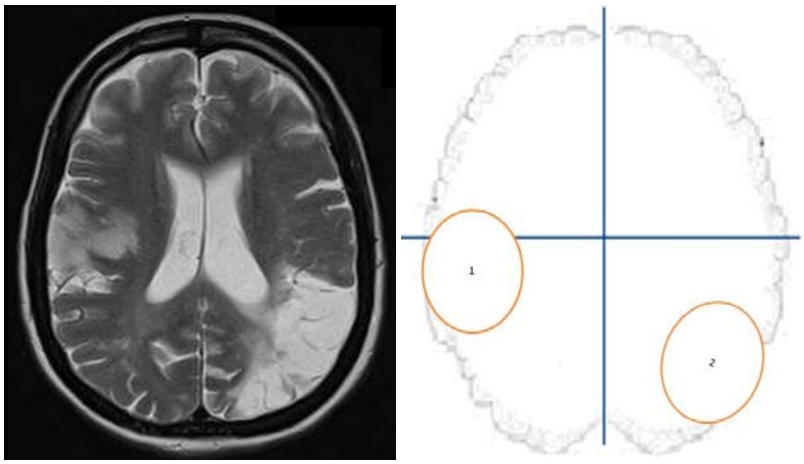


Figure 2.6 The original acute MRI image (left) with two abnormalities present and, the resultant study scoring criterion for the case, demonstrating which abnormality is the primary and which is the secondary, on the right.

2.3.3 Scoring criteria to Area of Interest (AOI)

The following figure 2.7 provides an insight into how the scoring criterion for each case was translated into an area of interest within the Tobii system to capture and record the eye movements of the observers.

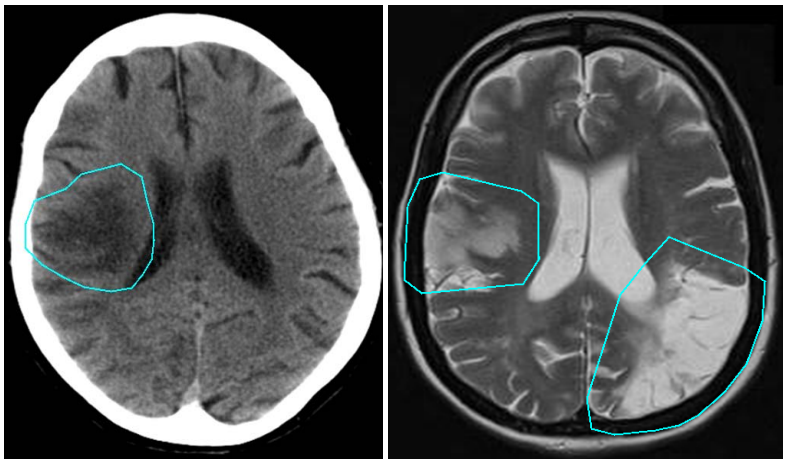


Figure 2.7 Demonstrate how the AOI has been applied to the original images within the Tobii system.

2.4 Data Analysis Organization

2.4.1 Participant Rating Sheets and Marking Criterion

As previously discussed in section 2.2, participants were required to examine each case thoroughly and reach a decision regarding the presence or absence of an abnormality. Upon completion of all cases by participants, the following categories; true positive, true negative, false positive or false negative result were applied to the locations plotted on the observer reporting sheet and examples of the scoring criteria are as follows (see section 2.4.3 for further clarification of decision categories). The following figure highlights a true positive rating scale by an observer who has correctly identified and located not only presence of the primary abnormality, but also the presence of small vessel changes within the original image.

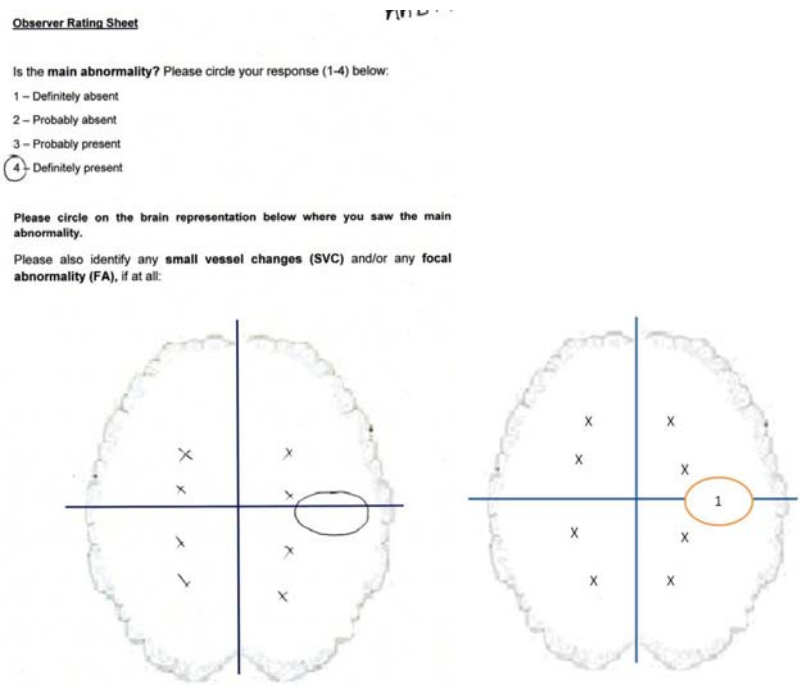


Figure 2.8 An example of a true positive selection.

Where observers correctly ruled out the presence of an abnormality, there were no location markings on the brain atlas and an overall rating of 'probably absent' or 'definitely absent'.

The following figure 2.9 highlights a false negative rating by an observer who has missed or not reported the main abnormality. The figures also highlight a false positive location has been scored by the same observer who has incorrectly circled a suspicious region which is not abnormal.

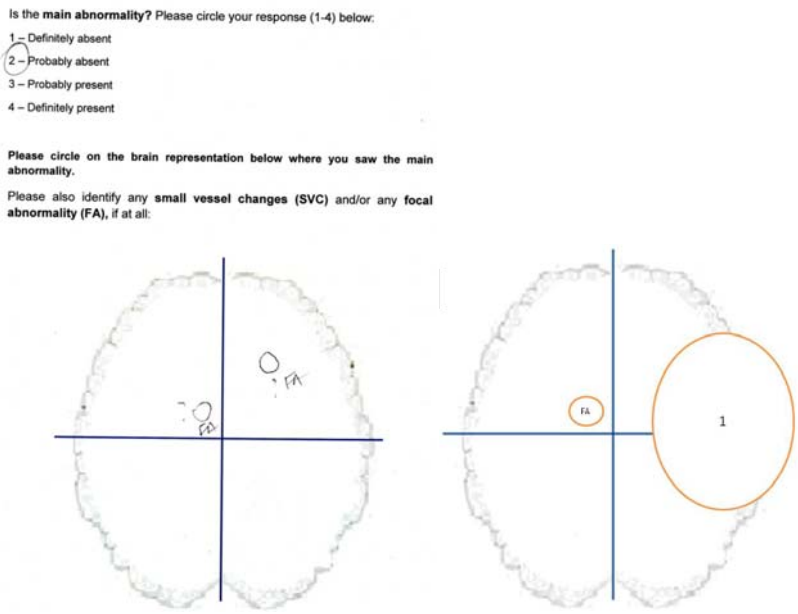


Figure 2.9 Represents an observer who has scored a correct focal abnormality location but has missed the primary (false negative). This observer has also scored an additional focal abnormality false positive.

2.4.2 Scoring of Misclassification Errors

As stipulated at the beginning of the experiment, participants were only required to circle the location of an abnormal area and indicate the degree to which they felt it was definitely present through to definitely absent of a primary infarct. When multiple abnormalities were present, observers indicated which was the most recent or primary infarct. On occasion, trainee or consultant readers made an error of classification i.e. they incorrectly scored a secondary as a focal abnormality or small vessel change as the abnormality was very subtle or confusion was evident regarding what focal abnormalities or small vessel changes were within an image. When very small secondary abnormalities were classed as a focal abnormality, or visa-versa, the observer was still classified as correct. When participants rated a subacute or chronic infarct as a focal abnormality, due to the large differences between the infarct appearance and descriptions, they were classed as incorrect and a misclassification error was attributed. In clinical practice this error of judgement would have resulted in the main abnormality being missed and possibly, the patient being sent home untreated.

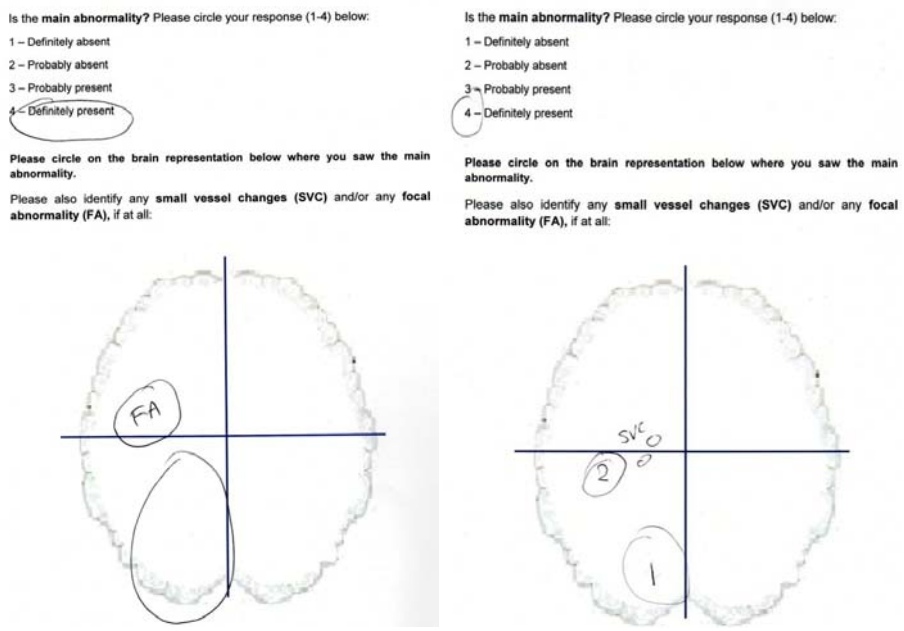


Figure 2.10 Represents an observer who has scored a correct focal abnormality location but has missed the small vessel changes in the case.

The above figure 2.10 demonstrates confusion over a case where the secondary abnormality has been classified as a focal abnormality and/ or small vessel changes have been attributed. As the differences between the abnormalities were very subtle, both observers were classified as correct. In clinical practice, both ratings would have resulted in the identification and correct treatment of the patient.

It is important to note that all scores were entirely dependent upon the location of an infarct, not the confidence rating without the presence of an infarct i.e. if a case was rated 'definitely present' but there was no location indicated or an incorrect area was outlined, the participant was rated as incorrect or attributed a false negative in the latter case.

2.4.3 Accuracy and its' derivation

Throughout this thesis, accuracy is defined as the percentage of correct decisions in the case, stroke type or entire experiment depending on what is being considered in the text or table. Accuracy is

calculated in a number of ways depending upon the question at hand. Primarily however, every correct decision i.e. a true negative decision or a true positive decision for either the primary, secondary or tertiary abnormality (e.g. focal abnormality or small vessel change detection), is either assigned a '1' or a correct decision or a '0' for an incorrect decision. Therefore, for each patient case, accuracy can be derived by adding all correct decisions in a participant group, dividing by the number of participants and then multiplying this number by 100. When accuracy is being derived for a stroke type (i.e. normals), collective average scores are combined, divided by the number of cases in the stroke type category (i.e. 6 normal cases) and then multipled by 100.

When accuracy is being derived for the entire set of cases and for one participant group i.e. novice accuracy in CT, the average scores for each stroke type subset are combined, divided by the number of cases (i.e. 24) and multiplied by 100, providing an overall percentage score of the combined case average for the group and thus is considered baseline 'accuracy' before sensitivity and specificity for the modality is then considered. The obvious drawback of calculating a single figure considered 'accuracy' in this manner is that it does not consider the false positive trade-off i.e. that if 6 locations are specified on the image, it is likely one will be correct, hence the decision to consider *performance* on a number of different dimensions: confidence, number of true positive, false positive, true negative and false negative decisions per case, as well as receiver operating characteristics and sensitivity and specificity cumulative values throughout this thesis.

2.4.4 Confidence and its' derivation

When calculating confidence between and within participants, it is important to consider that high confidence is reflected in low scores i.e. 1=definitely absent, whereas 4=definitely present. For this reason, the weighting of scores was inverted at the final stage to ensure comparability with abnormal cases, whereby '4' is a high confidence scores (i.e. 4=definitely present). Specifically, to calculate the total confidence score for normal and abnormal cases, the individual participant scores are combined to reach a total group score, which are then added for each patient case, and then combined to reach a total for the stroke type category. This figure is then divided by the total number of cases in the category itself and then by the number of participants in the group. Finally, the figure is then divided by the total number of cases and multiplied by 100 to reach the overall confidence figure for the participant group and stroke type. In normal cases where a low score

reflects high confidence, this number is then inverted by subtracting the total number from 100, yet it is not necessary to do this for the abnormal cases where a high score naturally reflects high confidence.

2.4.5 Receiver Operating Characteristic (ROC) Analysis Method

The Receiver Operating Characteristic (ROC) Analysis Method was originally borne out of signal detection theory – the ability to detect a 'signal' amongst background of 'noise'. The basic principles of signal detection theory were applied to armed force advantage in the 1940's Second World War to detect enemy ships where a 'signal' e.g. submarine, was detected against a noisy background such as the ocean.

In 1959, L. Lusted applied this approach to medical reasoning and formulated statistical-decision-theory, which compares actual performance with optimal performance when all relevant perceptual, cognitive and medical information is available and utilised. In 1966, this approach of theorem was introduced into psychology practice (Green & Swets, 1966). To quantify this theory in medical imaging observer studies, participants are required to decide whether a 'signal' i.e. a pathology is present within an image amongst the 'noise' i.e. background anatomical features and give a confidence rating about how certain he or she is regarding the actual presence of the detected pathology. Whether the observer was correct or incorrect in their medical decision determines where their rating is placed within one of the following categories; true positive, true negative, false positive or false negative.

1. A True Positive (TP) decision is attributed when an abnormality does exist within the image and the reader correctly locates its presence.

2. A True Negative (TN) decision is attributed when no abnormality exists within the image and the reader correctly rules out its presence,

3. A False Positive (FP) decision is attributed when an image is rated as being abnormal but no abnormality exists.

4. A False Negative (FN) decision is attributed when abnormalities are present within the image but the image is reported as being normal.

The (continuous) measure of confidence that accompanies the presence or absence of pathology from these four categories, i.e. definitely present through to definitely absent, is then plotted to form a Receiver Operating Characteristic (ROC) curve. Each point in ROC space, from 0-1, is plotted by calculating the sensitivity (True positive fraction or TPF, which forms the Y axis) and specificity (False positive fraction or FPF, which forms the X axis) of each classifier. In accordance with previous work by Metz (1986), overall rates of sensitivity and specificity for each abnormality type, and observer group, were calculated within the present series of studies as follows;

$$\textbf{\textit{Sensitivity}} = \frac{\textit{True Positives}}{\textit{True Positives + False Negatives}}$$

$$\textbf{\textit{Specificity}} = \frac{\textit{True Negatives}}{\textit{True Negatives + False Positives}}$$

The curve created provides a graphical summary of performance, akin to a cost/ benefits ratio of medical decision-making, regarding whether abnormalities were detected, which can be used to compare groups and/ or technical performance (Fawcett, 2006). ROC curves also provide information regarding standard deviations of error (Kundel, 2006). Optimal performance is where the curve is at its highest and curves towards the top-left hand corner i.e. the closer to '1', leaving little room for 'error' within the ROC space. Poor performers are plotted closer to the diagonal line, which represents performance is dependent upon 50/ 50 chance and with 'better than average' to 'good' performers falling between the two curves as plotted in the figure below;

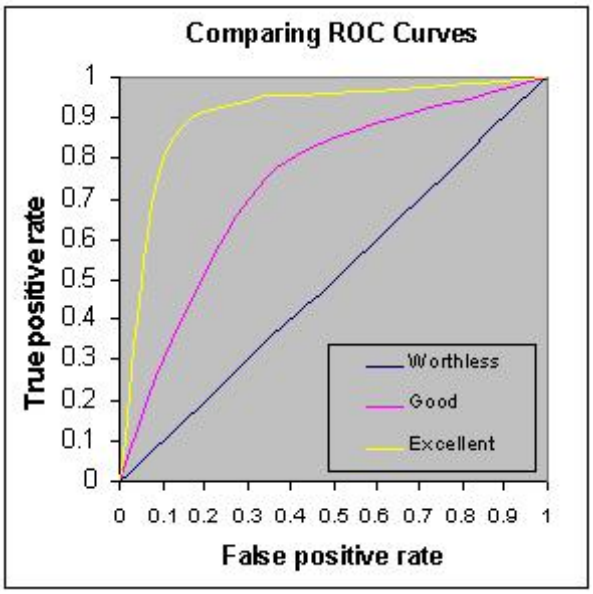


Figure 2.11 Comparing ROC Curves reproduced from Metz, C.E. 1978.

The area under the curve, or AUC, reflects the ability to define clinical manifestations as 'normal' or 'abnormal' (generally taken as a measure of accuracy), whereas the area over the curve relates to signal to noise properties, the observers' task at hand and their decision-making processes, which can be wide and varied, and thus more complicated to 'plot' statistically. The matched-pair ratio of TPF-FPF can be calculated as an AUC single figure, which can be useful when comparing modality performance in an observer task or in engineering to optimise technical performance but can also be considered too reductionist as it does not reflect how sensitivity and specificity can interrelate within the curve itself. For example, Metz (2006) highlighted that ROC curves can cross denote an interaction between rates of sensitivity and specificity between performers or technical assessments.

The ROC method has been widely applied to observer performance studies and is still regarded as a robust measure. The ROC method was converted into a computer program in 1992 by Dorfman, Berbaum and Metz, and subsequently a myriad of programs have been developed dependent upon experimental design and particular clinical circumstances. For example, the original ROC paradigm has been extended and manipulated to form LROC, FROC, AFROC and JAFROC paradigm based on differing assumptions of underlying distributions and to incorporate the correct identification of the abnormal location within the study. In the present series of studies, the decision was made to apply Binormal ROC Curves to the detection of primary abnormalities. To consider the respective detection rates of secondary abnormalities, focal abnormalities and small vessel changes,

the spread of all decisions has been discussed in detail throughout the following chapters. Further implications of this decision have been discussed in the final chapter 9 under limitations and recommendations for future work.

2.4.6 Eye movement Analysis Method

The eye movement data were analysed separately and combined with performance data to examine relationships between accuracy and visual search. The most important visual search behaviours considered in this series of studies were time to reach the AOI (an indication of attention being captured), time spent within an AOI within the first slice where it appears, and time in AOI throughout the whole image stack (as an indication of recognition and decision-making), time spent out of the AOI (as an indication of a thorough appraisal of the surrounding cortical areas), and time spent on each image and case (as an indication of the speed of decision-making) for each level of participant experience.

In this series of studies, the 'first appearing AOI in case' is defined as the slice where the AOI first appears in the image stack. As a rule, it is either the first or second slice in the image for primary AOI's. In summary, the following formulae were applied to the raw gaze data derived from the Clear View software;

1. Time to hit AOI = Time stamp of slice when it first appears on the monitor - start of first fixation within each AOI.

2. Time spent within each AOI in case = sum of all fixations captured within the AOI in each slice.

3. Time spent out of AOI = total sum of fixations – total sum of fixations captured within AOI(s).

4. Total fixations per case = sum of all fixation durations across all 5 images.

5. Total time spent within each case = time stamp of first slice (when it first appears on the monitor) – time stamp of final slice (when it is removed from the monitor).

6. Mean fixation or time to hit calculations is the sum of participant eye movement data divided by the number of observers in each group.

The following Figure 2.12 is a visual representation of how these formulae were applied to the eye movement data within the results sections that follow. A pilot study was conducted to inform the methodology, design, sample size and eye movement analysis which has been discussed in detail in the present chapter.

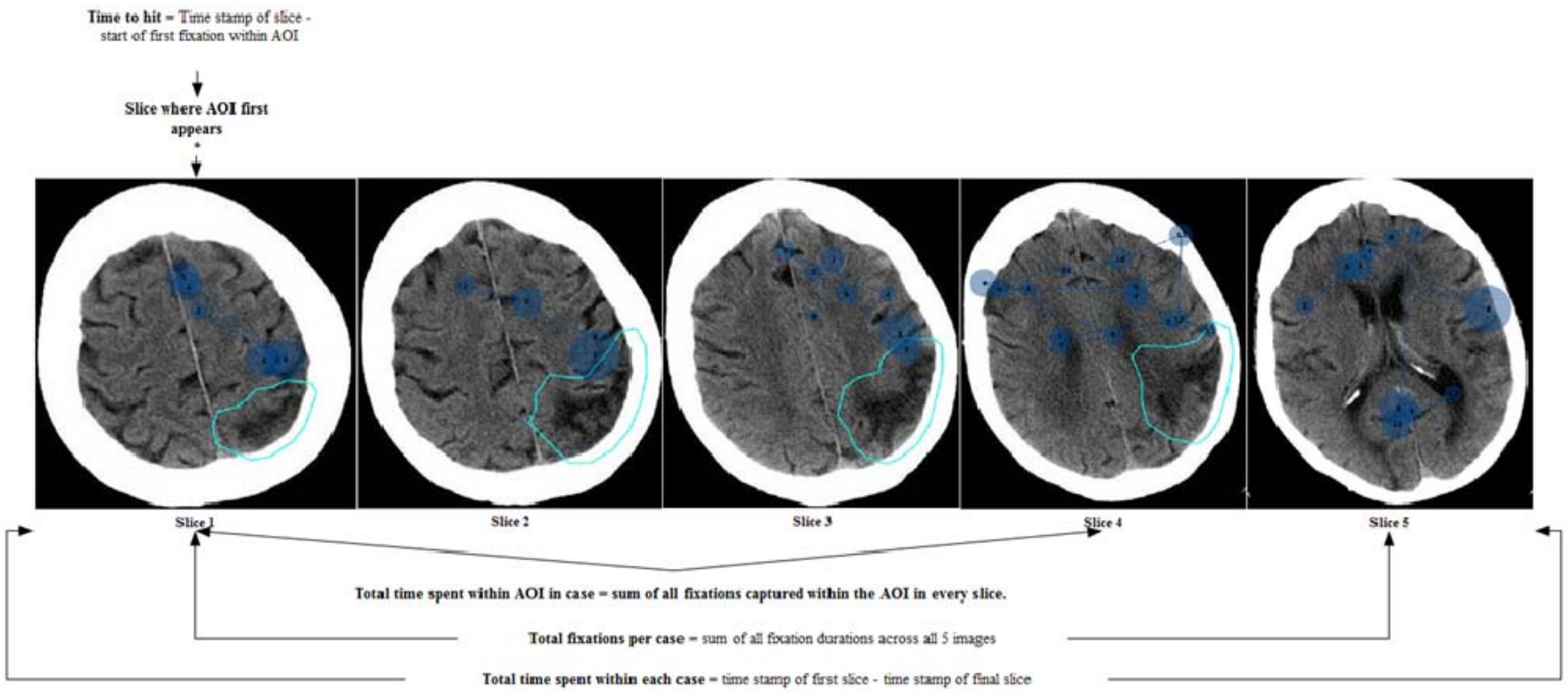


Figure 2.12 Represents how quantitative analysis was performed on the eye movement data within each AOI (defined in blue), image and case for each participant.

Pilot Study: Radiology Image Perception and Observer Performance

How does expertise and clinical information alter interpretation? A feasibility study to explore eye movements in stroke image analysis.

From the foregoing literature, this chapter presents an initial investigation into the visual interpretation of multidimensional from stroke medical images, in relation to observer performance. Observers with differing levels of radiological knowledge were presented with a series of CT and MRI brain images and their performance examined through their visual search behaviour. The difference between levels of expertise in relation to neuroradiological image appraisal was examined, including the influence of clinical information on image perception, interpretation, and diagnostic accuracy.

3.1 Study Aims and Objectives

The specific aims and objectives of this chapter were to conduct a feasibility study to; a) design and develop a full computer-based, eye-tracking experiment using stroke CT and MR images, b) ensure that accurate data could be recorded using the Tobii system, c) collect preliminary data surrounding expert versus novice interpretations of stroke medical images (CT and MRI), when clinical information was provided and withheld, d) ensure that any data collected could be analysed and produce meaningful scientific outputs, and e) gather preliminary data and expert knowledge to support an ethical approval submission to the National Research Ethics Service (NRes).

3.2 Method

3.2.1 Participants

Eight participants were recruited, namely: a) four novice participants, with no prior knowledge or experience of radiological images, who were recruited from staff and research students at Loughborough University; and b) four experienced readers with varying degrees of expertise; one second year radiology trainee based at the Norwich Radiology Academy, and c) three experienced

readers (two neuroradiology registrars, one approaching consultant status and the final participant; a Consultant Neuroradiologist) employed by Norfolk and Norwich University Hospitals NHS Trust.

3.2.2 Design

A computer-based study was designed to assess diagnostic accuracy and interpretation in stroke CT and MR imagery. Eight predetermined clinical cases were selected from a bank of clinical cases; four acute stroke cases; i) 'AST': typical acute stroke, ii) 'ASF': acute stroke with fluent aphasia), iii) 'ASA': acute stroke (alexia without agraphia), and iv) 'ASS': acute stroke with speech arrest), two sub-acute stroke cases; i) 'SSA': subacute with aphasia, and ii) 'SSS': subacute with loss of sensation), one fatal chronic case 'CSF' and one normal control 'NSN' representing normal ageing. Of the eight cases, five axial slices were selected per case were selected and rated by an experienced radiologist to represent between 75 & 100% of the lesion, totalling forty images overall. Axial slices of the control case were selected to match the abnormal areas in the other seven.

For each case, key descriptive statements regarding the patients' age, sex and health status prior to image acquisition were extracted by the researcher from the relevant radiology reports and converted into lay terminology, to assess the impact of clinical information upon medical image appraisal. Clinical information had already been condensed by expert radiologists to accompany the clinical cases; however, information only accompanied half the clinical cases and was withheld for the other half.

The independent variables of availability of clinical information (offered, withheld) and case severity (acute, subacute, chronic or normal aging control) were assessed in a within-participants design. Four experimental conditions were designed to counterbalance the order effects of the medical images, severity of each case and clinical information received. The order of patient case and clinical information given in each experimental condition is outlined in table 3.1. The dependent variable considered in the present analysis was inspection strategy.

Table 3.1 To represent the counterbalancing of experimental order effects.

Experimental Condition 1	Info given?	Experimental Condition 2	Info given?	Experimental Condition 3	Info given?	Experimental Condition 4	Info given?
Acute: Type 1	Yes	Acute: Type 1	X	Chronic	Yes	Chronic	X
Acute: Type 2	Yes	Acute: Type 2	X	Subacute: Type 2	Yes	Subacute: Type 2	X
Subacute: Type 1	Yes	Subacute: Type 1	X	Acute: Type 4	Yes	Acute: Type 4	X
Control Case	Yes	Control Case	X	Acute: Type 3	Yes	Acute: Type 3	X
Acute: Type 3	X	Acute: Type 3	Yes	Control Case	X	Control Case	Yes
Acute: Type 4	X	Acute: Type 4	Yes	Subacute: Type 1	X	Subacute: Type 1	Yes
Subacute: Type 2	X	Subacute: Type 2	Yes	Acute: Type 2	X	Acute: Type 2	Yes
Chronic	X	Chronic	Yes	Acute: Type 1	X	Acute: Type 1	Yes

3.2 Apparatus

Visual search behaviour of each participant was monitored using a Tobii X50 remote eye tracker, mounted below the computer monitor, which permits unobtrusive recording of saccadic eye movements. Images were viewed on an LG Flatron L2000, black LCD monitor (Display area: horizontal; 16.8”, vertical; 12.0”, diagonal; 20.1”). Additional apparatus included a pen, paper and brain atlas sheets for recording location data.

3.2.1 Procedure

The computer-based experiment was designed in a similar format to a PowerPoint presentation; participants were requested to press the space bar to move onto each slide when they were ready. All participants received a short introduction to stroke prevalence in the general population, why this condition is under scrutiny and the research aims and objectives. Participants were made aware they could ask questions at any time during the experiment and they could withdraw their participation at any time, without giving a reason. Participants were informed they could adjust the lighting conditions within the room to suit individual preference.

Each participant was first calibrated on the eye movement system; participants were requested to focus upon a white cross in the centre of the screen (to regulate the initial gaze point

between participants). All participants, irrespective of experience, received a short training on abnormality detection in stroke imaging; highlighting key features of normal and abnormal anatomy. Participants were informed that abnormal anatomy usually affected brain symmetry and an infarct appeared as light or dark 'patches' (dependent upon image acquisition), of varying size and shape, depicting the severity and age of the stroke (acute, subacute or chronic). During the training slides, participants were shown one normal brain image, three textbook examples of stroke images and three abnormal 'test' cases to facilitate detection in the experimental condition.

Once the experiment commenced, participants were requested to examine the eight clinical cases thoroughly and reach a decision regarding the presence or absence of an abnormality. Each participant saw the image slices for each case in the same order and could only scroll down through these slices. To clarify, although participants navigated down through five images, they only viewed one image at one time thus all five images were not on the same screen at the same time. Participants were asked to rate each case on a five-point Likert scale, namely; 1) abnormality definitely present, 2) abnormality probably present, 3) Unsure, 4) abnormality probably absent, 5) abnormality definitely absent. If an abnormality was considered present, participants were asked to circle the affected area on a separate observer rating sheet, provided by the research team (copies of the rating scale can be found in page 242).

After examination of all eight cases, participants were informed of the objectives of the study and informed their data would be made anonymous and held in strict confidence. Participants were thanked for their time, co-operation and if interested, were shown their gaze replay, gaze plot and hot-spot results.

3.3 Results

3.3.1 Image Analysis Results

Case study 1. Acute stroke: The following gaze-tracker images highlight the differences between readers' (novice, trainee and expert) visual inspection strategies when appraising images of acute stroke (AST). N.B. When reading and reporting CT and MRI scans it is important to remember that the left side of the image represents the right side of the individual, and vice versa due to the acquisition process.

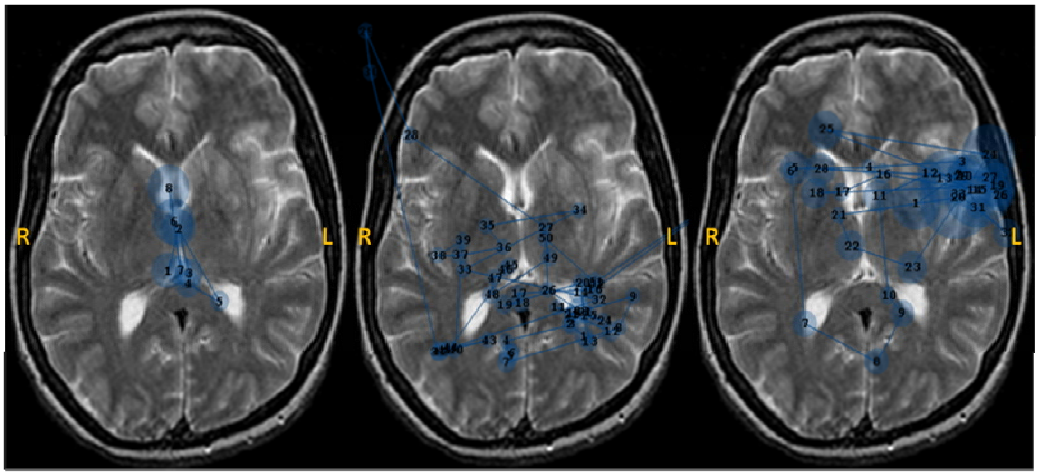


Figure 3.1 highlights the fixation patterns of a novice (a), a trainee (b) and an expert (c) reader when appraising an acute stroke case.

Figure 3.1 demonstrates how novice readers appraise normal, ventricular anatomy to the neglect of cortical tissue. In this image, the trainee reader perceived the AOI (located in the left patient hemisphere) with the 34th fixation and spends a total of 50 fixations on the image, yet the expert fixates upon the AOI with the first fixation from a total of 20 with a structured visual search pattern.

Case study 2. Subacute stroke. The following gaze-tracker images highlight the differences between readers' (novice, trainee and expert) visual inspection strategies when appraising images of subacute stroke with aphasia (SSA).

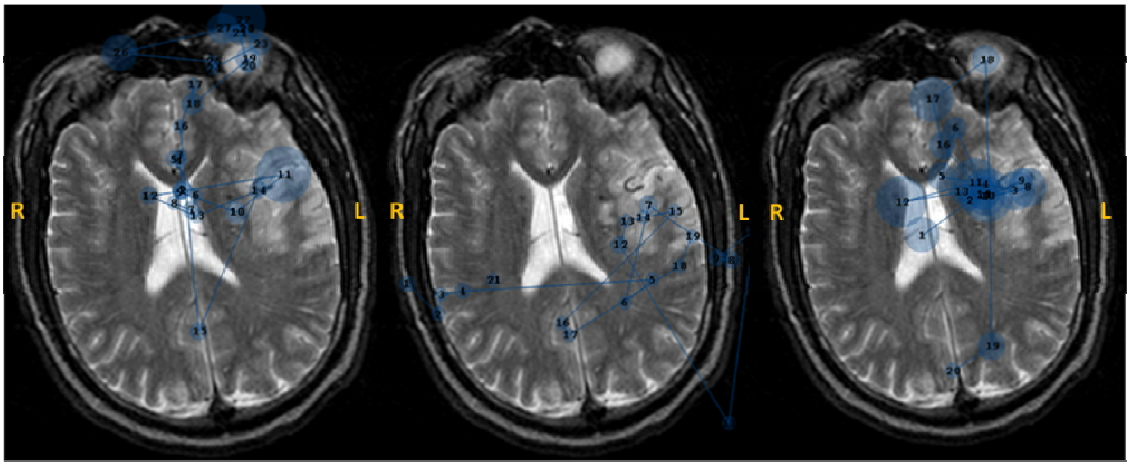


Figure 3.2 highlights the fixation patterns of a novice (a), a trainee (b) and an expert (b) reader when appraising a subacute stroke case.

Figure 3.2 highlights the novice unstructured search pattern over 28 fixations; the reader appears to 'fixate' upon the abnormality once detected and spends a large amount of time inspecting the left eye. Here the trainee fixates upon the image 17 times and 'hits' the lesion by the 12th, whereas the expert 'hits' the lesion by the second fixation.

Case study 3. Chronic stroke (CSF): The following gaze-tracker images highlight the differences between readers' (novice, trainee and expert) visual inspection strategies when appraising images of chronic stroke.

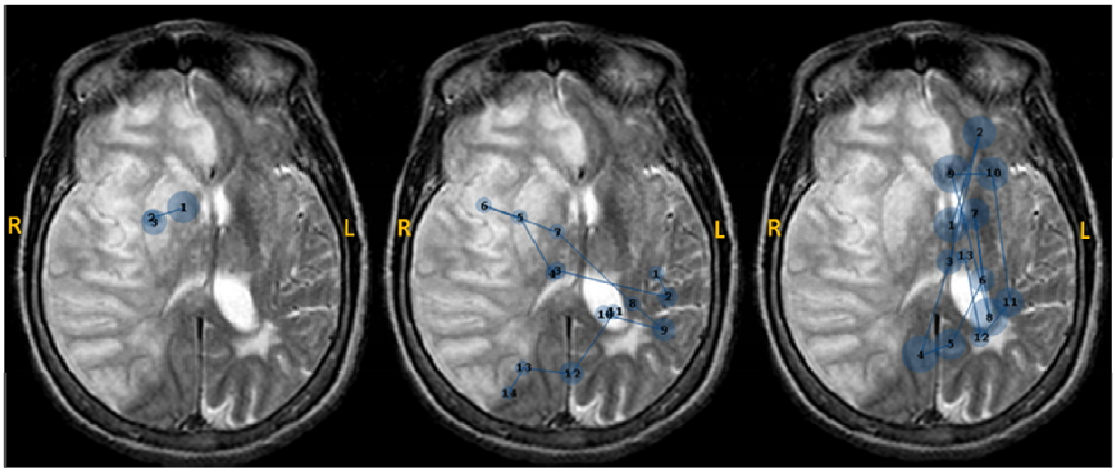


Figure 3.3 highlight the fixation patterns of a novice (a), a trainee (b) and an expert (b) reader when appraising a chronic stroke case.

This case study highlights interesting features of visual search whereby the expert concentrates upon affected tissue having 'seen' the abnormality upon the first fixation and attention then moves towards other affected tissue on the left hemisphere (c). The trainee, as represented by b, also spends more time appraising the affected hemisphere, yet the midline of the brain is not examined as thoroughly compared with the expert. Contrary to both the expert and the trainee, the novice only spends a short amount of time in the abnormal area before rating its presence (a), which might indicate a quick recognition but a lack of consideration for he clinical implications of the patient case.

3.3.2 Confidence Rating Data: Quantitative Analysis

The confidence rating data revealed that the novices detected 65% of abnormalities, the trainee 87% and the experts 96% as being broadly present or absent. As chronicity of case and size of lesions increased then abnormalities were more likely to be reported as present. Overall, participants were

most confident about acute case 2 (ASF) and the chronic case, with all participants reporting true positive results. Participants were most unsure about the first acute case (AST, 3 false negative reports, 1 'unsure' rating) and the control case (NSN; 1 false positive, 2 'unsure' ratings). Expertise was significantly positively correlated with judgements on acute case ASA ($p<.03$) and negatively correlated with the control; NSN ($p<.05$). Novice participants reported the most 'unsure' and false negative responses.

3.3.3 Location Data: Brain Atlas Task

When the location data was examined, the accurate response rate fell by 24%, 23% and 20% respectively for each level of expertise. As table 3.2 below highlights, the tertiary lesion in the second subacute case (SSS) was most challenging to detect, followed by the secondary lesion in the same case, primary lesion in acute case, AST and secondary lesion in acute case, ASA. The primary lesion in the chronic case was the easiest to locate across all three groups. Secondary and tertiary abnormalities in each of the three cases were either not recognised as abnormal and/ or were not reported by some experts; this finding appears to imply there were simple processing errors or they reached a satisfaction of search (Berbaum, 1990). One naïve participant correctly located an abnormality on the reporting sheet, but did not report its presence, whereas most naive participants appeared to examine and rate normal anatomy as abnormal i.e. ventricular anatomy.

Table 3.2 Correctly reported and located lesions (in percentages), by case and participant group.

Case	AST	ASF	ASA		ASS	SSA	SSS			CSF	
Lesion number	1	1	1	2	1	1	1	2	3	1	2
Expert	67	100	100	67	67	100	100	33	33	100	67
Trainee	0	100	100	0	100	100	0	100	0	100	0
Novice	0	75	75	0	75	75	75	0	0	100	75
Lesion detection %	25	86	63	25	75	88	75	25	13	100	38

3.3.4 Diagnostic Accuracy and Clinical Information

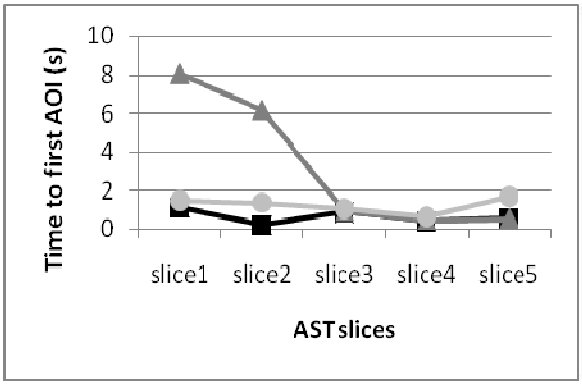
When the trainees' results were combined with the experts and examined with a 2-way ANOVA, there was a highly significant difference between novices and experts on accuracy tests, with a large effect size of experience (*df* 1, $p<.007$, F=16.2, Partial Eta Squared .73). Unsurprisingly, the more

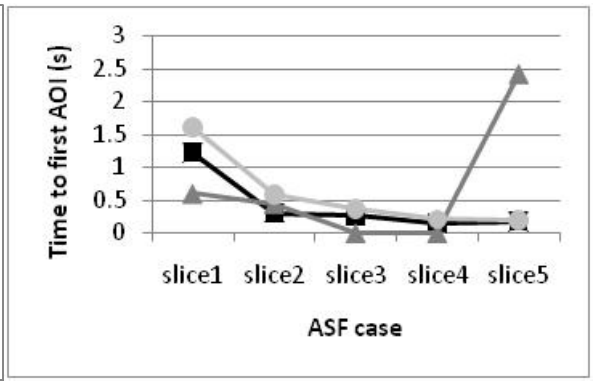
experienced readers were the optimal performers on both tasks, yet errors were still made on the location task, in particular the subacute stroke cases.

Overall, the presence of clinical information did not improve performance, either within-subjects or between them; the within-subject effect was small-moderate (Partial Eta Squared .04) but did not reach statistical significance (p<.620, F=.273). There was a large effect when we examined the interaction of case severity, expertise and information processing on accuracy, which was approaching significance (*df* 1, p<.071, F=4.8, Partial Eta Squared .44).

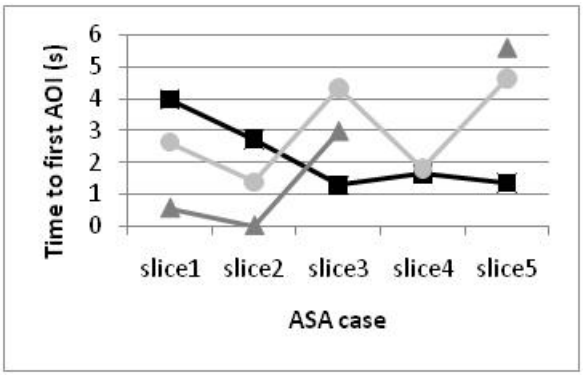
3.3.5 Eye-movement Data: Quantitative Analysis

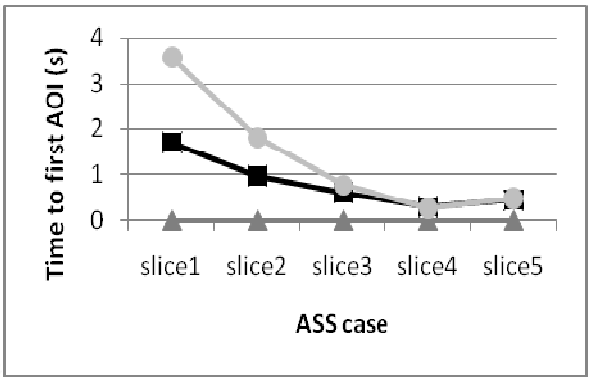
The following figures represent the mean time (ms) for each participant group to fixate the Area of Interest (AOI) around each primary lesion across each image 'slice' for each of the cases. The expert group is represented by a black line with squares; the trainee individual by a grey line with triangles and the novice group by a light grey line with circles.



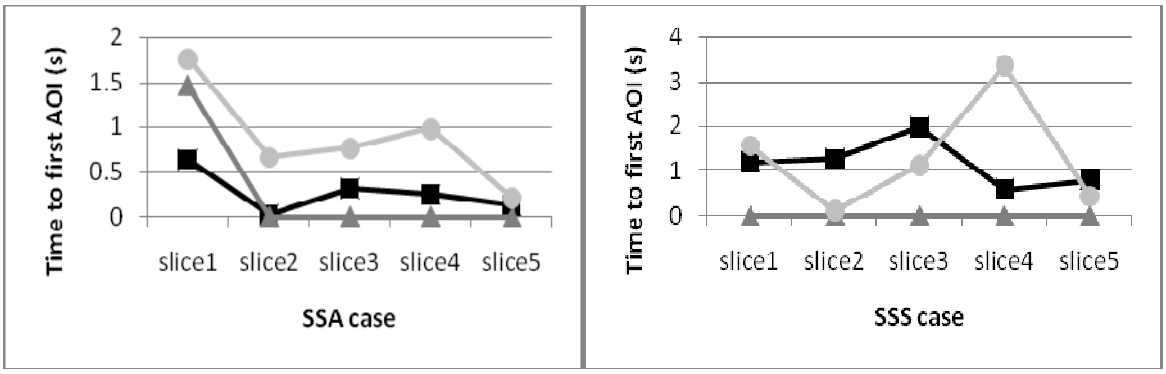


Figures 3.4.1 & 3.4.2 demonstrate mean time to lesion for cases AST and ASF.

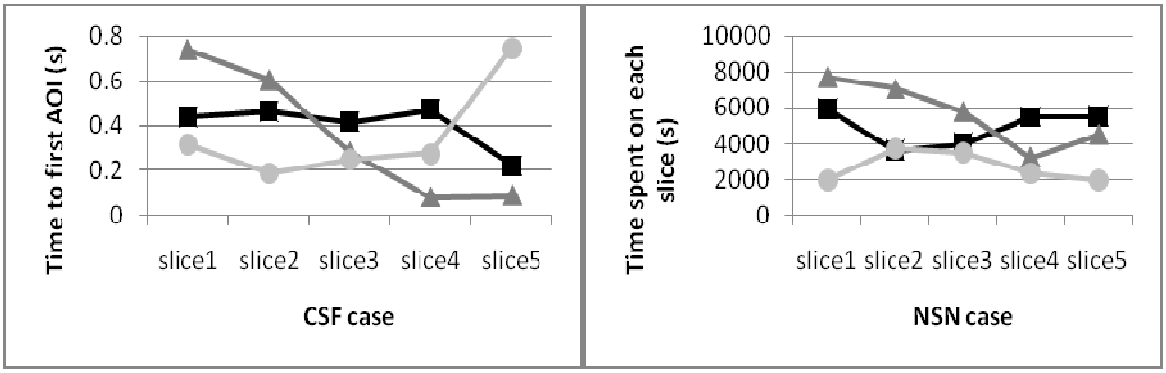




Figures 3.4.3 & 3.4.4 demonstrate mean time to lesion for cases ASA and ASS.



Figures 3.4.5 & 3.4.6 demonstrate mean time to lesion for cases SSA and SSS.



Figures 3.4.7 & 3.4.8 demonstrate mean time to lesion for cases CSF and NSN.

Figures 3.4.1 to 3.4.8 represent mean time (ms) to primary lesion across each image by participant group for AST-NSN cases. N.B. Trainee slice 4 was excluded (29.3ms) for ease of graph interpretation for case ASA (figure 3.4.3). The figures demonstrate that experts were more consistent both within and between cases than novices and trainees; experts operated within a smaller time range than novice participants in every case. Interestingly, novices took twice the amount of time to fixate upon a lesion in half of the cases than did the experts. For AST & SSA cases, the time to fixate the primary lesion was significantly more for novices than experts (*t-test;* AST: $p<.04$, SSA: $p<.02$). The trainee was most inconsistent and the range was largest across four of the seven cases, trainee scores were excluded from t-test analysis.

There appeared a sharp 'drop-off in time to lesion, at which point recognition may have occurred but can only be inferred. Experts appear to detect the lesion quickly in the first slice, whereas recognition appears to occur in the second or third slice for novices and trainees. Time to primary lesion across slices appeared to follow a pattern of longer fixation durations on the first and last slices with less time spent appraising the abnormality on the middle slice, reminiscent of an inverted 'U' effect for experienced observers.

3.3.6 Number of Eye Fixations in Area of Interest (AOI)

Experts fixated more in the AOI than trainees and novices, especially in cases AST and SSA which proved highly statistically significant ($p<.00$). However, this trend was reversed where the lesion was large and obvious; in these cases novices remain fixated on the primary lesion site to the neglect of additional affected tissue i.e. secondary and tertiary abnormalities. There was a trend towards the trainee fixating upon secondary lesions more than experts and novices, which may indicate a degree of uncertainty surrounding additional abnormal features.

Table 3.3 The mean number of fixations within an AOI, mean total fixations per case and percentage of fixations in the AOI.

Case		**AST**	**ASF**	**ASA**		**ASS**	**SSA**	**SSS**			**CSF**
Lesion number		**1**	**1**	**1**	**2**	**1**	**1**	**1**	**2**	**3**	**1**
Expert	AOI	12.4	3.2	1.9	1.3	5.1	9.5	4.6	2.5	2.2	9.0
	Total	21.1	9.1	19.4	-	9.3	15.2	20.1	-	-	13.8
	AOI%	58.8	35.2	16.5	-	54.8	62.5	46.3	-	-	76.8
Trainee	AOI	11.8	3.4	2.4	3.2	0.0	1.6	0.2	4.0	1.4	3.8
	Total	71.4	13.4	55.4	-	7.0	5.8	17.6	-	-	6.8
	AOI%	16.5	25.4	10.1	-	0.0	27.6	31.8	-	-	88.2
Novice	AOI	3.9	3.0	1.2	0.9	3.6	2.9	3.4	2.6	1.8	12.7
	Total	16.2	6.7	13.6	-	12.5	8.8	14.8	-	-	15.2
	AOI%	24.1	44.8	15.4	-	28.8	33.0	52.7	-	-	92.1

In terms of percentage data, experts spent more time in the AOI's than novices and trainees, with the exception of cases ASF, SSS and CSF. In the latter two cases, the trend of novices fixating on the primary lesion was again observed.

3.3.7 Time in AOI versus Total Slice Time

On average, experts spent more time fixating within the AOI than trainees and novices, especially in cases ASA and ASS, which proved significant ($p<.03$ & $.05$, respectively) and AST and SSA cases, which both proved highly statistically significant ($p<.00$). However, this was reversed where the case lesion was large and obvious e.g. novices remain fixated on the primary to the neglect of additional affected tissue such as secondary and tertiary lesions.

Table 3.4 The mean time (seconds) spent fixating within an AOI, mean total fixations per case and percentage of time in AOI.

Case		AST	ASF	ASA		ASS	SSA	SSS			CSF
Lesion number		**1**	**1**	**1**	**2**	**1**	**1**	**1**	**2**	**3**	**1**
Expert	AOI	3.7	1.1	0.6	0.5	1.4	2.6	1.2	0.5	0.5	1.9
	Total	7.3	3.0	6.5	-	3.4	4.7	6.3	-	-	4.1
	AOI%	50.8	35.3	16.4	-	41.9	55.7	34.9	-	-	55.7
Trainee	AOI	1.0	0.3	0.2	0.4	0.0	0.1	0.0	0.4	0.2	0.3
	Total	15.6	3.6	16.3	-	3.6	5.4	2.8	-	-	1.4
	AOI%	6.7	8.4	4.1	-	0.0	2.1	19.2	-	-	29.3
Novice	AOI	0.8	0.7	0.2	0.1	0.7	0.6	0.7	0.5	0.4	3.4
	Total	6.7	2.6	5.5	-	5.2	4.0	4.9	-	-	5.9
	AOI%	11.5	26.7	5.0	-	13.2	13.9	33.5	-	-	62.0

In terms of percentage data, experts spent more time in the AOI (versus time on the slice itself) than novices and trainees, with the exception of case CSF. In the latter case, the trend of novices fixating on the primary lesion was again observed. When total time was considered, the trainee spent the most time appraising each case; 34.8 seconds, on average per case, followed by experts (25.3 seconds) and novices (24.8 seconds).

3.3.8 Eye-movements and Reported Abnormalities

Two experts rated the first acute case (AST) as a true positive and time to first fixation on the first AOI was within 0.7 seconds, whereas correct novice readers took up to 3 seconds. Incorrect (false negative) trainee and expert false negative ratings were both over 2.6 seconds. All rated the second acute case (ASF) a true positive with experts and trainee fixating on the first AOI within 1.5 seconds of seeing the case, novices were within 2.3 seconds to foveal fixation. The trainee and experts rated the third acute case (ASA) a true positive and all fixated upon both AOIs within 5 seconds. Correct novices took up to 10 seconds to 'see' both AOIs. Correct ASA decisions were accompanied by fixations to the AOI within 2.2 seconds, novices were within 4.2 seconds. Fixation time to subacute (SSA) lesion was 0.8 seconds for experts, 1.5 seconds for trainee and up to 3.5 seconds for novices.

Experts took up to 10 seconds to spot all three AOIs in subacute case, SSS, whilst 3 of 4 novices reported this case was a true positive, not all lesions were identified on the atlas task and thus data times cannot be accurately compared with expert data. Novice readers fixated upon the true positive, primary chronic lesion within 0.9 seconds, yet the secondary abnormality was barely fixated upon and not reported. Experts and trainee also perceived the abnormality within 0.9 seconds and took up to 2.8 seconds more to appraise the secondary. Overall, it appears that novices

took twice as much time to reach an AOI than the experts, unless the lesion was large and/ or unambiguous. For the control case, participants who reported a lesion was not present (true negative), spent on average 5 seconds per image and 26 seconds per case. Participants who reported they were 'unsure' or rated the case as 'positive', spent much less time appraising the image (1.9 seconds & 1 seconds) and case (9.3 seconds & 5.1 seconds) than those previous.

Overall, experts detected a primary lesion within the first four fixations on an image (mean fixation time to lesion: 1.5 seconds & average dwell time: 240 milliseconds) a secondary within 13 fixations (mean fixation time to lesion: 4.1 seconds & average dwell time: 261 milliseconds), and tertiary lesions within the first 18 (mean fixation time to lesion: 4.7 seconds & average dwell time: 146 milliseconds). Conversely, novices perceived a primary lesion within the first seven fixations on an image (mean fixation time to lesion: 1.9 ms & average dwell time: 324 milliseconds). In addition, foveal fixations in an AOI were twice as likely to be recorded on the second slice than the first. Only one novice reported a secondary lesion on the chronic case and this was by the second fixation, on the third slice, within 0.7 seconds. Collectively, experts 'saw' but did not report the location of 8 (out of a possible 36 across individual experiments) lesions and fixated upon these areas for an average of 219 milliseconds. Where recognition did occur and was reported, the first fixation within an AOI was slightly higher at 242 milliseconds.

3.4 Discussion

3.4.1 Image Analysis and Eye Movements

As supported by the image analysis and eye-movements results, it was apparent that experts employed a strategic visual search pattern when identifying abnormalities in these stroke medical images; they also spent more time looking in the area of interest surrounding a lesion than novices and trainees, unless the abnormality was large and unambiguous. Experts detected an area of interest as being abnormal, directed their attention and quickly moved on to appraise other affected areas of tissue, the midline and anatomical symmetry, as also reported by Manning *et al.,* in chest radiograph images (2006). Thus, experts appeared to operate a system of deduction; ruling out certain areas very quickly, enabling visual attention to be redirected to more interesting clinical features.

Novices on the other hand, spent more time visually examining normal anatomy such as ventricles and/ or primary lesions following recognition. These findings appear to support previous research by Rogers (1995) and Garlatti and Sharples (1998) that novices may not generate enough

clinical hypotheses to consider complex decisions owing to a lack of biomedical knowledge and clinical problem-solving. The radiology trainee appeared to be less structured than the experts and spent more time on secondary abnormalities than novices, which may indicate uncertainty in making a final decision, also owing to a lack of caseload experience. These novice and trainee findings regarding the influence of caseload experience also appear to be in line with previous research by Nodine and Krupinski (1998).

3.4.2 Diagnostic Accuracy, Location Data and Clinical Information Processing

It is unsurprising that experts were the optimal performers on decision-making and the reporting 'brain atlas' location tasks. Experts were more confident in their decisions to refute the presence of abnormalities in the control case and more able to detect challenging abnormalities in acute cases than novices and trainees. Novices also made more false positive decisions and these findings are in agreement with much previous research. An interesting finding emerged as one novice participant located a lesion on the brain atlas task, however reported that the abnormality was unlikely to be present. This case appears to demonstrate that perception and cognition have occurred yet a lack of confidence meant the abnormality was not reported.

A few errors were made by experts on the lesion detection tasks, with secondary and tertiary abnormalities in scans being missed or not reported. This finding could potentially be linked with the satisfaction of search phenomenon (Berbaum, 1990), and will be explored in more detail in future studies with larger numbers of patient cases. In addition, there was a trend towards clinical information affecting decision-making, although this was not statistically significant, supporting research by Good (1990) and Tudor (1997), and therefore, the effect of clinical information and will be explored in future studies with improved statistical power.

3.4.3 Expertise and Eye-movements

Overall, experts were quicker to detect the primary lesion, and with the exception of the challenging ASA case, foveal fixation occurred within 2 seconds on the first slice of each case image 'stack'. For this acute case, time to lesion dropped to 2.7 milliseconds on the second slice from 3.9 milliseconds, and to 1.2 milliseconds following appraisal of the third. Experts were more consistent and operated within a smaller time range than novices and trainee. It was interesting to evaluate the 'shoulder effect' i.e. the sharp reduction in fixation time that occurred following, what may be inferred as an indication of lesion recognition and an apparent levelling out of fixation time thereafter within the

five image slices. Such observed trends will be examined in more detail within and across different types of stroke cases in future studies.

The number of foveal fixations in a predefined area of interest increased with level of expertise. Experts also had more fixations per case than novices and the trainee, however, the total number of fixations per slice was variable between cases, owing to differences in lesion type, size and number and therefore, a clear conclusion regarding an association between the number of fixations per image and level of experience could not be drawn in this study. Novices spent less time in an AOI, per image and less time overall than trainee and experts. The trainee spent the most time appraising all patient cases, followed by the experts, and finally novices.

This was an exploratory study investigating aspects of neuroradiology, particularly concerning taking a visual search perspective to observer performance. In doing so it is recognised that certain limitations apply; namely only a small number of experienced and naïve observers were studied as they examined a set of carefully selected images. In normal clinical practice radiologists would scroll up and down image stacks for each case, which was not permitted here for experimental reasons.

3.5 Conclusions and Recommendations

In this initial study of observer performance in examining CT and MR stroke images significant differences in visual search behaviour and performance were found. This included differences in search patterns and image coverage between novice, trainee, and expert observers including differences in foveal fixations, dwell and overall case time in the detection and interpretation of these exemplar stroke images. In addition, the study examines observer behaviour through multiple images per patient case rather than appraising single images.

In terms of whether the study met its original aims and objectives, a full computer-based, eye-tracking experiment was designed and conducted using stroke CT and MR images. Accurate data was recorded using the Tobii system and preliminary data was collected, analysed and discussed, which did produce meaningful scientific outputs that highlight interesting areas for further investigation. The results also informed the overall submission to the National Research Ethics Service (NRes), which was subsequently approved.

Study 1: The Influence of Expertise in Stroke CT Interpretation and Eye-Tracking

The imaging of neuroradiological conditions has received much attention from radiologists to compare rates of sensitivity and specificity within modalities. At present the sensitivity and specificity rates of CT performance between studies are variable; at 85% and 91% (Haughton *et al.*, 1986), 45% and 100% (Gonzalez, 1999), and other studies have demonstrated a similar sensitivity of less than 50% for CT (Mohr, 1995; Lansberg, 2000; Wintermark, 2007). Whilst stroke presentation has received attention from researchers to compare sensitivity and specificity rates (Mohr, 1995; Lansberg, 2000; Mullins, 2002 [b]; and Wintermark, 2007), observer performance and visual search in the examination of stroke images has not been previously explored.

Until very recently eye movement analysis across multidimensional images in radiology had not been reported (Phillips *et al*., 2005) despite Computed Tomography (CT) being the modality of choice for emergency and out-patient identification, treatment and stroke intervention. The previous chapter assessed the feasibility of a multislice study to examine observer performance and visual search within this area and the present study aims to uncover whether the feasibility study findings remain consistent when patient cases and participants are increased.

Study Aims and Objectives: To explore the visual interpretation of brain computed tomography images; observer detection and omissions of infarctions using eye-tracking, including, how experts appraise an image compared with radiology trainee and non-experts.

4.1 Methods

4.1.1 Participants

This study required a selection of participants with a comprehensive range of experience of reading medical images; therefore, three groups of participants i.e. novice, trainee and expert readers were identified and recruited. In total, 28 participants were recruited; ten novice readers, ten Specialist Radiology Registrars (trainees) and eight Consultant Radiologists (experts). 50% of novice participants were female and 50% were male. Most novice participants were between 20-30 years of

age. Half of the novice group were Ph.D. students based at Loughborough University, 30% were Research Associates and 20% were in managerial roles; head of research and design and managing director. These participants had no prior experience in medical image assessment and were therefore naïve of both the study aims and the medical reading task itself

Of the radiology trainees, 60% were male and 30% female. Most participants were between 30-40 years of age. Five participants were in the first year of the radiology training programme, one participant was in the second year, three participants were in the third year and one trainee was in the final year. All trainees in their second, third and final years had received training regarding the reading of neuroradiological images, however, only the third and final year students had completed their 3-month neuroradiology training. Those in the first year had only received the neuroradiology lecture series. The total number of CT cases assessed by trainees within the year prior to the study was 838. The range of CT cases assessed was between 4 and 253 cases per person, with the median number of cases being 54.

Of the expert reader group, 75% were male and 25% female. Most participants were between the ages of 40-50 years of age. 50% of consultants were neuroradiologists and the other 50% had a special interest in neuroradiology but performed other types of image assessments on a more frequent basis i.e. breast and chest. Radiologists had been qualified for an average of 15 years. The total number of CT cases assessed by radiologists within the year prior to the study was 1,318. The range of CT cases assessed was between 53 and 591 cases. The median number of cases was 170.

4.1.2 Design

In line with the overall research design, as detailed in chapter 2, one-hundred and twenty single CT clinical images were selected and made anonymous from a bank of twenty four predetermined clinical cases (five single images were selected from each case). The clinical cases selected represented a spread of six normals (controls), eight acute cases, six subacute cases and four chronic cases. The cases were selected by the Specialist Registrar on the basis of the radiology report information, which issued the final diagnosis and stroke classification. Following image selection and abnormality classification, a computer-based, eye-tracking study was subsequently developed to assess diagnostic accuracy and interpretation in stroke CT imagery.

4.1.4 Procedure

Prior to case assessment, a short training exercise was conducted regarding infarct location to provide baseline knowledge regarding the clinical features of stroke, as presented by CT imagery. When case assessment did commence, participants viewed each of the five images in sequence – from the top of the head to the bottom, with only one image being viewed at any one time. Each clinical case was then rated on a four-point Likert scale, namely whether a primary abnormality (i.e. stroke) was; 1) definitely present, 2) probably present, 3) probably absent, or 4) definitely absent. If an abnormality was considered present, participants were required to confirm the location of the infarct on a separate brain atlas task (please see page 241 for the observer reporting sheet). In addition, radiology trainees and consultant readers were encouraged to mark on the brain atlas if they were also aware of the presence of small vessel changes with a single or multiple 'x' on the reporting sheet. For further information regarding image and case type selection, group allocation and experimental procedure, please refer to chapter 2.

4.2 Results

All patient and participant personal identifiers were removed before data entry and analysis. Data were analysed to investigate; i) qualitative image analysis ii) accuracy and confidence ratings of performance, iii) quantitative eye movement analysis, and iv) stroke, expertise accuracy and visual search in CT imagery.

4.2.1 Qualitative Image Analysis in CT imagery

The following four case study examples illustrate the qualitative differences in visual search between the three reader groups.

Case study 1. Normal control case (NBH): The following gaze-tracker figure highlights the differences between readers' (a novice, a trainee and an expert) visual inspection strategies when appraising images of normal control cases in CT. N.B. When reading and reporting CT scans it is important to remember that the left side of the image represents the right side of the individual, and vice versa due to the acquisition process.

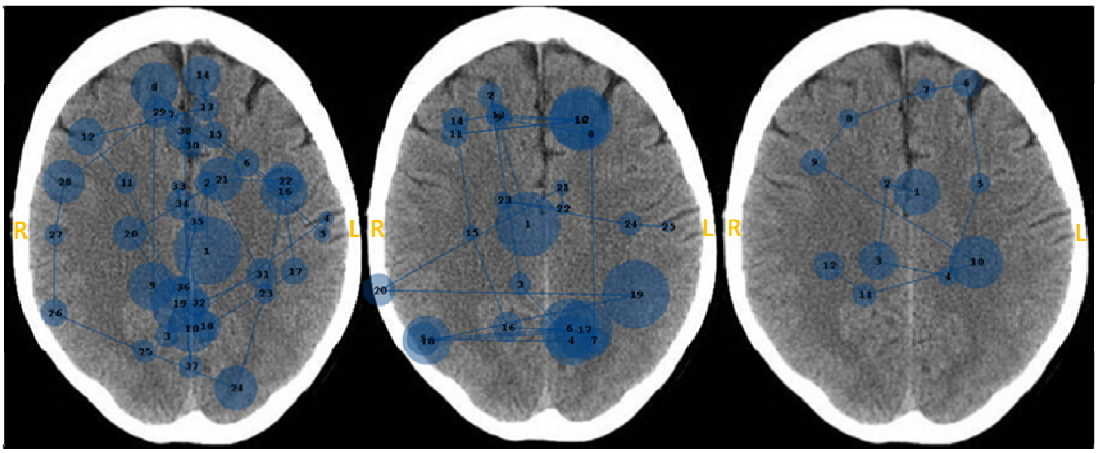


Figure 4.1 highlights the fixation patterns of a novice (a), a trainee (b) and an expert (c) reader when appraising a normal case.

The first, novice image (a), demonstrates a saturated appraisal of the image, which based upon the number and duration of total fixations, may indicate considerably decision-making and/ or confusion by the reader. The second, trainee image (b), highlights an approach which seems more structured than the novice reader as the trainee examines nearly all areas of the cortex, yet the third, expert image (c) is marked by fewer and quicker fixations indicating a quick true negative decision compared with the novice and trainee. Differences evident particularly between trainee and expert readers appear to indicate individual trainee fixations are longer in each spot than experts; to view all gaze tracker images for this case, please refer to pages 5, 6 and 7 of the appendix.

Case Study 2. Acute stroke (AAB): The following gaze-tracker figure highlights the differences between readers' (novice, trainee and expert) visual inspection strategies when appraising images of acute stroke in CT;

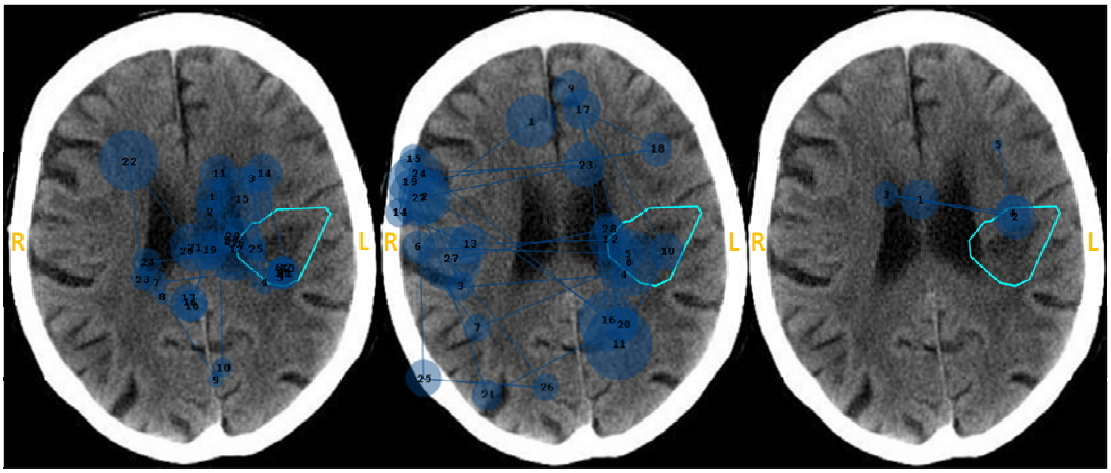


Figure 4.2 highlights the fixation patterns of a novice (a), a trainee (b) and an expert (c) reader when appraising an acute stroke case.

The gaze-plot images highlight differences between readers' visual inspection strategies when appraising images of acute stroke. The first, novice image demonstrates an approach which clusters around the normal anatomy i.e. appraisal of the normal areas of the ventricles, whereas the second figure (trainee reader) highlights much cross comparison to examine differences between hemispheres, despite the abnormality being 'hit' by the fourth fixation. The third image represents an accurate (i.e. true positive) decision made by an expert radiologist who fixated upon the abnormality with the second fixation and paid little attention to the ventricular regions and/ or cross comparing hemispheres for secondary abnormalities and brain symmetry. To view all gaze tracker images for this case, please refer to pages 9, 10 and 11 of the appendix.

Case study 3. Subacute stroke (SDP): The following gaze-tracker figure highlights the differences between readers' (novice, trainee and expert) visual inspection strategies when appraising images of subacute stroke in CT;

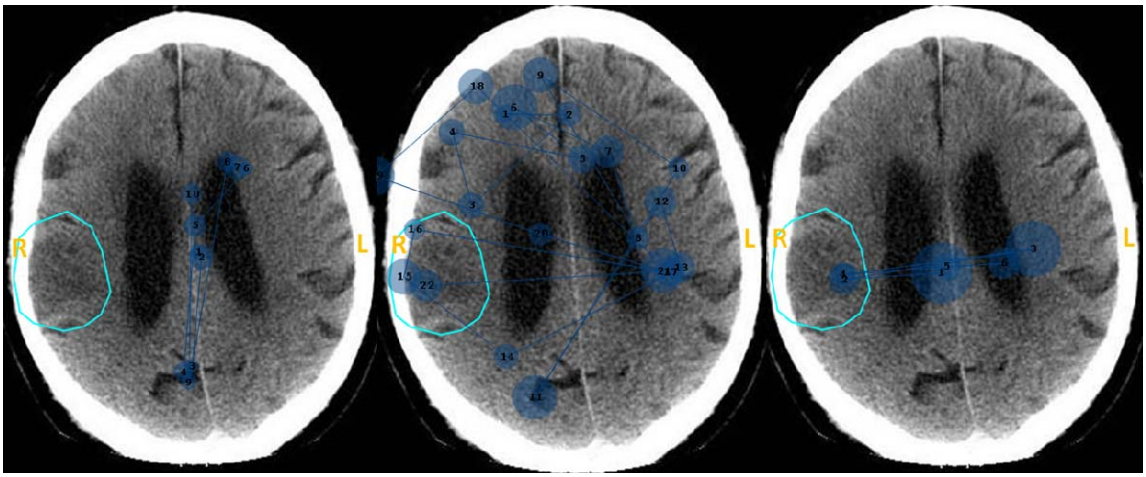


Figure 4.3 highlights the fixation patterns of a novice (a), a trainee (b) and an expert (c) reader when appraising a subacute stroke case.

Figure 4.3 demonstrates that the novice reader (a) barely fixates upon the image and misses the subacute infarct entirely as they look at a normal area of sulci, whereas the trainee compares both hemispheres and many suspicious areas of the cortex as demonstrated by an increased number of fixations and areas examined in image b. The expert reader fixates upon the abnormality with the second fixation and cross compares before making a decision after 5 fixations on the image. To view all gaze tracker images for this case, please refer to pages 13, 14 and 15 of the appendix.

Case study 4. Chronic stroke (CRW): The following gaze-tracker figure highlights the differences between readers' (novice, trainee and expert) visual inspection strategies when appraising images of chronic stroke in CT;

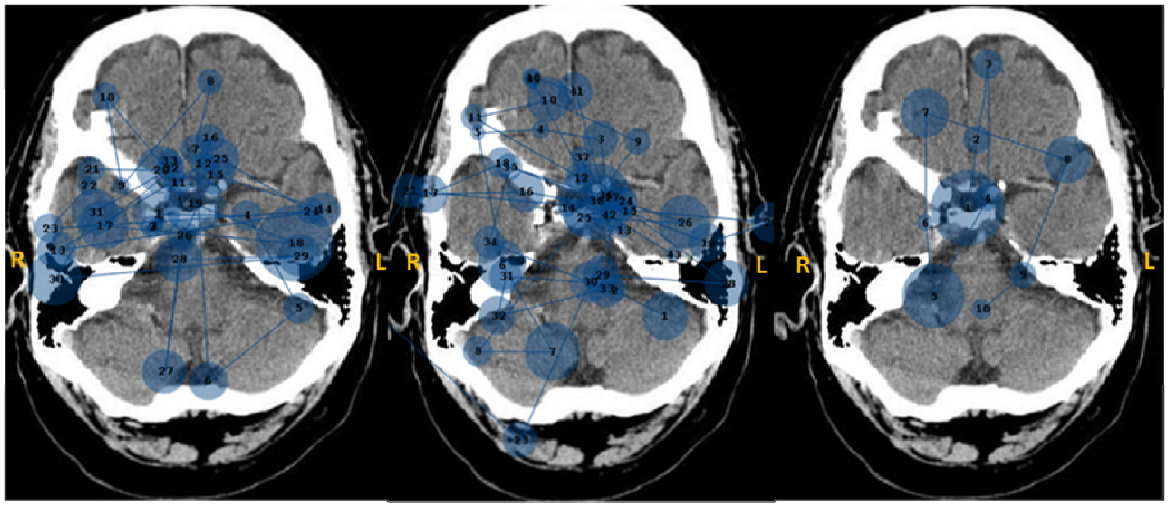


Figure 4.4 highlights the fixation patterns of a novice (a), a trainee (b) and an expert (c) reader when appraising a chronic stroke case.

There is no AOI in this slice to briefly examine differences in search patterns between participant groups where no abnormality exists in the case. Figure 4.4 demonstrates a different search pattern between participants; in this chronic case and image slice, the gaze patterns of the novice and trainee (a & b), do not appear very different. Both participants examine the brainstem; although the trainee spends more time appraising this area, and both participants cross compare both hemispheres for symmetry. The expert reader spends much less time examining the auditory anatomy and the image overall compared with novices and trainees and goes on to focus upon the brainstem, cerebellum and frontal lobes. In this case study, the experts' visual search pattern follows a clockwise pattern around the clinical features accompanied by 10 fixations compared with 38 by the trainee and 33 by the novice, highlighting a structured visual search pattern, which was evident and consistent between images for this particular reader. As a group, experts appear to focus on the brainstem and cortical tissue in fewer fixations than the other groups. To view all gaze tracker images for this case, please refer to pages 16, 17 and 18 of the appendix.

Summary

- Novice visual search patterns demonstrated that they fixated upon interesting high contrast image features i.e. the ventricles, the eye balls or auditory anatomy, even though they may not be aware of what anatomical features they are actually examining. Some novice eye movements appeared in line with trainee radiologist visual search patterns, whereas other either over or under examined the image.

- Trainee visual search patterns highlighted much cross comparison between hemispheres; a behaviour that was apparent in very few novice readers. Trainees also fixated much more than expert readers and the fixations tended to be of longer duration. Trainees spent much more time appraising the anatomy, which might suggest a desire to perform well in the task or it might indicate a level of uncertainty about the clinical features in the image.

- Expert visual search patterns highlighted a quick time to reach the abnormality, often in these cases within 2 fixations and less time spent appraising surrounding anatomy as demonstrated by fewer fixations than novices and trainees. Fixations were also of shorter duration and in the basal slices, fixations were clustered in the brainstem. Experts also examined fewer cortical regions, which might suggest that experts are reliant upon abnormal features to 'pop out' from normal tissue and draw their attention. Experts also appeared to appraise white and grey matter borders, where the texture of the anatomy changes and needs to be examined to ensure it isn't indicative of an underlying abnormality, particularly in case NBH.

4.2.2 Accuracy and Confidence Rating Data: Quantitative Analysis of Observer Performance

The following section examines the accuracy of readers in determining the location of the abnormal areas, their confidence when reporting the abnormal location, the perceived difficulty of the individual cases and differences in performance both within and between groups when comparing stroke case types overall. Please refer to chapter 2 for further details regarding how measures of accuracy and confidence between and within participants were derived.

4.2.2.1 Localisation and Case Difficulty

Table 4.1 compares all readers collectively to gain an indication of individual case and overall perceived case difficulty. Overall percentage of correct ratings indicates that normal and acute cases were judged to be the most difficult by all participants (69.9%), followed by chronic (90.4%) and then subacute stroke types (92.8%). Within the top ten most difficult cases were all six normal cases, three acute cases (ADH, AMW and AAB) and one chronic stroke case (CRW), with case ADH being rated as the most difficult amongst all participants.

Table 4.1 Lists all CT cases with average accuracy scores for all participants and an overall performance rating per stroke type to provide an overall case difficulty rating.

Case TYPE	Individual patient case	% CORRECT	Overall mean case % correct
Normal	NAL	71.7	
	NBH	65.8	
	NDG	75.0	
	NJM	60.8	
	NPA	75.8	
	NYK	70.0	69.9
Acute	AAB	55.0	
	ADD	100.0	
	ADH	30.8	
	AJM	76.7	
	AMW	36.7	
	ASC	86.7	
	ARH	90.0	
	AVB	83.3	69.9
Subacute	SAS	93.3	
	SBD	100.0	
	SCD	80.0	
	SCS	100.0	
	SDP	83.3	
	SGR	100.0	92.8
Chronic	CAJ	95.8	
	CDT	96.7	
	CJF	93.3	
	CRW	75.8	90.4

4.2.2.2 CT Receiver Operating Characteristics for Primary Infarction Detection

The following Conventional Binormal ROC Curves represent confidence ratings for primary abnormality detection by participant group; novice, trainee and expert readers when assessing all cases in CT.

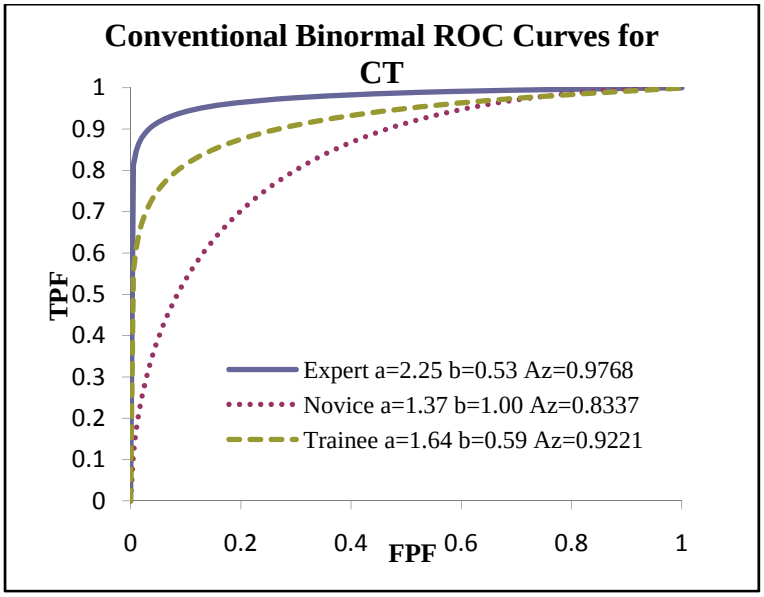


Figure 4.5 ROC Curve for novice, trainee and expert readers in CT

Figure 4.5 demonstrates that the expert readers provided greater accuracy score i.e. higher sensitivity and specificity rates, compared with trainee and novice when examining these CT cases, within this experimental study on the basis of their confidence rating scores that accompanied the cases. Within the ROC space the parameters 'a', 'b' and 'Az' represent the vertical intercept, the slope of the fitted curve and the overall accuracy value as calculated by the conventional Binormal curve process, respectively.

Table 4.2 Diagnostic accuracy of participant groups between normal, acute, subacute and chronic stroke types.

Stroke TYPE	Normal	Acute	Subacute	Chronic	Overall performance
Novice	53.3	58.8	90.0	87.5	72.4%
Trainee	75.0	65.0	88.3	90.0	79.6%
Expert	81.3	85.9	100.0	93.8	90.2%

When considering the exact number of primary abnormalities reported by participants in CT, participants' overall performance and their performance dependent upon stroke case type, table 4.2 demonstrates that experts were most accurate when ruling out or identifying the presence of an abnormality across every case category. Novice participants had more difficulty ruling out the presence of an infarct than detecting the presence of all other stroke types, but detected 31.2% more subacute strokes than acute ones and 1.7% more subacute strokes than trainees. Trainee

participants encountered more difficulty identifying the presence of acute stroke (65.0%) than ruling out an abnormality in control cases (75.0%), and found chronic stroke types easier to detect than subacute cases 90.0% compared with 88.3% respectively, which is contrary to expert performance.

When ANOVA tests were performed to examine between group difference of performance, there were significant group differences for primary infarct detection in CT (*df* 2, *p*<.000, F=17.43, Eta Squared=.05). Games-Howell posthoc comparisons demonstrated that the differences were evident between novices and trainees (*p*<.000), and novices and experts (*p*<.000), but differences were not found to be significant between trainees and experts for CT (*p*<.363).

4.2.2.3 Confidence Rating Data: Quantitative Analysis of Primary Infarctions.

When reporting participant confidence in their decision-making in CT (i.e., 1=definitely absent, 2=probably absent, 3=probably present and 4=definitely present), descriptive statistics show overall percentage of cases correct indicating that acute cases were perceived to be most challenging by all participants (79.1%), followed by normal cases (79.2%), then chronic (93.1%) and subacute stroke types (96.4%). For further information regarding how confidence scores were derived, please refer to chapter 2.

Table 4.3 highlights the confidence of participant groups when making decisions regarding patient cases i.e. normal, acute, subacute and chronic stroke types.

Stroke TYPE	Normal	Acute	Subacute	Chronic	Overall confidence
Novice	48.8	70.6	87.5	83.8	73.3%
Trainee	52.5	83.1	98.3	95.0	81.0%
Expert	63.0	92.2	98.4	98.4	81.5%

When ANOVA tests were performed to examine between group differences of confidence, there were also significant differences between all groups (*df* 2, *p*<.001, F=7.12, Eta Squared=.02), although the effect size was even smaller than diagnostic accuracy differences. Games-Howell posthoc comparisons demonstrated that significant differences were evident between novices and trainees (*p*<.003), and novices and experts (*p*<.003), but not trainees and experts (*p*<.940), therefore, experts are the most confident when performing the inspection task.

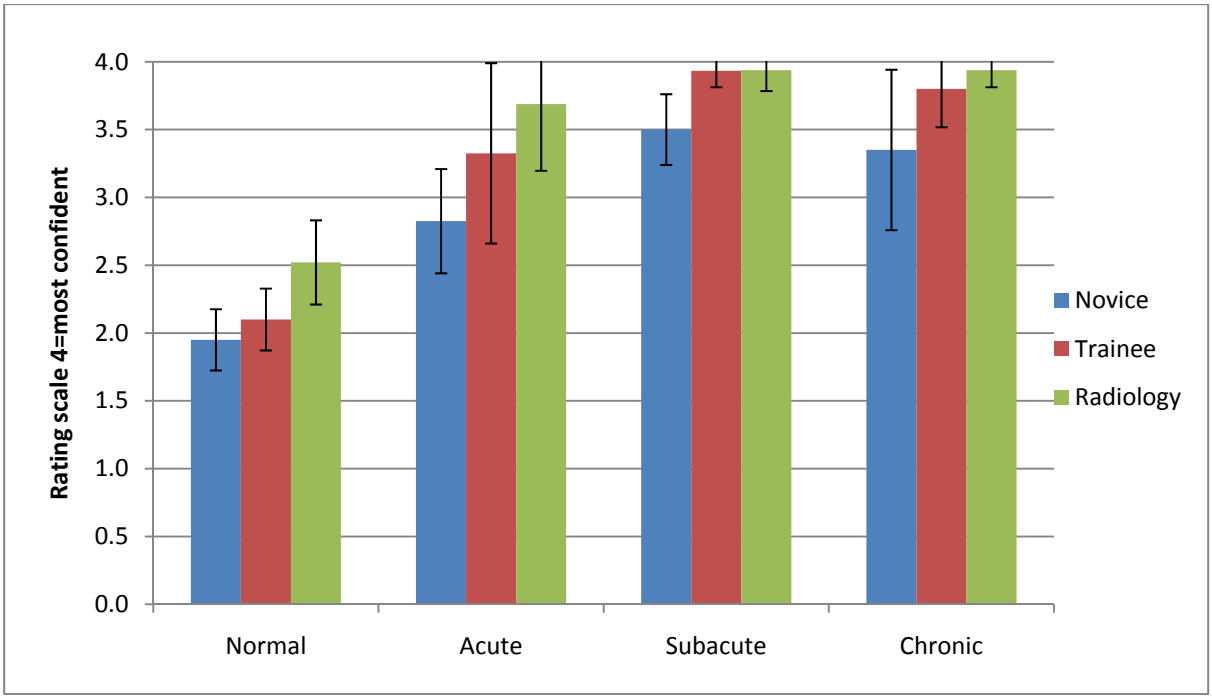


Figure 4.6 represents the mean confidence score per case by participant groups when making decisions regarding patient cases i.e. normal, acute, subacute and chronic stroke types.

The above table 4.3 and figure 4.6 demonstrate that all participants were most cautious when ruling out abnormalities and that the expert readers were most confident across all cases. Trainee readers were equally confident in subacute cases despite the lower overall accuracy levels but were much less confident at ruling out abnormalities than experts, with a result more comparable to novice readers.

4.2.3 Diagnostic Decision-making (i.e. TP/FP/TN/FN)

As demonstrated above, experts are the optimum performers in CT. However, the above data does not highlight the number of false positive and/ or false negative decisions which may be reported in a case. Whilst the above measure of 'accuracy' is essentially the number of true positive and true negative decisions in a case, table 4.4 highlights the spread of all possible decisions in a case to gain a more in-depth insight into their decision-making processes and performance dependent upon level of experience.

Table 4.4 demonstrates observer group decisions (i.e. TP/TN/FP/FN) across all CT cases.

CT Average	TP	TN	FP	FN
Novice (n=10)	13.6	3.2	4.1	5.0
SpR (n=10)	14.8	4.0	3.7	1.4
Radiol (n=8)	17.3	4.3	1.8	0.8

Unsurprisingly, radiologists had more true positive and true negative ratings than novices and trainees. Radiologists also had fewer false positive and false negative results also. Overall sensitivity rates between groups were; 82%, 91% and 96% for novice, trainee and expert readers. Specificity rates were 42%, 50% and 71%, respectively. Thus highlighting that whilst trainees were better at detecting abnormalities than novices, not much separated the two groups when ruling out the presence of abnormalities.

4.2.4 Eye-movements and Experience: Quantitative Analysis.

The primary aim here is to identify how visual search differs quantitatively between novice, trainee and expert readers in CT images of normal patients and those who have suffered a stroke. The secondary aim here is to identify statistically significant links between visual search behaviours (e.g. total time spent on the task, time to reach an AOI, time spent within and out of AOI's) and reported accuracy within and between participant groups in CT to gain an insight into the way inexperienced through to experienced readers appraise these images.

4.2.4.1 Total Case Viewing Time per Group

When considering the total viewing time per reader group within this study, there were highly significant differences between all groups when comparing total case fixation durations (*df* 2, $p<.000$, F=21.88, Eta Squared .77), with large differences evident between novice and trainee readers ($p<.016$), trainee and expert readers ($p<.000$) and novices and experts ($p<.000$); with novices fixating for longer than trainees, but much longer than experts. Trainees also fixated for longer than experts over the entire case selection. The prior qualitative image analysis results indicated that as experience increases, the number of fixations per case declines. The following ANOVA results agree with this assumption; there were highly significant difference between groups (*df* 2, $p<.000$, F=11.341, Eta Squared.03) with novices fixating much more, and for longer, than trainees ($p<.001$) and experts ($p<.000$).

4.2.4.2 Visual Search Behaviour throughout the Image 'Stack'.

The following figures 4.7-4.10 illustrate the visual search behaviour of the participant groups by investigating how the mean fixation duration of gaze differs between each axial slice throughout the five image 'stack' by case type (control, acute, subacute and chronic) and level of experience (novice, trainee and expert).

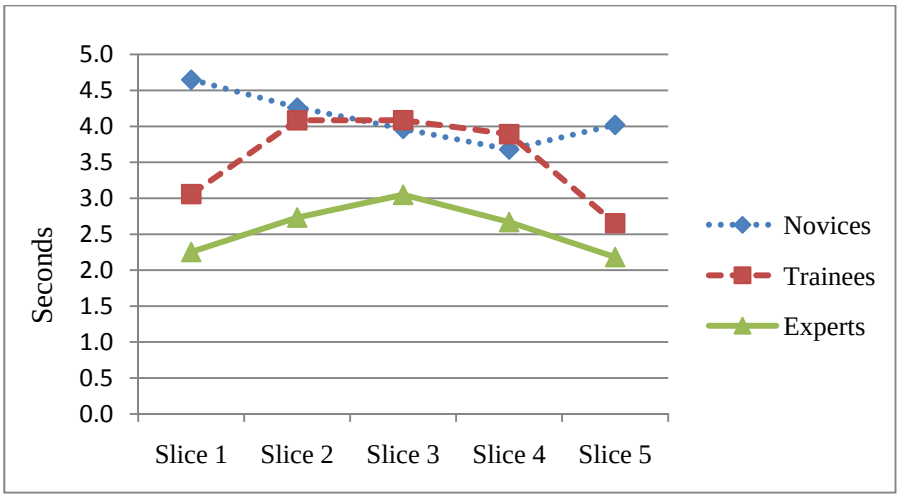


Figure 4.7 Mean fixation time per axial slice for all readers of normal patient CT images.

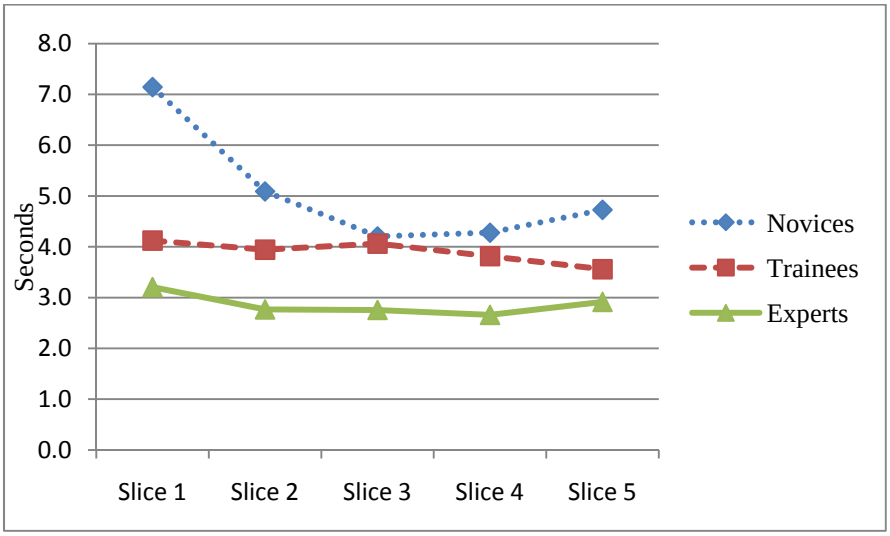


Figure 4.8 Mean fixation time per axial slice for all readers of acute patient CT images.

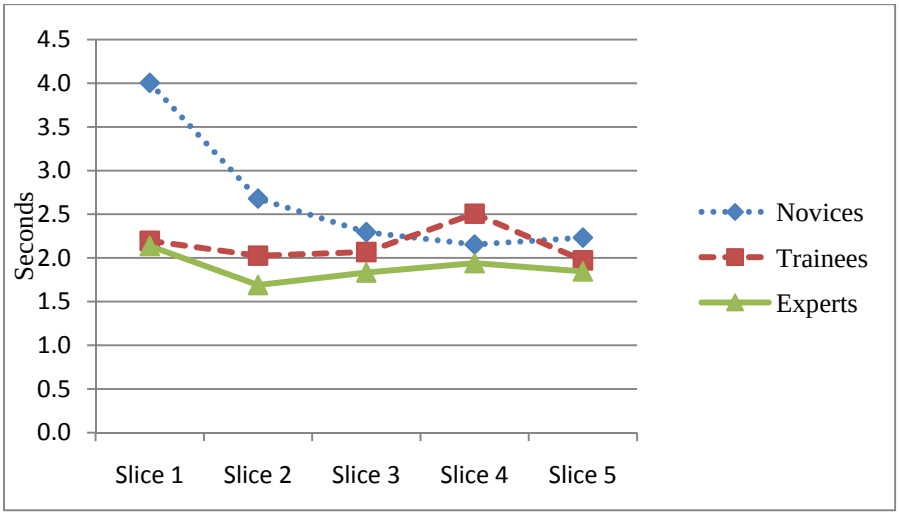


Figure 4.9 Mean fixation time per axial slice for all readers of subacute patient CT images

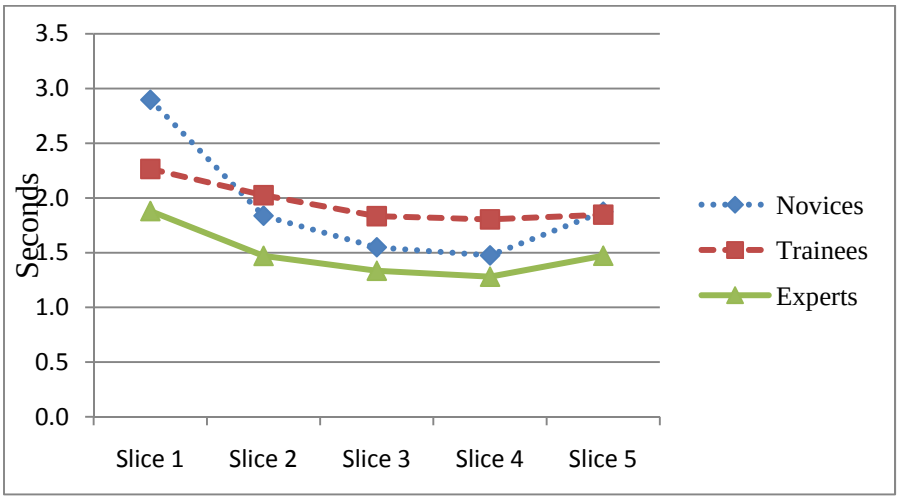


Figure 4.10 Mean fixation time per axial slice for all readers of chronic patient CT images.

A clear trend appears within the above figures towards trainee and expert readers following very similar fixation times throughout the image 'stack', despite the difference in stroke type. Although these readers follow a similar trend, in normal, acute and chronic cases in particular there appears a difference of up to 1 second between the fixation times within each axial slice. In figure 4.7 (control images), trainee and expert readers fixate for longer in slice three than slice 1 and 5 respectively, whereas novice readers spend more time appraising slice 1 than 5 and 3. In figure 4.8 (acute images), trainee and expert fixation patterns fall within the 2.8-4.2 second mean duration band throughout the image stack, whereas novice readers fixation patterns only converge with those of trainees and experts by the third slice. In figure 4.9 (subacute images), despite the trend being fairly similar between trainees and experts, trainee readers spent longer appraising each image slice than experts readers did, particularly slice 4. The mean fixation time for novice readers decreased steadily from slice 1 (4 seconds) to slice 5 (2.2 seconds). Finally figure 4.10 (chronic images), demonstrates another

similar pattern between trainee and expert readers, yet experts move quicker through the stack than trainees and novices. Overall the novice visual search trend followed a pattern of spending the most amount of time on slice 1 followed by a steady diminishing of time to slice 4 with a slight increase in slice 5, whereas trainee and expert readers' visual search pattern appeared to differ dependent upon stroke type but invariably followed a pattern of increased or reduced time in slice 3 compared with the other four slices.

Table 4.5 demonstrates the mean fixation time across per axial slices within each case type and per participant group in CT images.

CT mean total fixation duration across slices		Slice 1	Slice 2	Slice 3	Slice 4	Slice 5	Mean total fixation duration per case type
Normal	Novices	4.6	4.3	4.0	3.7	4.0	4.1
	Trainees	3.1	4.1	4.1	3.9	2.7	3.6
	Experts	2.3	2.7	3.0	2.7	2.2	2.6
Acute	Novices	7.1	5.1	4.2	4.3	4.7	5.1
	Trainees	4.1	3.9	4.1	3.8	3.6	3.9
	Experts	3.2	2.8	2.8	2.7	2.9	2.9
Subacute	Novices	4.0	2.7	2.3	2.2	2.2	2.7
	Trainees	2.2	2.0	2.1	2.5	2.0	2.2
	Experts	2.1	1.7	1.8	1.9	1.8	1.9
Chronic	Novices	2.9	1.8	1.5	1.5	1.9	1.9
	Trainees	2.3	2.0	1.8	1.8	1.8	2.0
	Experts	1.9	1.5	1.3	1.3	1.5	1.5

In terms of mean fixation duration over the 5 slices, table 4.5 demonstrates that, on average; novices spend much longer appraising normal, acute and subacute cases than trainee and expert readers, particularly acute cases (mean total fixation per case = 5.1s). Trainees spent marginally longer appraising chronic cases than novices and experts; 2.0 seconds compared with 1.9 and 1.5 respectively.

Overall, ANOVA significance testing uncovered that experts were significantly quicker to reach the primary AOI than novices ($p<.000$) and so were trainees ($p<.031$). Although there were no significant differences between groups when considering the time spent looking at primary infarcts, there were, highly significant differences (with a large effect size) when comparing total time out of abnormal areas (*df* 2, $p<.000$, F=35.61, Eta Squared .92), with novices spending longer than trainees ($p<.000$) and both novices and trainees spending longer appraising additional features than experts ($p<.000$ & $p<.000$).

4.2.5 Stroke, Expertise, Accuracy and Visual Search

4.2.5.1 Diagnostic Decision-making and Eye-movements i.e. TP/FP/TN/FN

The aim of this section is to examine whether certain visual search behaviours are associated with diagnostic accuracy. For example, it is hypothesized that a quick time to fixate upon an abnormal area and time spent within the AOI indicates a true positive result. It is also hypothesised that when participants are incorrect, their eye movements can provide clues to perceptual errors, and/ or underlying cognitive process such as training and experience, which may determine why the abnormality was missed (i.e. not fixated upon or a recognition error) or 'seen' and not reported (i.e. a decision error). Therefore, the following section aims to examine mean time to hit the primary AOI when it appears in the image stack for the first time, the time spent within that AOI, time spent out of the AOI and the mean fixation time within the same slice overall, in an attempt to uncover visual search patterns associated with true positive and false negative decisions.

4.2.5.2 True Positive Decisions

This section considers the eye movements of participants who provided a true positive decision i.e. a correct rating to indicate the abnormality was either probably or definitely present, which accompanied a correct location of the primary abnormality on the 'brain atlas' task.

Table 4.6 highlights true positive eye-movement data by group and stroke type (in seconds) in CT.

CT True Positive eye-movement data by group and modality (in seconds) in first appearing AOI.		Mean time to hit Primary AOI	Mean time spent in Primary AOI	Mean time out of AOI (in same slice)
Acute	Novice	3.2	1.6	7.9
	Trainee	1.6	1.4	3.8
	Expert	1.4	1.4	2.7
Subacute	Novice	2.5	1.7	5.2
	Trainee	1.1	1.4	2.3
	Expert	0.9	1.4	2.3
Chronic	Novice	1.5	1.4	4.6
	Trainee	2.4	0.9	3.7
	Expert	1.5	1.1	3.2

Table 4.6 demonstrates that in true positive decisions, experts were the quickest to reach the primary AOI in acute stroke cases (1.4 seconds) compared with trainees (1.6 seconds) and halving the mean time to hit of novice participants (3.2 seconds). Experts and trainees spent an average of 1.4 seconds within the AOI but experts spent much less time appraising surrounding cortical tissue than

trainees and novices; 2.7 seconds by experts compared with 3.8 and 7.9 seconds respectively. These results, and overall time in slice, indicate a quick recognition and swift examination of the image slice by experts.

In subacute cases, trainees reached the AOI in just over 1 second and experts under a second, whereas novice readers took 2.5 seconds on average to reach the same abnormalities. In these cases, trainees and experts spent the same amount of time within the AOI and outside of it, indicating a clear pattern of visual search between the trainee and expert groups. In chronic stroke cases, experts and novices reached the AOI in 1.5 seconds, whereas correct trainees took 900 milliseconds longer. However, once trainees had reached the AOI they spent less time appraising the abnormal tissue than experts and novices.

Overall, true positive decisions appear to be characterised by a quick time to hit, often around 1.5 seconds or less for experienced readers and accompanied by an appraisal of abnormal tissue for 1.4 seconds or less. Experienced readers in this study spent 3.9 seconds on average in the slice where the abnormality first appeared and a correct decision followed. On average, trainees spent 5.5 seconds and novices 7.4 seconds to reach a correct decision.

When results were examined using a one-way ANOVA, there were significant differences between reader groups in term of time to 'hit' the abnormal area (*df* 2, *p*<.000 F=9.49, Eta Squared.05), total fixations in first AOI slice (*df* 2, *p*<.000 F=19.920, Eta Squared.09), time out of AOI (*df* 2, *p*<.000 F=17.28, Eta Squared.79), total case fixation durations (*df* 2, *p*<.000 F=10.03, Eta Squared.44), and the number of fixations within a case (*df* 2, *p*<.045 F=3.123, Eta Squared.01), all decreasing with experience. Time spent within the area of interest, indicating that as experience increases, readers become quicker at recognising an abnormality, was approaching significance (*p*<.065).

4.2.5.3 False Negative Decisions

This section considers the eye movements of participants who provided a false negative decision i.e. an abnormality was present and it was either missed, not recognised as an abnormality or was not correctly reported.

Table 4.7 highlights false negative eye-movement data by group and stroke type (in seconds) in CT.

CT False Negative eye-movement data by group and modality (in seconds) in first appearing AOI.		Mean time to hit Primary AOI	Mean time spent in Primary AOI	Mean time out of AOI (in same slice)
Acute	Novice	2.7	1.9	6.9
	Trainee	2.3	0.6	3.5
	Expert	0.0	0.0	2.9
Subacute	Novice	1.3	3.2	8.0
	Trainee	0.02	0.7	4.2
	Expert	-	-	-
Chronic	Novice	0.02	2.3	3.8
	Trainee	-	-	-
	Expert	4.5	0.5	4.5

Table 4.7 examines the eye movement data of participants that provided false negative accounts of abnormal cases. On average, novices took 2.7 seconds and trainees took 1.9 seconds to reach a primary AOI. In the same cases, experts did not spend any time in an acute primary AOI at all and spent less than 3 seconds on average within the slice overall indicating a search error. Interestingly, novices were quicker to fixate upon an AOI when a false negative decision was made rather than a true positive. Novices also gazed at an AOI for 300 milliseconds longer when they made an incorrect decision regarding acute cases but spent less time in the slice overall than true positive decisions. On the contrary, trainees took 700 milliseconds longer to reach the AOI when incorrect decisions were made and spent much less time appraising the abnormal areas, indicating a recognition error.

In subacute stroke cases, experts did not make any false negative decisions; they were all true positive decisions. Novice eye movements differed in false negative decisions in that they spent 1.5 milliseconds longer appraising the abnormality compared with true positive decisions. Novices spent 3.7 seconds longer in the slice overall than those that made correct decisions, which might indicate a continued search for an abnormal area as the primary abnormality was either missed or not originally recognised, even with 3.2 seconds of appraisal time. Trainees took 20 milliseconds to reach the AOI, which might indicate they were fixating within that area when the image appeared on the screen and especially as the AOI only captured their attention for an average of 700 milliseconds. Trainees spent 1.3 seconds longer on FN images than where TP decisions were made regarding the same stroke type.

In chronic cases, incorrect novices also reached the AOI very quickly and they also spent longer appraising the area than trainees in subacute cases. Incorrect novices also spent more time within the AOI than incorrect experts in the same stroke type. Experts took 4.5 seconds to reach the

AOI, indicating it did not capture their attention until they were ready to move onto the next slice and when they did view the AOI, they only spent 500 milliseconds on it, indicating the abnormality was barely fixated upon (recognition error). Trainees did not make any false negative decisions; but incorrect trainees made false positive decisions in subacute stroke types instead (please refer to false positive table for these results in table 4.4). Overall, it appears that false negative decisions were characterised by either long fixation times within an AOI than true positive decisions, (indicating decision errors) or were barely fixated upon and/ or missed altogether (recognition and/ or search errors). Mean fixation times to reach an AOI and time spent within the slice overall appeared to be just as variable between the groups.

When results were examined using a one-way ANOVA, differences existed between groups for confidence (*df* 2, F=5.58, $p<.007$, Eta Squared .2), with trainees being more confident in their incorrect decisions than novices ($p<.014$) and as previously described, experts were more likely to err on the side of caution and report 'probably absent' than novices and trainees in CT images, although this finding did not reach significance.

There were differences between groups when considering the number of fixations in the first AOI appearing slice (*df* 2, F=4.9, $p<.011$, Eta Squared .2) and time spent viewing features outside of the AOIs (*df* 2, F=5.5, $p<.008$, Eta Squared .2), with incorrect novices spending more time in the first AOI slice, and more time out of the area of interest entirely than incorrect trainees ($p<.015$ & $p<.009$) and experts ($p<.014$ & $p<.026$).

4.2.5.4 False Positive and True Negative Decisions

As a number of false negative decisions were identified between participant groups and stroke types, the following section aims to examine whether the errors could be attributed to recognition and/ or decision errors of normal features by further exploring the participant reports and their eye movements. Further analysis was applied to examine false positive decisions; specifically those made by experts and compare their visual search patterns with true negative decisions by other experienced readers in the study.

Before proceeding onto the next section, it is important to consider that only abnormal areas can be predefined by creating an AOI before the experiment took place, therefore, the exact time and slice where an individual false positive decision was made (based on certain clinical features) cannot be completely determined retrospectively, owing to the 2D reporting strategy of a 3D anatomical structure. However, eye movements throughout the image stack in these cases often

infer the clinical features that were implicated in the decision error, as demonstrated in the following CT images and overlaid gaze patterns.

4.2.5.5 False Positive Decisions

As seen in table 4.4 of this chapter, novice readers made more false positive decisions than trainee and expert readers; 41, 37 and 14 respectively. Table 4.8 demonstrates where readers made incorrect decisions and their accompanying eye movements.

Table 4.8 highlights mean false positive eye-movement data by group and stroke type (in seconds) across all CT images in the case stack. N.B. experts did not make any errors in subacute cases.

Pptn Group	**FP Location**	**No of FP's**	**Slice 1** (time in s)	**Slice 2** (time in s)	**Slice 3** (time in s)	**Slice 4** (time in s)	**Slice 5** (time in s)	**Mean total Case time** (s)	**Total group time** (s)
Novice	Normals	33	8.4	8.0	7.6	7.5	8.6	40.1	157.2
	Acute	11	9.9	9.5	9.3	8.5	10.3	43.1	
	Sub-acute	2	3.8	5.6	3.2	2.6	3.1	18.2	
	Chronic	1	15.6	9.6	9.5	6.6	14.4	55.7	
Trainee	Normals	16	6.2	7.9	7.2	5.5	3.9	30.7	119.5
	Acute	19	6.8	6.3	7.0	6.1	5.1	31.3	
	Sub-acute	12	3.3	3.3	4.0	5.1	4.6	20.3	
	Chronic	9	6.6	6.8	7.0	7.8	8.9	37.1	
Expert	Normals	9	3.4	3.6	5.2	4.7	4.0	20.8	54.0
	Acute	5	4.2	2.9	3.6	3.1	3.0	16.8	
	Sub-acute	0	-	-	-	-	-	-	
	Chronic	1	2.9	2.0	2.7	4.1	4.7	16.4	

Novices, trainees and experts made the majority of their false positive decisions in normal cases. Novices spent the most time across all FP decisions than trainees and experts. The novice participant who gave the chronic case a FP rating, spent much more time appraising this case than other participants who made decisions in other cases e.g. the mean time spent in subacute cases by incorrect novices was 18.2 seconds compared with 55.7 seconds. Although novices made more FP marks in normal cases, they spent more time considering those in acute cases; 43.1 seconds compared with 40.1 seconds. Trainee readers also spent more time viewing this chronic case than any of the other stroke types.

In terms of significance testing between groups in false positive decisions, there were differences between groups in terms of confidence in their decisions (*df* 2, F=6.2, *p*<.003), the number of fixations in the slice where the AOI first appears (*df* 2, F=5.5, *p*<.006), the time out of the AOI (*df* 2, F=5.3, *p*<.007), the total number of fixations within each case (*df* 2, F=5.3, *p*<.007) and the fixation durations overall (*df* 2, F=5.3, *p*<.007). Novices were least confident in these decisions but trainees were overly confident, compared with experts who tended to state 'probably present'. Although most eye movement measures decreased with experience, even when incorrect, experts fixated more than trainees in false positive decisions.

4.2.5.6 True negative decisions

In line with true positive decisions, confidence in true negative decisions increases with experience (*df* 2, *p*<.000 F=8.98, Eta Squared.14) and whilst it would be expected that the number of fixations and time viewing the case overall declined with experience, there were no statistically significant differences within these decisions.

Table 4.9 highlights a comparison of true negative and false positive ratings and the accompanied mean eye-movement data by participant group and fixations through the image slices (in seconds) in normal CT cases.

Participant Group	Participant Decision	Slice 1	Slice 2	Slice 3	Slice 4	Slice 5	Mean total Case time
Novice	TN	7.5	6.6	6.0	5.2	5.5	28.7
	FP	8.4	8.0	7.7	7.5	8.6	40.3
Trainee	TN	4.8	6.3	6.7	6.9	4.6	29.2
	FP	6.2	7.9	7.2	5.5	3.9	30.7
Expert	TN	3.8	4.8	5.1	4.4	3.6	21.6
	FP	3.4	3.6	5.2	4.7	4.0	20.8

The above table highlights a comparison of fixation duration between true negative and false positive decisions throughout the 5 image slices of just normal cases. In the novice group, true negative decisions are accompanied by less gaze time spent within each slice and thus, less time within the case overall than when false positive decisions are made. In fact, novices spend an average of 11.6 seconds longer on false positive decisions than true negative ones. Trainees also spent longer on false positive decisions, yet the difference was much smaller at 1.5 seconds. When expert readers made false positive ratings regarding a case, they spent less time on these cases than true negative ratings, but yet again, the difference was smaller at 800 milliseconds

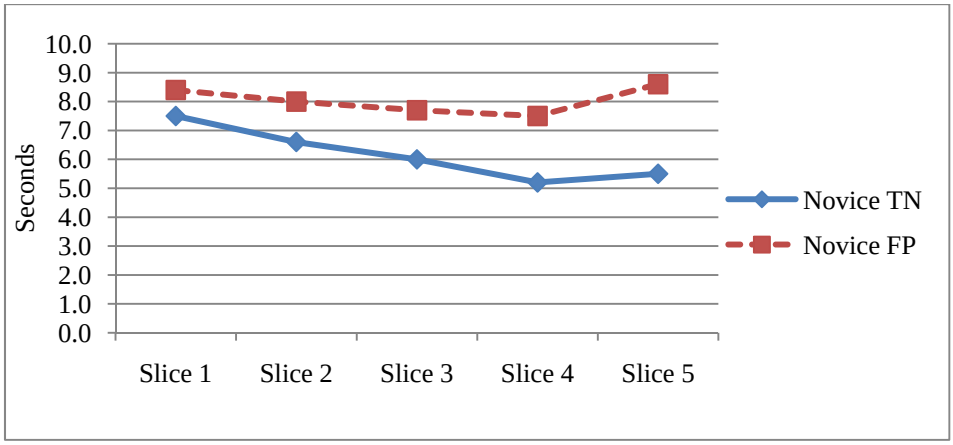


Figure 4.11 Mean fixation time per axial slice of novice readers who made TN & FP decisions in normal CT cases

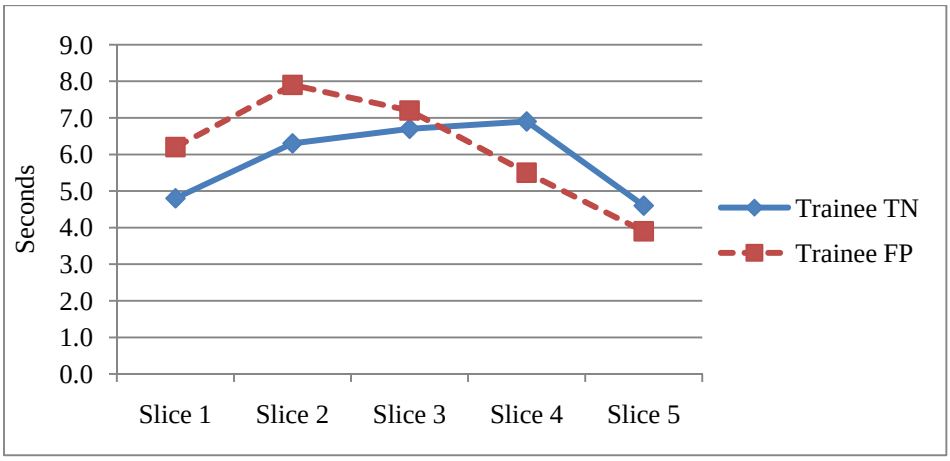


Figure 4.12 Mean fixation time per axial slice of trainee readers who made TN & FP decisions in normal CT cases

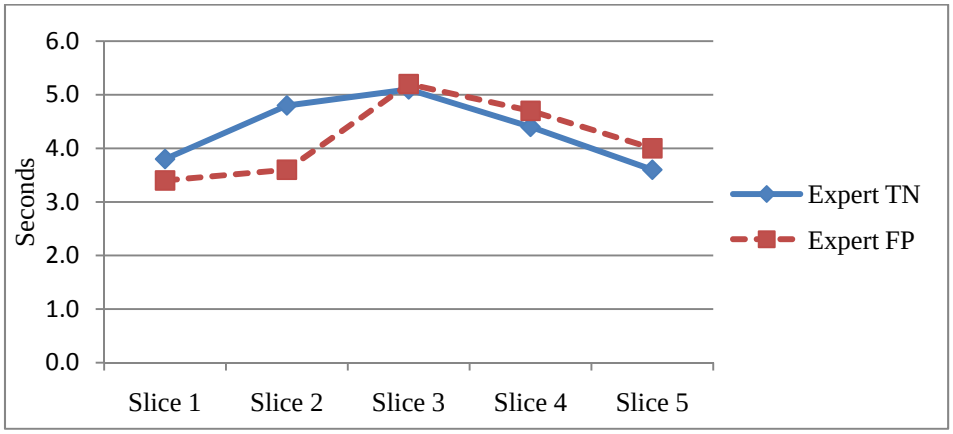


Figure 4.13 Mean fixation time per axial slice of expert readers who made TN & FP decisions in normal CT cases

The above figure 4.11-4.13 consider the mean fixation time per axial slice (not just overall time spent in each case) of novice, trainee and expert readers when making true negative and false positive decisions regarding a case. Figure 4.11 highlights that not only do novice spend more time over the case but incorrect novices appear to spend the most time (up to 10 seconds) appraising the final slice, where abnormalities are least likely to occur in stroke cases (due to blood vessels being more likely to be occluded higher up in the brain), and with true negative decisions making a steady decline of fixations through the slices. Trainee who made incorrect decisions appeared to spend more time appraising the second and third slices than trainees who made true negative decisions and spent more time in slices 3 and 4 than slice 2. Whilst expert true negative eye movements follow a similar trend to trainees, false positive decisions do not appear to follow a clear 'trend' as they spend the least amount of time in the first two slices compared with slices 3-5.

4.2.5.7 An In-Depth Examination of Expert FP Decisions

This section aims to examine the false positive decisions made explicitly by experts to gain an in-depth insight into why and when experts make mistakes in CT imaging.

Table 4.10 to demonstrate the total duration of fixations in a retrospectively defined FP AOI.

Expert Reader	Case	Slice 1	Slice 2	Slice 3	Slice 4	Slice 5	Total fix time in FP
1	NJM	2.6	1.2	3.4	3.5	0.3	11.1
2	NJM	0.1	0.0	0.3	0.0	0.0	0.4
3	NJM	0.0	0.0	0.7	0.4	0.3	1.4
4	NDG	1.4	0.8	0.7	0.8	1.6	5.3
3	NDG	0.6	0.7	1.6	0.4	0.4	3.7
5	NAL	0.4	0.0	2.7	2.3	0.8	6.1
3	NAL	0.2	0.2	0.7	0.4	0.0	1.5
6	NBH	0.5	0.6	0.0	0.6	1.6	3.4
5	NPA	0.9	0.3	1.5	0.2	1.5	4.4
7	CDT	0.0	0.0	0.0	0.0	0.9	0.9

The above table demonstrates that, of the ten FP decisions made by experts, cases NJM, NDG and NAL caused the most confusion. In case NJM, readers 2 and 3 defined fairly small FP regions, which appears to be correlated with the relatively few fixations within the region across the 5 slices, with reader 2 spending only 400 milliseconds in the false positive region and reader 3 spending 1.4 seconds in the region. It appears therefore, that the false positive decisions of readers 2 and 3 were

made hastily on consideration of a very subtle feature that was easily misinterpreted, whereas reader 1 spent a large time assessing the area and considered it to be much more important than it truly was. In previous research false positive decisions have been shown to be either made hastily or following a long period of deliberation.

In case NDG, a relatively small area was predefined yet readers, revisited the region throughout the image stack highlighting the awareness of the reader to a suspicious region and thus, total fixations within that area came to 3.7 and 5.3 seconds. In case NAL, two very different regions were defined as abnormal by the expert readers and as in case NDG; the abnormal sites were revisited throughout the stack. These results demonstrate that in CT, FP decisions accompany a wide variation in eye movements; from 100 milliseconds to 3.5 seconds in a single slice and up to 11.1 seconds across a whole case consisting of 5 axial slices. The following gaze tracker and hot spot images demonstrate where each FP location was plotted on the brain atlas task and has been retrospectively overlaid across each image slice to gain a qualitative appraisal of the participants' decision-making process.

Figure 4.14 Case NJM: Reader 2 False positive decision.

Figure 4.15 Case NJM: Reader 1 False positive decision.

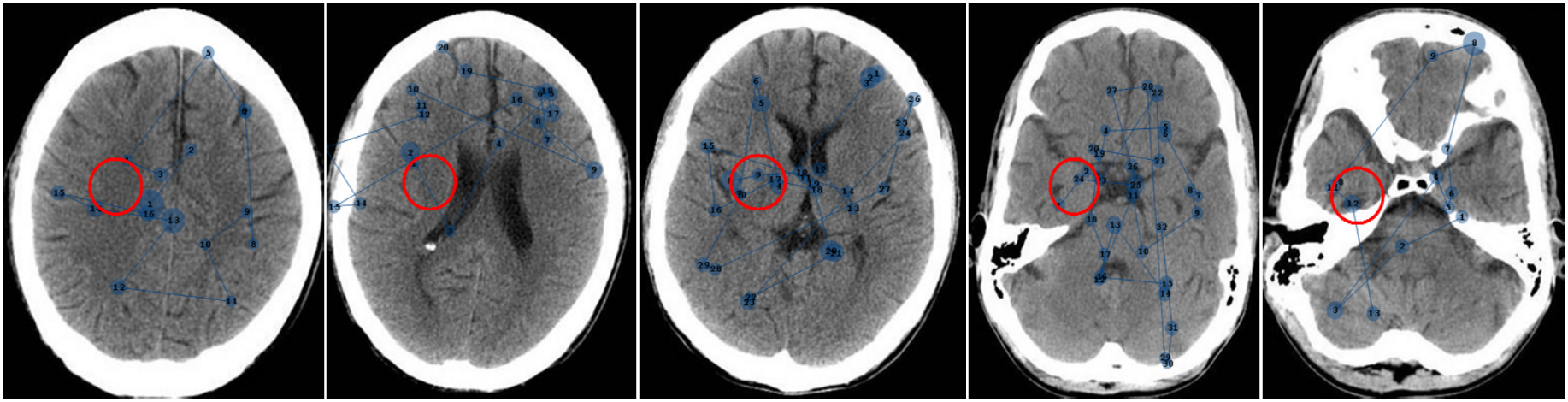


Figure 4.16 Case NJM: Reader 3 False positive decision.

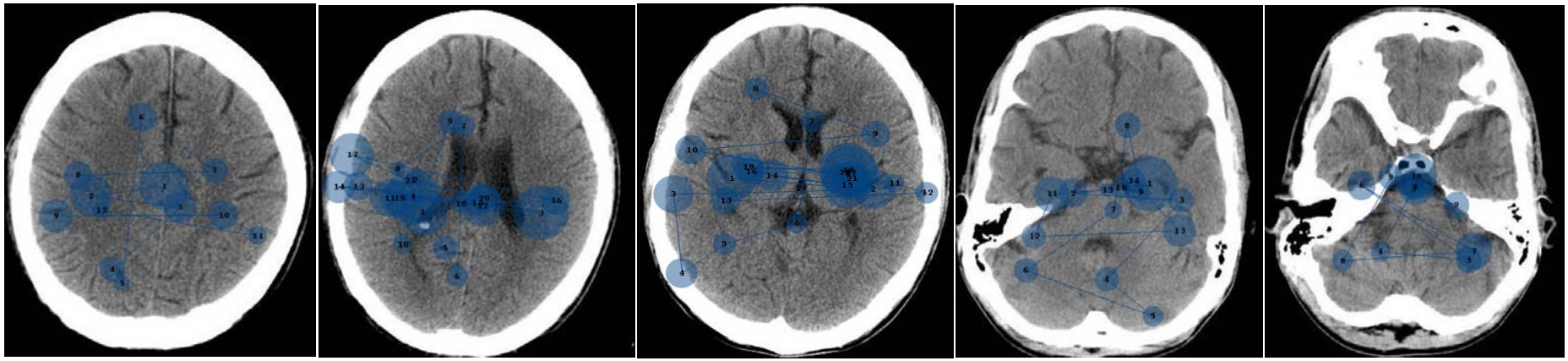


Figure 4.17 Case NJM: Reader 8 True negative decision.

Figures 4.14-4.17 highlight very different visual search patterns between readers of FP and TN decisions. Primarily, reader 2 fixates comparatively few time per slice, there are also very few fixations within the FP region, which was defined on the brain atlas task. Reader 1 spends much more time cross comparing the hemispheres and defines a rather large 'wedge' of the middle cerebral to classify as abnormal. The third reader defines a region similar in size to reader 2, yet on the adjacent hemisphere. It is clear in this case that a tissue fold, or sulci, within the cerebral tissue has been mistaken for an infarction. The true negative appraisal of reader 8 highlights a thorough appraisal of the middle cerebral regions of both hemispheres, with a particular focus upon the sulci and brainstem, which had been considered abnormal by readers 1, 2 and 3, yet an infarction has been ruled out.

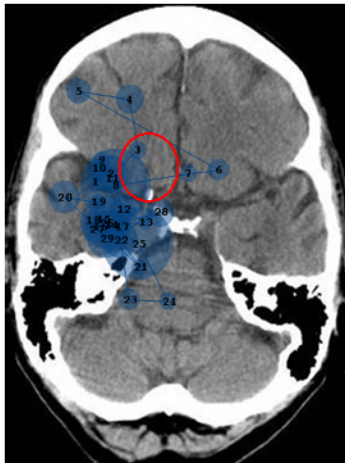


Figure 4.18 Case NDG: Reader 4 False positive decision.

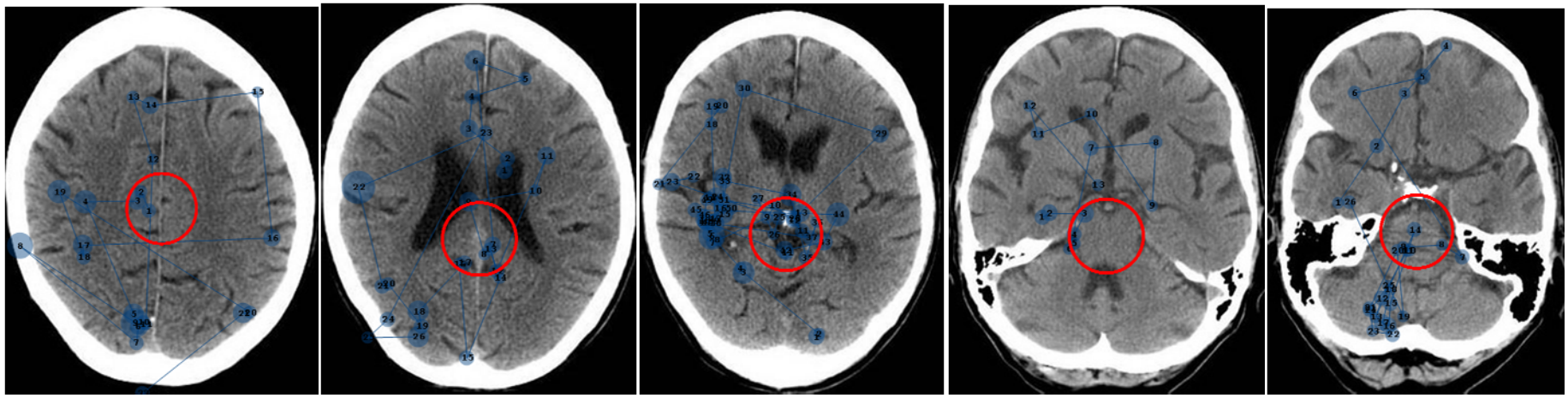


Figure 4.19 Case NDG: Reader 3 False positive decision.

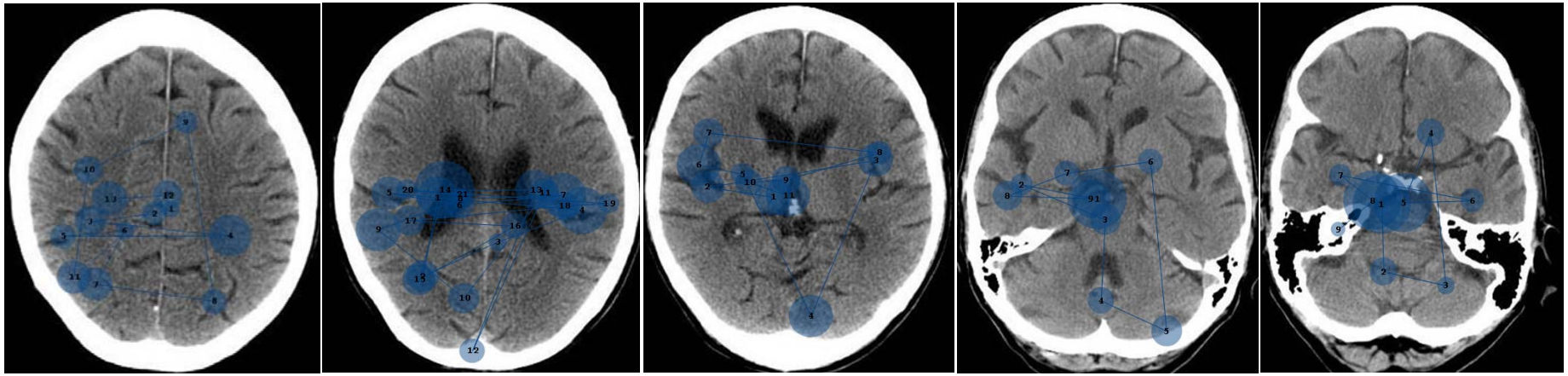


Figure 4.20 Case NDG: Reader 1 True Negative decision.

In figure 4.18, case NDG, reader 4 plots the false positive in the upper quadrant of the left hemisphere and reader 3 implicates the brainstem as being abnormal. As in case NJM, correct reader 1 appraises the same areas yet does not consider them suspicious. Whilst reader 1 has a similar number of fixations as reader 4 in slices 1-3, reader 4 has an increasing number of fixations in slices 4 and 5. Reader 3 spends a large amount of time fixating upon the brainstem region and has made a decision error regarding the features therein.

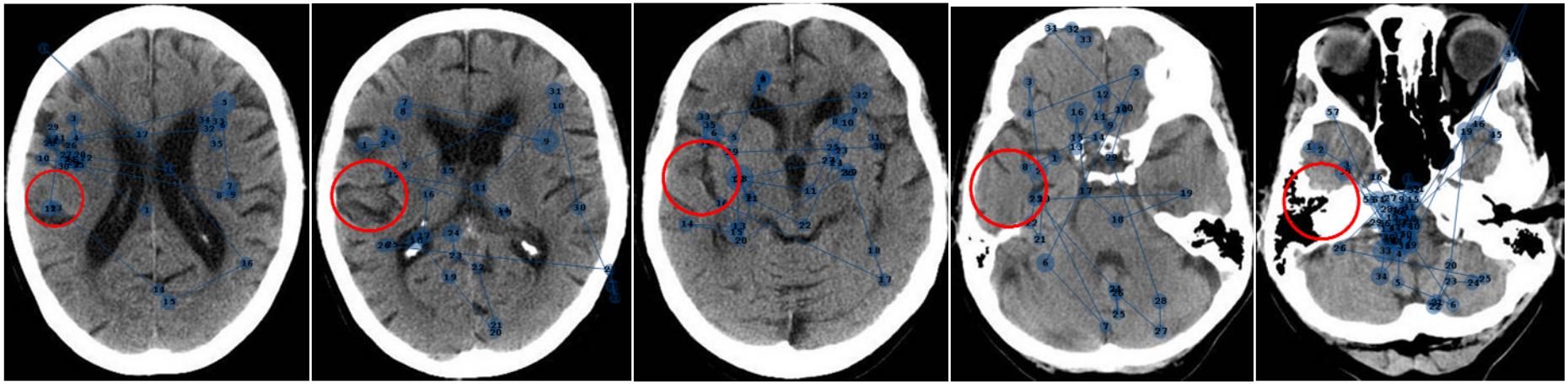


Figure 4.21 Case NAL: Reader 3 False positive decision

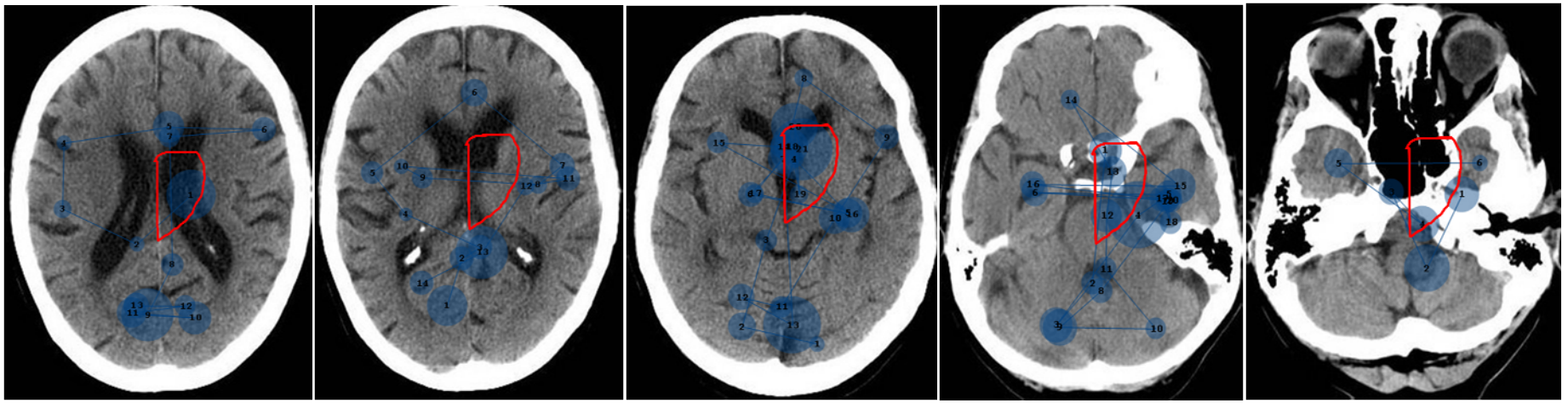


Figure 4.22 Case NAL: Reader 5 False positive decision

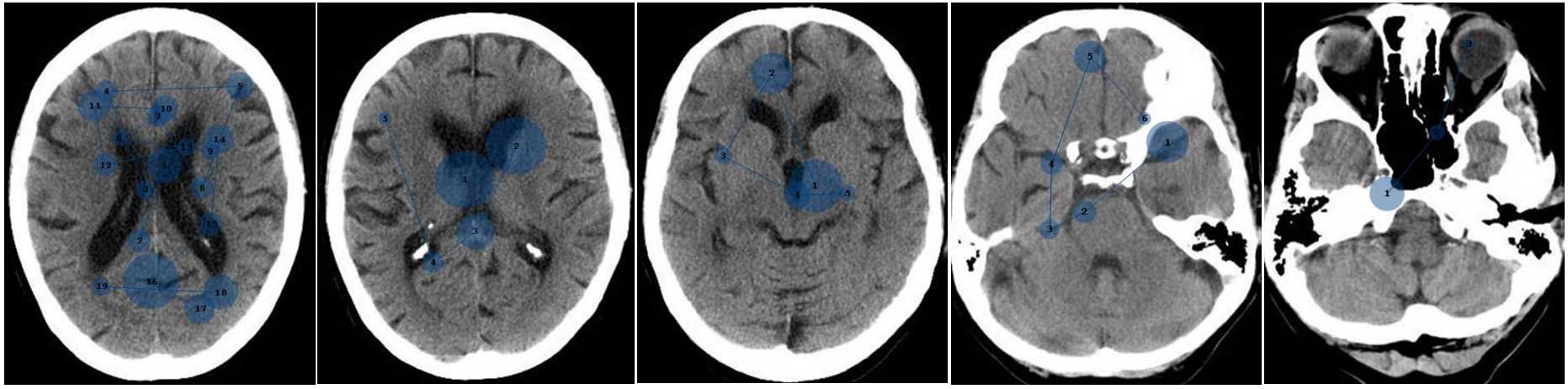


Figure 4.23 Case NAL Reader 1 True negative decision

In case NAL, it is clear that reader 3 has mistaken a fold in cortical tissue as an abnormality, which was also apparent in case NJM, whereas reader 1 barely fixates upon this region. Reader 3 also spends much time in the brainstem region as previously stated in case NDG. The false positive region that reader 5 has defined appears to bear no relevance to the features within the image; this could be down to incorrect reporting, which might have been more clearly defined if the reader was able to mark the image itself.

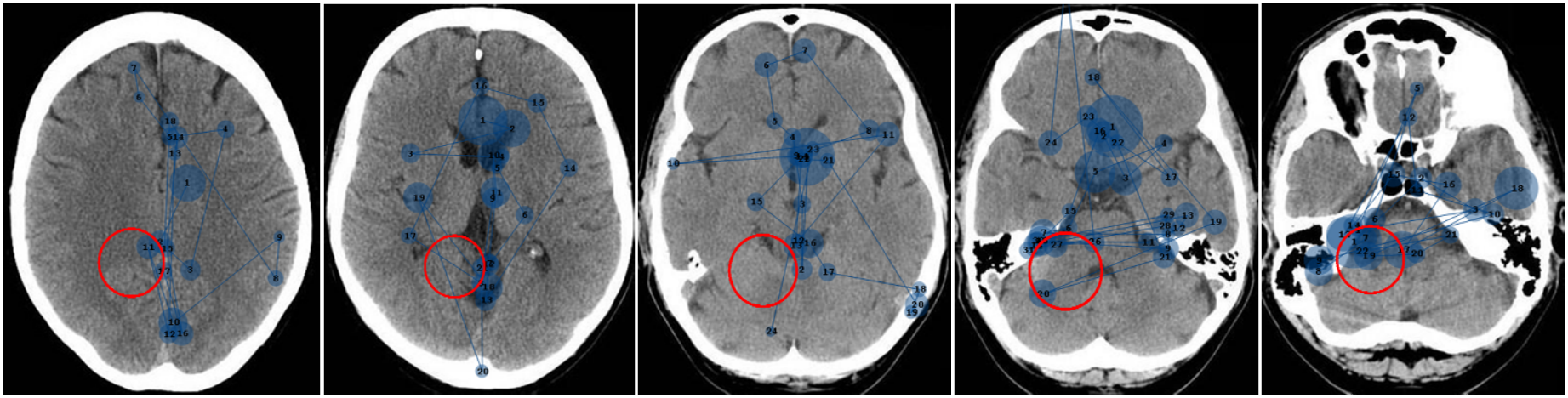


Figure 4.24 Case NBH: Reader 6 False Positive decision

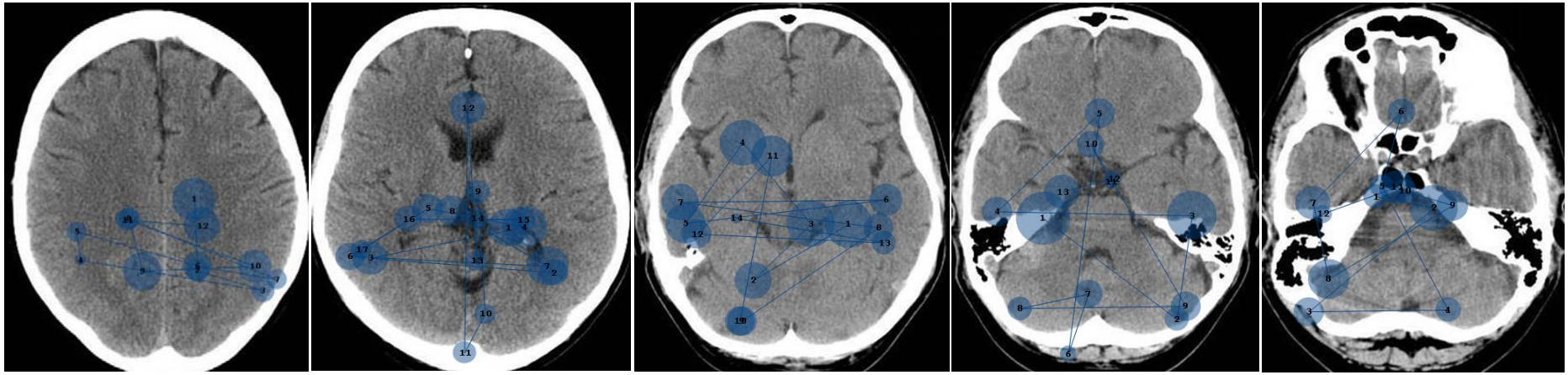


Figure 4.25 Case NBH: Reader 8 True negative decision

Case NBH provides an interesting cross comparison between true negative and false positive visual search patterns; reader 6 appears to focus upon the midline, whereas reader 8 makes horizontal cross comparisons between the hemispheres. Although not completely clear, it appears that reader 6 has either mistaken a fold of cortical tissue for an infarction (slice 2) or implicated the area of low attenuation in the base of the cerebellum (slice 5) as demonstrated by the clustering of eye movements in both areas and much diagonal cross comparison within slice 5. Whilst reader 8 has appraised the brainstem region, this region did not draw the attention of reader 8 who correctly ruled out the presence of an abnormality in this case.

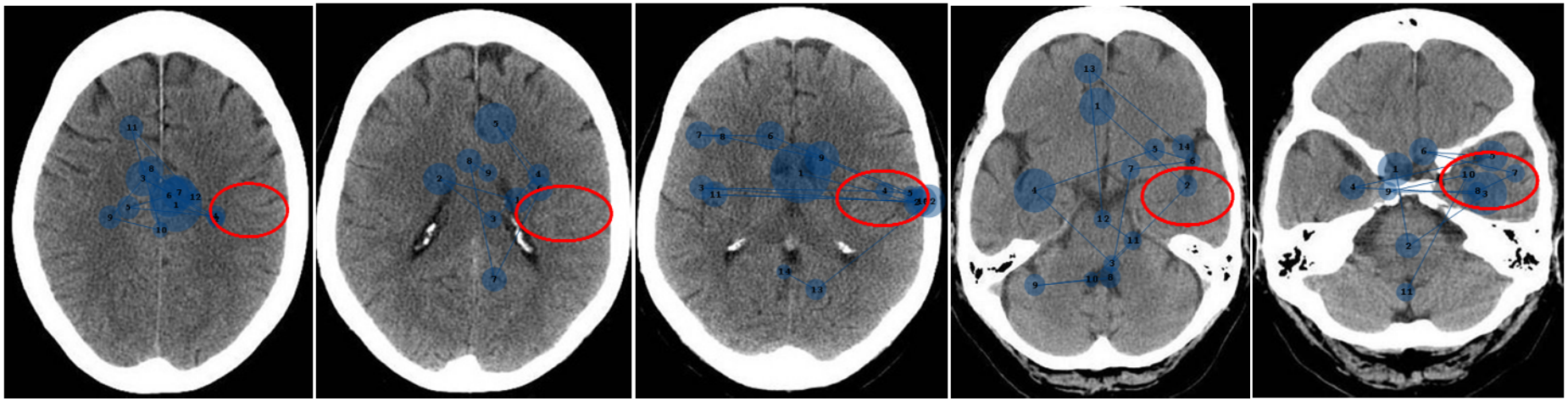


Figure 4.26 Case NPA: Reader 5 False Positive decision

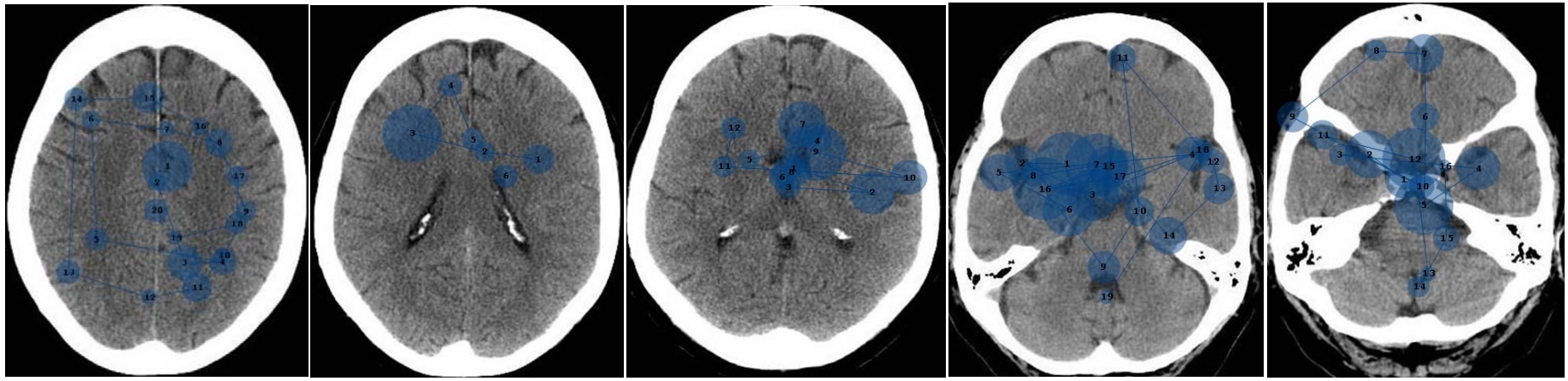


Figure 4.27 Case NPA: Reader 1 True Negative decision.

Both readers in this final control case NPA demonstrate very similar visual search patterns between slices 2-5 with identical regions of interest being analysed. However, whilst reader 1 fixates upon the sulcus in the right hand side of the middle slice and moves on, reader 5 interprets this region as suspicious and reappraises the area in slice 5, therefore indicating an identification error similar to those previously discussed in cases NBH (reader 6) and NAL (reader 3).

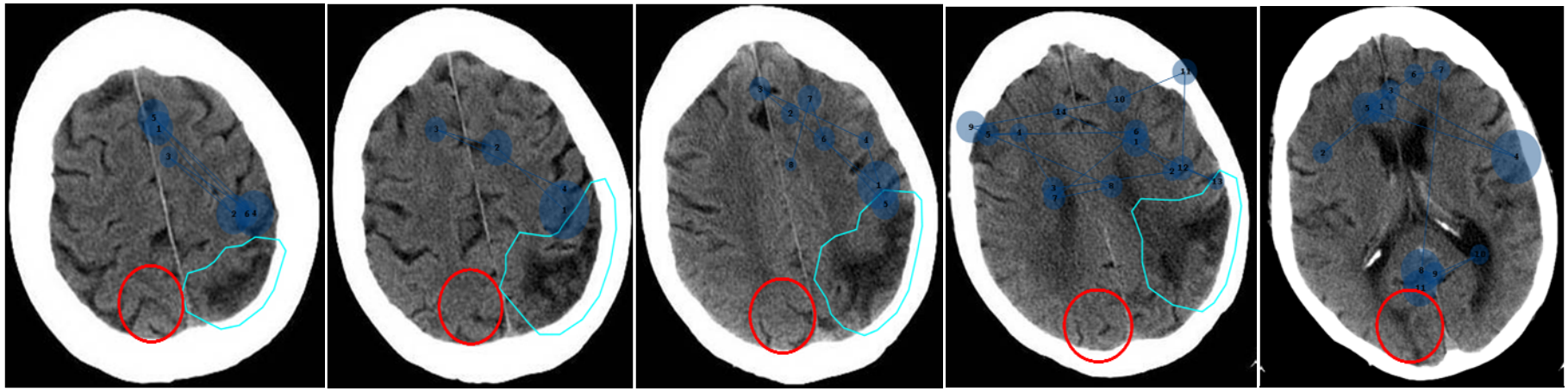


Figure 4.28 Case CDT: Reader 7 False Positive decision.

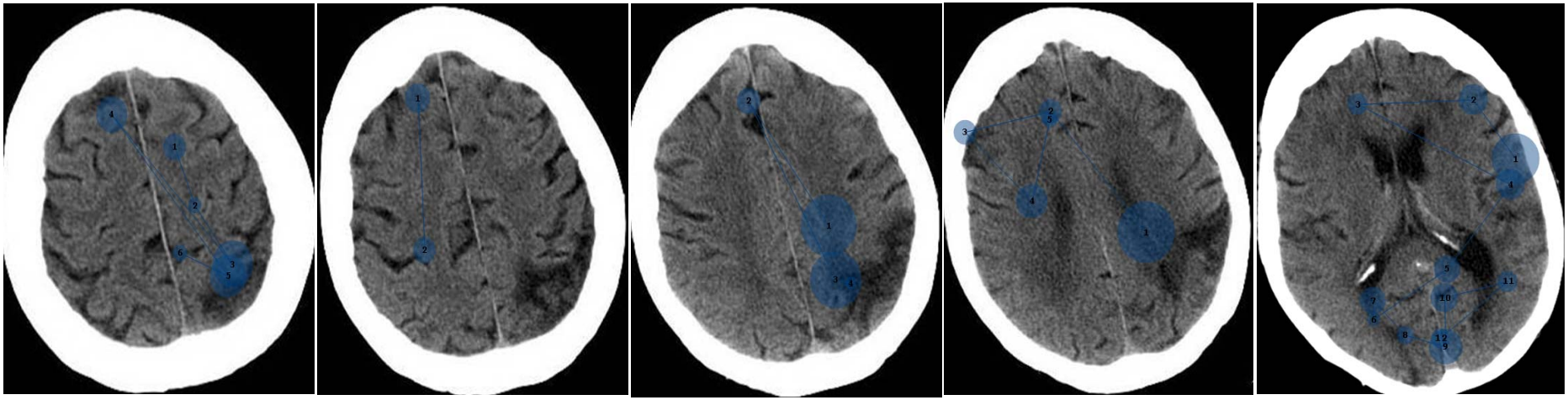


Figure 4.29 Case CDT: Reader 1 True Positive decision.

The false positive decision in case CDT appears very different from the others in this chapter. Primarily, there are no eye movements within the FP decision area and the eye movements between readers appear very similar i.e. the fixations are similar in terms of anatomical features appraised and time to reach the chronic abnormality. The number of fixations within each image and between readers is also very similar i.e. FP Fixation count; 6, 4, 8, 13 & 11 versus TN Fixation count; 6, 3, 5, 7 & 12. Therefore, it appears likely that this decision was down to a reporting error rather than a decision error e.g. the reader circled the wrong hemisphere on the reporting sheet.

4.2.6 Secondary Abnormalities

Within this study there were two secondary locations, as depicted below. There was one secondary located in subacute case SGR and one in chronic case CAJ, with the latter being slightly smaller in size than the former.

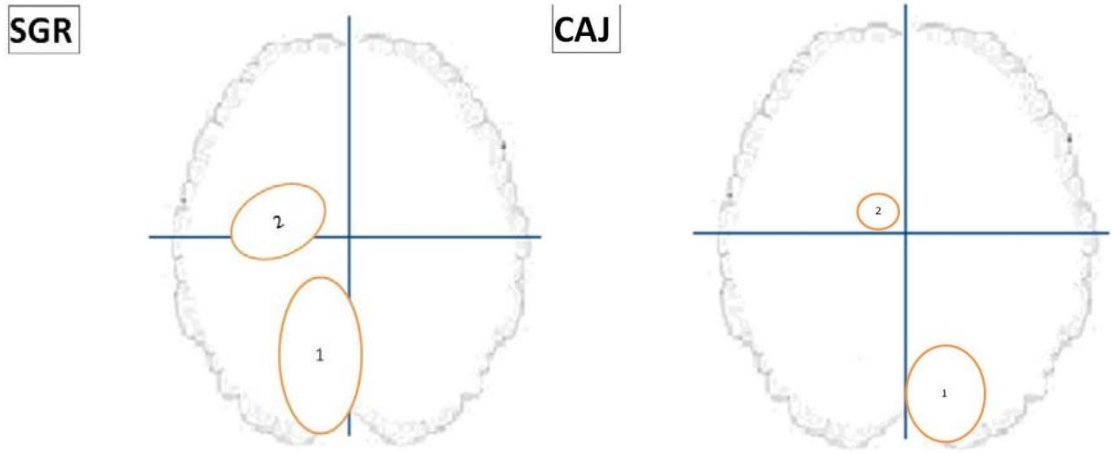


Figure 4.30 The scoring criteria of cases SGR and CAJ.

In terms of accuracy for the above secondaries, once again experts were most accurate when identifying secondary abnormalities, followed shortly by trainees and then novices. Figure 4.31 highlights the percentage of secondaries correctly located and rated with a comparison of primary and overall detection performance, which also demonstrates increasing accuracy for both infarct types by level of experience.

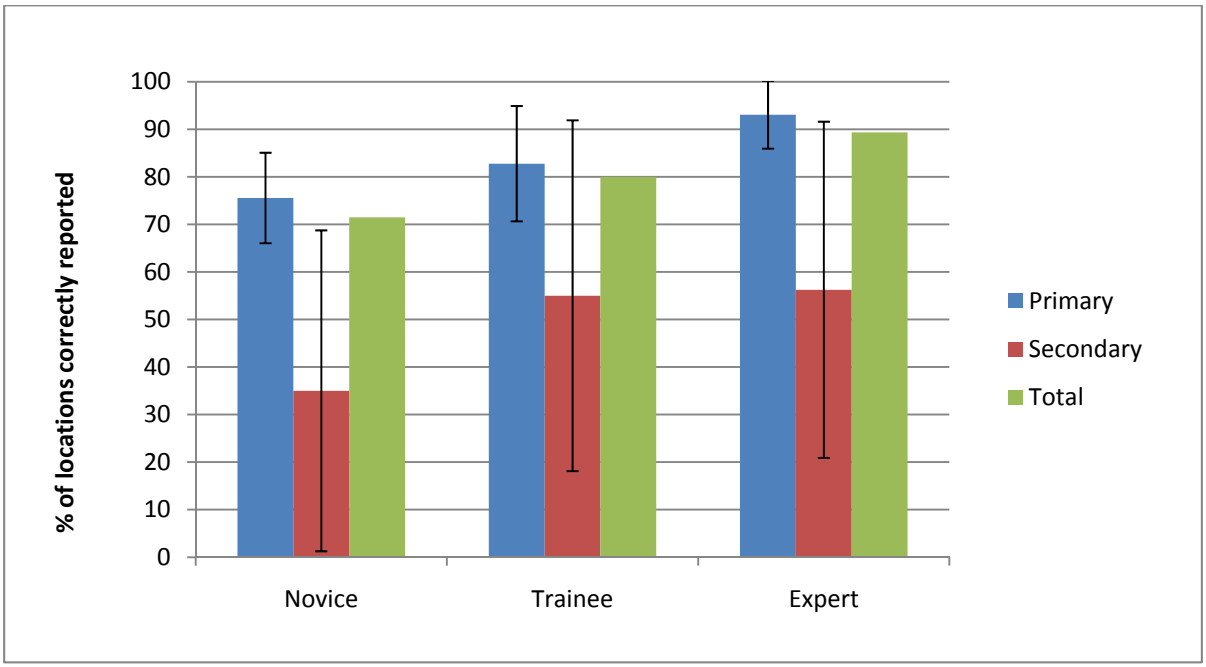


Figure 4.31 Percentage of primary and secondary locations correctly reported by participants in CT.

As previously stated, there were significant group differences for the detection of primary abnormalities and there were also significant differences in secondary infarct detection (*df* 2, *p*<.002, F=6.87, Eta Squared=.15). The most notable differences were evident between novices and experts (*p*<.001) but not trainees and expert readers (*p*<.172) when Games-Howell posthoc tests were performed.

4.2.6.1 Secondary Abnormalities and Eye Movements

When considering the time spent within AOI compared with secondary AOI's, novices and trainees spent the majority of their time in the primary AOI, but experts spent more time in the secondary AOI in the subacute case. In the chronic case, the same trend was also apparent but novices spent much longer in the chronic primary than the subacute one. Experts also spent much longer in the chronic primary than the secondary compared with the subacute case. Despite there being no significant differences between groups when looking at primary infarcts, there were significant differences between the groups when comparing the time spent within the secondary abnormalities (*df* 2, *p*<.000, F=8.66, Eta Squared .18). The main differences were between novices and experts (*p*<.023) and trainees and experts (*p*<.038) for total time within the secondary AOI.

4.2.6.1.1 True Positive Secondary Decisions

When considering the time to reach the subacute secondary AOI, trainees were the quickest and spent a mean time of 907 milliseconds appraising the abnormal region (see table 4.11). Whilst experts spent marginally longer appraising the AOI (56.7 milliseconds), it also took them 1.9 seconds longer to reach the AOI than trainees. Experts also took 500 milliseconds longer than novices, although novices only fixated upon the area for an average of 295 milliseconds. Overall, correct experts spent longer appraising this image than novices and trainees. The chronic secondary abnormality, in comparison, highlights a reversal of the subacute trend for experts spending more time within the image overall as correct novices and trainees spent an average of 7.5 seconds on the image, compared with 6.1 seconds by the experts.

Whilst trainees were quickest to reach the AOI in the subacute case, they took 3.4 seconds to reach the chronic secondary but over both image slices, spent 916 milliseconds appraising the features. Novices spent 937 milliseconds appraising the AOI over both slices, yet correct experts did not spend any time within the AOI over either slice indicating they may have appraised the anatomy surrounding the infarct, as previously seen in the pilot study. Despite the differential size of the secondary abnormalities, it appears mean time within an AOI was not affected by novice or trainee readers but experts did not fixate directly upon the chronic abnormality. Time to hit was increased for trainee readers, who took 2.7 seconds longer to reach the chronic infarct; however, this may have been due to its location and appearance rather than infarct size. Unfortunately, the effect of size on eye movements cannot be clearly defined within this study.

Table 4.11 highlight eye-movement data by correct participants (in seconds) in the first appearing secondary AOI in CT.

CT True Positive eye-movement data by group and modality (in seconds) in first appearing SECONDARY AOI.		Mean time to hit Secondary AOI	Mean time spent in Secondary AOI	Mean time out of AOI (in same slice)	Mean total fixation duration in first AOI slice overall
Subacute (SGR)	Novice	2.1	0.3	3.1	3.3
	Trainee	0.7	0.9	2.8	3.0
	Expert	2.6	1	3.6	4.4
Chronic (CAJ)	Novice	2	0.9	6.6	7.5
	Trainee	3.4	0.4	7.3	7.5
	Expert	0	0	6.1	6.1

4.2.6.1.2 False Negative Secondary Decisions

The eye movements that accompanied false negative decisions highlight that incorrect trainees took the longest to reach the AOI and when they only spent 179 milliseconds within it (see table 4.12). Novices spent a similar amount of time viewing the abnormal region (219 milliseconds) but reached it in half the time of trainees. Experts, on the other hand, spent 3 seconds on average appraising the AOI and reached it within 100 milliseconds, these participants also spent around 7 seconds within the image overall. In comparison with true positive decisions, incorrect trainee and expert readers spent much longer within an image and whilst experts dwelt upon the AOI for longer in incorrect decisions, trainees spent longer over true positives; indicating experts dwell upon a feature for much longer than true positives ones, which may indicate cognitive dissonance regarding the decision to rule it out or not, whereas trainees do not appear to recognise a feature at all when they gaze over it.

In the chronic case, incorrect experts take longer to reach the AOI than novices and trainees and only gaze over the feature towards the end of the duration of time within the image. All participants spent longer over these images than the subacute images (on average), but whilst novices spent a total time of 519 milliseconds in the AOI over two slices, trainees took 1076 milliseconds and experts 1319 milliseconds, indicating once again that experts had enough time to appraise the clinical features but did not consider the region suspicious enough to classify as an infarct.

When false negative decisions were made, the same pattern was observed between participant groups and abnormality type as true positive decisions, although in the subacute case, novice and trainees spent much less time on the secondary whilst experts spent up to two seconds longer appraising the same feature than readers who correctly reported the abnormality.

Table 4.12 Highlights eye-movement data by incorrect participants (in seconds) in secondary CT AOI.

CT False Negative eye-movement data by group and modality (in seconds) in first appearing SECONDARY AOI.		**Mean time to hit Secondary AOI**	**Mean time spent in Secondary AOI**	**Mean time out of AOI (in same slice)**	**Mean total fixation duration in first AOI slice overall**
Subacute (SGR)	Novice	2.9	0.2	1.8	1.8
	Trainee	6	0.2	4.7	4.8
	Expert	0.1	3	4	7.1
Chronic (CAJ)	Novice	5.2	0.5	6.8	6.9
	Trainee	3.1	0.6	8.4	8.7
	Expert	8.3	0.6	6.1	6.4

4.2.7.1 Focal Abnormality

Within this study there was one focal abnormality present in acute case AMW, as depicted below.

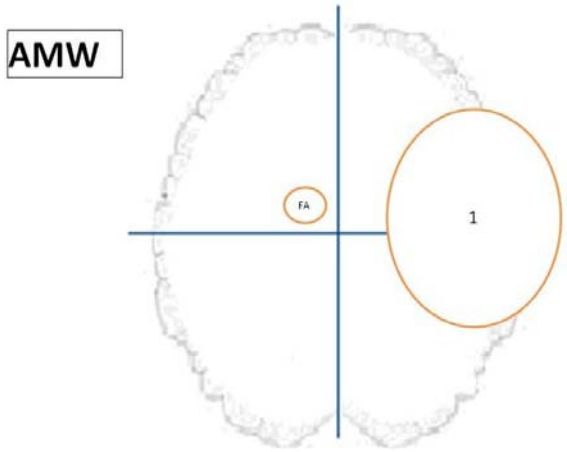


Figure 4.32 The scoring criteria of acute case AMW.

Novice participants did not detect this feature and therefore this section only applies to trainee and expert readers. Experts were more sensitive to the location of focal abnormalities than trainees with 87.5% of radiologists detecting the focal compared with 50% of trainees.

4.2.7.1.1 Focal Abnormality Qualitative Analysis of Eye Movements

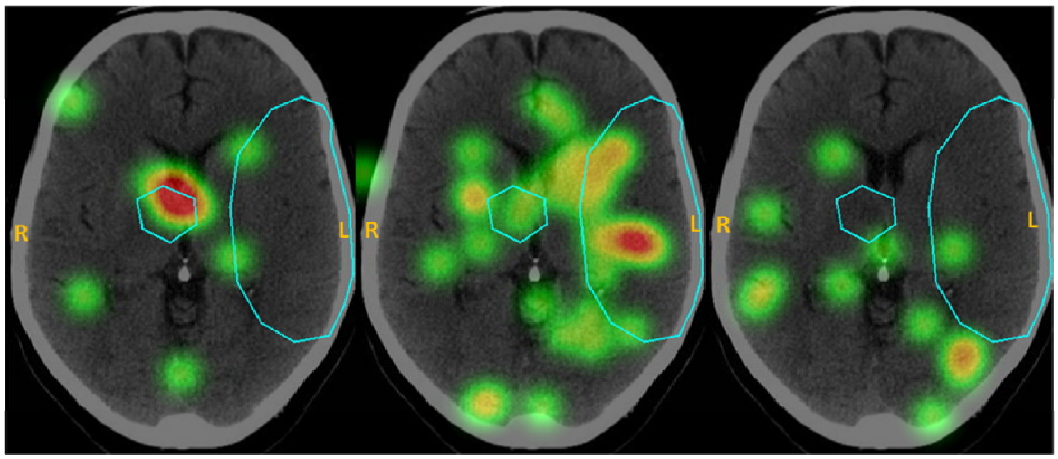


Figure 4.33 represents the eye movements as depicted by heat map images of a correct trainee, a trainee who saw the focal but did not report it, and a trainee who did not fixate on the focal area at all.

The first image in figure 4.33 is taken from case AMW from a fourth year trainee reader who correctly located and identified the primary and focal abnormality – the eye movements within this case were incredibly similar to expert readers. The eye movements from the second image are from a second year trainee, pre-neurology training who clearly saw the focal, but did not recognise it to be a focal abnormality, which was not subsequently reported and the third image demonstrates a false negative case by a first year trainee who did not fixate within the AOI at all.

When considering time spent fixating upon the primary versus the focal abnormality, **n**ovices and trainees spent the majority of their time in the primary AOI; 1.7 and 3.1 seconds, compared with the secondary focal AOI (500 milliseconds and 1.5 seconds respectively) but experts spent more time examining the focal AOI than the primary; 3.8 seconds compared with 1.7 seconds, respectively. Overall, trainees spent the most time within this image, averaging 13.6 seconds of fixation time per trainee reader.

4.2.7.1.4 Eye Movements and Focal Abnormality Accuracy

Where readers correctly circled the location of the focal abnormality, they spent up to 3 seconds appraising the features (table 4.13). The mean time to hit in this case was between 2.4 and 2.6 seconds, which is clearly down to the large subacute abnormality capturing the readers' attention. Fortunately, satisfaction of search did not take place within these secondary true positive cases e.g. where the secondary is missed due to the observer rating the primary and satisfactorily moving onto the next slice or patient case.

Table 4.13 highlights eye-movement data by correct participants (in seconds) in the focal abnormality in CT.

CT True Positive eye-movement data by group and modality (in seconds) in the FOCAL ABNORMALITY.		**Mean time to hit Secondary AOI**	**Mean time spent in Secondary AOI**	**Mean time out of AOI (in same slice)**
Acute (AMW)	Novice	-	-	-
	Trainee	2.4	1.1	11.9
	Expert	2.6	3.0	4.2

Table 4.14 highlights eye-movement data by incorrect participants (in seconds) in the focal abnormality in CT.

CT False Negative eye-movement data by group and modality (in seconds) in the FOCAL ABNORMALITY.		Mean time to hit Secondary AOI	Mean time spent in Secondary AOI	Mean time out of AOI (in same slice)
Acute (AMW)	Novice	5.7	0.5	5.8
	Trainee	7.7	0.4	5.5
	Expert	7.1	0.8	7.8

Where readers missed or ruled out the abnormality, between groups they viewed it for between 500 and 800 milliseconds (table 4.14). As no novice participant rated the focal present, it appears that a mean value of 500 milliseconds indicates a lack of time within the AOI to process the feature, or there is no knowledge of this type of abnormality within the hypothetical cognitive 'schema' to back up the visual processing within this AOI.

4.2.7.2 Satisfaction of search

In terms of satisfaction of search within CT images and between reader groups, 66.7% of novices saw the primary first when compared with the secondary abnormality. Of those who saw the primary first, 40% of these readers detected the secondary. However, of the 17% who saw the secondary first, only 17% of these readers saw the primary afterwards. Therefore, satisfaction of search was likely to occur in this group of readers, particularly when a secondary was seen first. Readers were also more likely to miss the primary when the secondary was large and more obvious. In the trainee group, 70% of trainees saw the primary first and 33% of these readers also spotted the secondary subsequently. Of the 16% who saw the secondary first, 100% of these readers went onto see the primary. Of the 67% of experts who saw the primary first, over half also identified the secondary abnormality (56%). Only 21% spotted secondary first and 67% of these went onto spot the primary subsequently.

4.2.7.3 Small Vessel Changes

As previously discussed in chapter 1, small vessel changes are common in the brains of elderly people and can be incidental. Within this study there were eight cases that contained small vessel changes; NAL, AAB, AJM, SBD, SCD, CDT and CJF. Sensitivity rates between trainee and expert groups for small vessel changes were; 30% and 65%, respectively. Specificity rates were 95% and 89% for the same observers. Highlighting that experts were more sensitive to small vessel changes. The following image slices are from acute case AAB and contain very discrete signs of small vessel changes.

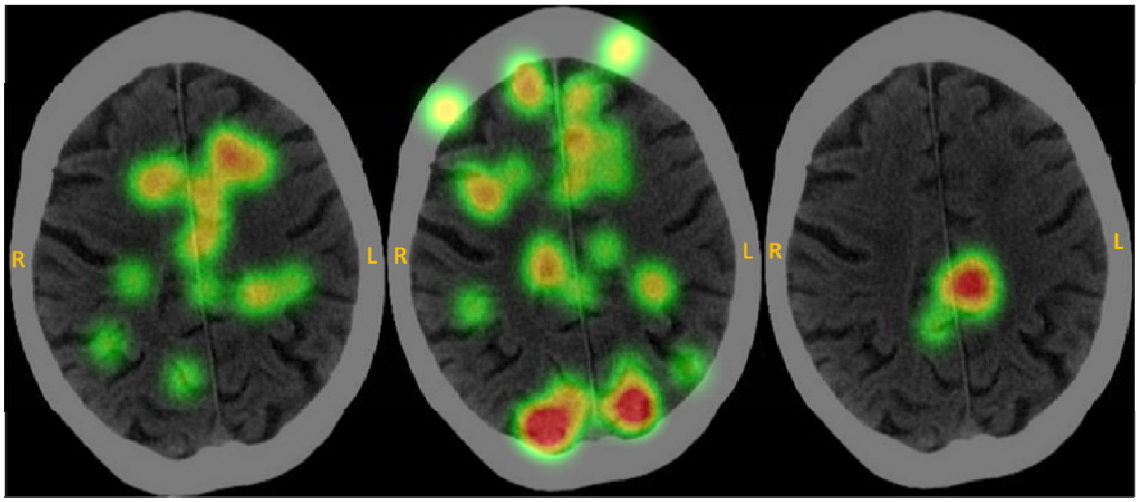


Figure 4.34 represents the eye movements of an expert, a trainee and a novice reader when searching for small vessel changes in CT.

All expert readers detected the changes and 60% of trainees did also. As all trainee and expert participants spent much time searching within this image, even those who ruled out the presence of SVC's, the third figure in this sequence is from a novice reader to demonstrate the lack of visual search compared with trainees and experts, as novices were unaware of the presence of small vessel changes or their appearance within an image. It is apparent that much visual search has taken place within the second figure, despite the SVC's being falsely ruled out.

As abnormalities became more discrete, trainees were less able to distinguish not only between the features of normal and abnormal anatomy but also between what constitutes a focal and what feature types constitute a small vessel change classification. Problems of classification between feature types will be further explored within the discussion section.

4.3 Discussion

This study aimed to explore the visual interpretations of multidimensional CT images of stroke when viewed by readers with varying levels of experience; from novice to experienced consultant radiologist. It also aimed to explore observer detection and omissions of infarctions using eye movement analysis in much detail; by observer group, case type, decision and location accuracy, including visual search by image, case, and experience level paired with the clinical decision made regarding each abnormality type commonly featured within the stroke population.

4.3.1 Diagnostic Accuracy and Confidence by Group

Across all cases, it was unsurprising that experts were the optimum performers regarding their accuracy and confidence when completing this study. Experts were the optimum readers and decision-makers when it came to ruling out or ruling in the presence of an infarct, and when reporting the exact location of both primary and secondary infarcts. The experts' consistent performance also extended to the detection of discrete focal abnormalities and/ or small vessel changes, which were present in some of the cases. Although previous research does not refer to stroke, the differences in experience levels and accuracy are in agreement with much previous literature discussed in chapter 1 e.g. Rogers (1995), Nodine and Krupinski (1998), Manning *et al.*, (2005 & 2006) and Donovan (2006).

When comparing confidence scores across all participants, it was apparent that participants were most cautious when ruling out abnormalities and that the expert readers were very confident in their decisions across all abnormal cases. In control cases, all cases were reported to be challenging, between every reader group and thus, overall confidence was reduced within these cases, when compared with abnormal cases. Over and above this finding was the trend whereby experts consistently reported that normal cases were 'probably absent' or devoid of any abnormal areas, rather than explicitly stating 'definitely absent'. This emergent trend is a reporting behaviour which is likely to transcend from clinical practice; it is preferable to rule out abnormalities with caution to ensure a patient isn't sent home prematurely or to avoid an infarction being missed by a sole reader with either negative consequences for the patient (lack of, or delayed treatment, which may result in disability or death) and/ or the care team (e.g. patient complaint submitted to clinical governance, poor individual/ departmental performance history, and/ or a litigation battle ensues).

Therefore, whilst experts were the most confident group at ruling out abnormalities, more caution was exercised when ruling out clinical problems than when they were clearly detected.

4.3.2 Diagnostic accuracy, Expertise and Eye-movements

While it is comparatively obvious that experts would be the optimum performers across all cases and between reader groups, what is less obvious and least well understood is how cognitive and perceptual processing differed between the groups when viewing separate images of the same and different case types. Not forgetting the relationship between these factors and the resultant decision that is made.

When considering visual search between the groups overall, there were clear differences; not only within images but between them also. Novice visual search patterns were highly variable; from very few fixations per image, to a saturation of fixations across every image within every case. Trainee readers spent much time appraising each case, but appeared to have much more structure than the novices. As trainees became more experienced i.e. fourth year residents were more experienced than first year trainees, their visual search patterns appeared to emulate those of the consultants; a quick time to 'hit', few fixations per image yet key features were consistently explored between readers. In addition, the type of stroke appeared to influence the manner in which the reader group appraised each case. Common trends highlighted an increased or decreased appraisal of the middle slice dependent upon whether a case was normal or chronic for more experienced readers, whereas novices spent more time appraising the first and last images, irrespective of stroke type. This finding may indicate that experts have a predefined knowledge of where the problems are likely to be, whereas novices grasp at the final slice to find problems if they have overlooked or not interpreted earlier findings within prior slices.

As novices were 'naive' readers, appraisal of normal anatomy i.e. the ventricles, eye balls or auditory canals demonstrated a lack of anatomical knowledge, and/ or, general interest regarding these features differed compared with their medical counterparts who have regular exposure to medical imagery. Most trainees were much more thorough than some novices and frequently cross compared hemispheres, implying an understanding of how to detect 'stroke like features' within the image, even if most trainees lacked the wealth of experience to back it up. Whilst these conclusions appear comparatively 'clear-cut', visual search patterns were highly variable both within and

between groups dependent upon the clinical decision made and not even expert readers were free from all errors within this study.

When true positive decisions were made, readers quickly saw the abnormality and gazed directly over its features for approximately 1.4 seconds – a consistent between group finding. The more experienced the reader, the quicker they reached the abnormality coupled with a quick recognition, based upon observed time to 'hit' rating from eye movement data, and swift appraisal of the surrounding cortical regions. Where trainee behaviour differed from an expert, it was more often down to the time spent appraising surrounding tissue rather than time appraising abnormal regions, therefore implying uncertainty regarding the case overall or a lack of confidence to move on quickly separated these readers. Novice readers demonstrated a clear lack of confidence in their decisions by frequently scoring the correct locations but they did not provide a correct rating i.e. they correctly located the abnormality but it was accompanied by a 'probably absent' rating. However, in line with the pilot study, as abnormalities became larger in size and more obvious to detect (such as large chronic or subacute cases), the more correct novice reader eye movements appeared to converge with those of the experienced readers i.e. when the problem was large and obvious the more the novice reader accuracy, confidence and visual search mimicked the expert readers.

When abnormalities were correctly ruled out in true negative cases, all readers spent less time over these decisions than when suspicious areas were falsely judged to be abnormal i.e. readers consistently spent less time per image and case overall than when false positive decisions were made, which is in line with previous research by Manning (2004). However, correctly ruling out the presence of an infarct is not just about economising effort and spending the least amount of time possible on the case, as experts who moved too swiftly through the images made more incorrect decisions than those who were slightly more thorough, even by increasing their reading time by a second over 5 images.

Where experts did make false positive decisions, it was likely more inexperienced consultants would make them. For instance in a few cases, more inexperienced consultants in the area of neuroradiology made feature mistakes whereby a fold in cortical tissue was mistaken for an infarct. In addition, some consultants placed more importance (both visually and when reporting locations) on suspicious regions in the cerebellum and brainstem rather than middle slices when

middle cerebral arterial infarctions are more likely, indicating problems with feature recognition and decision errors rather than missing crucial features altogether, unlike novices. In this study, novices and trainees appeared to have a desire to opt in suspicious areas rather than risk ruling them out as demonstrated by a large number of false positives in these groups. Whilst experts might expect there to be a high proportion of abnormalities in CT patients, the more experienced consultants in neuroradiology have the knowledge and experience to rule out which clinical features are a problem and those which are not, or are indicative of normal ageing.

Previous studies which examined false positive decisions in single patient images, uncovered that where single or clustered fixations appeared to be 1000 milliseconds in duration or over, their accompanying decisions are likely to be incorrect (Manning, 2005). In this multi-slice study it is more difficult without designing a specific programme to assess exactly where and when a FP decision takes place to examine the exact fixation duration which accompanies an incorrect location, although frequently the region can be implied by the gaze-tracker results. In addition, the opportunity existed to revisit the suspicious location in another image slice, and therefore, multi-slice image analysis might have further implications for these types of decisions, which has not been previously exposed in single image experiments and further multi-slice studies will have to unravel whether the 1000 ms 'rule' still applies or whether a new threshold is more appropriate.

Where readers displayed the opposite behaviour i.e. incorrectly ruling out problem areas with false negative ratings, their eye movements were accompanied by either long fixation times within an AOI than true positive decisions, (indicating decision errors) or were barely fixated upon and/ or missed altogether (indicating recognition or search errors). Mean fixation times to reach an AOI and time spent within the slice overall appeared to be just as variable between the groups. As a group, novices had more detection and recognition errors than any other readers, which may indicate eagerness to completely rule out abnormalities and finish the task, or maybe it is more likely their lack of knowledge regarding the prevalence of stroke within the experimental cases prevented them from making the right decisions. In addition, novices might underrate the number of abnormal cases within those who go for a CT scan, whereas trainees and experts are more likely to err on the side of caution when ruling out abnormalities or expect a high proportion of individuals who had a CT scan to have an abnormality present. When experts made false negative decisions, it was because they did not see the abnormal area at all, which was much more likely when the infarcts were acute and/ or discrete in appearance.

When experts correctly detected secondary abnormalities, they spent much more time appraising the secondary than the primary and therefore did not reach a satisfaction of search in these cases. In addition, incorrect experts often spent enough time to appraise the clinical features, rather than miss it altogether and/ or move onto the next case, rather it appears they did not consider the region suspicious enough to classify as an infarct. As secondary abnormalities are less common than a single primary infarct, an increased fixation duration is linked with an increase in cognitive processing in experts also, compared with novices and trainees who appeared to become 'fixated' by the primary and reached a satisfaction of search.

Inexperienced readers who correctly located a secondary may have done so genuinely, or more likely by a 'best guess' as indicated by the increased number of false positives by these readers *per se*. However, some novice readers did appear to perform equally as well as some trainee readers, as not much separated their performance in terms of accuracy ratings and visual search patterns, which, although drawing comparisons with a different clinical group, is in line with findings from a previous study comparing novices with pre-trained radiographers (Manning, 2006). Whilst novice participants had no prior experience reading neuroradiological images, those who were optimal performers within their group were from either a design engineering profession or from a research background and may have a preference for visual tasks over other cognitive performance tasks.

Expert performance for focal abnormalities and small vessel change detection did not fall below 60% accuracy yet trainee performance fell below 30%, indicating that as abnormalities became more discrete trainees were less able to distinguish not only between the features of normal and abnormal anatomy but also between what constitutes a focal and/ or small vessel change classification. Many trainee and expert readers mistook focal abnormalities for small vessel changes and vice-versa in these CT cases. This misclassification error could be because focal abnormalities, and particularly small vessel changes, are more difficult to detect and classify in CT images, or it could be that readers were unable to distinguish between the two with their current knowledge base.

Misclassification errors were not solely found in reporting practices but much less frequently, a few general reporting errors were made by participants i.e. circling the wrong hemisphere on the reporting sheet or highlighting an area which bore no relevance to clinical features or their eye movement data results. These perceived errors could be down to a number of factors; a) the location

was recalled retrospectively, rather than during the case analysis, b) consultant time pressure resulted in a desire to complete the task and continue on with their working day, c) natural fluctuations in observer concentration which impacts upon performance and or d) a limitation of the study reporting design i.e. having to report a 3D feature on a 2D reporting sheet. If the latter hypothesis is true, the reporting could have been more clearly defined if participants were able to mark the area within the image itself. Unfortunately due to design constraints this was not achievable within this CT or the following MRI study at the time of completion. Other limitations included the inability of readers to scroll up and down image stacks, and use pan and zoom functions at will, which is possible within normal clinical practice. For experimental design reasons this could not be permitted here.

Where recommendations could be made for optimising performance or for future training packages, it appears that an increased knowledge of neuroanatomy coupled with knowledge of cerebrovascular pathways, including areas most susceptible to arterial occlusion might increase performance when searching for abnormal features in CT. In addition, a targeted training programme to assess recognition of normal sulci with infarct features, including which areas of low attenuation are normal or suspicious, could increase performance from an above average level, to exceptional, within the trainee group and as a quick but intensive training session for some experts who appeared less familiar with neuroradiology.

4.4 Conclusions

Overall this study highlighted that experts performed with more accuracy and confidence than novices and trainees. Experts were quicker to identify abnormal areas, spent more time in challenging areas of interest (AOI) and were more consistent throughout the reading task. Conversely, trainees spent a large amount of time searching for secondary abnormalities and had more foveal fixations per image than experts. In line with the previous pilot study, novices performed optimally when lesions are large and obvious but did not recognise focal abnormalities. Novices took longer to reach an AOI and also took longer to complete the study overall compared with trainee and expert radiologists.

In terms of the search patterns accompanying different clinical decisions, true positive decisions were characterised by a quick time to lesion whereas true negative decisions were characterised by a thorough, yet not hasty image appraisal. Both decisions were associated with confidence and experience. False positive decisions were more likely as experience decreased and where the image was not fully appraised, or where the reader was inexperienced, abnormalities may be missed altogether or go unrecognised. These findings are in line with much research, even from differing disease manifestations such as chest or breast.

4.5 Study Reflections

To-date few studies have explored observer performance in neuroradiology and the present study is the first to fully examine multi-slice image appraisal of stroke representation and detection performance in CT imagery. The findings demonstrate that visual search patterns within this study are in line with findings from other researchers within this field, when single images are considered. Although the present study could be considered comparable to mammography studies that make medio-lateral and oblique images of the same patient available to observers, this study remains different from other visual search tasks as it considered visual search within and between multiple images (i.e. five images) of the same organ, imaged in the same orientation, and within the same patient i.e. not two opposing images of the same organ, within the same organism that are essentially different owing to the orientation of the images. In future studies, it is planned to enhance this difference further between studies by offering hundreds of images of the same patient and of multiple orientations to compare differences in visual search in an attempt to demonstrate brain imaging examination by the radiologist which is as close to the clinical task as possible, within experimental constraints.

One of the interesting findings from this study was the differential search patterns between readers of different stroke types and in this study, radiology trainee and expert visual search patterns appeared to be characteristic of a particular case type i.e. control, acute, subacute and chronic, when viewing the small cross-section of clinical images, so indicating that images from each case type in stroke are examined in much the same way - a unique finding which has not been uncovered in previous literature and will be further explored in the following study into MRI images of stroke cases.

Chapter 4

All visual search patterns and observer performance dimensions explored within this study will be compared with MRI to unravel whether, and how, modality type has an influence over observer performance and visual search. In addition, as many misclassification errors were uncovered, these errors will be explored within the following chapter and comparisons will be drawn in chapter 6 (CT versus MRI) regarding whether classification errors were still evident in MRI.

Study 2: The Influence of Expertise in Stroke MRI Interpretation and Eye-Tracking.

The previous chapter uncovered differences in performance and visual search within CT images of stroke. Magnetic Resonance (MR) imaging is also frequently relied upon to identify, diagnose and recommend treatment for cardiovascular disease, but particularly the identification of subtle abnormalities as well as regular stroke cases. When MRI has been explored to compare rates of sensitivity and specificity, variable rates were also reported between studies; 88% sensitivity for spinal abnormalities and 82% overall, whilst specificity was reported to be 81% when 100% when cases were viewed by two neuroradiologists and one research fellow (Haughton *et al.*, 1986). Specific stroke rates have been reported to be 18% sensitivity and 100% specificity in a study by Gonzalez (1999), but 58% sensitivity and 100% specificity by Mullins *et al.* (2002)[b].

Despite modality performance research, visual search in neuroradiological images has not been extensively explored, highlighting a paucity of research from a human observer perspective in this clinical field. In addition, whilst stroke presentation has received attention from researchers to compare sensitivity and specificity rates (Mohr, 1995; Lansberg, 2000; Mullins, 2002[b]; and Wintermark, 2007), observer performance and visual search in the examination of stroke images has not been previously explored, particularly in MRI. Therefore, a study was undertaken to explore eye-movements and observer performance across differing levels of expertise for MR, multi-slice imaging in the interpretation of stroke.

Study Aims and Objectives: To explore the visual interpretation of magnetic resonance brain images i.e. observer detection and omissions of infarctions using eye-tracking, including, how experts appraise an image compared with radiology trainee and experts, as per chapter 4. The hypothesis in this chapter is that further group differences will be evident in MRI owing to the increase in anatomical detail and structural noise.

5.1 Methods

5.1.1 Participants

This study required a selection of participants with a comprehensive range of experience of reading medical images; therefore, novice, trainee and expert readers were identified and recruited. To allow further investigation of MR compared with CT performance, the same 28 participants were recruited, with participant demographics the same as chapter 4.The total number of MR cases assessed by trainees within the year prior to the study was 422 patient cases. The range of MR cases assessed was between 1 and 141 cases per person. The median number of cases was 12. The total number of MR cases assessed by radiologists within the year prior to the study was 2,144. The range of MR cases assessed was between 11 and 1,309 cases per person. The median number of cases was 94.

5.1.2 Design

In line with the overall research design, as detailed in chapter 2, one-hundred and twenty single MR clinical images were selected and made anonymous from a bank of twenty four predetermined clinical cases (five single images were selected from each case). The clinical cases selected represented a spread of six normal (controls), eight acute cases, six subacute cases and four chronic cases. The cases were selected by the SpR on the basis of the radiology report information, which issued the final diagnosis and stroke classification. Following image selection and abnormality classification, a computer-based, eye-tracking study was subsequently developed to assess diagnostic accuracy and interpretation in stroke MR imagery.

5.1.3 Procedure

Prior to case assessment, a short training exercise was conducted regarding infarct location to provide baseline knowledge regarding the clinical features of stroke, as presented by MR imagery. When case assessment did commence, participants rated each case on a four-point Likert scale, namely whether a primary abnormality (i.e. stroke) was; 1) definitely present, 2) probably present, 3) probably absent, or 4) definitely absent. If an abnormality was considered present, participants were required to confirm the location of the infarct on a separate brain atlas task. In addition, radiology trainees and consultant readers were encouraged to mark on the brain atlas if they were also aware of the presence of small vessel changes with a single or multiple ‘x’ on the reporting sheet. For further information regarding image and case type selection, group allocation and experimental procedure, please refer to chapter 2.

5.2 Results

Study data was analysed to investigate; i) qualitative image analysis ii) accuracy and confidence ratings of performance, iii) quantitative eye movement analysis, and iv) stroke, expertise accuracy and visual search in MR imagery.

Case study 1. Normal control case (NHG): The following gaze-tracker images highlight the differences between readers' (a novice, a trainee and an expert) visual inspection strategies when appraising images of normal control cases in MRI. N.B. When reading and reporting MRI it is important to remember that the left side of the image represents the right side of the individual, and vice versa due to the acquisition process.

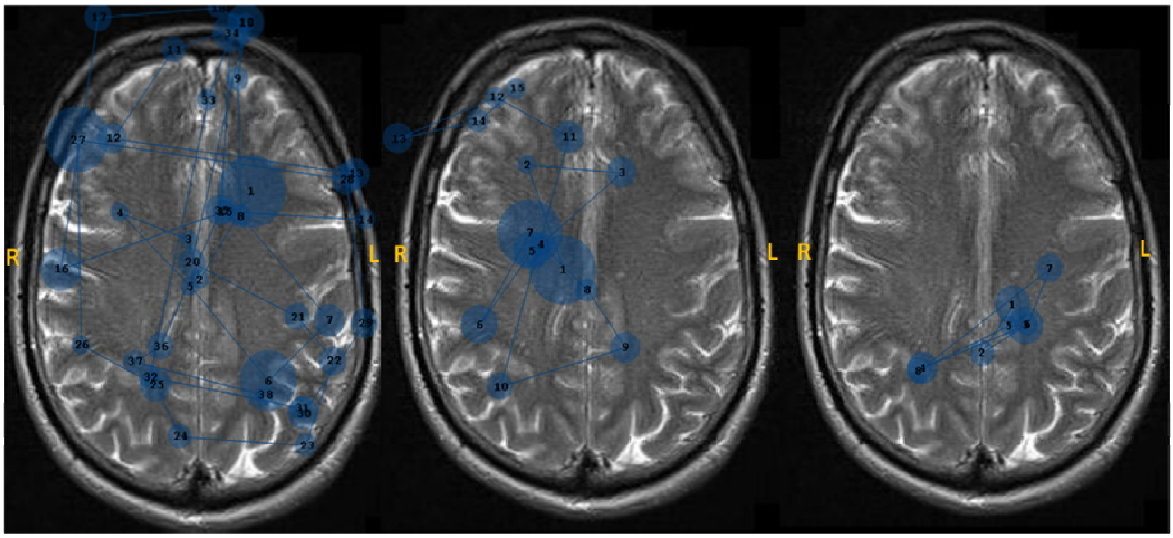


Figure 5.1 represents the eye movements of a novice, a trainee and an expert reader of a normal MR case.

Figure 5.1 highlights a pattern of reduced fixation with increase level of expertise. An increased number of fixations appear to be indicative of increased searching, decision-making and/ or confusion, although the official reason cannot be ascertained without discussing the behaviour with the reader. The expert spends very little time on the image and may well be investigating the possibility of small vessel changes as many subtle details are evident on the original image. Experts may well 'zone out' of an image, focussing on a small region yet using global visualisation and possibly recognition, yet this phenomenon cannot be proven in the present thesis. When considering group behaviour for this case, novices appear to examine most regions of high signal (i.e. normal sulci), which could be considered abnormal with little consistency within the group. Trainee readers focus on the midline and make many cross-hemisphere comparisons, whilst experts appear to make

comparisons between the front and the back of the brain; to view all gaze tracker images for this case, please refer to pages 19, 20 and 21 of the appendix.

Case Study 2. Acute stroke (ACW): The following gaze-tracker images highlight the differences between readers' (novice, trainee and expert) visual inspection strategies when appraising images of acute stroke in MRI;

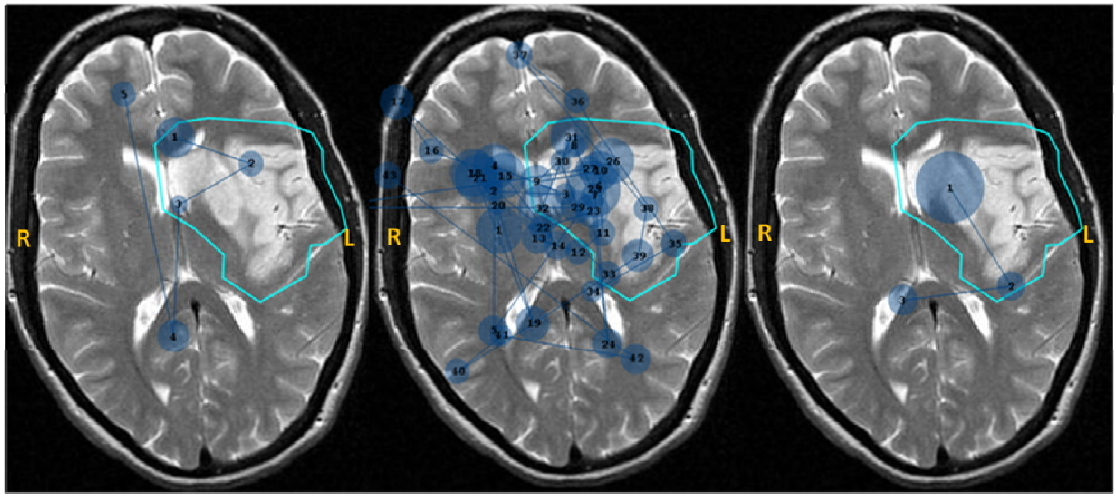


Figure 5.2 represents the eye movements of a novice, a trainee and an expert reader of an acute MR case.

Figure 5.2 highlights the fixation patterns of novice, trainee and expert readers. These images demonstrate how an experts' attention is quickly drawn to the acute stroke and they move on after 3 fixations. The novice participant also perceives the infarct quickly and moves on. Conversely the trainee 'hits' the infarct on the third fixation but continues to appraise the rest of the image. In terms of group behaviour, novices spend rather little time on this slice, trainees, as a group spend more time than novices and expert fixation numbers resemble the novices more than the trainees (in terms of number of fixations). To view all gaze tracker images for this case, please refer to pages 23, 24 and 25 of the appendix.

Case study 3. Subacute stroke (SAC): The following gaze-tracker images highlight the differences between readers' (novice, trainee and expert) visual inspection strategies when appraising images of subacute stroke in MRI;

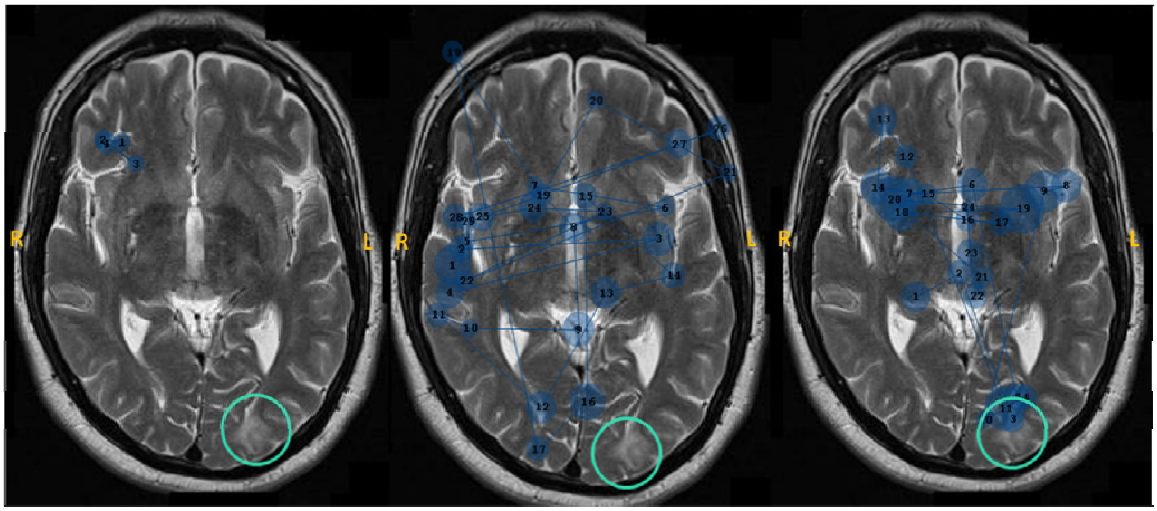


Figure 5.3 represents the eye movements of a novice, a trainee and an expert reader of a subacute MR case.

Novice and trainee readers miss the subtle infarct (FN) whilst the expert reaches it within the third fixation. As a group, only three novices looked at the infarct but only one recognised it – others glanced over it without returning for a further fixation. Most trainees looked at the abnormality but not all recognised it as abnormal. Experts again appeared to spend time drawing comparisons between the hemispheres in this case and there is a clear indication of less cognitive overload than the trainee image results with fixations being much shorter in duration than novices and trainees. To view all gaze tracker images for this case, please refer to pages 27, 28, and 29 of the appendix.

Case study 4. Chronic stroke (CSS): The following gaze-tracker images highlight the differences between readers' (novice, trainee and expert) visual inspection strategies when appraising images of chronic stroke in MRI;

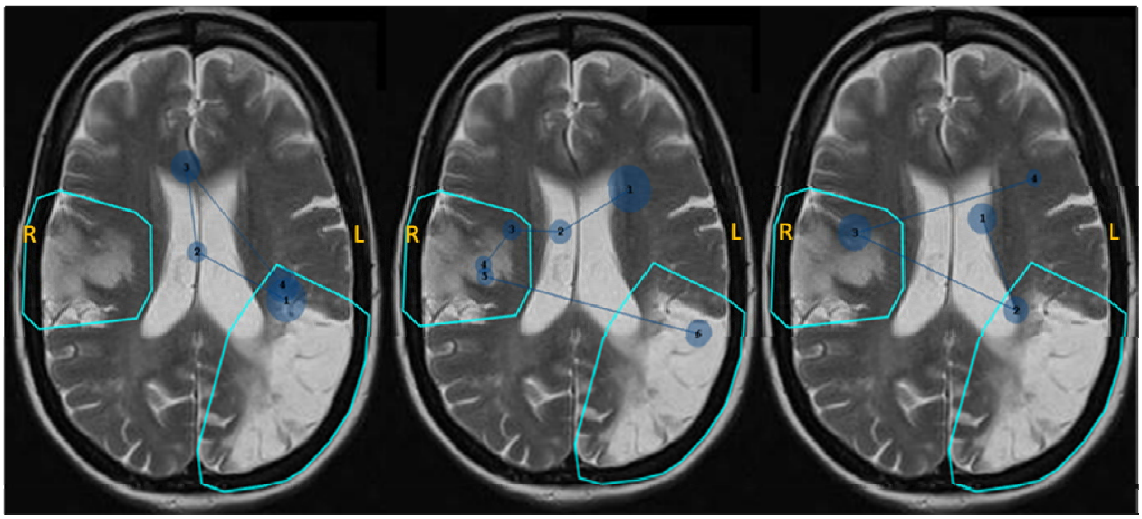


Figure 5.4 represents the eye movements of a novice, a trainee and an expert reader of a chronic MR case.

This case study highlights experts are quickest to both lesions (TP), although the 4th year trainee was just a fraction slower than the expert. The novice reader is distracted by the high signal of the secondary, and completely misses the primary in the right hemisphere. In terms of group behaviour, novices examined this case less than trainees and experts and in this case trainee and expert visual search patterns were comparatively similar. To view all gaze tracker images for this case, please refer to pages 30, 31 and 32 of the appendix.

Summary;

- In line with the previous CT study, novices were inconsistent; eye movements either over or under appraised the images and spent more time examining normal anatomy.
- Some trainee eye movements did not differ extensively from novices, although trainees did appear to cross compare hemispheres more than novices.
- Experts were efficient and consistent in their visual search patterns, spending most of their time within area of interest with fewer, and shorter, fixations than both novices and trainees.

5.2.1 Accuracy and Confidence Rating Data: Quantitative Analysis of Observer Performance

Please refer to chapter 2 for further details regarding how measures of accuracy and confidence between and within participants were derived.

5.2.1.1 Localisation and Case Difficulty

The following table compares all readers collectively to gain an indication of individual case and overall perceived case difficulty in magnetic resonance imaging.

Table 5.1 lists all MR cases with average accuracy scores for all participants and an overall performance rating per stroke type to establish, which cases were perceived to be most challenging.

Case TYPE	Individual patient case	% CORRECT	Overall case % correct
Normal	NCK	83.3	
	NGB	75.8	
	NHG	65.8	
	NJH	76.7	
	NTH	90.0	
	NPM	75.8	77.9%
Acute	ACG	51.7	
	ACW	100.0	
	AED	100.0	
	AJC	70.0	
	AJE	100.0	
	AMB	66.7	
	ARJ	56.7	
	ASC	100.0	80.6%
Subacute	SAC	38.3	
	SCD	56.7	
	SEB	93.3	
	SMC	96.7	
	SNE	86.7	
	SSA	100.0	78.6%
Chronic	CAD	90.0	
	CBA	80.0	
	CCW	93.3	
	CSS	80.0	85.8%

Overall % of cases correct indicates that control cases were perceived to be most difficult by all participants (77.9%), followed by subacute (78.6%), acute (80.6%) and then chronic stroke types

(85.8%). Within the top ten most challenging cases were four acute cases, four normal control cases and two subacute cases, with the most challenging case being SAS.

5.2.1.2 MR Receiver Operating Characteristics for Primary Infarction Detection

The following Conventional Binormal ROC Curves represent confidence ratings for primary abnormality detection ratings by participant group; novice, trainee and expert readers in MR. Within the ROC space the parameters 'a', 'b' and 'Az' represent the vertical intercept, the slope of the fitted curve and the overall accuracy value as calculated by the conventional Binormal curve process, respectively.

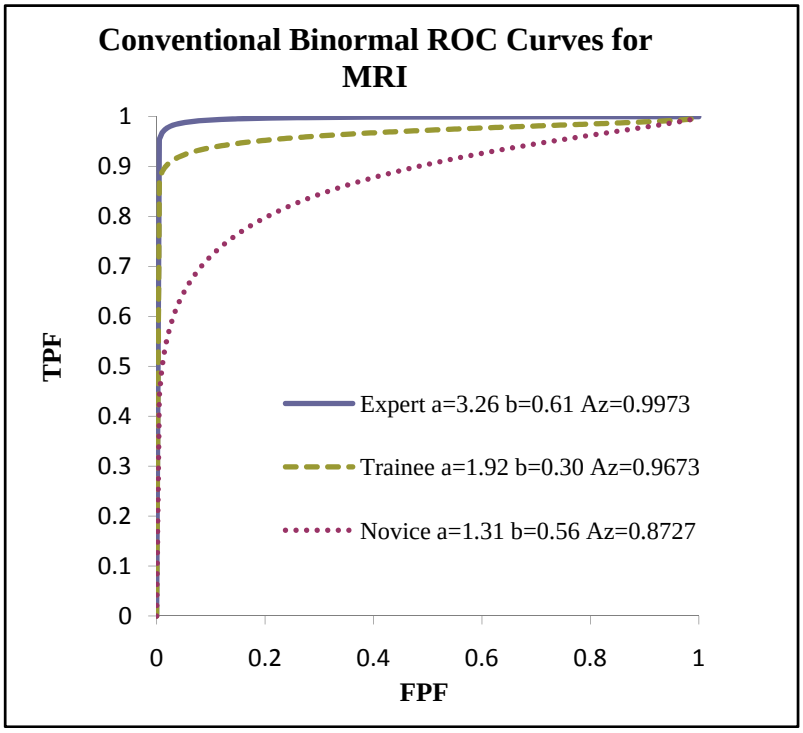


Figure 5.5 ROC curves for novice, trainee and expert readers in MRI.

The above ROC curve demonstrates that the expert readers provided greater accuracy score compared with trainee and novice when examining these MR cases, within this experimental study on the basis of their confidence rating scores that accompanied the cases. When considering accuracy of localising primary abnormalities in MRI, novice participants correctly reported the absence of, or abnormal locations in 64.3%, trainees 81.4% and experts 96.6% of patient cases.

Table 5.2 highlights the correct diagnostic accuracy of participant groups between patient cases i.e. normal, acute, subacute and chronic stroke types.

Stroke TYPE	Normal	Acute	Subacute	Chronic	Overall performance
Novice	56.7	58.8	61.7	80.0	**64.3%**
Trainee	83.3	86.3	78.3	77.5	**81.4%**
Expert	93.8	96.9	95.8	100.0	**96.6%**

Table 5.2 demonstrates that experts were the most accurate when detecting or ruling out the presence of an abnormality across all four case types. Novice participants had more difficulty ruling out the presence of an infarct than detecting all other stroke types, whereas trainee participants had the most difficulty with chronic cases (77.5%).

When ANOVA tests were performed to examine between group difference of performance, there were significant group differences for diagnostic accuracy in MR (*df* 2, *p*<.000, F=42.69, Eta Squared=.11). Games-Howell posthoc comparisons demonstrated that the there were significant differences between all group in MRI appraisal i.e. between expert and novice participant groups (*p*<.000), trainee and expert readers (*p*<.000), and novice and trainee readers (*p*<.000) with experts being the optimum performers when pinpointing the exact locations of an infarct, than novices and trainees.

5.2.1.2 Confidence Rating Data: Quantitative Analysis of Primary Infarctions.

When reporting participant confidence in their decision-making, descriptive statistics show overall percentage of cases correct indicates that experts were the most confident in their clinical decisions (82.3%), shortly followed by trainees (80.1%) and then novices (73.8%).

Table 5.3 highlights the confidence of participant groups when making decisions regarding patient cases i.e. normal, acute, subacute and chronic stroke types.

Stroke TYPE	Normal	Acute	Subacute	Chronic	Overall confidence
Novice	51.7	81.3	75.0	90.6	**73.8%**
Trainee	60.4	95.9	91.7	93.1	**80.1%**
Expert	66.1	99.6	96.4	99.2	**82.3%**

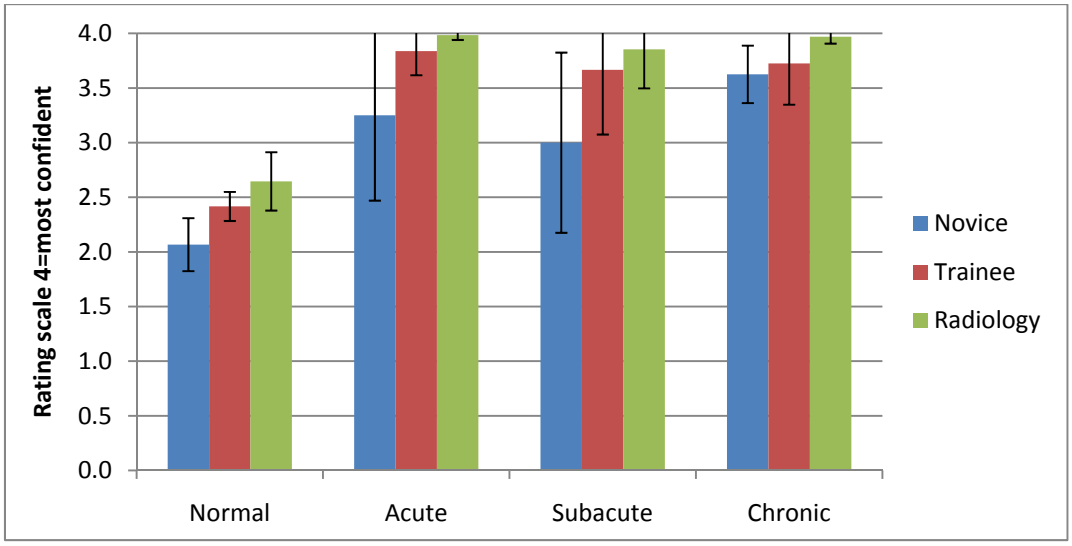


Figure 5.6 Mean confidence score per case by participant groups when making decisions regarding patient cases i.e. normal, acute, subacute and chronic stroke types.

Table 5.3 demonstrates that the Radiologists were most confident at ruling out abnormalities with their score of 66.1%, followed by the trainees (60.4%) and novice readers (51.7%). The table also indicates that Radiologists were most confident at rating the presence of abnormalities in MR with a high average score of 82.3%, followed closely by the trainees (80.1%) and then the novice readers (73.8%). Whilst trainees were very confident in their decisions, the previous accuracy section highlights that their decisions were not necessarily accurate. Experts were significantly more confident in their decisions than novices ($p<.004$), and trainees were more confident than novices ($p<.023$) in Tukey posthoc comparisons (df 2, $p<.001$, F=6.85, Eta Squared=.002). Overall % of cases correct indicates that control cases were perceived to be most difficult by all participants (78.5%), followed by subacute (80.3%), acute (81.0%) and then chronic stroke types (91.0%).

5.2.2 Diagnostic Decision-Making (i.e. TP/FP/TN/FN)

As demonstrated above, experts are the optimum performers in MR. The following table 5.4 highlights the spread of true and false positive ratings and true and false negative per group to gain a more in-depth insight into the decision-making processes of participants throughout the patient cases. The figures were derived by totalling all individual decisions for the primary infarction (or lack thereof) for each case and each group, followed by dividing the figure by the number of participants in each group to reach an average figure per decision. It is important to remember that there were 6 control and 12 abnormal cases per experiment and therefore the original spread of cases was not equal.

Table 5.4 demonstrates average reader group decisions (i.e. TP/TN/FP/FN) in MRI.

MR Average	TP	TN	FP	FN
Novice **(n=10)**	11.8	3.5	4.4	6.2
SpR **(n=10)**	15.5	5.0	2.6	2.5
Radiol **(n=8)**	17.5	5.6	0.6	0.5

In line with CT chapter 4, the above table demonstrates that, radiologists had more true positive and true negative ratings than novices and trainees. Radiologists also had fewer false positive and false negative results than novices and trainees. Novices once again had the most false positive and false negative results i.e. they ruled out the presence of an abnormality more than trainees and experts. Overall sensitivity rates between groups were; 74%, 92% and 98% for novice, trainee and expert readers. Specificity rates were 41%, 55% and 92%, respectively. Highlighting that, overall, not much separated the trainee and novice groups.

5.2.3 Eye-movements and Experience: Quantitative Analysis.

The primary aim of this results section is to identify how visual search differs quantitatively between novice, trainee and expert readers in MRI of normal patients and those who have suffered a stroke. The secondary aim of this results section is to identify statistically significant links between visual search behaviours (e.g. total time spent on the task, time to reach an AOI, time spent within and out of AOI's) and reported accuracy within and between participant groups in MRI to gain an insight into the way inexperienced through to experienced readers appraise medical images.

5.2.3.1 Total Task Viewing Time per Group

When considering total viewing time per reader group within this modality, there were significant differences between total case fixation durations and reader group with trainees fixating more over the 5 image slices than novices ($p<.000$) and experts ($p<.000$). There were also significant differences between novices and experts ($p<.042$). In terms of the number of fixations between groups in MRI; trainees fixated much more than experts ($p<.000$) and novices ($p<.000$), but there were no differences between novice and experienced readers.

5.2.3.2 Visual search behaviour throughout the image 'stack'

Having uncovered that there were significant differences between reader groups when viewing each case image, the following figures (5.7-5.10) aim to demonstrate visual search behaviour of participant groups by investigating how mean fixation duration of gaze differs between each axial slice throughout the five image 'stack' by case type (control, acute, subacute and chronic) and level of experience (novice, trainee and expert).

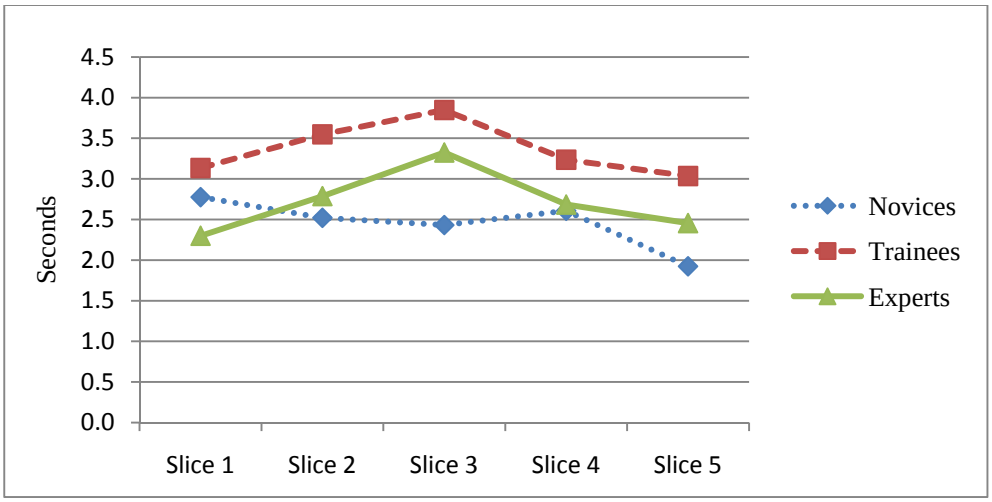


Figure 5.7 Mean fixation time per axial slice for all readers of normal patient MR images.

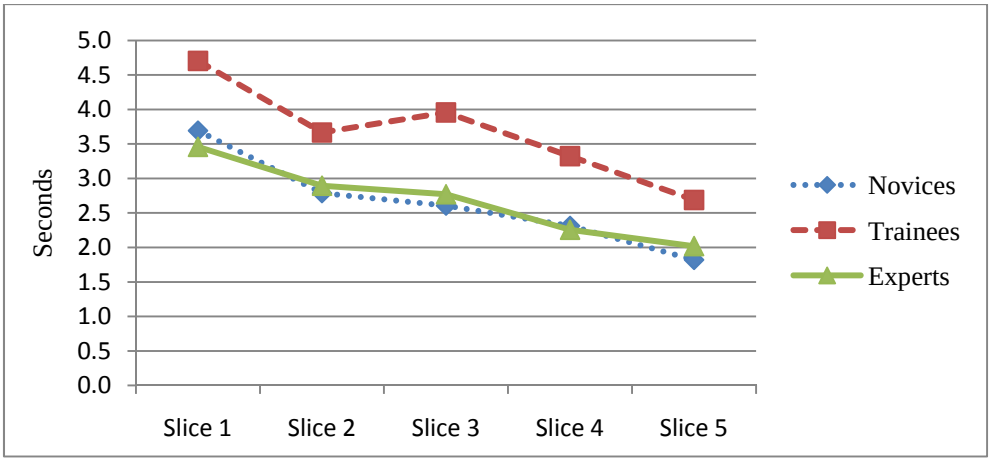


Figure 5.8 Mean fixation time per axial slice for all readers of acute patient MR images.

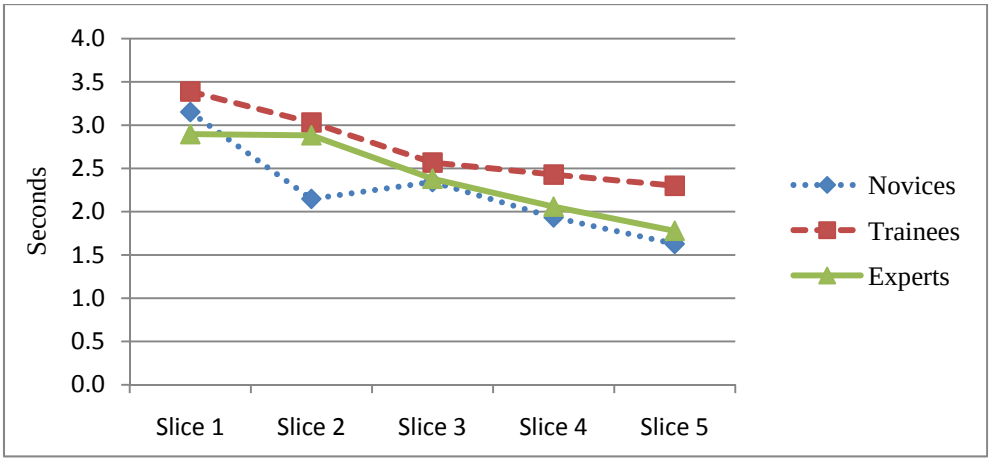


Figure 5.9 Mean fixation time per axial slice for all readers of subacute patient MR images.

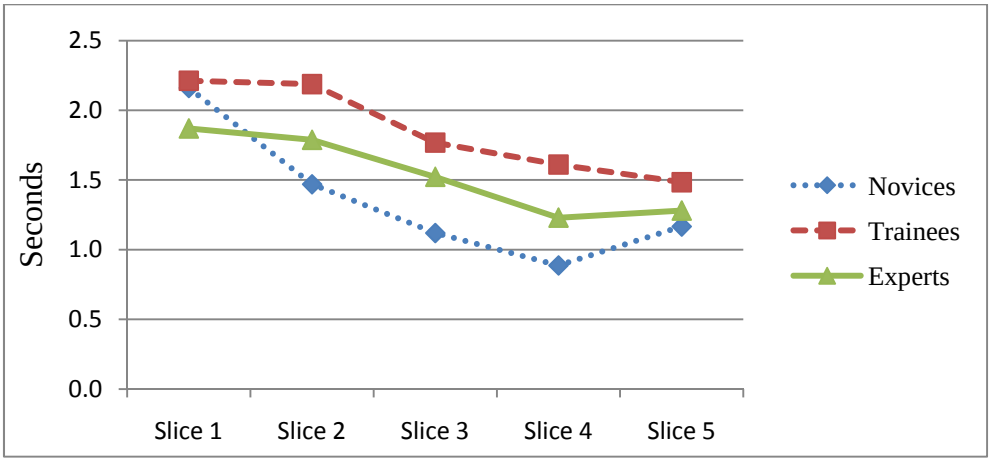


Figure 5.10 Mean fixation time per axial slice for all readers of chronic patient MR images.

Figure 5.7 demonstrates a trend towards trainee and expert readers following a similar fixation pattern when viewing normal patient images, yet trainees spent up to a second longer within each slice than experts. In figure 5.8 (acute images), novice and expert readers follow a similar fixation trend with the most time in slice 1, which steady decreases across the remaining four slices. In this case, trainees spend up to 1.5 seconds longer appraising each slice than novice and expert readers, although appearing to follow the same trend. In figure 5.9 (subacute images), there is not a clear trend despite all participants spending the most amount of time in slice 1 and the least in slice 5, as novices spent much less in slice 2 than experts In chronic images (graph 4), whilst trainee and expert times across slices gradually diminishes to below 1.5 seconds, novice reader time from the previous slice increases from 0.9s to 1.2s. In figures 5.7-5.9 novice readers followed a similar pattern

throughout i.e. spent the most amount of time in slice 1 and the least in slice 5, unless they were appraising chronic cases, in which case they spent the least amount of time in slice 4.

Table 5.5 demonstrates the mean fixation time per axial slices within each case type and per participant group in MRI.

MRI mean fixation duration across slices		Slice 1	Slice 2	Slice 3	Slice 4	Slice 5	Mean total fixation duration per case type
Normal	Novices	2.8	2.5	2.4	2.6	1.9	2.5
	Trainees	3.1	3.5	3.8	3.2	3.0	3.4
	Experts	1.8	2.2	2.7	2.1	2.0	2.2
Acute	Novices	3.7	2.8	2.6	2.3	1.8	2.6
	Trainees	4.7	3.7	4.0	3.3	2.7	3.7
	Experts	2.8	2.3	2.2	1.8	1.6	2.1
Subacute	Novices	3.2	2.1	2.3	1.9	1.6	2.2
	Trainees	3.4	3.0	2.6	2.4	2.3	2.7
	Experts	2.9	2.9	2.4	2.1	1.8	2.4
Chronic	Novices	2.2	1.5	1.1	0.9	1.2	1.4
	Trainees	2.2	2.2	1.8	1.6	1.5	1.9
	Experts	1.5	1.4	1.2	1.0	1.0	1.2

In terms of mean fixation duration over the 5 slices, table 5.5 demonstrates that trainees spent longer appraising all case types in MR images, especially acute stroke types, than both experts and novice readers. Overall and prior to considering decisions type (i.e. TP/FP/FN/TN), ANOVA significance testing uncovered that there were no significant differences either within or between groups when comparing time to hit rates and time in the area of interest between readers in MRI. There were, however, highly significant differences between groups when comparing the time spent outside of the area of interest (*df* 2, $p<.000$, F=10.25, Eta Squared=.41), with trainees ($p<.000$) and novices ($p<.007$) spending much more time searching the surrounding areas of the image than experts in MRI.

5.2.4 Stroke, Expertise, Accuracy and Visual Search

5.2.4.1 Diagnostic Decision-Making and Eye-Movements i.e. TP/FP/TN/FN

In line with the CT chapter, the aim of this section is to also examine whether certain visual search behaviours are indicative of diagnostic accuracy but in MRI. The section also aims to examine mean

time to hit the primary AOI when it appears in the image stack for the first time, the time spent within that AOI, time spent out of the AOI and the mean fixation time within the same slice overall, in an attempt to uncover visual search patterns associated with true positive and false negative decisions and whether the same findings are uncovered when the MR modality is considered.

5.2.4.2 True Positive Decisions

This section considers the eye movements of participants who provided a true positive decision i.e. a correct rating to indicate the abnormality was either probably or definitely present, which accompanied a correct location of the primary abnormality on the 'brain atlas' task.

Table 5.6 highlights true positive eye-movement data by group and stroke type (in seconds) in MR.

MR True Positive eye-movement data by group and modality (in seconds) in first appearing AOI.		Mean time to hit Primary AOI	Mean time spent in Primary AOI	Mean time out of AOI (in same slice)
Acute	Novice	1.3	1.6	2.4
	Trainee	1.0	1.6	3.4
	Expert	1.3	1.3	2.5
Subacute	Novice	1.6	1.4	3.9
	Trainee	1.3	1.5	3.5
	Expert	1.8	1.2	3.9
Chronic	Novice	1.2	1.3	2.0
	Trainee	2.1	1.4	2.6
	Expert	1.3	1.3	2.5

The above table descriptively demonstrates that in true positive decisions, trainees were quickest to reach the primary AOI in acute stroke cases (1 second), shortly followed by both novice and expert readers(1.3s). When experts did reach the AOI, they spent an average of 300 milliseconds less appraising the abnormal tissue than novices and trainees. Trainees also spent one second longer on the image than experts but whilst time spent on this image was significantly different between trainee and expert reader (*df* 2, *p*<.001 F=7.42, Eta Squared.36), time to 'hit' was not.

In subacute cases, trainee readers were again the quickest to fixate within an AOI; however, they also appraised the feature for 1.5 seconds, which was longer than novices and experts. Expert readers took an average of 1.8 seconds to fixate within the AOI, and as comparable with acute cases, spent less time appraising the anatomy than novices and trainees. In chronic stroke cases, novices were quickest to hit the AOI within 1.2s on average, 900 milliseconds quicker than trainees. Novices

also spent the same amount of time in the AOI as experts but not the same time appraising surrounding tissue, indicating that novice true positive results regarding primary chronic abnormalities were very much in line with their far more experienced counter participants yet they lacked the additional search for secondary problems.

Overall, true positive decisions appear to be characterised by a quick time to hit, often around 1.3 seconds or less for experienced readers and accompanied by an appraisal of abnormal tissue of around the same time (1.2-1.3 seconds). Experienced readers in this study spent 4.4 seconds on average in the slice where the abnormality first appeared and a correct decision followed. On average, trainees spent 4.6 seconds and novices slightly less at 4.2 seconds, to reach a correct decision.

When these results were compared overall using ANOVA statistical tests, there were no significant differences in terms of time to hit, but the groups did differ significantly in terms of the total time outside of the AOI (*df* 2, *p*<.000 F=13.00, Eta Squared.61), total number of fixations (*df* 2, *p*<.000 F=14.32, Eta Squared.07), and total fixation durations within a case (*df* 2, *p*<.000 F=16.91, Eta Squared.78), with trainees fixating more and for longer than novices (*p*<.000) and experts (*p*<.000). Novices spent much less time appraising the surrounding cortical areas than trainees (*p*<.000) and trainees spending more time than experts (*p*<.001) over true positive decisions.

5.2.4.3 False Negative Decisions

This section considers the eye movements of participants who provided a false negative decision i.e. an abnormality was present and it was either missed, not recognised as an abnormality or was not correctly reported.

Table 5.7 highlights false negative eye-movement data by group and stroke type (in seconds) in MR.

MR False Negative eye-movement data by group and modality (in seconds) in first appearing AOI.		Mean time to hit Primary AOI	Mean time spent in Primary AOI	Mean time out of AOI (in same slice)
Acute	Novice	1.5	0.6	3.9
	Trainee	1.4	1.1	3.1
	Expert	0.8	0.3	1.3
Subacute	Novice	0.1	0.9	2.6
	Trainee	0.6	0.3	4.3
	Expert	0.7	0.3	2.1
Chronic	Novice	0.0	0.7	0.0
	Trainee	3.5	0.7	4.6
	Expert	-	-	-

Table 5.7 examines participants that provided false negative accounts of acute cases. On average, incorrect decisions in acute cases appear to be characterised by either an even quicker time to hit for experts or much longer than true positive decisions for trainees and novices, yet much less time appraising the AOI than in true positive decisions by all other participants within their group. Novices took 200 milliseconds longer to reach the AOI but also spent a second less time appraising the abnormal area. Whilst novices spent marginally longer in the slice overall than in true positive decisions, trainees and experts spent about a second less, indicating they did not spot the abnormality and moved onto the next axial slice.

In subacute cases, all participants reached the AOI very quickly but barely fixated within the AOI, indicating it did not retain their attention once it had been gazed over. In chronic cases, expert readers did not make any false negative decisions. Of those novices and trainees who did make false negative decisions, once again novices reached the AOI quickly but trainee readers took 3.5 seconds on average. Trainees also gazed over the AOI towards the end of the image appraisal indicating a recognition error, prior to moving swiftly onto the next axial image. Neither group spent much time appraising the AOI as seen in the acute and subacute cases.

Overall, it appears that false negative decisions made by participants in this MR study, are characterised by either an even quicker time to hit or much longer than true positive decisions, yet much less time appraising the AOI than in true positive decisions by other participants within their group. Experienced readers who made a false negative decision spent 2.9 seconds on average in the slice where the abnormality first appeared, whereas trainees spent 4.4 seconds, and novices 3

seconds which are slightly less time than in true positive decisions. ANOVA results demonstrated that the trends in false negative decisions were not significantly different between groups.

5.2.4.4 False Positive & True Negative Decisions

As a number of false negative decisions were identified between participant groups and stroke types, the following section aims to examine whether the errors could be attributed to recognition and/or decision errors of normal features by further exploring the participant reports and their eye movements. Further analysis was applied to examine false positive decisions; specifically those made by experts and compare their visual search patterns with true negative decisions by other experienced readers in the study.

Before proceeding onto the next section, it is important to consider that only abnormal areas can be predefined by creating an AOI before the experiment took place, therefore, the exact time and slice where one, or more, false positive decisions were made (based on certain clinical features) cannot be completely determined retrospectively owing to the 2D reporting strategy of a 3D anatomical structure. However, eye movements throughout the image stack in these cases often infer the clinical features that were implicated in the decision error, as demonstrated in the following MR images and overlaid gaze patterns.

Table 5.8 highlights false positive eye-movement data (in seconds) by group and stroke type in MR

Pptn Group	FP Location	No of FP's	Slice 1	Slice 2	Slice 3	Slice 4	Slice 5	Mean total Case time	Total group time
Novice	Normals	28	4.0	4.1	3.9	5.0	3.2	19.9	84.4
	Acute	13	5.8	4.9	4.7	4.7	3.4	22.9	
	Sub-acute	7	6.5	5.5	5.4	5.2	3.8	26.4	
	Chronic	1	3.6	4.4	2.8	2.3	2.2	15.2	
Trainee	Normals	11	4.6	5.0	5.0	5.4	3.4	23.4	106.5
	Acute	8	5.3	3.5	3.7	4.8	3.0	20.3	
	Sub-acute	12	7.3	6.9	4.9	4.6	4.1	27.7	
	Chronic	7	7.9	7.9	7.1	6.3	6.0	35.1	
Expert	Normals	3	5.0	5.9	7.4	3.9	4.8	27.0	36.7
	Acute	1	1.6	1.5	2.8	2.4	1.5	9.7	
	Sub-acute	0	-	-	-	-	-	-	
	Chronic	0	-	-	-	-	-	-	

As seen in table 5.8 of this chapter, novices made more false positive decisions than trainee and expert readers; 4.9, 3.8 and 0.5 respectively per person and totalling 49, 38 and 4 per group respectively. Novices and trainees made FP decisions across all cases whereas experts made no FP decisions in subacute or chronic cases. In terms of number of false positives by group and case type; all participants made the most FP decisions in normal cases (novices: 28, trainees: 11 and experts: 3) and the least in chronic as seen above. The eye movement data highlights that despite the most FP decisions being made regarding normal cases, novices and trainee spent the most time appraising subacute cases, whereas experts spent the most time appraising normal cases.

5.2.4.5 True negative decisions

In line with true positive decisions, confidence in true negative decisions increases with experience (*df* 2, *p*<.007 F=5.1, Eta Squared.07).

Table 5.9 highlights a comparison of true negative and false positive ratings and the accompanied mean eye-movement data by participant group and fixations through the image slices (in seconds) in MR.

Participant Group	Participant Decision	Slice 1	Slice 2	Slice 3	Slice 4	Slice 5	Mean total Case time
Novice	TN	5.0	4.3	4.2	4.0	3.5	20.8
	FP	4.1	4.3	4.1	5.0	3.1	20.0
Trainee	TN	5.5	6.2	6.8	5.5	5.4	29.3
	FP	4.6	5.0	5.0	5.4	3.4	23.4
Expert	TN	2.9	3.6	4.2	3.8	3.2	17.5
	FP	5.0	5.9	7.4	3.9	4.8	27.0

Table 5.9 demonstrates that, unlike the results in chapter 4, novices spend approximately the same amount of time in true negative as false positive case decisions. Trainees spend an average of 5.9 seconds more over true negative decisions than false positive decisions when appraising MR images, which is a reverse of the trend demonstrated in the CT study. Trainees also fixated significantly more and for longer novices (*p*<.010 & *p*<.039) and experts (*p*<.028 & *p*<.001). In MR, experts spent 9.5 seconds longer over false positive decisions than true negative ones, which is 6.2 seconds longer than false positive decisions in CT image appraisal.

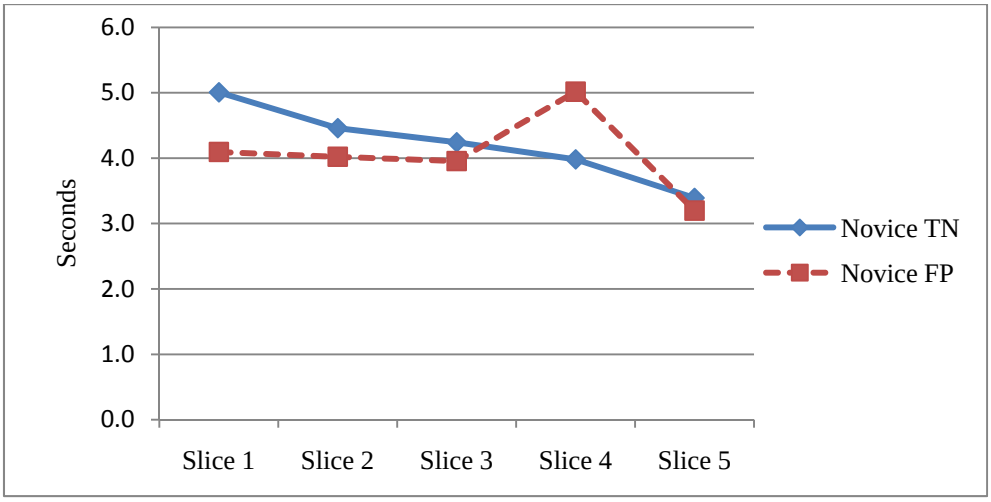


Figure 5.11 Mean fixation time per axial slice of novice readers who made TN & FN decisions in normal MR cases.

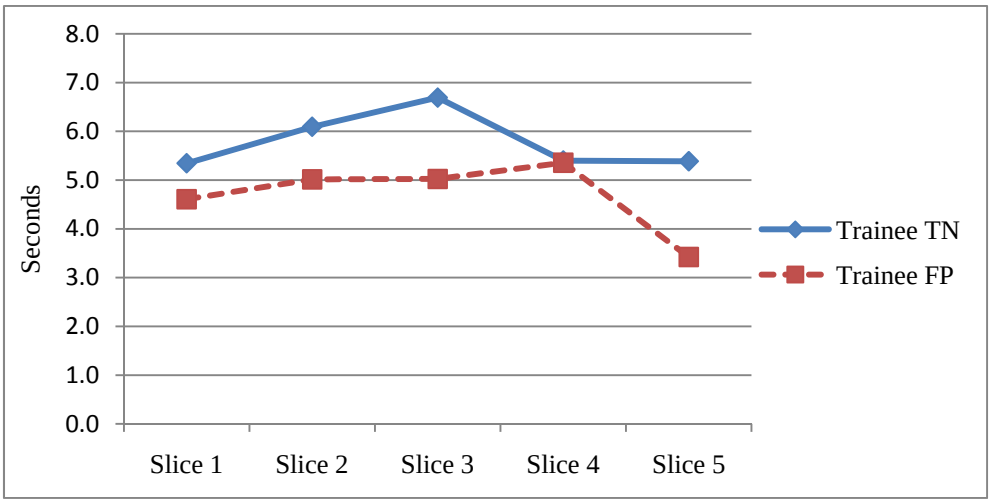


Figure 5.12 Mean fixation time per axial slice of trainee readers who made TN & FN decisions in normal MR cases.

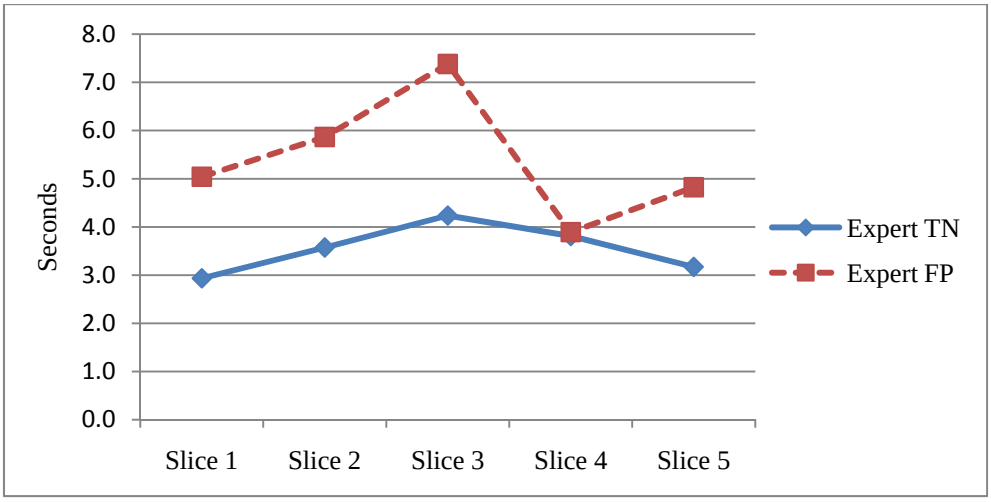


Figure 5.13 Mean fixation time per axial slice of expert readers who made TN & FN decisions in normal MR cases.

The above figures 5.11-5.13 consider the mean fixation time per axial slice (not just overall time spent in each case) of novice, trainee and expert readers when making true negative and false positive decisions regarding a case. The first graph highlights a similar trend, which was observed in chapter 4; true negative decisions made by novices make a steady decline of fixations through the slices, whilst false positive eye movement through the stack are less consistent. Trainee and expert true negative decisions are also accompanied by a clear trend that is both consistent within this study and also between modalities; more time is spent within the middle slice than first and last axial slices. On the contrary, false positive decisions do not appear to follow any clear or consistent pattern through the image stack unlike true negative decisions. In comparison with true negative trends, false positive decisions appear more erratic in nature, even between groups.

5.2.4.6 An in-depth examination of expert FP decisions

This section aims to examine the false positive decisions made explicitly by experts to gain an in-depth insight into why and when experts make mistakes in MR imaging.

Table 5.10 demonstrates the total duration of fixations in a retrospectively defined FP AOI.

PPTn	Case	Slice 1	Slice 2	Slice 3	Slice 4	Slice 5	Total Fix time in FP
A	NPM	0.0	3.0	0.0	0.0	0.0	3.0
B	NGB	1.6	0.0	0.0	0.0	0.1	1.7
C	NHG.1	0.0	3.0	0.5	0.3	0.0	3.8
	NHG.2	0.0	0.2	0.0	0.0	0.0	0.2

Table 5.10 demonstrates the time spent viewing the anatomical region that was considered abnormal. The result highlights that in cases NGB the FP decision appears to have taken place on the first image slice, whereas case NPM, NHG 1 and 2 have take place on the second slice. These results also indicate that FP regions are dwelt upon for very little time i.e. between 200 milliseconds and 3 seconds of visual search in MR. The following gaze tracker and hot spot images demonstrate where each FP location was plotted on the brain atlas task and has been retrospectively overlaid across each image slice to gain a qualitative appraisal of the participants' decision-making process. The false positive region was retrospectively defined after assessing the readers' reported location on the brain atlas task.

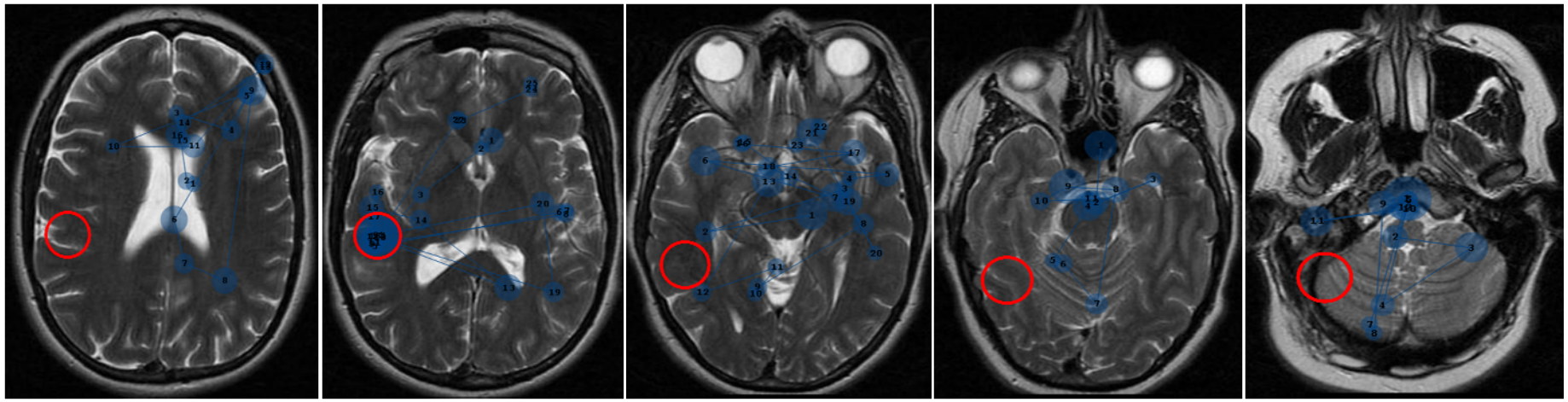


Figure 5.14 Case NPM: Reader A False positive decision

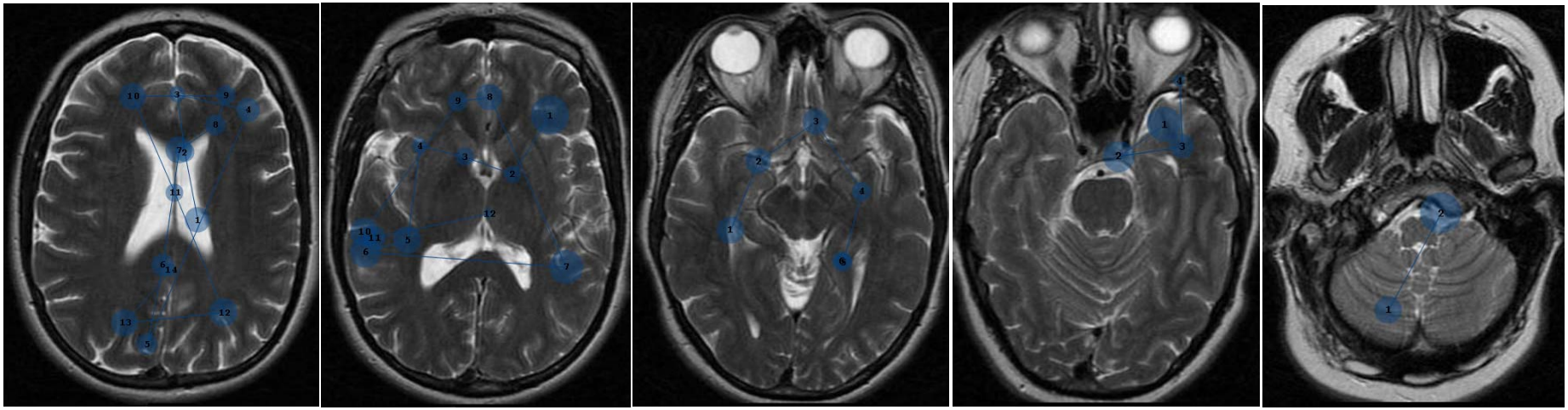


Figure 5.15 Case NPM: Reader D True Negative decision

In the above case, NPM, the visual search across slices 1-5 appear similar between both readers in terms of features examined, location and duration of gaze, yet reader A has highlighted a suspicious region on the left hand side of the image. It is clear that the decision was made in slice two as no other slices contain eye movements within this area. Reader D also appraised this area, including an additional region in slice 4 behind the right eye, but ruled both out as being abnormal.

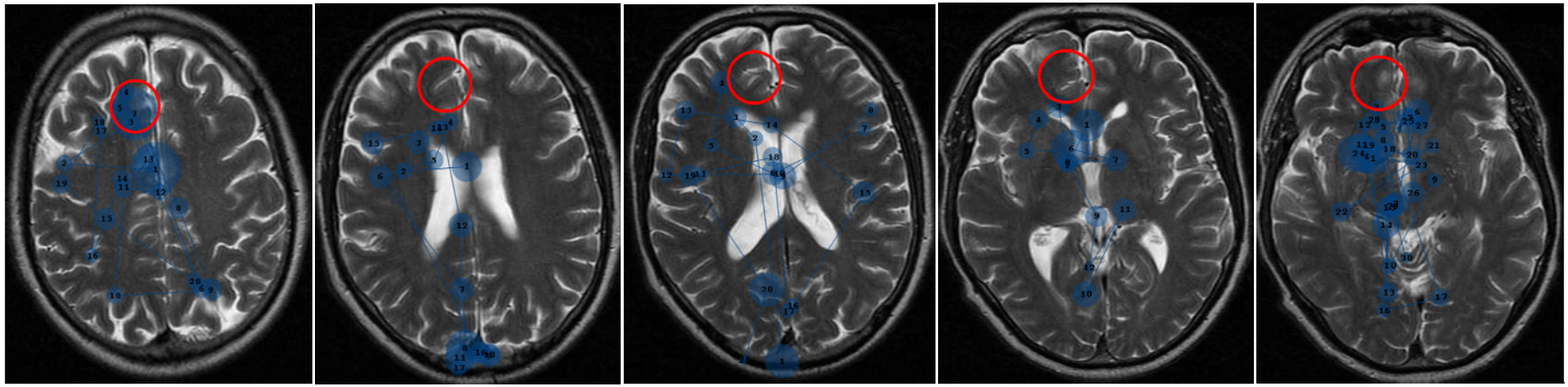


Figure 5.16 Case NGB: Reader B False positive decision

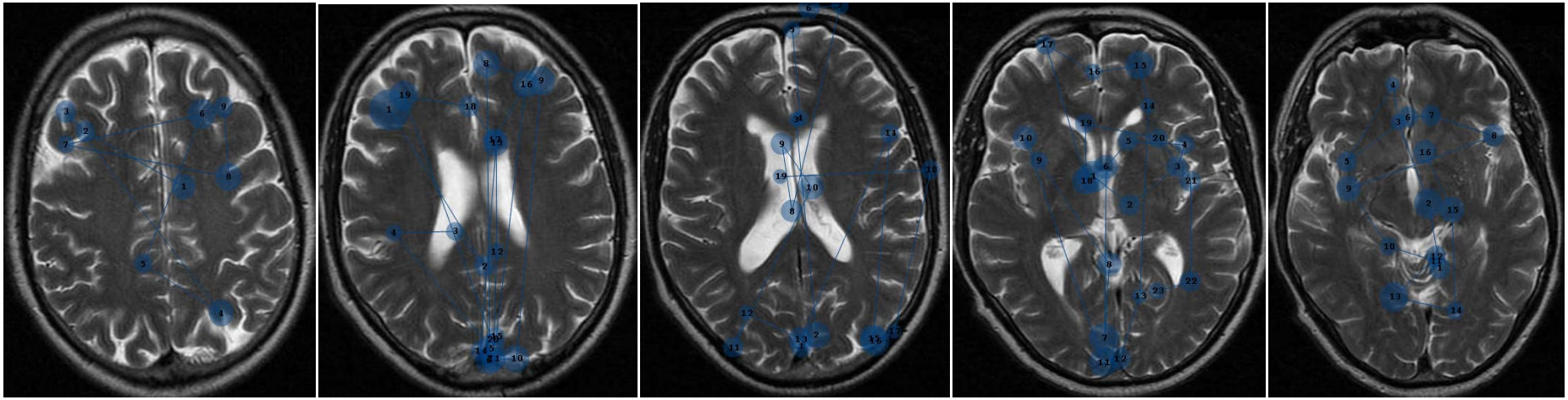


Figure 5.17 Case NGB: Reader D True Negative decision

In the above case NGB, it is clear that the area defined as abnormal was gazed upon by reader B on the first slice as the region was not returned to between slices 2-5. Both readers spent time appraising the back of the brain between slices 2 and 3 but reader D spends more time appraising this area and continues the appraisal onto slice 4. The hot spot data representing reader D suggests a much more thorough search throughout the medical image compared with the gaze tracker data in the former slices by B indicating true negative decisions are accompanied by a more thorough appraisal, as indicate by an increased number of fixations, of not only the entire slice but a full examination of suspicious regions over many slices.

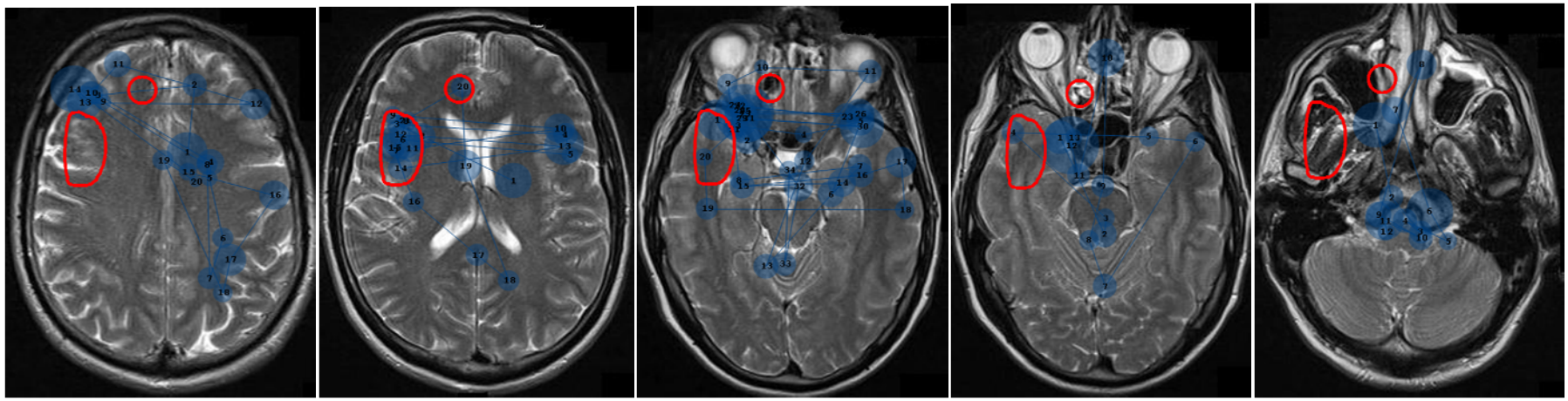


Figure 5.18 Case NHG: Reader C False Positive decision

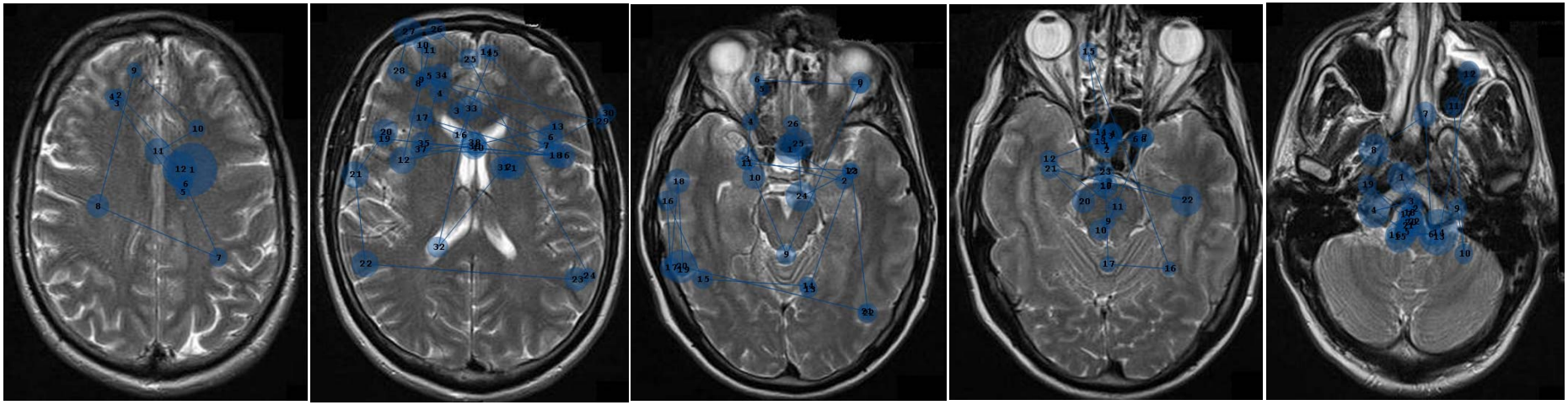


Figure 5.19 Case NHG: Reader A True negative decision

Case NHG demonstrates two false positive scores within one case by reader C. Once again it appears the decision was made in the second slice owing to the eye movements within these regions. However, it is surprising that this reader has made a false positive decision considering they have cross compared the hemispheres eight times; therefore, incorrect decisions appear to have been dwelt upon and revisited a number of times. On the contrary, the second false positive decision was only dwelt upon once for 200 millisecond duration. Comparing both readers, it is apparent that the same region has been investigated yet reader A appears to spend longer on single fixations than reader C. Hotspot images appear to represent more clearly how fixation durations and coverage differ between readers. Despite the table in section '5.9' indicating that longer is spent in FP cases than TN, this qualitative analysis appears to indicate than TN decisions are characterised by an increase in the number of fixations and regions investigated.

5.2.5 Secondary Abnormalities

Within this study there were three secondary locations, as depicted below. There were two secondaries located in acute cases ACG and ASC and one in chronic case CSS, with all infarcts being roughly the same size even thought they are in different regions of the brain.

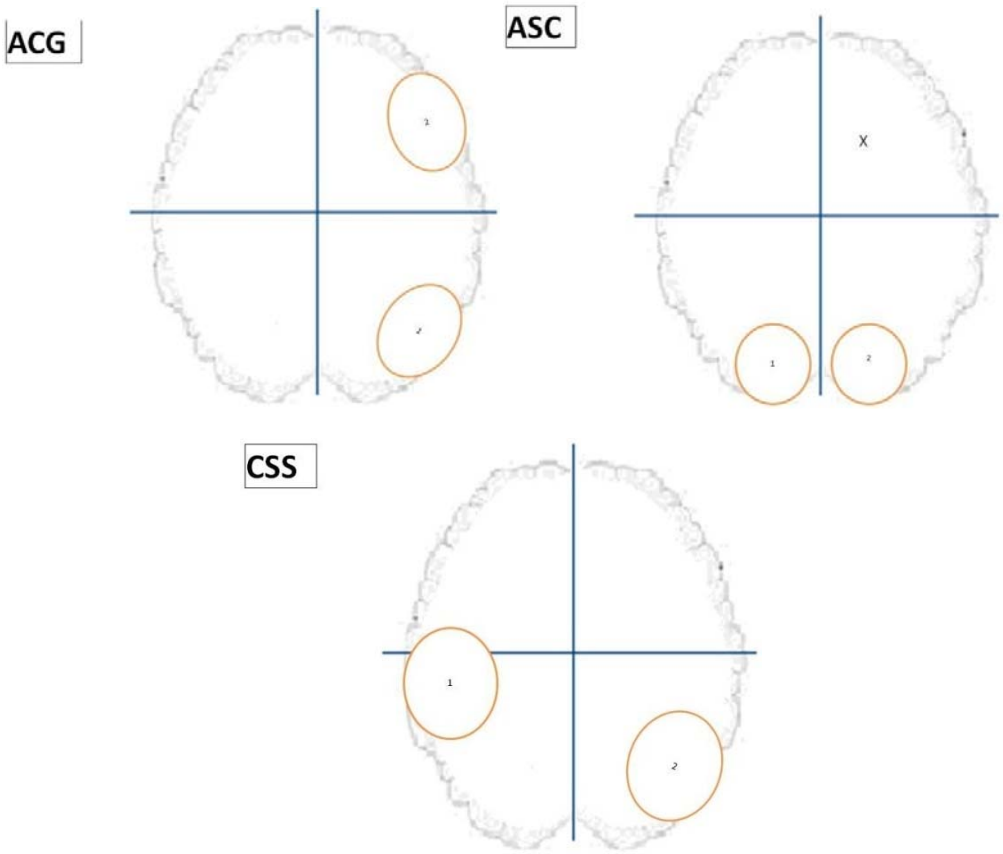


Figure 5.20 demonstrates the scoring criteria for cases; ACG, ASC and CSS, where secondary abnormalities are present.

In terms of accuracy for the above secondaries, once again experts were most accurate when identifying secondary abnormalities in MR, followed shortly by trainee and then novices. Although, there were no statistically significant differences between groups when comparing secondary abnormality detection in MRI.

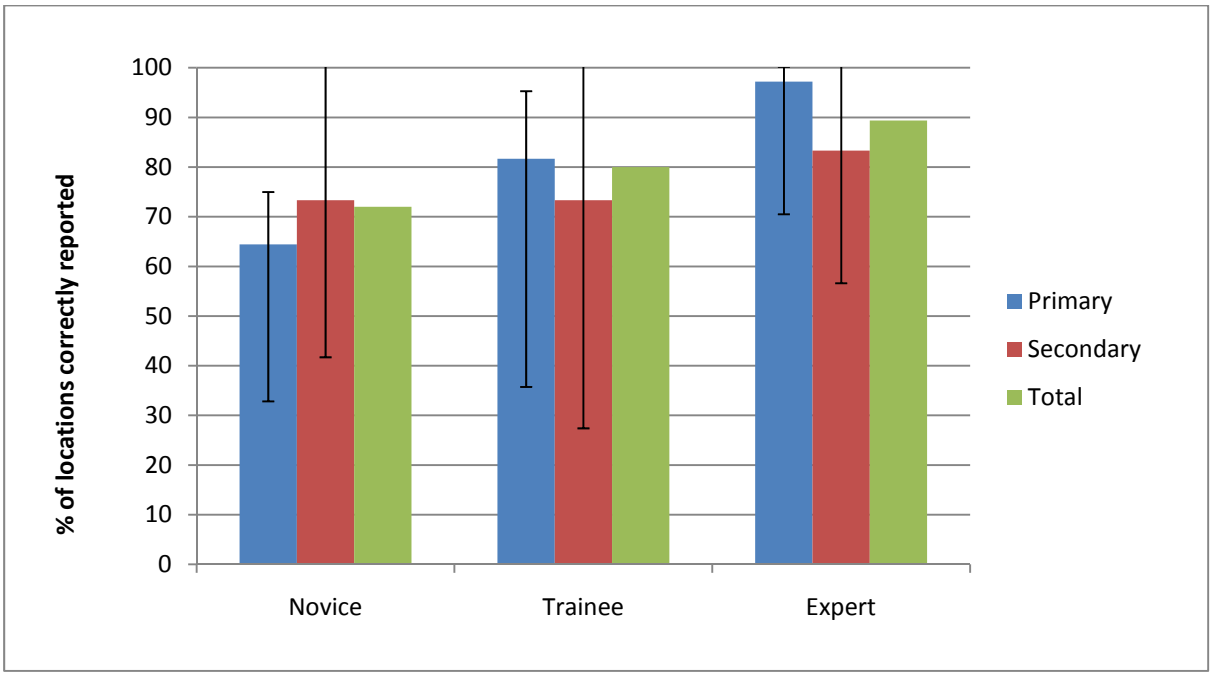


Figure 5.21 Percentage of primary and secondary locations correctly reported by participants in MR.

5.2.5.1 Secondary Abnormalities and Eye Movements

When considering time spent fixating upon the primary versus the secondary abnormality, novices and trainees spent the majority of their time in the primary AOI compared with the secondary AOI, which is in line with previous findings in CT. Experts spent the most time in the secondary AOI compared with novices and trainees in all cases. Trainees spent the most time in each case overall; 9.7, 15.8 and 7.3 seconds in each respective case.

5.2.5.1.1 True Positive Secondary Decisions

When considering the time to reach the secondary AOIs in the present study, experts were quickest in all three cases. They also spent the most time within the secondary AOI, as seen in table 5.11.

Table 5.11 to highlight eye-movement data by correct participants (in seconds) in the first appearing secondary AOI in MR.

MR True Positive eye-movement data by group (in seconds) in first appearing SECONDARY AOI.		Mean time to hit Secondary AOI	Mean time spent in Secondary AOI	Mean time out of AOI (in same slice)
Acute case: ACG	Novice	0.0	0.0	4.5
	Trainee	1.7	0.9	2.8
	Expert	0.9	0.9	1.4
Acute case: ASC	Novice	0.0	0.0	4.1
	Trainee	1.8	0.4	3.6
	Expert	0.8	0.6	3.7
Chronic case: CSS	Novice	0.0	0.0	4.2
	Trainee	1.4	0.4	2.6
	Expert	0.6	1.4	2.7

Although some novices did locate the presence of the infarct, they did not fixate within any of the AOIs in these cases, as seen in table 0.24. The time spent within the secondary by the trainees and experts was between 400 milliseconds and 1.4 seconds, which was also quite variable between the groups and within each case.

6.2.5.1.2 False Negative Secondary Decisions

Table 5.12 highlights that when the secondary abnormality in acute case ACG was ruled out it was because observers did not see it at all. In acute case ASC, novice and trainee readers were quick to reach it but only gazed within the AOI for between 800-900 milliseconds. In chronic case CSS, all novice and trainee readers detected the secondary and the region was only missed by one expert reader, who took 4.2 seconds to reach the AOI and only viewed it for 100 milliseconds.

Table 5.12 to highlight eye-movement data by incorrect participants (in seconds) in the first appearing secondary AOI in MR.

MR False negative eye-movement data by group (in seconds) in first appearing SECONDARY AOI.		Mean time to hit Secondary AOI	Mean time spent in Secondary AOI	Mean time out of AOI (in same slice)
Acute case: ACG	Novice	0.0	0.0	2.1
	Trainee	0.0	0.0	4.8
	Expert	0.0	0.0	1.2
Acute case: ASC	Novice	1.5	0.8	4.3
	Trainee	0.8	0.9	11.1
	Expert	3.3	0.1	4.5
Chronic case: CSS	Novice	-	-	-
	Trainee	-	-	-
	Expert	4.2	0.1	0.4

It appears that, when the secondary abnormality in acute case ACG was ruled out it was because observers did not see it at all. In acute case ASC, novice and trainee readers were quick to reach it but only gazed within the AOI for between 800-900 milliseconds. In chronic case CSS, all novice and trainee readers detected the secondary and the region was only missed by one expert reader, who took 4.2 seconds to reach the AOI and only viewed it for 100 milliseconds.

5.2.5.1.3 Satisfaction of search

In MRI, novices spotted the primary and/ or the secondary in equal numbers (43%), although those who spotted the primary first, were more likely to see the secondary (44%), than if the secondary was seen first (32%). 27% of trainees saw the primary first (57%), compared with the secondary (30%) and those who saw the primary first were also more likely to spot the secondary afterwards; 37% compared with 15% detection. In a similar trend, experts who spotted the primary first (50%) were more likely to spot the secondary (44%) than those who saw the secondary first (42%), with 36% seeing the primary after the secondary.

5.2.5.1.4 Small Vessel Changes

Within this study there were three cases that contained small vessel changes; ASC, SEB and SNE. Trainee sensitivity for small vessel changes was only 23%, with experts being 19% better at detecting these subtle changes. In terms of specificity, not much differentiated these readers with trainee specificity being 89% and experts, 88%.

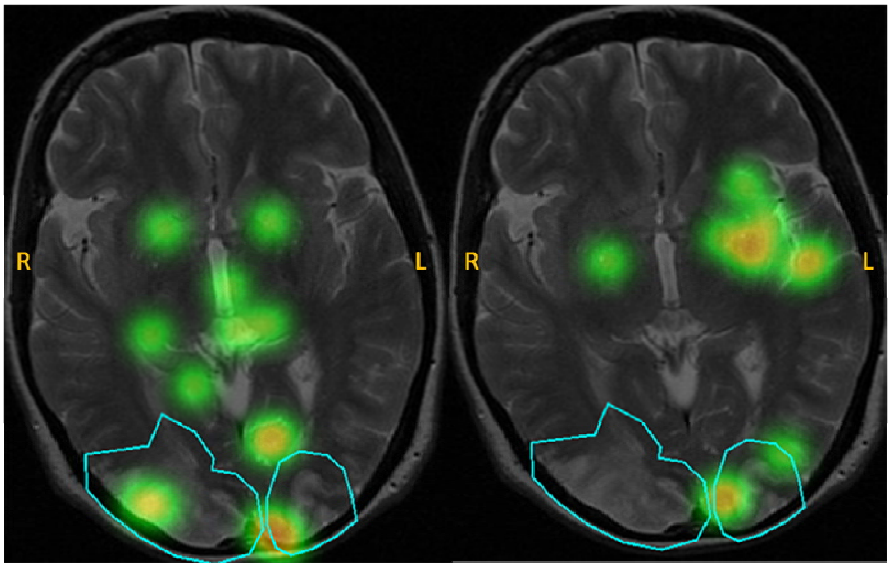


Figure 5.22 demonstrates a hotspot image of two trainee readers' eye movements when searching for small vessel changes in case ASC.

Figure 5.22 and the respective image slices are from acute case ASC and contain very discrete signs of small vessel changes. 37.5% of experts detected the SVD and 60% of trainees did also. The first figure is from a second year trainee (TML), who not only detected both abnormalities but clearly detected the SVD on both sides of the hemispheres towards the frontal part of the brain. The second image is from another second year trainee (TAK) who also detects both abnormalities but rules out the SVD, despite much appraisal of the features present.

Whilst experts were the optimum detectors of primary and secondary abnormalities, it appears not much separated trainees and experts when searching for secondary and focal abnormalities and discrete features were challenging to detect in this study. Problems of classification between feature types were still present in this study and will be further explored within the discussion section.

5.3 Discussion

5.3.1 Diagnostic Accuracy and Confidence by Group

This study aimed to explore the visual interpretations of multidimensional MR images of stroke when viewed by readers with varying levels of experience; from novice to experienced consultant radiologist. The aims, objectives and results from this study are most certainly in line with previous findings; experts were still the optimum performers across all dimensions of accuracy and decision-making when viewing MR images, i.e. small vessel change, primary, secondary and focal abnormality detection. The differences between groups were also highly statistically significant.

Whilst experts were the most confident at detecting and ruling out abnormalities, trainees were also very confident in their decisions, yet their decisions regarding features in MRI were not necessarily accurate. Trainees also made more false positive scores than experts, although both groups made fewer false positive scores in MRI than the previous CT study and contrary to CT findings, trainees did not generate more false positive results than novices, indicating enhanced clarity facilitates the ability to rule out suspicious areas than rule them in, which occurred in CT.

5.3.2 Diagnostic accuracy, Expertise and Eye-movements

The general trends of visual search between the groups did not differ between the present MR study and the previous CT study; a) novices still under or over-appraised the images and spent more time appraising normal anatomy, b) trainee visual search patterns did not differ extensively from novices, or they spent much time on the task, suggesting a degree of uncertainty or a desire to perform well and c) experts were still the most efficient, if not more accurate and economical, in the present study, which is in line with much previous work particularly Manning *et al.,* (2005).

Within this study, however, trainees spent longer appraising all case types in MR images, especially acute stroke types, than both experts and novice readers. In this study it appeared that novices often emulated expert readers in terms of their time between slices of the 'stack'. As previously discussed in the literature review, MRI provides higher clarity of the surrounding clinical features, which might explain the additional time spent by trainee readers both within and outside of the main abnormality. Higher clarity might result in more features to systematically appraise and cognitively process. It may also indicate that trainees spend longer to generate and refute decisions

than experts, who have knowledge and experience, and novices who do not, as previously discussed by Garlatti and Sharples (1998).

It was interesting to observe that once again, search patterns throughout the stack were dependent upon the case type being examined even though the cases were not the same; experts and trainees followed the same trend when appraising normal cases in CT as well as MRI. Expert trends did not appear to differ between CT and MRI cases, however, they spent much less time in the latter image slices in MRI than in CT, which suggests their case decision had been made in early slices and they moved quickly onto the next case. MRI, therefore, appears to quicken the decision-making process by experienced readers but not trainee decisions, which appear much quicker in CT – a modality which they may be more familiar with in the early stages of their careers.

When true positive decisions were made by observers, experts consistently appeared to spend 1.3 seconds or less appraising an AOI in MRI to ensure a correct recognition, which is quicker than in the CT study. It could also be suggested that expert readers had already recognised the presence of the abnormality in their peripheral vision prior to direct visual attenuation of the abnormality, as they were frequently the last participant group to reach the AOI. Trainee readers appeared to spend 1.6 seconds on average within an AOI and this figure may have important implications for the recognition of abnormal features by more inexperienced readers over a number of image slices.

Akin to experienced readers, novices also spent less time within the AOI to make a true positive decision. Less time spent by novices could be a result of limited background knowledge of what to expect and thus these readers are more reliant upon their instinct regarding normal and abnormal feature types and therefore, do not have the ability to generate specific hypotheses based on cognitive 'schemata' regarding what is normal and abnormal. This hypothesis appears to be backed up by the search behaviour in chronic cases: correct novices were quickest to hit the AOI (1.2s on average) and spent the same amount of time in the AOI as experts. However, novices did not spend the same time appraising surrounding tissue, which indicates a distinct satisfaction of search for the primary abnormality. Experts, on the other hand, searched the rest of the image for secondary problems to ensure additional abnormal features were not missed.

Leading on from true positive results, novices were also most eager to rule out the presence of primary abnormalities (false negative ratings); a finding which was not solely limited to the neglect

of secondary abnormality searches but was also consistent with CT. Therefore, where clarity is enhanced in MR, novice readers still fail to recognise and detect important abnormal features. False negative eye movements were characterised by either an even quicker time to hit or much longer than true positive decisions within and between all participant groups. However, much less time was spent appraising the AOI suggesting the abnormality simply was not recognised or the feature did not retain their attention once it had been gazed over – a finding which was consistent behaviour between all groups (although false negatives were less likely for expert readers). Referring back to the eye movement data, it appears that 1 second or less within an AOI, does not allow enough time for recognition of an abnormal area or a correct decision to be made regarding an abnormal feature type across multiple slices.

False positive decisions were mostly made within normal cases and experts spent the most amount of time appraising the features within these cases. Novices and trainees spent the most amount of time appraising subacute cases, indicating a degree of uncertainty regarding features in these cases, which could be explained by the low mass effect of subacute strokes i.e. the boundaries of the infarct are less clear than acute or chronic infarcts. It also highlights a readiness to classify normal features within control cases as abnormal by novice and trainees, without giving them enough detailed inspection or cognitive consideration, which might have resulted in a more thorough reasoning into whether they truly were abnormal. Therefore, more clarity can often confuse what appears to be normal and abnormal by the inexperienced observer.

In terms of patterns within expert false positive decisions, results indicate that not only were false positive decision made by a select few readers, but similar mistakes were made between and within these individuals i.e. normal anatomical features were considered suspicious such as brain tissue folds known as sulci, or within the brainstem region where anatomy frequently appears a darker shade of grey than in other axial slices towards the top of the brain, although there were much fewer false positive decisions made in the MRI study than CT study. This insight points towards the more inexperienced consultants in neuroradiology i.e. consultants who are more familiar with another area of radiology (e.g. breast or musculoskeletal anatomy) appear to make more mistakes than radiologists who view neurological cases on a more regular basis, which is also in line with previous findings. As few neuroradiologists are present in most hospitals, there were other radiologists recruited from other disciplines to the present series of studies. Whilst recruiting other radiologists from other specialities might not seem ideal, all radiologists go through assessed training

in neuroradiology and therefore, all radiologists in this study had the same baseline, 'expert' experience.

Upon consideration of eye movements associated with the false positive locations, false positive decisions appear bipolar in nature i.e. either false positive regions have been dwelt upon (e.g. between 1.5 and 3 seconds) and revisited a number of times, which is line with the 'over 1000ms rule' applied by Manning (2005) but not adhered to in the CT study, or false positive locations are visited once for a very small fixation duration (between approximately 100 and 200 milliseconds) and still subsequently reported. In addition the pattern of fixation duration throughout the image stack is less consistent than where true negative decisions have been made. Therefore, false positive decisions in MRI appear more clear-cut than in the CT study. In the present study, decisions well over 1.5 seconds are considered for much too long to be correct or are under 200 milliseconds, where a decision has not been thoroughly considered. The optimum time threshold for making correct decisions, therefore, appears to be more than 200 milliseconds but less than 1.5 seconds – a time bracket of up to 1.3 seconds in MRI.

The visual search patterns of true negative decisions once again suggest a more thorough appraisal of not only the entire slice with more short duration fixations, but a full examination of suspicious regions over many slices. As some hotspot images highlight, true positive cases often appear more saturated with fixations than (false positive) gaze tracker images, which owing to the clinical knowledge base, indicates a thorough image appraisal rather than the cognitive overload associated with novice readers in the early part of this chapter. In addition the mean time spent within each image of the stack, highlights a consistency of visual search behaviour within each reader group throughout the stack, even though the patterns differ between each reader group.

When considering the visual patterns surrounding the secondary AOIs, correct experts were quickest in all three cases and spent the most time appraising the infarct. The time spent within the secondary by the trainees and experts was between 400 milliseconds and 1.4 seconds; a time band which appears slightly longer than the primary abnormalities (i.e. 0.2-1.3seconds). When the secondary abnormalities were ruled out it was frequently because the observers did not see it at all. In acute case ASC, novice and trainee readers were quick to reach it but only gazed within the AOI for 800-900 milliseconds, which is enough time to fully appraise the problem are but it appears they did not recognise it as a secondary infarct. Although it would be likely that MRI would enhance the

detection of secondaries, unfortunately it did not and the detection rate was highly similar to that observed in the CT study. Performance for small vessel change detection in MRI highlighted that these abnormalities were far more challenging to detect in MR than CT, which may also be a result of the enhanced acuity – increased clarity of the anatomy might mask the very discrete abnormalities, which become hidden within the image.

The present study appeared to worsen novice performance by reducing the number of true positive scores and increasing the number of false positive and false negative ratings. The enhanced clarity of MRI did increase the number of true negative ratings, however, it also increased the number of false positive ratings; indicating that when novices were not making incorrect decisions regarding features, they were better at ruling out the presence of abnormalities in MRI than CT. Trainees accuracy was enhanced in MRI as the number of true positive and true negative ratings were increased compared with CT. Whilst the number of false positive locations were also reduced with enhanced clarity, the number of false negative ratings was increased in the present study, indicating they had more decision errors regarding the features, which could be seen more clearly.

In terms of case difficulty, normal cases were perceived to be the most difficult cases to define among all observers, as previously seen in the CT study. However, in MRI the second most difficult case category was the subacute stroke type. The present study highlights that as stroke size becomes larger and the boundaries naturally less clear, enhanced clarity is not necessarily preferable or conducive to enhanced performance. As acute stroke types were found to be less challenging in MRI than CT, it does highlight that for smaller, 'younger' or more discrete stroke types, the enhanced clarity is much preferred. Therefore, if a patient is imaged quickly following infarction and admission to the hospital, MRI might be more preferable than CT. Following this early period of up to one week, CT performance might be equally as preferable, if not more preferable for subacute strokes than MRI, which was also in line with previous research conducted by Mullins *et al.*, 2002 [b].

Overall, it is unequivocal that experience impacts upon visual search and decision-making. It appears the less experienced and confident the reader, the increased likelihood of false positives, even within the expert radiologist group. As radiologists become more experienced compared with their peers, or more in tune with neuroradiology, the more the visual search pattern appears to emulate the other more experienced radiologists. Experienced readers seem to appraise the same

clinical feature but have the confidence to allow abnormal areas to 'pop out' of the image and draw their finely tuned attention.

Problems regarding misclassification also existed within this study, but mostly within the trainee observer group. Despite the enhanced clarity, some trainee readers consistently reported the correct locations but applied an incorrect classification to the infarction. In one case, a first year trainee reader scored a chronic primary as small vessel changes, which is quite disconcerting considering their differential clinical implications. Whilst some trainees could be forgiven for scoring a discrete acute infarct as a focal abnormality, confusing a chronic infarct with a focal abnormality or small vessel changes requires addressing. It was most likely that the selection of trainee readers who misclassified the abnormalities had not received their head and neck imaging specialist training module, which was predetermined at the start with 50% having received the training and 50% not and the head and neck training module would most likely reduce the number of these misclassification errors, however, targeted training should also extensively explore and address misclassification errors in future packages over and above the suggestions already laid out in chapter 4 (i.e. increased knowledge of neuroanatomy coupled with knowledge of cerebrovascular pathways, including areas most susceptible to arterial occlusion), which would be most likely to augment overall performance. It also appears that trainees require more exposure to MR images in stroke detection as they appear quicker and more confident in CT.

The limitations of the present study were the same as those discussed in the CT study, however, in this study there were fewer reporting errors i.e. circling the wrong hemisphere on the reporting sheet or highlighting an area which bore no relevance to clinical features or their eye movement data results. The recommendations which arise out of the present study appear to infer that, in addition to the training requirements put forward in this MR study and in the CT study, different modalities might be preferable for different stroke types, which is dependent upon how much time had elapsed since the infarct occurred. Since exact time to infarct has not been explored within the present study, only considered with reference to acute, subacute and chronic stroke types, these factors and modality type should be considered in more detail in future work.

5.4 Conclusions/ Summary

Results from this study, highlight differences in visual search patterns amongst novice, trainee and expert observers in MR images; the most marked differences occurring between novice readers and experts. Differences in search patterns, image coverage, foveal fixations, dwell and overall case time were observed in the detection and interpretation of these stroke MR images. To-date few studies have explored observer performance in MRI and the present study also examined multi-slice image appraisal.

The trends observed within this MR study were also observed in the previous CT study with reference to true positive, true negative, false positive and false negative decisions, yet the enhanced detail of the MRI study appeared to alter the performance of all observers, particularly trainees. The enhanced detail in MRI appears to quicken expert reader and novice readers, who either have a vast knowledge of radiological images or none at all, but appears to generate uncertainty in the trainees, who took much more time over this study than the CT study.

5.5 Study Reflections

This study produced results very similar to those of chapter 4; however, there were subtle differences in visual search and performance throughout the two studies, which may not have been overtly apparent. The results of chapters 4 and 5 will now form the basis of chapter 6, to firmly establish whether MRI search strategies are, in fact, different from CT owing to the increase in structural noise and to uncover how and why the modalities affect performance.

Study 3: Does Modality Preference Enhance Observer Performance?

(The assessment of stroke multidimensional CT and MR imaging using eye movement analysis)

As previously discussed, both Computed Tomography (CT) and Magnetic Resonance (MR) imaging modalities are frequently used in the prevention, identification, diagnosis and treatment of people who are predisposed to, or have suffered an acute or chronic neurological deficit. Stroke presentation has received attention from researchers to compare modalities in terms of sensitivity and specificity to meet patient requirements (Gonzalez, 1999; Mullins, 2002[b]) and have demonstrated a similar sensitivity of less than 50% for both conventional modalities, with neither modality proving superior to the other from an observer performance perspective (Mohr, 1995; Lansberg, 2000; Wintermark, 2007). Whilst the previous chapters discussed findings within each modality, to the authors' knowledge, an in-depth examination of visual search and observer performance analysis between CT and MR multidimensional images in neuroradiology has not been reported. In this present chapter, although two differing modalities are adopted, the cases therein have been matched for difficulty level, perceived noise, the number and type of abnormalities presented and where possible, the location of the cerebral infarct, to ensure that accurate comparisons could be reliably made between experiments and chapters.

6.1 Study Aims and Objectives

The present chapter compares the performance and visual search findings between reader groups and modalities. Explicitly, the objectives of this chapter are as follows;

- Explore observer performance between participant groups when viewing conventional CT and MR, multi-slice imaging of stroke.
- Explore visual search behaviour between participant groups when viewing conventional CT and MR images.
- Assess visual search behaviour within and between 'stack' images of CT and MR stroke cases.

- Re-examine pilot findings, with a larger sample size (n=28) and an increase in patient cases from 8 to 48.

6.2 Method

6.2.1 Participants

As per chapters 4 and 5, 28 participants were recruited; ten novice readers, ten Specialist Radiology Registrars and eight Consultant Radiologists with the same participants completing both CT and MR experimental studies.

6.2.2 Design

Two-hundred and forty clinical images were selected and made anonymous from a bank of predetermined clinical cases at Norfolk and Norwich University Hospital Picture Archiving and Communications System (PACS) with the assistance of a resident Specialist Registrar in Radiology. Of these images, 120 were acquired using a CT scanner and 120 acquired using MR. All axial image slices were extracted from forty-eight predetermined clinical cases. Both CT and MR cases were represented by a spread of six normal controls, eight acute cases, six subacute cases and four chronic cases, which were matched for each modality. All cases were ischaemic, not haemorrhagic, stroke. The cases were selected by the Specialist Registrar in Radiology on the basis of radiology report information, which issued the final diagnosis and stroke classification. The independent variables of modality (CT, MRI) and case severity (acute, subacute, chronic or normal aging control) were assessed in a within and between participant design.

6.2.3 Procedure

The procedure for the present study did not differ from chapters 4 and 5. For further information regarding image and case type selection, group allocation and experimental procedure, please refer to chapter 2.

6.3 Results

Study data was analysed to investigate; i) qualitative image analysis ii) accuracy and confidence ratings of performance, iii) quantitative eye movement analysis, iv) stroke, expertise accuracy and

visual search and v) diagnostic accuracy and clinical information availability: within and between participant results and vi) all results compared between CT and MR imagery.

6.3.1 Qualitative Image Analysis: Visual Search Behaviour between Groups and Modality

The following two case study examples illustrate the qualitative differences in visual search between the three reader groups by examining a single image to highlight differences in visual search patterns between readers and modalities in detection of acute stroke types. N.B. When reading and reporting CT and MR images, it is important to remember that the left side of the image represents the right side of the individual, and vice versa due to the acquisition process.

CT and MRI acute stroke comparisons: The following gaze-tracker images highlight the differences between readers' (a novice, a trainee and an expert) visual inspection strategies when appraising images of Acute CT and MR stroke images.

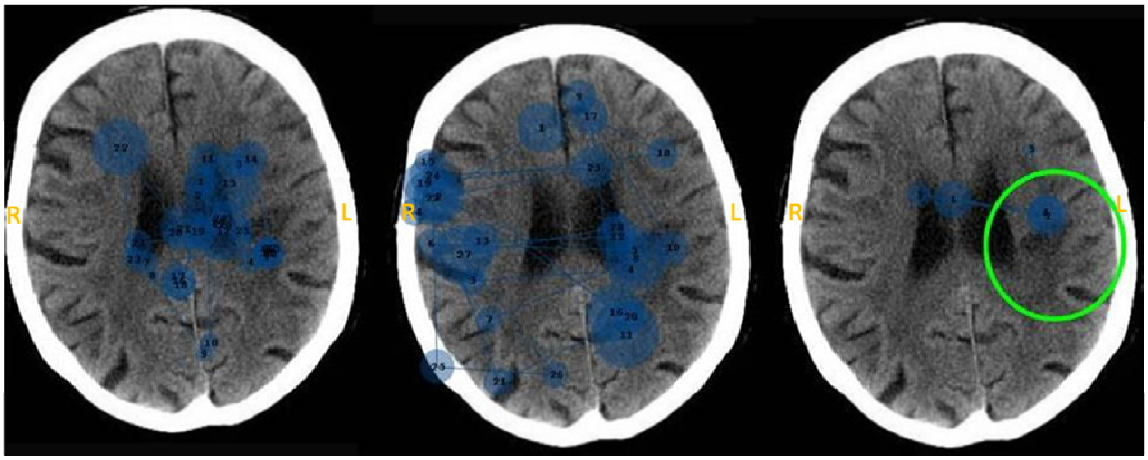


Figure 6.1 demonstrates the eye movements of a novice (a), a trainee (b) and an expert reader (c) in an acute CT stroke case.

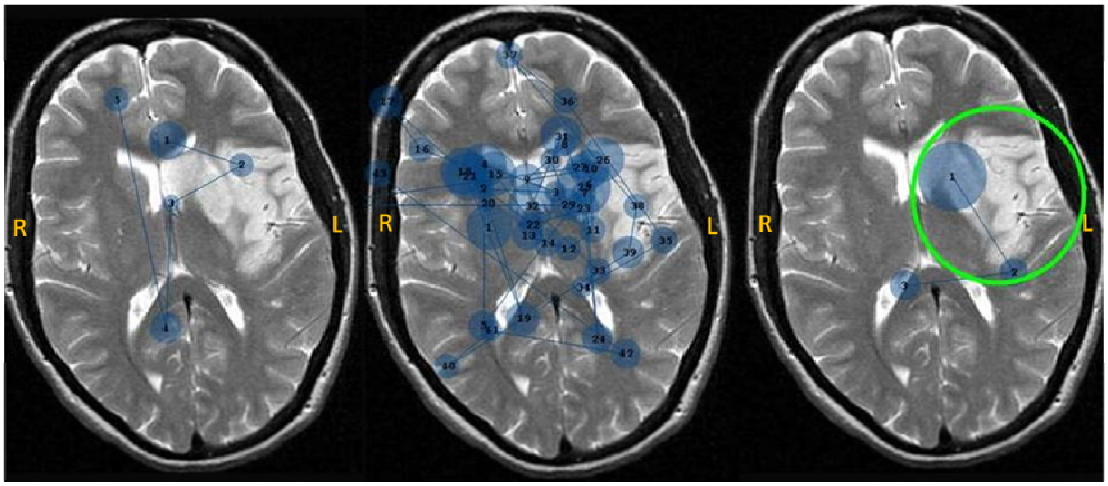


Figure 6.2 demonstrates the eye movements of a novice (a), a trainee (b) and an expert reader (c) in an acute MRI stroke case.

Figures 6.1 and 6.2 highlight the fixation patterns of a novice (a), a trainee (b) and an expert (c) reader when examining CT case AAB and MRI case ACW. Figure 6.1 demonstrates how a novice reader appraises normal, ventricular anatomy to the neglect of cortical tissue. In the second image, the trainee's gaze entered the AOI (located in the left patient hemisphere) with the 4th fixation and spends a total of 28 fixations on the image. The expert fixates upon the AOI with the second fixation from a total of 5, with anecdotally, an efficient visual search pattern accompanied by a true positive rating.

Figure 6.2 highlights a quick time to 'hit' over 5 fixations; this readers' eye movements appear similar to those of the expert in figure 6, yet novices are known to be unaware of the potentiality of secondary problems and small vessel changes, which may influence patient status. The trainee spends further time within the slice, thoroughly appraising peripheral sulci, cortical tissue and ventricles. The trainee's gaze is quickly drawn to the abnormality, as indicated by the third fixation, and a confident rating, but owing to their in-depth training should also be aware of the need to thoroughly examine the surrounding tissue, unlike the novice. The expert makes a quick, confident and accurate decision, demonstrated by the eye movements and accompanying true positive rating.

Collectively, these figures demonstrate the differing visual search patterns between a novice, a trainee and an expert observer between different patient cases and modality. For instance, although the infarct in the MR patient is slightly larger than the CT patient, the CT novice reader has more fixations around the ventricular area than the MR novice reader, who fixates within the AOI immediately denoting quick identification. The trainee readers demonstrate similar search patterns between CT and MR cases; detailed inspection including much cross comparison is evident on both,

however there are an increased number of fixations on the MR (43 fixations) image compared with the CT image (28 fixations). This finding may be indicative of uncertainty when searching for the possibility of secondary abnormalities such as small vessel disease or lacunar infarcts, as the primary abnormality appears much less challenging for nearly all participants. Expert eye movements appear to demonstrate consistency between readers, in terms of a quick time to fixate the infarct and few fixations between modalities overall. To view all gaze tracker images for this case, please refer to pages 16, 17 and 18 of the appendix.

6.3.2 CT and MR Receiver Operating Characteristics

The following Conventional Binormal ROC Curves represent primary abnormality detection ratings by participant group (i.e. novice, trainee and expert readers) and images per modality (i.e. CT and MR) of confidence rating scores per case. Within the ROC space the parameters 'a', 'b' and 'Az' represent the vertical intercept, the slope of the fitted curve and the overall accuracy value as calculated by the conventional Binormal curve process, respectively. ROC curve 1 demonstrates that novice sensitivity is superior in MR than CT, yet specificity is marginally improved in CT. When modality performance was explored using RocKit, novices performed significantly better in the MR condition (p<.050) when examining these cases in this study.

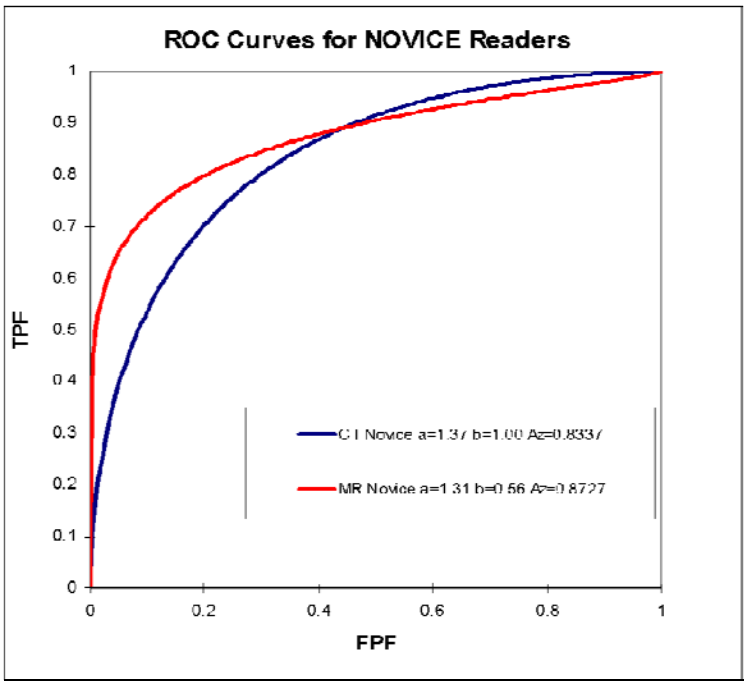


Figure 6.3 ROC Curve 1: Represents primary abnormality detection scores for novice readers between CT and MR modalities.

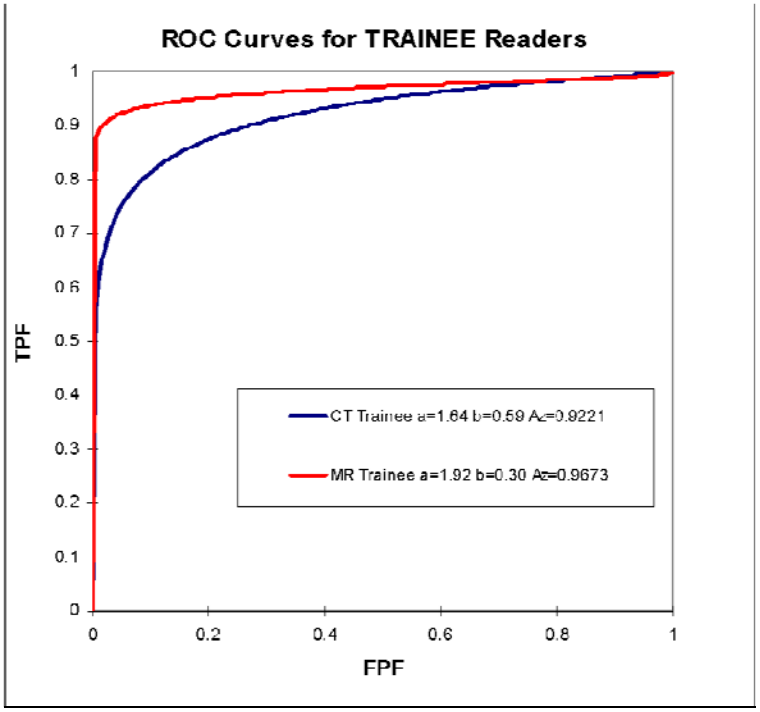


Figure 6.4 ROC Curve 2: Represents primary abnormality detection scores for trainee readers between CT and MR modalities.

ROC curve 2 demonstrates that trainees are more accurate in the MR than CT condition in terms of both sensitivity and specificity when examining these cases in this study. When this difference was explored using RocKit, the modality difference was highly statistically significant for trainee readers ($p < .008$).

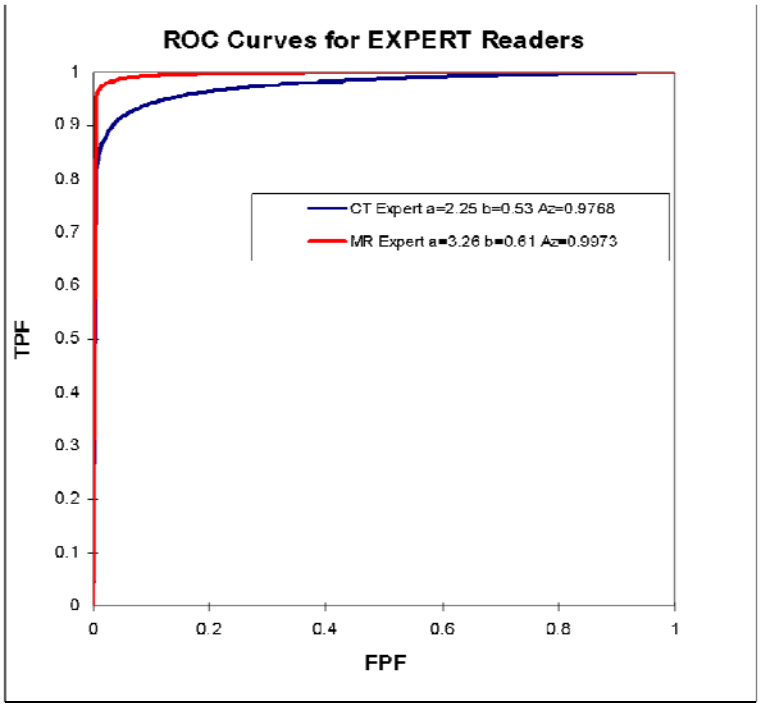


Figure 6.5 ROC Curve 3: Represents primary abnormality detection scores for expert readers between CT and MR modalities.

ROC curve 3 demonstrates that experts are highly accurate in both MR and CT condition in terms of both sensitivity and specificity when examining these cases in this study. When this difference was explored using RocKit, the difference was approaching statistical significance for expert readers (p<.058). To examine differences between the groups, Tukey posthoc comparisons were applied and demonstrated that the differences were most evident between novice and expert groups for CT (p<.000) and MR (p<.000). Comparisons between novice and trainee reader were significant in MR (p<.000) and CT (p<.010), and comparisons between trainee and expert readers were also significant in MR (p<.007) and CT (p<.036) when examining these cases in this study.

6.3.3 Accuracy and Confidence Rating Data: Quantitative Analysis of Observer Performance

When considering diagnostic accuracy in CT, novice participants correctly reported the absence of or abnormal locations in 70.83%, trainees 81.25% and experts 90.1% of all patient cases. In MRI, novice participants correctly reported the absence of or abnormal locations in 63.8%, trainees 85.4% and experts 96.4% of patient cases. Novice accuracy was significantly better in CT (CT=.72, MRI=.63, t=2.22, *df* =239, p<.028) when these results were compared using a paired samples t-test. Conversely, expert accuracy was significantly better enhanced in the MR condition (CT=.91, MRI=.96,

t=-2.13, *df* =191, *p*<.033). Whilst trainees appeared to perform better in MR, this difference was not statistically significant.

When reporting participant confidence in their decision-making in CT, descriptive statistics (table 6.1) show overall percentage of cases correct indicates that acute cases were perceived to be most difficult by all participants (79.1%), followed by controls (79.2%), chronic (93.1%) and then subacute stroke types (96.4%). In MRI, overall percentage of cases correct indicates that control cases were perceived to be most difficult by all participants (78.5%), followed by subacute (80.3%), acute (81.0%) and then chronic stroke types (91.0%). Therefore, MR control cases may be considered to be the most difficult cases when all participant results were combined shortly followed by acute and control CT cases.

Table 6.1 represents confidence scores (in percentages) by participant group and stroke type.

Confidence scores (%) by stroke type	**Modality**	**Novice**	**Trainee**	**Expert**
Normal Control Score	CT	48.8	52.5	63.0
	MR	51.7	60.4	66.1
Acute	CT	70.6	83.1	92.2
	MR	81.3	95.9	99.6
Subacute	CT	87.5	98.3	98.4
	MR	75.0	91.7	96.4
Chronic	CT	83.3	95.0	98.4
	MR	90.6	93.1	99.2
Average Confidence Score	**CT**	**73.3%**	**81.0%**	**81.5%**
	MR	**73.8%**	**80.1%**	**82.3%**

Novice descriptive statistics show a higher confidence when rating chronic cases in MR. Trainees show higher confidence levels when rating subacute cases in CT and experts show a higher confidence when rating acute cases in MR. All participants show lower confidence levels when ruling out abnormalities in CT cases. There were significant differences between participant groups for confidence in both MR (*df* 2, *p*<.001, F=6.85, Eta Squared=.02) and CT (*df* 2, *p*<.001, F=7.123, Eta Squared=.02) in ANOVA statistical tests. Tukey posthoc comparisons demonstrated differences in confidence between novice and trainee groups (MR *p*<.013, CT *p*<.005), and expert and novice groups in both conditions (MR *p*<.013, CT *p*<.005). However, there were no significant differences between trainee and expert confidence scores in CT (*p*<.926) and MR (*p*<1.00) scores.

6.3.4 Diagnostic Decision-Making (i.e. TP/FP/TN/FN) between Modalities

As demonstrated above, experts are the optimum performers in both CT and MRI. The following table 6.2 highlights the spread of true and false positive ratings and true and false negative per group and modality to gain a more in-depth insight into the decision-making processes of participants.

Table 6.2 represents mean spread of decisions between observer group and modality type.

Average ratings	Modality	TP	TN	FP	FN
Novice **(n=10)**	CT	13.6	3.2	4.1	5.0
	MR	11.8	3.5	4.4	6.2
SpR **(n=10)**	CT	14.8	4.0	3.7	1.4
	MR	15.5	5.0	2.6	2.5
Radiol **(n=8)**	CT	17.3	4.3	1.8	0.8
	MR	17.5	5.6	0.6	0.5

In both CT and MRI, radiologists were the better performers as previously discussed in chapters 4 & 5. The number of false positive ratings by experts was reduced from 14 in CT to 4 in the MR study. False negative ratings were also halved in the MR reading task; from 6 in CT to 3 in MR. Novices perform better in CT than MR when the overall spread of decisions are considered and clarity is reduced. Trainee readers did not differ substantially between modalities.

Overall sensitivity and specificity rates demonstrate that novice sensitivity and specificity did not differ substantially between modalities (CT; Sensitivity: 82%, specificity: 42% & MR; Sensitivity: 74%, specificity: 41%), neither did trainee sensitivity and specificity (Sensitivity; CT: 91% & MRI: 92%. Specificity; CT: 50% & MR: 55%). Whilst expert sensitivity did not differ substantially between modalities (CT: 96% & MR: 98%), specificity was much enhanced in MRI (92%) compared with CT (71%).

6.3.5 Eye-movements and Experience: Quantitative Analysis.

This quantitative results section aims to identify statistically significant links between visual search behaviours (e.g. time to reach an AOI, time spent within an out of AOIs) and reported accuracy within and between participant groups and modality to provide insights into the way inexperienced through to experienced readers appraise visual stimuli.

6.3.5.1 Task Viewing Time per Modality

When assessing the differences in reader time viewing the images, in terms of number of fixations and the duration of these fixations, cumulatively, novices fixated much more (CT=83.91, MRI=63.09, t=5.60, *df*=239, *p*<.000) and for longer in the CT rather than the MR image condition (CT=28755.78, MRI=18119.31, t=6.98, *df*=239, *p*<.000), whereas the opposite was true for trainees (CT=67.96, MRI=85.62, t=-5.19, *df*=239, *p*<.000). Experts did not fixate significantly more or less in either image condition, however, they did take longer to view the images in CT compared with MR (CT=18364.21, MRI=1554.28, t=3.36, *df*=191, *p*<.001).

6.3.5.2 Visual Search Behaviour throughout the Image 'Stack'

The following graphs demonstrate mean time spent in each axial slice throughout the five image 'stack' by case type (control, acute, subacute and chronic), level of expertise (novice, trainee and expert) and modality (CT, MR) irrespective of decision type.

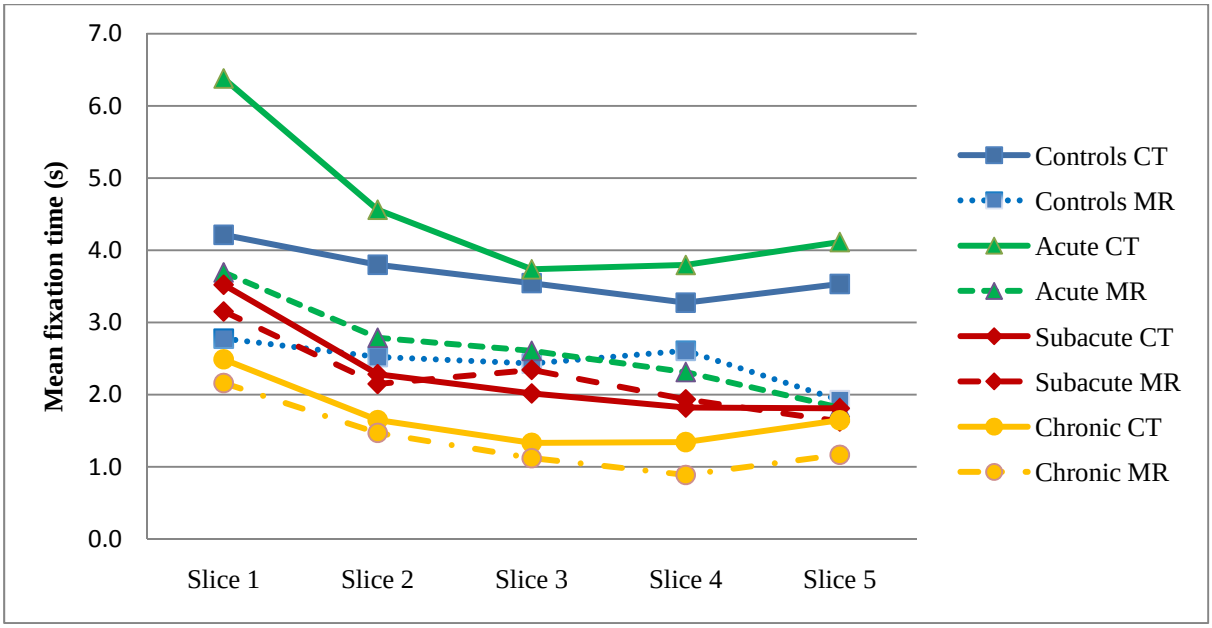


Figure 6.6 Demonstrates the mean fixation time per axial slice for novice readers across all CT and MR cases, irrespective of decision type.

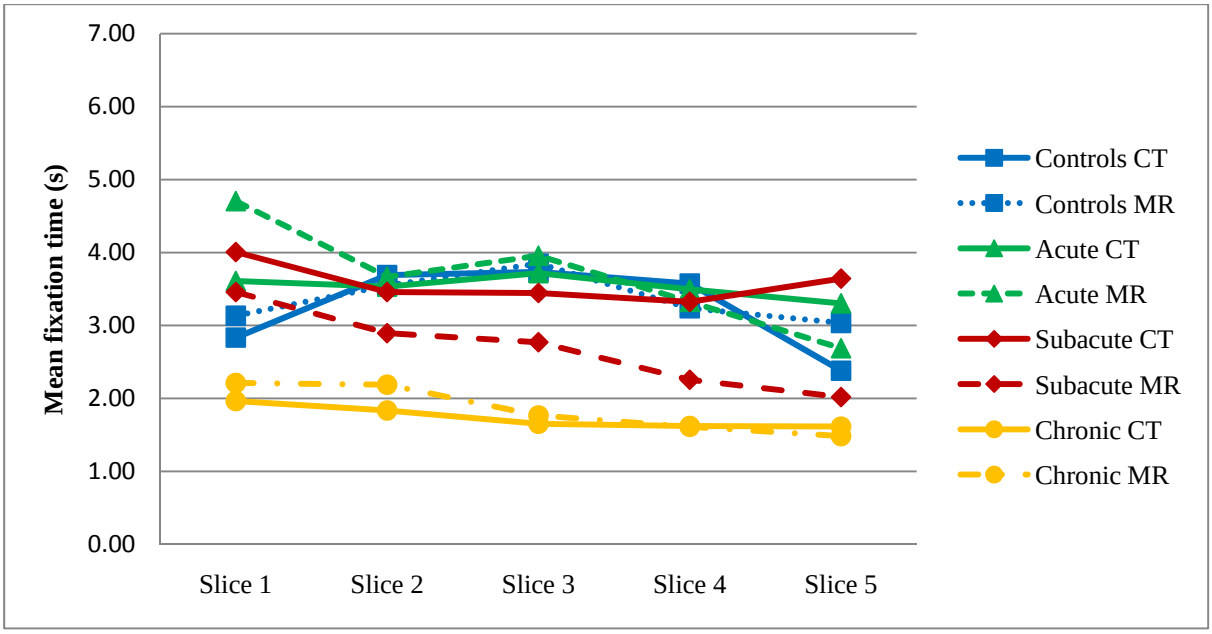


Figure 6.7 Demonstrates the mean fixation time per axial slice for trainee readers across all CT and MR cases, irrespective of decision type.

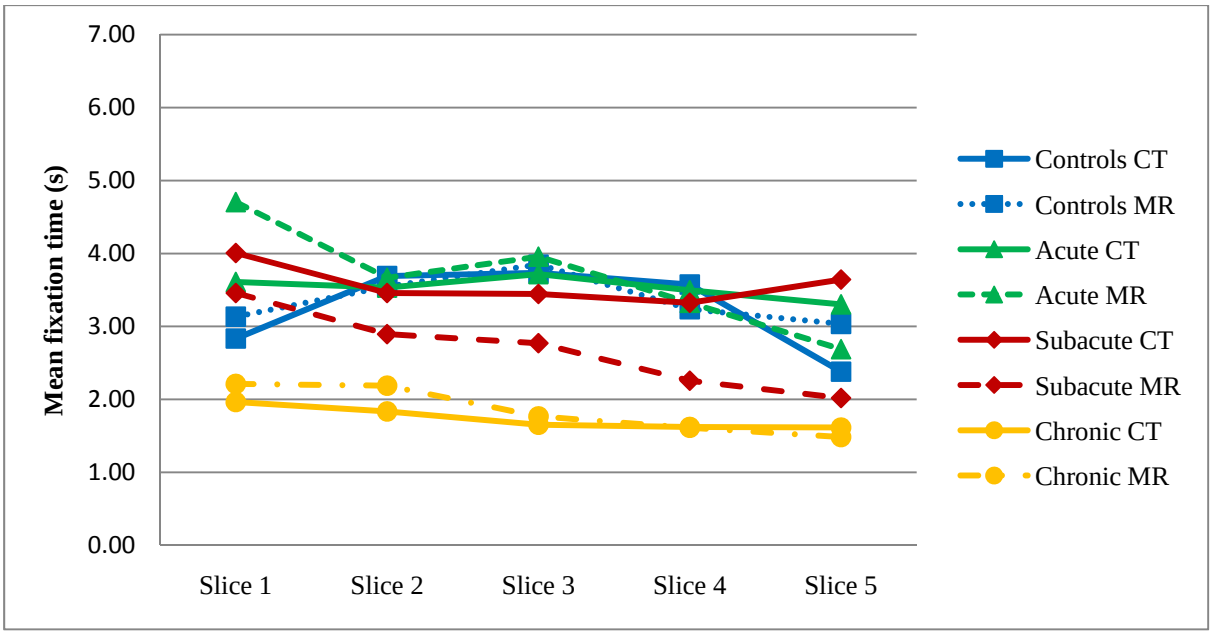


Figure 6.8 Demonstrates the mean fixation time per axial slice for expert readers across all CT and MR cases, irrespective of decision type.

Figure 6.6 demonstrates that novice readers spent the most time appraising control and acute cases overall and the least time on chronic CT and MR images. There appears a trend towards novices spending the most time on the first slice with a quick shoulder effect towards slice 3, which

sometimes reduced further in slice four and increased when viewing the final slice. It appears a sharp reduction in slice time may be indicative of lesion recognition, even if a correct clinical decision was not made following the viewing.

Radiology trainee and expert results (displayed in figures 6.7 and 6.8), highlighted that visual search patterns throughout the stack appeared to depend upon the stroke type; with some case groups following a trend towards spending more time within the middle slice e.g. acute and chronic CT, whereas other data followed a trend towards spending less time in the middle slice than first and last axial images e.g. control CT and MR. Trainees and experts appeared to spend more time in the middle slices of the image stack in both CT and MR control cases. This trend exhibited in normal cases may be indicative of experts examining middle axial sections for additional abnormalities in the absence of an obvious lesion, such as small vessel changes and lacunar infarctions, which frequently appear around the ventricular area. For other cases such as subacute and chronic MR, much time was spent on the first slice, followed by a steady decline of time spent examining the rest of the stack.

Experts, akin to novice results, spent the most amount of time in acute CT cases as indicated by our descriptive statistics results. In agreement with study 1, experts appeared to be more consistent both within and between cases than novices and the time range of visual search was smaller than novice participants in every case. Although qualitatively, there appeared to be more variability between trainees and experts in terms of visual search behaviour on individual axial images, trainees appeared to follow a similar visual search pattern to the experts, on a case by case basis. The above graphs also demonstrate that in the majority of cases, participants spent more time in CT than MR images.

6.3.6 Stroke, Expertise, Accuracy and Visual Search

6.3.6.1 Diagnostic Decision-Making and Eye-movements i.e. TP/FP/TN/FN

This section aims to compare true positive decisions (i.e. correct localisation of abnormality and accompanying confidence rating) with false negative decisions (i.e. where an abnormality was missed or not reported) with the expectation of identifying commonalities embedded within visual search behaviours, and the reported correct and incorrect decision-making between participant group and modality.

The following tables 6.3 and 6.4 demonstrate mean time to hit the primary area of interest (AOI), the mean time spent appraising features within the AOI, the mean time appraising cortical

features outside of the abnormal tissue/ AOI and the mean fixation duration in the image slice overall by participant group. As participant true positive and false negative decisions, particularly trainee and expert decisions, appeared largely dependent upon stroke type and modality, each decision category and stroke type was considered separately.

Table 6.3 represents the true positive eye movement data by group and modality, in the first appearing primary AOI.

True Positive eye-movement data by group and modality (in seconds) in first appearing PRIMARY AOI.			Mean time to hit Primary AOI	Mean time spent in Primary AOI	Mean time out of AOI (in same slice)
Acute	Novice	CT	3.2	1.6	7.9
		MR	1.3	1.6	2.4
	Trainee	CT	1.6	1.4	3.8
		MR	1.0	1.6	3.4
	Expert	CT	1.4	1.4	2.7
		MR	1.3	1.3	2.5
Subacute	Novice	CT	2.5	1.7	5.2
		MR	1.6	1.4	3.9
	Trainee	CT	1.1	1.4	2.3
		MR	1.3	1.5	3.5
	Expert	CT	0.9	1.4	2.3
		MR	1.8	1.2	3.9
Chronic	Novice	CT	1.5	1.4	4.6
		MR	1.2	1.3	2.0
	Trainee	CT	2.4	0.9	3.7
		MR	2.1	1.4	2.6
	Expert	CT	1.5	1.1	3.2
		MR	1.3	1.3	2.5

For true positive decisions regarding acute cases, trainees were the quickest to reach the AOI (denoting abnormal tissue) in MR images i.e. averaging 1 second. Whilst novices and trainees spent circa 1.6 seconds viewing the AOI in CT and MR images, experts spent less time in MR images – indicating slightly quicker lesion recognition with enhanced image clarity. Novices spent circa three times longer appraising the areas outside of the AOI in CT than experts and nearly twice the amount of time compared with trainees. All readers spent longer in acute CT than MR images and overall experts are much quicker to achieve a correct response than novices and trainees.

In subacute cases, experts and trainees were quicker to see the AOI in CT than MR and also spent the same amount of time within the AOI in both modalities. Conversely, novices took up to 2.5

seconds on average to reach the AOI in CT. In chronic cases, experts and novices were quickest to spot the lesion in MR images and also spent 1.3 viewing the abnormal area. Trainees were less quick to reach the chronic AOI in both CT and MRI (2.4 and 2.1 seconds, respectively). All participant groups spent the least amount of time appraising abnormal areas in chronic cases than acute and subacute, irrespective of modality.

When the results were analysed using paired samples t-tests to compare performance between modalities in a single reader group, irrespective of stroke type, correct novices took significantly longer to reach the AOI in CT than MRI (CT=2146.6, MR=1302.3, t=-2.15, *df* 89, *p*<.034) and once they'd seen the abnormal area, they spent longer viewing surrounding features in CT than MRI (CT=5244.0, MR=2728.9, t=-3.39, *df* 109, *p*<.001). Novices' overall fixated much more in the first AOI appearing slice (CT=6738.4, MR=4405.7, t=-3.19, *df* 113, *p*<.002), over the 5 images overall (CT=82.1, MR=55.3, t= -5.07, *df* 115, *p*<.000) and for longer in general in CT than MRI (CT=25469.2, MR=16070.6, t=-4.64, *df* 115, *p*<.000).

When trainee performance and visual search measures were compared between modalities, there were no significant differences in terms of visual search, there were however, differences in terms of their confidence ratings when rating the presence or absence of abnormalities in the images. Trainees were significantly more confident in MRI than CT (CT=3.63, MR=3.95, t=6.38, *df* 146, *p*<.000), and whilst their accuracy scores were enhanced in MRI, their false negative rate was actually increased in MRI.

When assessing differences in performance and visual search of experts, there were no significant differences in terms of confidence, time to 'hit', time in AOI, time out of the AOI and number of fixations, demonstrating that radiologists are consistent observers between modalities. Experts did spend slightly longer overall appraising CT rather than MR images (CT=16963.9, MR=14693.9, t=-2.65, *df* 139, *p*<.009).

Table 6.4 Represents the false negative eye movement data by group and modality, in the first appearing primary AOI.

False Negative eye-movement data by group and modality (in seconds) in first appearing PRIMARY AOI.			Mean time to hit Primary AOI	Mean time spent in Primary AOI	Mean time out of AOI (in same slice)
Acute	Novice	CT	2.7	1.9	6.9
		MR	1.5	0.6	3.9
	Trainee	CT	2.3	0.6	3.5
		MR	1.4	1.1	3.1
	Expert	CT	0.0	0.0	2.9
		MR	0.8	0.3	1.3
Subacute	Novice	CT	1.3	3.2	8.0
		MR	0.1	0.9	2.6
	Trainee	CT	0.02	0.7	4.2
		MR	0.6	0.3	4.3
	Expert	CT	-	-	-
		MR	0.7	0.3	2.1
Chronic	Novice	CT	0.02	2.3	3.8
		MR	0.0	0.7	0.0
	Trainee	CT	-	-	-
		MR	3.5	0.7	4.6
	Expert	CT	4.5	0.5	4.5
		MR	-	-	-

For false negative decisions regarding acute cases, descriptive statistics (table 6.4) showed that novices and trainees took longer to reach the AOI, spent longer outside of the AOI (associated with normal cortical tissue) and in the image overall in CT than MR acute images, yet time spent within the AOI was minimal for most groups except novices viewing CT images, which suggests novices saw the abnormality but did not consider it abnormal. In subacute cases, the same trend was apparent for novice readers. Trainees and experts, once again, spent minimal time gazing at the AOI in subacute cases, which suggests they were largely overlooked. There were no decision errors made by experts in subacute CT cases or chronic MR cases for both experts and trainees, and therefore, no gaze information was gathered.

When comparing TP versus FN decisions between modalities; false negative decisions made by novices regarding acute cases, were accompanied by a quicker time to hit, but less time spent around the AOI than TP decisions in CT images. Of the trainees that made incorrect decisions, the time to hit was longer for acute cases indicating they were not drawn quickly to the abnormality as they had been for TP decisions and they spent less time in the respective AOIs themselves.

When true negative results were compared using paired samples t-tests to compare performance between modalities in a single reader group there were no significant differences between trainee or expert performance between modalities. There were, however, significant differences in false negative decisions in novice visual search patterns. As before, novices looked at the CT images much more (CT=, MR=, t=-2.4, *df* 28, *p*<.023) and for longer than MR images (CT=33157.6, MR=17240.7, t=-3.68, *df* 28, *p*<.001). They were also much quicker to see the abnormality in MR than CT (CT=2702.2, MR=989.6, t=-3.08, *df* 13, *p*<.009).

6.3.7 Secondary Abnormality Comparisons between Modalities

When comparing secondary detection between modalities, novice performance was significantly better in MRI than CT (CT=.23, MR=.73, t=-4.35, *df* 29, *p*<.000). Trainee scores were also significantly enhanced in MRI than CT (CT=.47, MR=.77, t=-2.34, *df* 29, *p*<.026), yet experts were unaffected by modality, which demonstrates a consistency between image types for secondary as well as primary detection. Both novices and trainees spent much longer looking at secondary abnormalities in MR than CT (CT=205.6, MR=2404.2, t=-5.794, *df* 23 *p*<.000 & CT=246.9, MR=2091.95.2, t=-5.372, *df* 21 *p*<.000, respectively) indicating an increase in fixation time is necessary to detect secondary abnormalities within the same case detection. However, one abnormality in CT was slightly smaller than the other two which might have influenced these differences. Future studies should ensure that secondary abnormalities are the same sizes to allow 100% in cross comparisons between modality experiments.

6.3.8 Satisfaction of Search

In terms of satisfaction of search between modalities, novice were more likely to look at the abnormality in CT than MRI (66.7% compared with 43%), with 40% looking at the secondary afterwards in both modalities. Trainees were more likely to detect the primary in CT than MRI, with a third proceeding on to see the secondary afterwards in both modalities, indicating a degree of satisfaction of search between novice and trainee readers that did not differ between image types.

17% more experts spotted the abnormality first and were 12% more likely to go onto see the secondary in CT compared with MRI. Expert readers were 21% more likely to see the secondary first in MRI, but of these, were less likely to look at the primary afterwards compared with CT, which may indicate the primary was clear and could be visualised in peripheral vision owing to the high detection rate regardless of looking directly at the abnormality in this slice.

6.4 Discussion

6.4.1 Diagnostic Accuracy and Confidence by Group and Modality

This study aimed to explore observer performance between participant groups when viewing conventional CT and MR multidimensional images of stroke. For both modalities, ruling out the presence of an infarct and detecting small infarctions were perceived to be the most challenging tasks among all stroke cases. Highly significant differences were observed between groups and modalities when investigating accuracy, with all groups performing better in MRI, even if novice ratings were more apprehensive than their more experienced counterparts. It seems as image clarity enhances, a more experienced eye is required to confidently filter out what is normal from abnormal.

Whilst novice readers may prefer the enhanced detail of MRI, as reflected by enhanced confidence ratings attached to their reported abnormalities, they actually perform better in CT (i.e. their true positive ratings are enhanced and false positives are decreased), therefore, enhanced confidence does not necessarily mean they perform better in this condition. Conversely, trainee ratings appeared to favour CT over MRI, which may be indicative of a reduced amount of reading time trainees had previous to the study with MRI, yet their overall performance was superior in MRI, despite taking them longer to scan the images. Expert readers spent longer searching for abnormalities in CT but were confident and consistent within and between both modalities.

6.4.2 Image Analysis, Expertise and Eye Movements

Qualitative image analysis results highlighted a clear difference between participant groups when viewing all medical images, irrespective of modality; novices spent more time visually examining normal anatomy such as ventricles and/ or spent much time concentrating upon a large and unambiguous lesion. Owing to lack of experience when appraising medical images and clinical problem-solving, novices had to make decisions about clinical features with very little training and no feedback on performance throughout the experimental study. Novices, therefore, are reliant upon their own limited experience to ‘construct’ their own, or multiple ‘mental schema’ regarding what constitutes an abnormality and what might not.

Length of time viewing images appears to indicate either a need to be thorough with prior knowledge that certain image types may ‘mask’ a present abnormality, or may indicate a degree of

uncertainty when making a final decision. Expert readers demonstrate a more thorough appraisal of CT images, probably owing to the former rationale, whereas trainee visual search in MR images appears to conform to the latter reasoning. Trainee readers do, however, spend much more time cross comparing hemispheres than expert readers, which appears to demonstrate a clear intention to be thorough in their image appraisal yet their ability to recognise abnormalities in MR is likely to be limited by experience.

With much caseload experience experts were more capable of directing their attention to an abnormal clinical feature, as time to hit and accuracy reports reflect, make a quick confident decision and promptly move on to appraise the surrounding cerebral tissue. Compared with novice readers, experts appeared to operate a system of deduction; ruling out certain areas very quickly to economise time and effort as demonstrated by the direction and duration of fixations within the medical image. Expert readers may indeed have a mental schema of acquired information surrounding disease manifestation and how, due to known cerebrovascular pathways, different stroke types pervade different vascular routes, although this assertion is out of the remit of this series of studies. These qualitative and quantitative findings surrounding performance and visual search appear to be in agreement with literature previously discussed e.g. Rogers (1995), Nodine and Krupinski (1998), Garlatti and Sharples (1998), and Manning (2006).

6.4.3 Diagnostic accuracy, Modality, Expertise and Eye-movements

It is well known that technical limitations of the chosen modality can affect accuracy. Previous studies into modality preference have not rated conventional CT as superior to conventional MR or vice-versa (Mohr, 1995; Lansberg, 2000; Wintermark, 2007) and unenhanced CT remains the primary modality adopted when patients present with stroke-like symptoms (Kloska, 2004). The primary reason for adopting CT over MRI in clinical practice, is that CT is adequate for differentiating ischaemic from haemorrhagic stroke and secondly, because the use of CT scanners as a first port of call frees up the MRI resource for many other conditions. However, for these forty-eight cases, it appears experts performed this reading task much quicker and with more accuracy in the MR condition over CT.

Trainees outperformed novice readers in both CT and MRI but whilst they spent slightly less time viewing the CT images, trainees spent significantly more time viewing MR images than both novice and expert readers. As previously alluded, the enhanced anatomical detail and/ or lack of previous reading time with MRI may have led to uncertainty and an increase in task time for trainee readers. Specific reasons for this difference would need to be further explored with trainee

radiologists in a separate study. For novice readers, however, specificity was enhanced in CT where a more aggressive rating was adopted compared with a more conservative specificity rating in MRI. Sensitivity was enhanced for novice readers in MRI compared with CT and the overall area under the curve (AUC) values for both modalities, although slightly reductionist, indicate a higher accuracy value for MRI at .87 than CT at .83.

Overall, experts were quicker to detect the primary infarction in MR for acute (mean time: 1.1 seconds) and chronic stroke types (mean time: 1.1 seconds) but quicker in CT for subacute cases (mean time: 0.9 seconds). Experts were quicker to reach an AOI and spent less time appraising surrounding cortical tissues around the abnormality than novices and trainees in the most challenging cases, appearing to confirm the qualitative results indicating quicker, more accurate decisions and also confirming findings discussed in study 1. False positive decisions were characterised by more inconsistencies within and between groups in terms of eye-movements; either taking an unusually long time to reach an AOI (e.g. 6.7 seconds for an expert FP decision in chronic CT) or not enough time for perception to amount to recognition when viewing the abnormality itself. It appears that recognition of an abnormality, if the reader is experienced enough to recognise it, is circa 240 milliseconds or more.

In this study, radiology trainee and expert visual search patterns appeared to be characteristic of a particular case type i.e. control, acute, subacute and chronic, when viewing these clinical images, indicating a trend towards differential image appraisal and visual search dependent upon abnormality size, density and overall stroke type, even between modalities. This study also suggested that expert reading time is reduced and accuracy enhanced by adopting MR over CT imagery and trainees may benefit from further training in MR imagery of stroke.

6.5 Conclusions/ Summary

Novice and expert readers spent longer appraising CT images than MR, compared with trainees where the inverse was true. Diagnostic accuracy and confidence ratings were positively correlated with experience but in less experienced readers, confidence did not always equate to accuracy and performance.

Differences were observed between novice, trainee and expert visual search patterns. In particular, trainee and expert search behaviours through the image ‘stack’ appeared to depend upon stroke type being scrutinised. Image analysis trends did not appear to differ between modalities, but time spent within clinical images, accuracy and relative confidence performing the task did differ

between CT and MR reader groups. Eye movement results suggest that true positive decisions were marked by a quick time to infarct and a consistency within reader group, whereas false positive decisions were more inconsistent both within and between groups. Consistent between modalities was the finding that the number and duration of fixations decreases with experience.

6.6 Study Reflections

To-date few studies have explored observer performance in neuroradiology and the present study examines multi-slice image appraisal by comparing matched pairs of CT and MRI stroke cases between novice, trainee and expert readers. Consultants are the optimum performers in both the CT and MR image appraisal tasks, yet they performed much quicker in MRI, which may have implications for stroke scanning in the radiology department if these findings are evident at other radiology departments within the United Kingdom.

The following chapter moves on to examine the performance of neurologists in the CT and MR task compared with consultant radiologists. Consultant neurologists are specialists within the field of Neurology and can be matched for level of expertise with the radiologists at NNUH. Whilst neurologists request imaging procedures for patients and frequently review image findings either alone or with an accompanying radiologist, they have not had specific radiology training. In addition, the task of writing and reviewing radiology reports on patient status is not within their remit. Therefore, consultant neurologists provide an interesting comparison with consultant radiologists from a diagnostic accuracy and visual search paradigm.

Study 4: Comparing Consultants: How Does Radiology Performance Compare With A Matched Sample of Neurologists?

(An exploration of CT and MRI performance and eye movements when level of experience is controlled.)

As previously discussed in chapter 1, neurology is a medical specialty which deals with all diseases or disorders which affect the brain and the central nervous system. Neurologists are specialists within this area, who diagnose and treat patients, adults and/ or children, with an array of neurological problems. Neurologists consult radiologists and request a patient scan when it is necessary to gain further information regarding internal functioning to confirm or refute a queried medical diagnosis based upon the medical history and/ or presenting patient status. Radiologists facilitate neurologists, by providing accurate feedback on the patients' image findings. It is clear that the two specialties work in tandem in neuroradiology, however their roles are quintessentially different; radiologists interpret patient image findings, which allow neurologists to diagnose and treat.

In the case of stroke detection, radiologists play a key role in determining whether a set of suspicious symptoms, e.g. loss of vision, balance or speech, are a result of cardiovascular occlusion characteristic of a stroke or are symptomatic of another medical problem such as an intracranial haemorrhage or brain tumour, for example. When radiologists feed this essential information back to the neurologist in the form of a radiology report, neurologists can take direct action to appropriately treat the patient. Treating the patient without key information from a head scan, confirming early parenchymal signs of anatomical change can have catastrophic consequences. For instance, a patient who is given thrombolytic therapy for an intracranial haemorrhage will further bleed internally (Schriger *et al.,* 1998, Grotta, 1999), yet for an arterial blockage, the same therapy alleviates the clot, removing the build up of pressure on the area of the brain.

In previous studies, stroke detection and classification results demonstrated that radiology and neurology performance did not differ, with both groups averaging a score of 83% accuracy (Schriger *et al.*, 1998). 40% of neurologists and 52% of radiologists scored 100% sensitivity, and overall study sensitivity was 82%. Follow-on studies have discussed substantial variability both

between readers and between specialties when Cohen's Kappa was used to compare inter-rater reliability (Grotta, 1999).

The present study explores not only stroke detection between neurologists and radiologists working in an acute care trust, but also the visual search patterns between the groups when examining the images, which has not been studied before. Both consultant groups have undergone rigorous medical training and have chosen to specialise in one area or another; however, neurologists have some experience viewing medical images, as radiologists have an idea of triage and treatment. With a baseline of medical training and years of consultancy controlled for, what differences will exist (if any) in the visual search and accuracy patterns between the two expert groups from differing specialties? It is hypothesised that neurologists will not out-perform radiologists and their visual search patterns will be significantly different from radiologists. Neurologist performance and visual search is likely to be in line with trainee performance and visual search patterns.

Study Aims and Objectives: To explore the visual interpretation of brain MRI images between Radiologists and Neurologists, and examine differences in performance, confidence ratings, and visual search behaviours within and between CT and MR image modalities.

7.1 Methods

7.1.1 Participants

Two groups of participants were selected and recruited for this study; radiology and neurology consultants. Expert radiologists were defined as individuals who had completed their core training, were registered practitioners with either a practising special interest in neuroradiology or had completed the full neuroradiology specialty training via the masters in head and neck radiology route. Expert Neurologists were defined as individuals who had also completed their core medical training, were registered practitioners and had completed their specialty training in Neurology.

In total, 16 participants were recruited; eight Consultant Radiologists and eight Consultant Neurologists. The same consultant radiologists were incorporated into the present study. In terms of neurology participants, 87% were male and 12.5% female. Most participants were between 30-40 years of age. Neurologists had been qualified for an average of 13.5 years. Whilst neurologists do not assess the images, they do refer patients onto the radiologists. The total number of CT cases referred

onto radiologists by this subset of neurologists was 150 within the year prior to the study. The range of CT cases referred was between 2 and 57 cases. The median number of cases referred was 24. The total number of MR cases referred by this subset of neurologists was 796 cases and the range of MR cases assessed was between 134 and 268 cases. The median number of cases referred was 197.

7.1.2 Design and procedure

The design and procedure was identical to study 3, with the same participants completing both CT and MR experimental studies.

7.2 Results

Study data was analysed to investigate; i) qualitative image analysis ii) accuracy and confidence ratings of performance, iii) quantitative eye movement analysis, iv) stroke, expertise accuracy and visual search and v) diagnostic accuracy: within and between participant results and vi) all results compared between CT and MR imagery between radiology and neurology readers.

Case study 1. CT normal case: The following figure highlights the differences between consultant readers' (a radiologist and a neurologist) visual inspection strategies when appraising images of a normal case (NBH).

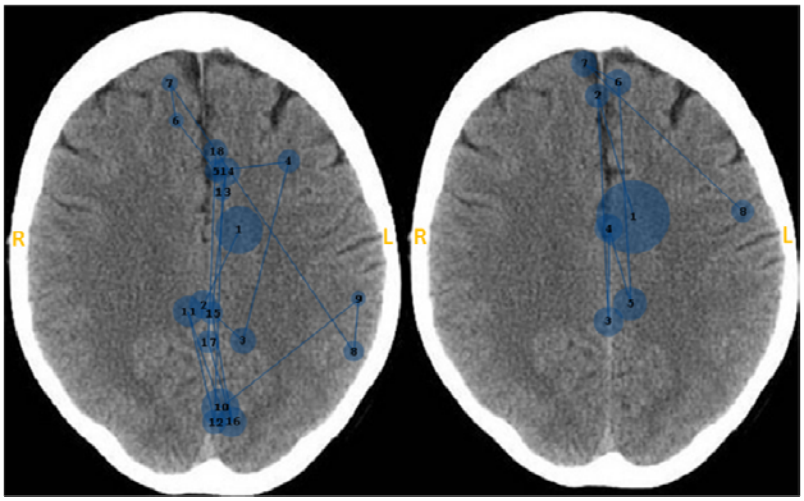


Figure 7.1 highlight the fixation patterns of a radiology and neurology reader when assessing a normal CT image slice.

The above hotspot images demonstrate how similar the search patterns of these two readers are, even though they come from different specialties; one consultant with search strategy training and one consultant without. To view all gaze tracker images for this case, please refer to pages 7 and 8 of the appendix.

Case study 2. CT acute stroke: The following gaze-tracker figure highlights the differences between consultant readers' (a radiologist and a neurologist) visual inspection strategies when appraising images of Acute CT stroke images (case AAB).

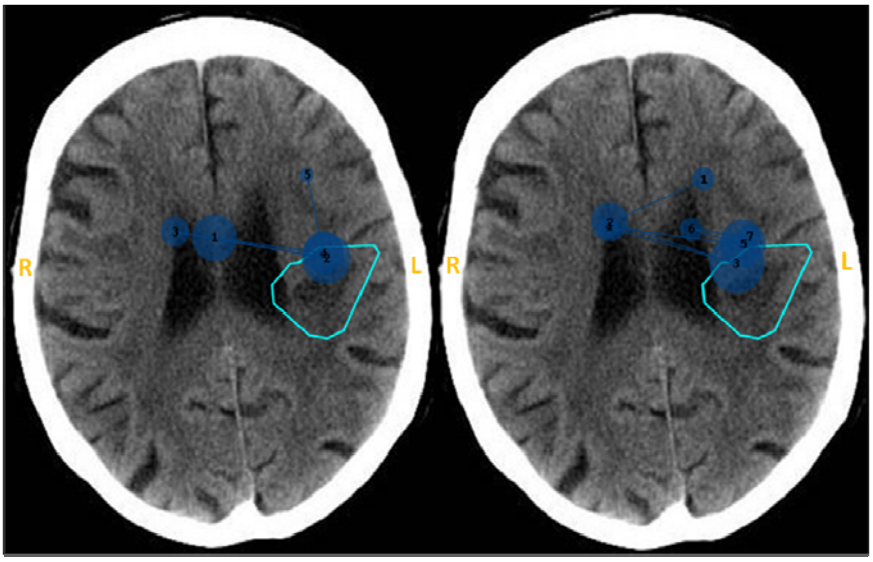


Figure 7.2 highlight the fixation patterns of a radiology and neurology reader when assessing an acute CT image slice.

These acute gaze-plot images are also very similar between readers but with only two fixations more by the neurology reader compared with the radiology reader. Both images highlight a quick time to hit and almost identical scan paths. To view all gaze tracker images for this case, please refer to pages 11 and 12 of the appendix.

Case study 3. MR normal case: The following figure highlights the differences between consultant readers' (a radiologist and a neurologist) visual inspection strategies when appraising images of a normal MRI case (NHG).

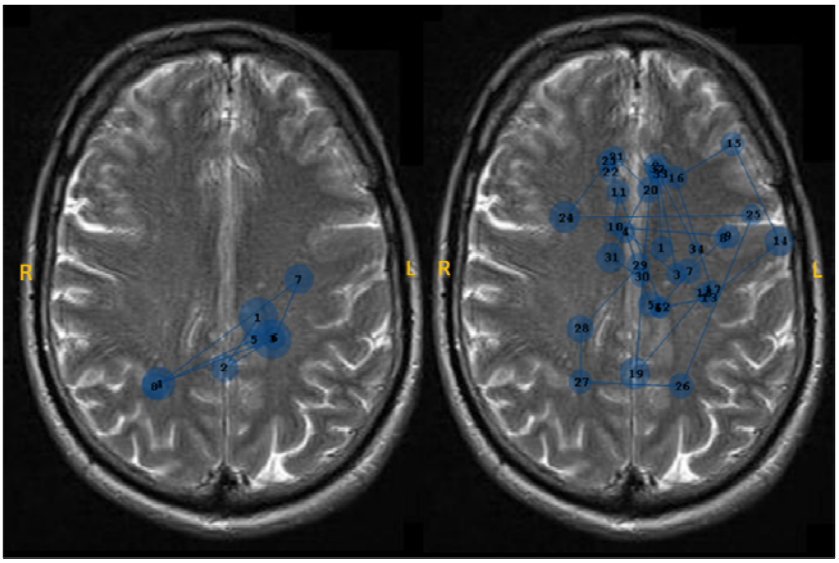


Figure 7.3 highlight the fixation patterns of a radiology and neurology reader when assessing a normal MRI slice.

When searching for abnormalities in the MR image, neurologists spend much more time appraising all the clinical features, as marked by an increase in fixations and saccadic coverage when compared with this radiologist, as seen in hotspot figure 7.3. To view all gaze tracker images for this case, please refer to pages 21 and 22 of the appendix.

Case study 4. MR acute stroke: The following gaze-tracker figure highlights the differences between consultant readers' (a radiologist and a neurologist) visual inspection strategies when appraising images of Acute MR stroke images (case ACW).

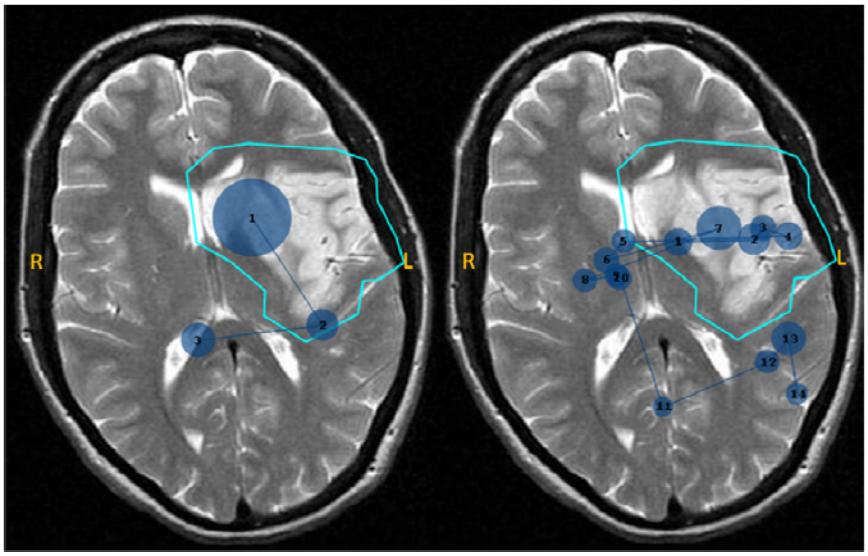


Figure 7.4 highlight the fixation patterns of a radiology and neurology reader when assessing an acute MRI slice.

Once again, there are very little differences between the visual search patterns of this radiology reader compared with the neurology reader. Although the neurologist fixated 11 more times than the radiologist, both saw the abnormal area in the first fixation. To view all gaze tracker images for this case, please refer to pages 25 and 26 of the appendix.

Although neurologists spent marginally more time gazing over the clinical features, collectively the above figures demonstrate very few differences in visual search between radiology ad neurology readers in both CT and MRI cases.

7.2.1 CT and MR Receiver Operating Characteristics between Consultants

The following Conventional Binormal ROC Curves represent primary abnormality detection ratings by participant group (i.e. Radiologists versus Neurologists) and images per modality (i.e. CT and MR) of confidence rating scores per case.

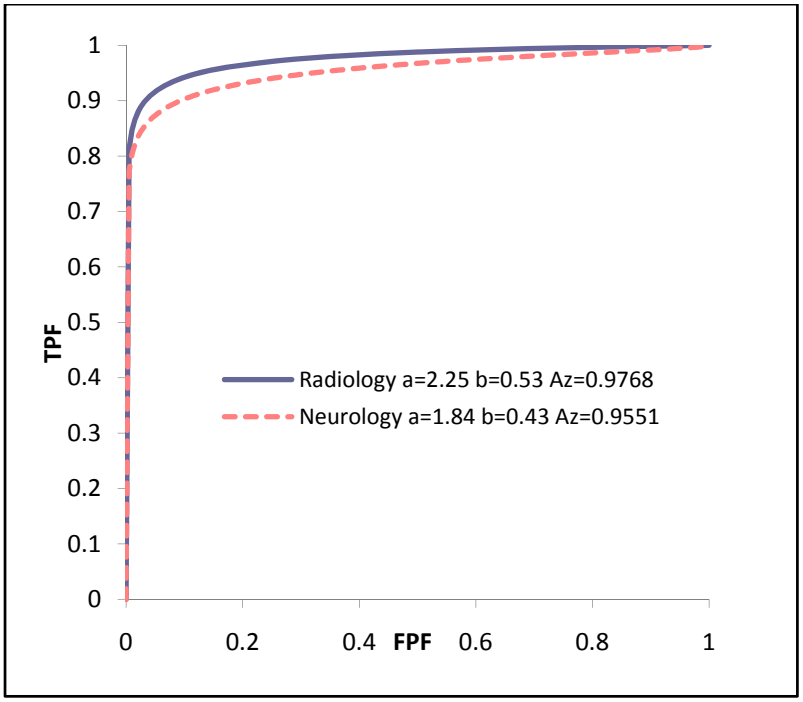


Figure 7.5 ROC Curve 1: Represents primary abnormality detection scores between radiologists and neurologists when rating CT images.

The above ROC curve demonstrates that the radiology readers were more accurate i.e. higher sensitivity and specificity rates than neurologists when examining these CT cases.

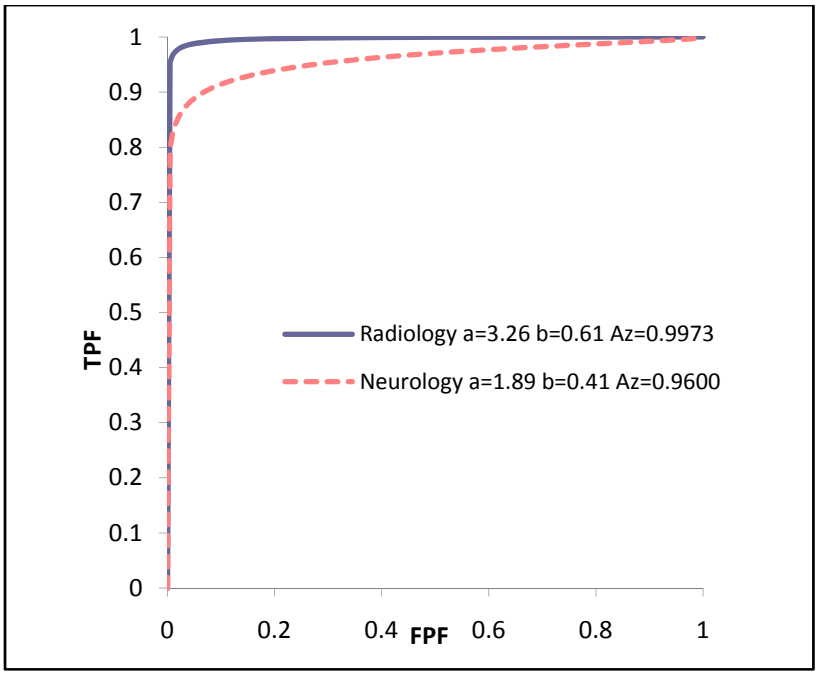


Figure 7.6 ROC Curve 2: Represents primary abnormality detection scores between radiologists and neurologists when rating MR images.

The second ROC curve in this chapter demonstrates, unsurprisingly, that radiology readers were also more accurate than neurologists when examining these MR cases. As outlined in previous chapters, within the ROC space the parameters 'a', 'b' and 'Az' represent the vertical intercept, the slope of the fitted curve and the overall accuracy value as calculated by the conventional Binormal curve process, respectively.

7.2.2 Accuracy and Confidence Rating Data: Quantitative Analysis of Observer Performance

The following section examines the accuracy of readers in determining the location of the abnormal areas, their confidence when reporting the abnormal location, the perceived difficulty of the individual cases and differences in performance both within and between groups when comparing stroke case types overall. Please refer to chapter 2 for further details regarding how measures of accuracy and confidence between and within participants were derived.

7.2.2.1 Diagnostic accuracy

When considering diagnostic accuracy between consultants and modalities, table 7.1 highlights that radiologists who use MR imaging are the most accurate compared with neurologists who are most accurate when using CT imaging. Neurologists were equal to, or outperformed radiologists when examining images of normal and chronic cases in CT but did not outperform the radiologists in any stroke type in MRI.

Table 7.1 highlights the diagnostic accuracy of participant groups between patient cases i.e. normal, acute, subacute and chronic stroke types. For further information regarding how accuracy scores were derived, please refer to chapter 2.

Diagnostic accuracy (%) by stroke type	Modality	Radiology	Neurology
Normal control score	CT	81.3	92.0
	MR	93.8	88.0
Acute	CT	86.0	78.0
	MR	96.9	94.0
Subacute	CT	100.0	98.0
	MR	95.8	88.0
Chronic	CT	94.0	94.0
	MR	100.0	91.0
Average accuracy score	CT	90.2%	90.4%
	MR	96.6%	89.8%

As previously discussed in chapter 6, radiologists were significantly more accurate in MR than CT (CT=.91, MRI=.96, t=-2.13, *df* =191, *p*<.033) yet, neurology performance did not differ between modalities. When ANOVA tests were performed to compare consultant performance within and between modalities, there were no significant differences between the radiology and neurology accuracy scores in CT (*df* 1, *p*<.496 F=.47, Eta Squared.00). There were, however, significant differences in MRI with radiologists outperforming neurologists (*df* 1, *p*<.015 F=6.00, Eta Squared.02).

7.2.2.2 Confidence Rating Data: Quantitative Analysis of Primary Infarctions.

When reporting participant confidence in their decision-making in CT, descriptive statistics show overall percentage of cases correct indicates that radiologists and neurologists were most confident in MR images (radiology: 82.3%. neurology: 80.4%), which is interesting as neurologists made more errors in MR than CT cases. Across all cases and between modalities, radiologists were the most confident in their decisions, although both groups were least confident in normal CT cases. In CT, Radiologists found case AMW the most difficult, and all of the subacute cases the least challenging. Neurologists found case ADH the most difficult, the most difficult and correctly reported 98% of subacute cases. In MRI, both groups found case SAC the most challenging and whilst radiologists performed optimally in chronic MR cases, neurologists performed better when rating acute stroke types.

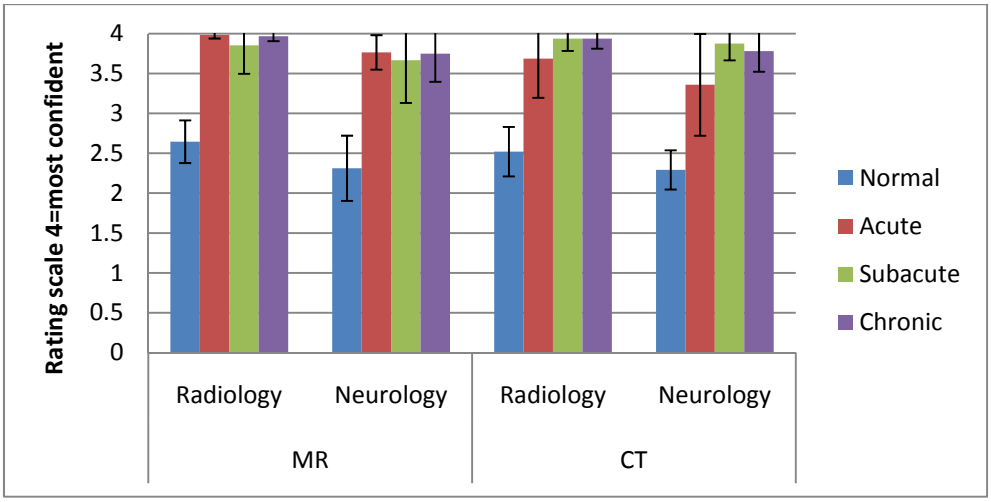


Figure 7.7 represents the mean confidence score per case type, participant group and modality.

N.B. Rating scale 4 = most confident.

The figure above demonstrates very few differences between the two specialties and there were no significant differences within or between consultant groups in terms of confidence scores either in CT or MRI. Between modalities and groups, neurologists are marginally less confident across all stroke case types.

7.2.3 Diagnostic Decision-making (i.e. TP/FP/TN/FN) between Modalities

As demonstrated above in the ROC curves, radiology readers are the overall optimum performers in both CT and MRI. The following table highlights the spread of true and false positive ratings and true and false negative per group and modality to gain a more in-depth insight into the decision-making processes of each reader group to highlight where subtle differences may be present.

Table 7.2 Represents spread of decisions between consultant group and modality type.

Overall ratings	Modality	TP	TN	FP	FN
Radiology (n=8)	CT	138	34	14	6
	MR	140	45	4	3
Neurology (n=8)	CT	127	43	10	12
	MR	130	40	11	11

Table 7.2 demonstrates that, unsurprisingly, radiologists had more true positive ratings and less false negative ratings than neurologists overall. However, although there were more true negative results reported by radiologists in MRI, neurologists were better at ruling out primary abnormalities in CT

than radiologists. Interestingly, whilst radiologists reported four more false positives in CT than neurologists, neurologists reported seven more false positives in MRI. Radiology sensitivity rates were 96% in CT and 97% in MRI, whereas neurology sensitivity was 91% in CT and 92% in MRI. Radiology specificity was much enhanced in MRI: 92% compared with 70% in CT yet neurology specificity was 75% in CT and 69% in MRI.

7.2.4 Eye-movements and Experience: Quantitative Analysis.

This quantitative results section aims to identify statistically significant links between visual search behaviours (e.g. total time viewing cases, time to reach an AOI, time spent within an out of AOIs) and reported accuracy within and between radiology and neurology readers to assess how similar observer groups appraise visual stimuli in neuroradiology.

7.2.4.1 Task Viewing Time per Modality

When assessing the differences in reader time viewing the images, There was a statistically significant difference within the radiology reader group (CT=18364.21, MRI=1554.28, t=3.36, *df*=191, $p<.001$), with radiologists spending much longer gazing over CT than MR images. The same difference was not found for neurology readers. There were also no significant differences between reader groups for time spent viewing either CT or MRI.

7.2.4.2 Visual Search Behaviour throughout the Image 'Stack'

The following figures 7.8-7.11 demonstrate mean time spent in each axial slice throughout the five image 'stack' by case type (control, acute, subacute and chronic), participant group (radiology and neurology) and modality (CT, MR), irrespective of decision type.

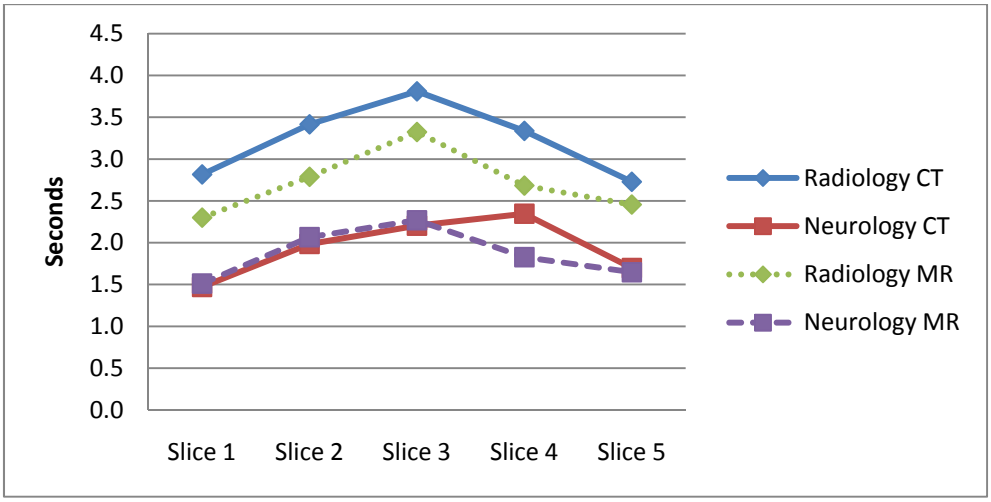


Figure 7.8 Mean fixation time per axial slice for expert readers across all CT and MR control cases

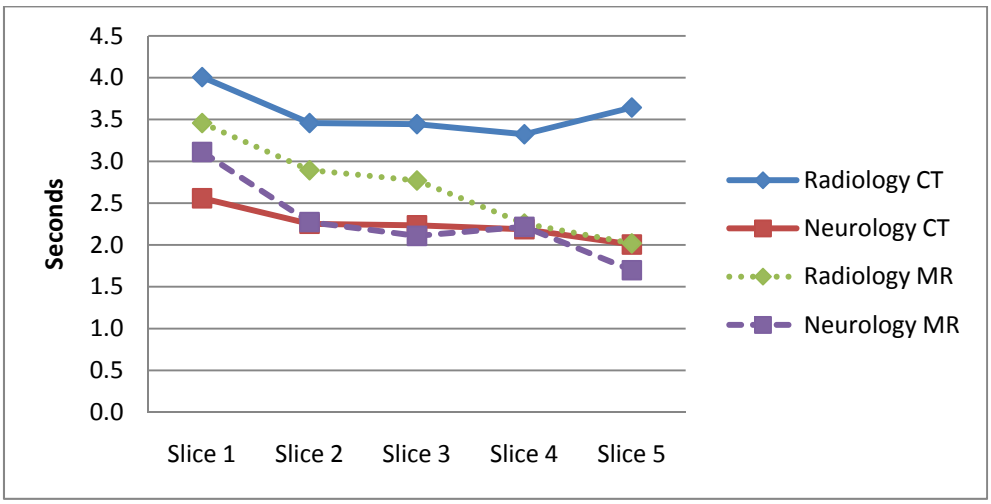


Figure 7.9 Mean fixation time per axial slice for expert readers across all CT and MR acute cases

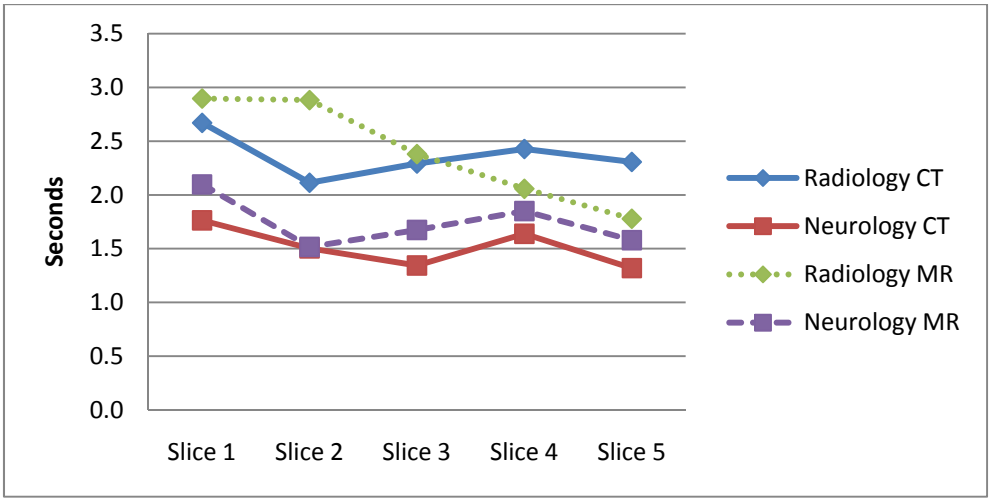


Figure 7.10 Mean fixation time per axial slice for expert readers across all CT and MR subacute cases

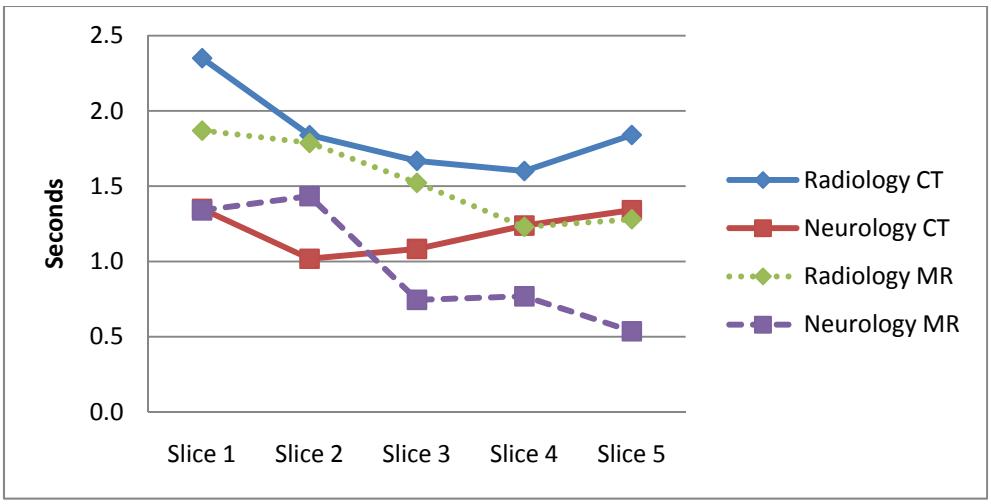


Figure 7.11 Mean fixation time per axial slice for expert readers across all CT and MR chronic cases

The above figures 7.8-7.11 demonstrate differences in visual search between consultant groups. When contrasting search between slices for CT and MR, neurology search patterns are remarkably similar for normal, acute and subacute cases, however, search patterns differ quite considerably in the chronic case type where neurologists either spend gradually less time across the slices in MRI, or slightly more towards the final slice in CT. Neurology visual search patterns do not appear too dissimilar from radiology patterns throughout the stack although, neurologists do not appear to spend more than 3.25 seconds in any one image compared with radiologists who spend

approximately two seconds longer within each image slice than neurologists.

7.2.5 Stroke, Expertise, Accuracy and Visual Search

7.2.5 1 Diagnostic Decision-making and Eye-movements i.e. TP/FP/TN/FN

As in previous chapters 4 (CT) and 5 (MRI), this section aims to compare true positive decisions with false negative decisions with the expectation of identifying commonalities embedded within visual search behaviours, and the reported correct and incorrect decision-making between participant group and modality. The following tables 7.3 and 7.4 demonstrate mean time to hit the primary area of interest (AOI), the mean time spent appraising features within the AOI, the mean time appraising cortical features outside of the abnormal tissue/ AOI and the mean fixation duration in the image slice overall by participant group. As participant true positive and false negative decisions, particularly trainee and expert decisions, appeared largely dependent upon stroke type and modality, each decision category and stroke type was considered separately.

Table 7.3 Represents true positive eye movement data by group and modality in the first appearing primary AOI.

True Positive eye-movement data by group and modality (in seconds) in first appearing PRIMARY AOI.			Mean time to hit Primary AOI	Mean time spent in Primary AOI	Mean time out of AOI (in same slice)
Acute	Radiology	CT	1.4	1.4	2.7
		MR	1.3	1.3	2.5
	Neurology	CT	2.2	1.3	2.9
		MR	2.3	1.3	3.4
Subacute	Radiology	CT	0.9	1.4	2.3
		MR	1.8	1.2	3.9
	Neurology	CT	0.9	1.4	2.9
		MR	1.5	1.5	3.0
Chronic	Radiology	CT	1.5	1.1	3.2
		MR	1.3	1.3	2.5
	Neurology	CT	1.4	1.2	2.5
		MR	3.3	3.3	2.0

For true positive decisions regarding acute cases, neurologists (irrespective of modality) took marginally longer to reach the AOI than radiologists, although this finding wasn't statistically significant. In subacute cases, the results between consultants for CT were nearly identical with both groups reaching the infarct under a second. In MR subacute images, neurologists were slightly quicker to reach the AOI than radiologists and spent less time in the slice than radiologists. In chronic

cases, it took neurologists two seconds longer (on average) to reach the AOI in MR images although they were very slightly quicker to reach the AOI in CT.

Descriptive statistics demonstrated that neurologists spent longer in the first appearing AOI slice and outside of the AOI than radiologists. In MRI, this finding was significantly different between consultant groups with neurologists spending an average of 885 milliseconds longer than radiologists in the first slice where the abnormality appears (*df* 1, *p*<.050 F=3.85, Eta Squared.01). There was also a large effect size observed for time spent outside of the area of interest with neurologists spending much more time viewing features surrounding the abnormality (*df* 1, *p*<.000 F=28.21, Eta Squared.83). Whilst neurologists spent twice as long appraising the primary abnormalities in CT imagery when compared with abnormalities in MRI (CT=3330.3, MR=1198.2, t=-6.66, *df* = 83, *p*<.000), there were no significant differences in expert eye movements within or between modalities and the number of fixations and/or fixation duration did not differ significantly between consultants or modalities.

Table 7.4 Represents false negative eye movement data by group and modality in the first appearing primary AOI.

False Negative eye-movement data by group and modality (in seconds) in first appearing PRIMARY AOI.			**Mean time to hit Primary AOI**	**Mean time spent in Primary AOI**	**Mean time out of AOI (in same slice)**
Acute	Radiology	CT	0.0	0.0	2.9
		MR	0.8	0.3	1.3
	Neurology	CT	1.6	1.0	2.4
		MR	0.2	0.9	0.4
Subacute	Radiology	CT	-	-	-
		MR	0.7	0.3	2.1
	Neurology	CT	-	-	-
		MR	1.0	0.6	3.2
Chronic	Radiology	CT	4.5	0.5	4.5
		MR	-	-	-
	Neurology	CT	0.0	0.0	3.0
		MR	0.9	0.8	3.8

For false negative decisions regarding acute cases, the above table highlights a mixed result; the radiologists either do not fixate within the AOI or they barely look upon it. Neurologists spend approximately one second within the AOI in the slice where it first appears, but the abnormal area is subsequently discounted as not suspicious. Where consultants made errors in subacute cases, they were only with regard to MR images and the AOI was fixated upon for less than 600 milliseconds

approximately. In chronic cases the pattern appears similar to the subacute cases, whereby the abnormality was either missed altogether or barely fixated upon, indicating a recognition error had occurred. When comparing TP versus FN decisions between modalities the results appear similar to those previously discussed in chapters 4, 5, and 6, yet with true positive decisions, neurologist results are much more in line with radiologists than the radiology trainees in previous studies i.e. quicker time to hit and similar decision times within and out of the AOI. Whilst there were subtle descriptive differences in table 4, ANOVA results only uncovered one significant difference when comparing false negative decisions between consultants and modalities, with neurology readers spending much longer in CT than MR cases (CT=22695.1, MR=10295.1, t=-3.45, *df* 11, $p<.005$) when making false negative decisions.

7.2.5.2 False Positive & True Negative Decisions

As a number of false negative decisions were identified between participant groups and stroke types in this study also, the following section aims to examine whether errors could be attributed to recognition and/or decision errors of normal features by further exploring the participant reports and their eye movements. Further analysis was applied to examine false positive decisions; specifically those made by experts and compare their visual search patterns with true negative decisions by other experienced readers in the study.

As in the previous chapters, it is important to consider that only abnormal areas can be predefined by creating an AOI before the experiment took place, therefore, the exact time and slice where an individual false positive decision was made (based on certain clinical features) cannot be completely determined retrospectively, owing to the 2D reporting strategy of a 3D anatomical structure. However, eye movements throughout the image stack in these cases often infer the clinical features that were implicated in the decision error, as demonstrated in the following images and overlaid gaze patterns.

Table 7.5 highlights mean false positive eye-movement data by group and stroke type (in seconds) across all case types between readers and modality.

	Reader Group	FP Location	No of FP's	Slice 1	Slice 2	Slice 3	Slice 4	Slice 5	Mean total Case time	Total group time
CT	Radio.	Normals	9	3.4	3.6	5.2	4.7	4.0	20.8	
		Acute	5	4.2	2.9	3.6	3.1	3.0	16.8	
		Sub-acute	0	-	-	-	-	-	-	
		Chronic	0	-	-	-	-	-	-	34.6
	Neuro.	Normals	5	2.5	2.7	4.4	4.2	4.2	18.0	
		Acute	5	2.5	2.6	2.6	3.3	2.3	13.3	
		Sub-acute	4	7.7	5.0	5.2	6.2	5.1	29.2	
		Chronic	0	-	-	-	-	-	-	60.5
MR	Radio.	Normals	3	5.0	5.9	7.4	3.9	4.8	27.0	
		Acute	1	1.6	1.5	2.8	2.4	1.5	9.7	
		Sub-acute	0	-	-	-	-	-	-	
		Chronic	0	-	-	-	-	-	-	36.7
	Neuro.	Normals	8	3.6	4.1	4.7	3.1	3.5	19.0	
		Acute	5	2.3	1.2	0.6	2.8	1.2	8.1	
		Sub-acute	1	2.3	0.2	0.6	0.3	5.7	9.1	
		Chronic	4	6.7	7.9	3.9	3.9	2.6	25.0	61.1

Table 7.5 highlights that these radiologists made the most false positive decisions in normal CT cases, whereas these neurologists made the most false positives in normal MRI cases. Neurology reader false positive decisions were characterised by spending twice as long as radiologists in these cases. Neurology readers spent the most time in chronic and subacute cases, whereas radiologists spent more time investigating normal cases, presumably searching for secondaries or subtle changes indicative of stroke.

Table 7.6 highlights mean fixation duration by reader group and decision type (TN or FP) throughout the image stack.

	Reader group	Participant Decision	Slice 1	Slice 2	Slice 3	Slice 4	Slice 5	Mean total Case time
CT	Radiology	TN	3.8	4.8	5.1	4.4	3.6	21.6
		FP	3.4	3.6	5.2	4.7	4.0	20.8
	Neurology	TN	3.1	4.2	3.6	5.0	3.5	20.3
		FP	2.5	2.9	4.8	4.6	4.7	19.5
MR	Radiology	TN	2.9	3.6	4.2	3.8	3.2	17.5
		FP	5.0	5.9	7.4	3.9	4.8	27.0
	Neurology	TN	3.4	4.8	5.3	4.4	3.9	19.6
		FP	3.6	4.3	4.6	2.9	3.2	18.6

When comparing fixation durations between correct (true negative) and incorrect (false positive) decisions throughout the 5 images slices of just normal cases, both radiologists and neurologists spend slightly longer over true negative decisions than false positive decisions in CT. Conversely, in MRI, radiologists spend an average of 9.5 seconds longer in MR false positive decisions than true negative ones – a trend which is also 8.4 seconds longer than neurology false positives.

It is important to note that radiologists were much more confident in ruling out abnormalities in CT than MRI (CT=1.1, MR=1.3, t=2.23, *df* 38, $p<.032$) and this difference was reflected in the amount of time they spent viewing each case in CT compared with MRI also (CT=70.5, MR=83.7, t=2.32, *df* 38, $p<.026$). Conversely, neurologists were more confident in MRI (CT=1.8, MR=1.5, t=-2.50, *df* 41, $p<.017$), although this perceived confidence did not influence their visual search patterns. In true negative decisions overall, radiologists were more confident in both modalities than neurologists (CT=*df* 1, F=34.41, $p<.000$, Eta Squared .29, and, MR= *df* 1, F=5.79, $p<.018$, Eta Squared.06). To examine the differences between the two groups between the slices, the following figures display mean fixation time between each axial slice by decision type;

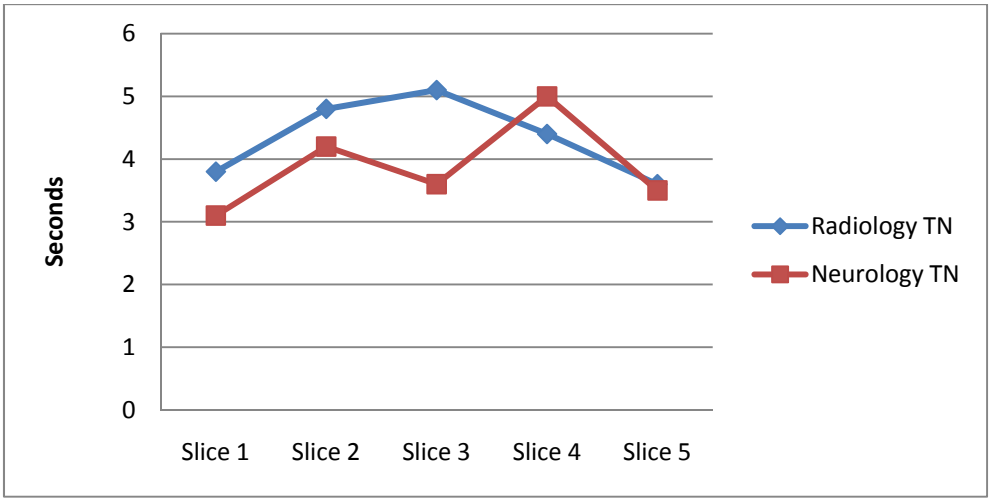


Figure 7.12 Mean fixation time per axial slice of expert readers who made true negative decisions in normal CT cases

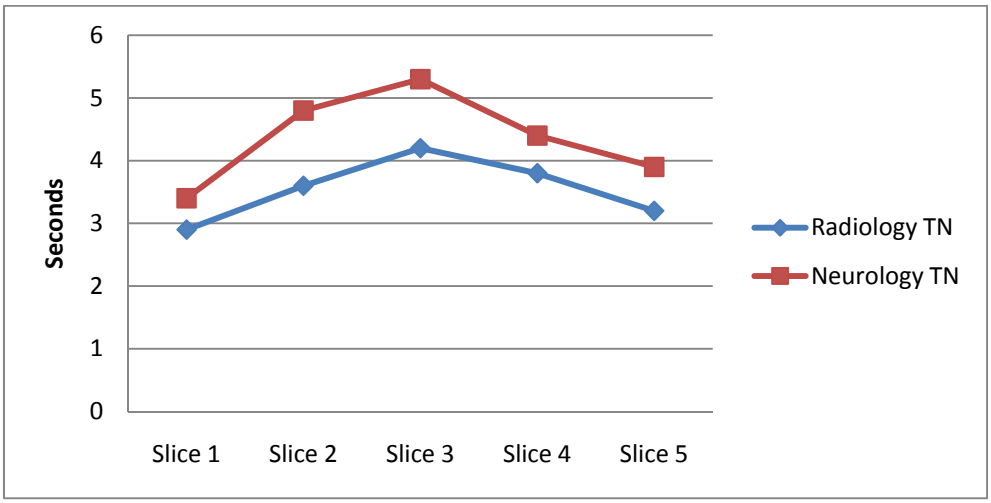


Figure 7.13 Mean fixation time per axial slice of expert readers who made true negative decisions in normal MR cases

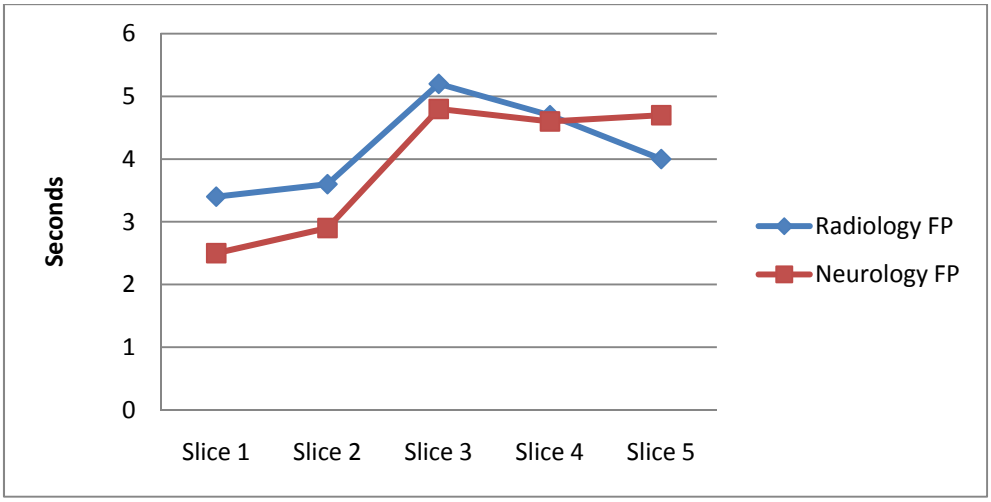


Figure 7.14 Mean fixation time per axial slice of expert readers who made false positive decisions in normal CT cases

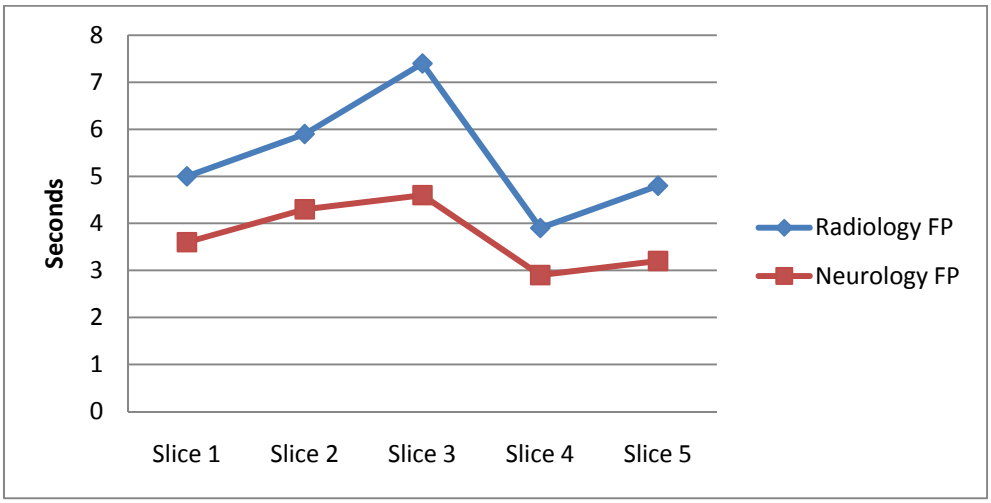


Figure 7.15 Mean fixation time per axial slice of expert readers who made false positive decisions in normal MR cases

The above figures 7.12-7.15 demonstrate that true negative decisions in MR and false positive decisions in CT and MR are marked by very similar patterns through the image 'stack'. The neurologists spend more time over true negative decisions in normal MR cases compared with the latter two figures for false positive decisions in both modalities. In the normal CT case figure for true negative decisions, whilst the radiology pattern of spending more time within the third slice is consistent between modalities, the neurologists spend much less time overall than the radiologists in every slice.

Summary;

Overall, once neurologists 'saw' the abnormal area, they spent twice the amount of time viewing it in CT than radiologists (*df* 1, *p*<.001, F=12.39, Eta Squared .01). This finding for neurologists was also a significant difference between modalities i.e. neurologists spent much longer in abnormal areas in the CT than the MR task (CT=3087.19, MR=1255.69, t=5.64, *df* =98, *p*<.000).

When comparing the total number of fixations in the first AOI appearing slice, there were significant differences between the neurology readers when viewing CT and MR images, with neurologists viewing the MR images more than CT images (CT=61.25, MR=71.40, t=-2.08, *df*=191, *p*<.039) but the same difference was not found for radiologists.

There were highly significant results when comparing the total time out of the area of interest for radiologists (CT=2969.39, MRI=1732.20, t=5.92, *df*=190, *p*<.000), with radiology readers spending longer in CT than MRI, but the opposing difference within neurology eye movements was only approaching significance (CT=2971.00, MRI=3744.45, t=-1.92 *df*=120, *p*<.057).

7.2.6 Secondary Abnormalities

Within this section all secondary abnormalities were compared between consultants i.e. SGR, CAJ and AMW from CT cases and ACG, ASC and CSS from MRI cases.

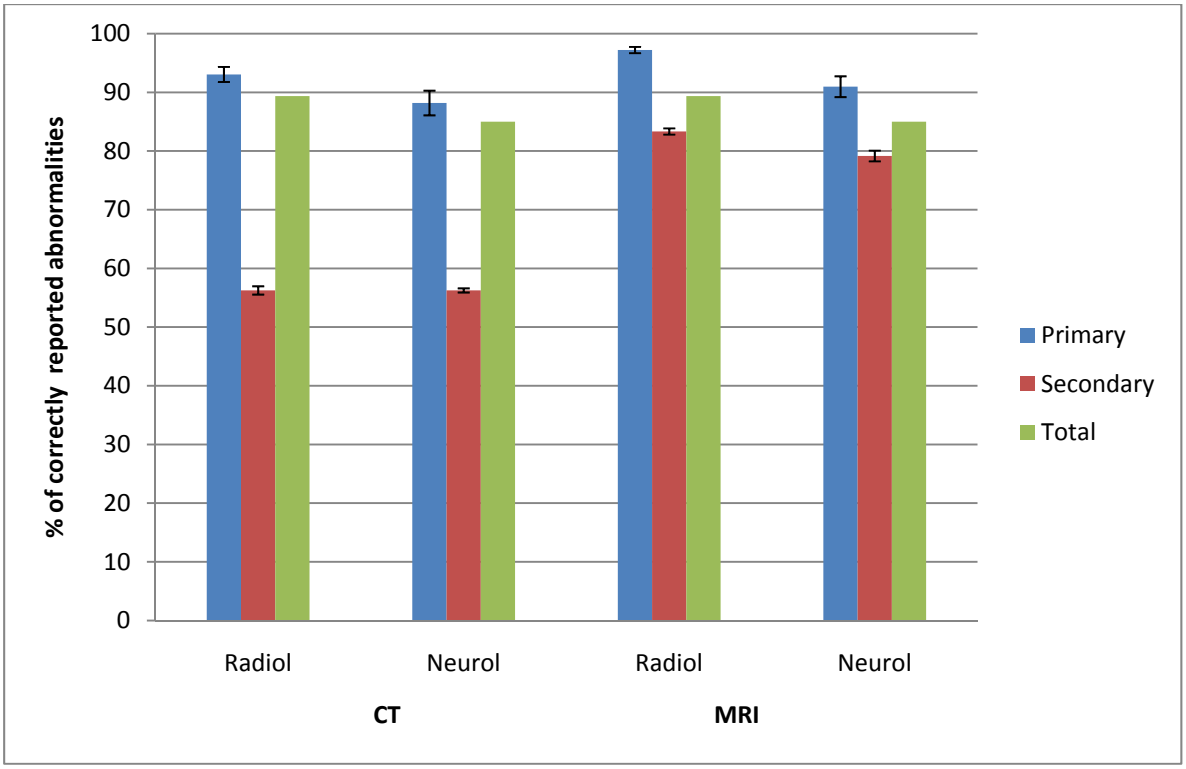


Figure 7.16 to represent percentage of primary and secondary locations correctly reported by participants.

Figure 7.16 demonstrates that radiologists and neurologists detect more primaries and secondaries in MRI compared with CT, with secondary sensitivity being substantially enhanced in MRI than CT. The error bars were also very small indicating both groups operated within a smaller range than novices and trainees in previous studies.

7.2.6.1 Secondary Abnormalities and Eye movements

7.2.6.1.1 True Positive Secondary Decisions

The following table highlights differences in visual search behaviours between reader groups when comparing the detection of secondary abnormalities. In true positive secondary decisions between modalities and consultant groups, radiologists were much quicker than neurologists to detect acute and chronic secondaries in MRI.

Table 7.7 to highlight eye-movement data by correct participants (in seconds) in the first appearing secondary AOI in CT and MRI.

	Patient Case	Reader group	Mean time to hit Secondary AOI	Mean time spent in Secondary AOI	Mean time out of AOI (in same slice)	Mean total fixation duration in first AOI slice overall
CT	Subacute case : SGR	Radiology	2.6	1	3.6	4.4
		Neurology	2.2	1.2	3.6	4.3
	Chronic case: CAJ	Radiology	0	0	6.1	6.1
		Neurology	0	0	8.1	8.1
MR	Acute case: ACG	Radiology	0.9	0.9	1.4	2.5
		Neurology	5.8	0.4	4.6	6.2
	Acute case: ASC	Radiology	0.9	0.6	3.3	4.2
		Neurology	3.9	1.1	4.6	6.2
	Chronic case: CSS	Radiology	0.6	1.4	1.0	4.1
		Neurology	1.8	0.9	2.5	4.3

However, as table 7.7 demonstrates, neurologists were quicker to reach the AOI in the subacute CT case. In the chronic secondary case, it appears no eye movements were captured within the secondary AOI despite the accurate reporting and much time being spent within the slice overall. All participants who reported a correct location for the secondaries, apart from the chronic case, spent up to 1.4 seconds appraising the abnormality.

7.2.6.1.2 False Negative Secondary Decisions

The following table 7.8 highlights differences in visual search behaviours between reader groups when comparing the ruling out of secondary abnormalities when an abnormality was present.

Table 7.8 to highlight eye-movement data by incorrect participants (in seconds) in the first appearing secondary AOI in CT and MRI

	Patient Case	Reader group	Mean time to hit Secondary AOI	Mean time spent in Secondary AOI	Mean time out of AOI's (in same slice)	Mean total fixation duration in first AOI slice overall
CT	Subacute case : SGR	Radiology	0.1	3.0	1.0	4.0
		Neurology	0	0	2.6	2.6
	Chronic case: CAJ	Radiology	0.0	0.0	6.1	6.1
		Neurology	0.6	1	4.0	6.4
MR	Acute case: ACG	Radiology	0	0	1.2	1.2
		Neurology	0	0	0.6	0.6
	Acute case: ASC	Radiology	-	-	-	-
		Neurology	0.5	0.2	1.4	1.8
	Chronic case: CSS	Radiology	4.2	0.1	0.3	0.5
		Neurology	-	-	-	-

Where abnormalities were ruled out, either the readers did not fixate within the AOI at all or they over or under appraised the cortical region i.e. they only spend an average of between 200 milliseconds and/ or circa three seconds. In the chronic case, judging by the eye movement results, neurologists do spot it quickly but can't have recognised compared with radiologists who (if they see it at all) spend a long time within the slice. Overall the patterns between true positive and false negative decisions appear rather similar throughout each of the studies.

7.2.6.1.3 Eye Movements and Focal Abnormality Accuracy

Within this study there was one focal abnormality present in acute CT case AMW. Radiologists were more sensitive to the location of focal abnormalities than neurologists with 87.5% of radiologists detecting the focal compared with 37.5% of neurologists.

7.2.6.1.4 Satisfaction of Search

67% of radiologists spotted the primary first compared with 58% of neurologists in CT. Of these, neurologists were 11% more likely to go onto look at the secondary, although radiologists were more likely to search for the primary if a secondary was seen first (67% compared with 14%). In MRI, radiologists were 21% more likely to spot the primary before the secondary, and were 27% more likely to search for the secondary, indicating radiologists were less likely to be satisfied with their search than neurologists.

7.2.6.1.5 Small Vessel Changes

When comparing small vessel change detection between consultants, radiologists were more likely to detect SVCs whereas neurologists were more likely to rule them out in both modalities. Sensitivity was 65% and 35% for radiologists and neurologists in CT, and 42% and 8% respectively for MRI. Specificity was 89% and 92% in CT and 88% and 93% in MRI.

7.3 Discussion

This study aimed to explore the overall performance and visual interpretations of computed tomography and magnetic resonance images of stroke when viewed by experienced readers from different specialties; radiology or neurology consultants. Whilst radiology readers were obviously, the better performers on both tasks, there were some interesting differences between the two consultant groups. It was interesting to uncover that both radiologists and neurologists were most confident in MR images and it could be that enhanced clarity of MR imaging makes neurology readers feel more confident about their decisions, even if they are ultimately, incorrect. This finding is similar to other inexperienced readers in chapters 4, 5, and 6.

In the Schriger study (1998), overall sensitivity (i.e. the ability to detect abnormalities) rates were reported to be 82% for radiologists, neurologists and emergency physicians with 52% of radiologists achieving 100% sensitivity. The Schriger study did not explicitly examine individual sensitivity rates but both radiology and neurology sensitivity rates in CT and MR were higher than 82% indicating that both groups were good at detecting the signs of infarctions within the cases. The enhanced sensitivity rates in the present study are likely to be attributable to the study population

emanating from the same acute care trust rather than a convenience sample of largely community neurologists and general radiologists. Specificity (i.e. the ability to rule out the presence of disease) in the present study was vastly enhanced for radiologists in the MR condition (20% increase in specificity), yet neurology specificity did not differ between modalities indicating only radiology specificity is facilitated by the enhanced detail of MR imaging. Neurologists on the other hand, do not appear to have enough training or experience for the enhanced detail to facilitate the ruling out of abnormalities and may even worsen performance.

To further corroborate this finding, radiologists found normal cases the most difficult in CT, yet neurologists found acute cases the trickiest in CT. This finding mirrors those of the sensitivity/ specificity findings; radiologists are keen to ensure they rule out the presence of abnormalities just as much as identifying them, whereas neurologists are more focussed on identifying features that are indicative of being abnormal. The easiest cases in CT were rated as subacute strokes by both groups, which is likely to be due to the large area of high density 'spread' characteristic of subacute infarctions. In MRI, radiologists still found normal cases the most challenging, yet neurologists found both normals and subacute cases the trickiest, further confirming that neurologists may not always recognise the abnormalities they are examining, especially when the detail is enhanced in MR. When neurologists are aware of the abnormal features such as acute stroke, the enhanced clarity in MR appears to enhance detection by 16%.

7.3.1 Eye Movements and Consultant Expertise

Visual search patterns throughout the 'stack' did not differ significantly between the reader groups but other image appraisal behaviours did, and were largely dependent upon the type of decisions being made regarding the case itself. When comparing total time viewing the images, radiologists were much more thorough when viewing CT than MR as indicated by the increased time spent examining these images and better sensitivity, specificity and confidence overall compared with the neurology reader group. Radiologists also spent longer outside of the abnormality than neurologists, which indicates an additional search for secondary abnormalities.

When neurologists spend significantly longer viewing CT than MR images, they are more likely to be incorrect, mainly because they appeared to glance over the abnormality itself for very little time without recognising it among the other features. Once neurologists recognise an abnormality, especially in CT, they took much longer over correct decisions than radiologists, which may indicate a degree of uncertainty and/ or a lack of experience when viewing images. This finding

is not surprising, after all, their main role is making decisions regarding patient care rather than frequently assessing the images themselves.

A surprising finding is that, neurologists performed much better than originally hypothesised. It was originally expected that neurologists would perform similarly to radiology trainees but as consultants gained experience; their performance became more similar to those of the early consultant radiologists. A follow-on question might be, if there is minimal difference between the two specialties within this area and if neurologists could perform as well as radiologists with additional training, why are the roles mutually exclusive? And what might this mean for radiology as a distinct discipline?

The answer to this would come in many forms; but primarily, although neurologists have access to head and neck images, they have chosen to specialise in making diagnostic decisions about the patient rather than the images themselves. Secondly, neurologists probably do not have time to complete the image assessments in their current jobs as well as their routine activities. Ultimately, the additional training on top of the already lengthy neurology training would mean that practitioners would spend more time training than practicing. Radiologists on the other hand have chosen to specialise in imaging and have chosen to specialise in head and neck imaging after doing the majority of clinical rotations, and have also chosen not to specialise in treating patients following their core medical training. Therefore, personal preference for career direction has led to these individuals to work in very different, but much aligned specialities. A follow-on study might consider the performance of neurosurgeons, neurologists and neuroradiologists, as neurosurgeons rely very heavily on feedback from images within their practice to deliver patient needs i.e. intra-operative neurosurgery (Seifert, 2003).

Although the differences between radiologist and neurologist were smaller than between consultant radiologist and trainee, this discussion is not promoting that neurologists can or should make decisions in the absence of a resident radiologist. It is, however, discussing the boundaries between the two specialities and questioning how much one specialty could competently achieve if the other were not available. For instance, if there were no radiologist on site, could a neurologist competently rule in the presence of acute infarction by assessing the images alone and/ or could a radiologist prescribe the correct drug treatment having understood the image findings, but being unsure when interpreting the patients' previous medical history? The present study cannot answer these questions and is not advocating either consultant group make decisions independent of the other and as such, is in agreement with the conclusions of Schriger (1998).

It is also difficult to ascertain just how much contact this group of neurologists had with medical images prior to this study; the neurologists here might be particularly well versed in image interpretation and might not only view the images themselves before receiving the reports from the radiologist, but might also discuss the findings with the radiologist in the reading room too. It is known that neurologists working in acute care settings order more imaging procedures than those in community or remote settings (Mitchell, 1996), and these neurologists might be particularly 'hands-on' compared with the neurologists who participated in the Schriger (1998) and Grotta (1999) studies, and those who work in environments where imaging is frequently outsourced due to few radiology resources. Unfortunately, this study only examined the performance of consultants within one centre, but future studies should examine neurology performance within and between multiple treatment sites to gain further insights into consultant performance differences.

As there is overlap between the radiology and neurology roles, some articles have urged the two distinct specialties to work together more frequently, even if they already work together cohesively in some departments, there is potential to further collaborate when training undergraduate and postgraduate medical students (Gunderman, 2003). Gunderman discussed the huge importance of the specialities working more closely together to ensure students get a 'real-world' understanding of disease manifestation by not only getting the information about an illness, but also seeing the patient images first hand. Gunderman states that radiology in the medical school curriculum can bring together fragmented teaching areas, such as anatomy, and students would benefit more from specialist groups, such as neurology and radiology, working more closely together rather than becoming defensive with "departments vying to defend their piece of an ever more hotly contested pie" (Gunderman, 2003, *pp*1239); namely budgets and resources. Additional benefits could be more students entering into both specialties as a result of high quality teaching and collaboration, which in turn would benefit consultant numbers, and maybe even reduce stress and burnout among co-workers. Even if students entered into neurology rather than radiology (if choosing between the two) following high quality training, they could be much more likely to order appropriate medical examinations in the first place, reducing time and money in both departments.

7.4 Conclusions/ Summary

In the present study, it is unsurprising that radiologists outperformed neurologists on every measure. However, neurologists still performed to a high standard, outperforming the majority of radiology trainees. Therefore, it was questioned whether neurologists, with further training in head and neck imaging could emulate the performance of radiologists. Whilst this question was out of the remit of

the present study, future work could tackle this possibility. Even though many neurologists performed well on this task, it does not mean that they should add image assessments to their 'to-do lists', and radiologists should not feel their roles could be over thrown; there are many reasons why neurologists performed well in this task and the neurologists within this particular trust could be more 'hands-on' and more experienced in reading images compared with other neurologists at other centres in different places within and/ or outside of the UK. In addition, as current opinion seems to suggest that the more specialists collaborate together, the stronger their respective specialties become, it may well be that the consultants at NNUH already work closely together and might suggest why performance appeared comparatively similar.

7.5 Study Reflections

As eluded in the discussion section, neurologists and radiologists could work together more closely in the area of medical school training. However, what about communication and cohesion between the two groups when in direct contact? For example, neurologists communicate with radiologists (and vice-versa) in many ways, but when image procedures are requested they are formally accompanied by a medical report regarding the patient's current status and previous clinical indications of stroke (in this area of study). When radiologists receive patient information, how is the information processed by the reader and does the content impact upon the way the image was perceived, if at all? The following chapter examines the clinical histories submitted by the neurologist to the radiologist, and whether the information communicated either enhances, worsens or has no affect at all on image appraisal and performance.

Study 5: The Effect of Clinical Information on Medical Image Appraisal and Accuracy

As previously discussed in chapter 1, communication regarding patients' examination and treatment occurs in a number of ways. Between neurologists and radiologists, a patient referral is accompanied by a report of the patients' current status and clinical history. As images are rarely viewed in isolation from other influences, whether prior information influences image assessment and decision-making process is an important question to consider, especially if this influence does indeed bias the reader in one way or another, leading them to over or under-report the presence of abnormalities in the scan.

Research studies across a number of reading tasks to date have had mixed findings. Whilst some studies report increasing true positive ratings following presentation of clinical information (Schrieber, 1963) and agree with this assumption, others report no significant increase in performance (Good *et al*., 1990). False positive decisions have also been said to either increase when information was present (Doubilet & Herman, 1981; Norman, Brooks, Coblentz & Babcock, 1992) or when information was withheld (Tudor *et al.*, 1997). Some authors have also alluded to reliance upon information when more than one image of the patient is present (Leslie *et al.*, 2000).

In studies pertaining solely to the detection of stroke, Mullins *et al.,* (2002)[a] uncovered a significant improvement in sensitivity for readers who were made aware of a likelihood of stroke in unenhanced CT cases. Where neurology and radiology consultant performance was compared in a focal abnormality detection task of two ambiguous CT scans, which were randomised amongst 7 other scans (up to 9 slices of the brain were accessible) neurologists were more likely to miss the focal abnormality than the radiologists, who were more sensitive, but that clinical information did not, in this instance, bias either reader group in either direction (Bonke *et al.*, 1989).

Aside from whether information influences decision-making, the text reading process itself is not an easy visual task (although most are able to read 'easily' once the learning process is over) being comprised of a number of 'jumpy' fixations (between 200 and 250 milliseconds), linked by a series of saccades made by the eyes, which either progress onto next word appraisal or return to re-examine previous words within the text (Rayner, 1998; Starr, 2001). Two select models consider reading to be a product of; a) how frequently a reader visits a word, with the reader being driven by

low-level oculomotor movements of the eye (oculomotor model) (McConkie, 1988), or b) the fixation duration and where the reader is driven by attentional processes that underpin the eye movements to decipher meaning (processing model) (O'Regan, 1992; Starr, 2001). Other theories surround the influence of word length, frequency, familiarity and the visual acuity of the parafovea in determining where and whether the proceeding word is fixated.

Whether word features or top-down/ bottom-up processes are responsible for understanding the process behind reading (i.e. whether eye-movements dictate meaning, or cognitive attention dictates eye movements, or both), eye-tracking remains a valuable research tool to interpret where our attention is guided to, or where cognition dictates what we read and when we read it (Rayner, 1998, Starr, 2001). Most current reading models agree that eye movements and attention are inseparable. Differing attentional processes might be dependent upon prior knowledge, expectations and understanding of what implications the information itself, without interpreting image findings, might have for the patient. With caution, limited attentional and cognitive inferences regarding fixations and word meaning can be made from eye tracking outputs within the context of experimental research.

Upon consideration of the aforementioned studies, the impact of clinical information and the process of reading within this context appears more complex than originally thought; information might indeed influence overall performance, might increase the number of true positives and/ or false positives, or it might not, depending upon the observers, the patient cases and/ or the study design itself.

Study aims and objectives: Owing to the mixed findings surrounding the effect of information on image appraisal, and very few papers reporting the effect of clinical information within CT and MRI, the present study aims to consider the overall impact of providing or withdrawing information regarding patient status on subsequent image assessment and decision-making. In addition, whilst oculomotor research literature exists regarding how we read, the present study will attempt to elucidate where observers attribute their attention in the reading process itself e.g. which clinical words attract the most attention and attempt to explain why. In summary, the research questions which underpin this chapter are as follows;

1. Does the presence or absence of information enhance or worsen performance? E.g. do true positive ratings increase or decrease following presentation of clinical information?

2. Do false positive ratings increase or decrease following presentation of clinical information?

3. Does the presence or absence of information enhance or decrease confidence?

4. If information does alter performance, which observer groups (i.e. novice, trainee and expert groups) does it affect? If any?

5. Is the effect of clinical information, if there is any, more likely in one modality more than the other? I.e. in CT or MRI?

6. Which clinical words were focussed upon the most, and by which observer groups?

7. Which categories of words were focussed upon the most and by which observer groups? i.e. patient age, sex (e.g. male or female), status upon arrival (e.g. 'was found lying on the kitchen floor'), presenting symptoms (e.g. persistent headache and facial weakness), anatomical regions (e.g. left and ride side), previous indications (e.g. known diabetic or previous coronary bypass graft), lifestyle factors (e.g. current or ex-smoker) and/ or reference to time (e.g. 'patient was found two-days ago').

8.1 Method

8.1.1 Design

The design of this study is identical to the experimental protocols predefined in chapters; 4, 5, 6 and 7: forty-eight predetermined clinical cases were selected from CT and MRI chapters and a computer-based, eye-tracking study was subsequently developed to assess diagnostic accuracy and interpretation in stroke CT and MR imagery. As per the previous studies, following eye movement calibration and participant instruction, observers were requested to either assess the clinical information prior to case examination, or proceed to case examination following a control image.

The present study differs from the previous studies as it only considers the aspect of the previous study protocols that concerns the clinical information, which was either provided or withheld, and the resultant decision i.e. this study examines the amount of clinical information seen and 'read' (from an eye movement and attention perspective), the information that was attended to most by each participant group (i.e. novices, trainees, expert radiologists and expert neurologists) as demonstrated by eye movement data, and finally, whether the outcome (i.e. the clinical decision) was affected by the presence of the clinical information in either a positive or negative direction i.e. does clinical information enhance or decrease true positive, true negative, false negative or false positive ratings, or indeed, whether its presence or absence has little effect on the radiology task at all?

As per previous study design, the independent variables of modality (CT, MRI), case severity (acute, subacute, chronic or normal aging control) and influence of clinical information (clinical information given or withheld) were assessed in a within and between participant design. The orders of presentation of the independent variables were counterbalanced within and between participants to control for order effects. A table providing an explanation of the conditions and the counter-balancing of order effects can be located in chapter 2.

8.1.2 Participants

For this study, all participant data were included from the following groups; novice, trainee, expert radiologists and expert neurologists. The same participants were included in this study as predefined in previous chapters. All participants were identified, recruited and briefed regarding the study aims as per prior studies, therefore, 36 participants contributed towards the experimental data for the present study.

8.1.3 Procedure

The procedure of this study is identical to those predefined in chapters; 4, 5, 6 and 7: Visual search behaviour of each participant was monitored using a Tobii X50 remote eye tracker, mounted below the computer monitor and participants' eye movements were calibrated on a 5-point scale. Prior to viewing patient axial slices, participants were either presented with the clinical history information regarding patient status on admission, presenting symptoms, patient age, sex and any existing conditions and/ or lifestyle factors that might facilitate accurate diagnosis.

Participants were only presented with clinical history information in 50% of cases. A control slide, with no information other than the patient number in the sequence, was inserted when information was withheld. As per previous studies, participants rated each case on a four-point Likert scale, namely whether a primary abnormality (i.e. stroke) was; 1) definitely present, 2) probably present, 3) probably absent, or 4) definitely absent. If an abnormality was considered present, participants were required to confirm the location of the infarct on a separate brain atlas task. Participants rated forty-eight predetermined CT and MRI clinical cases, as outlined in the previous chapters.

8.2 Results

8.2.1 Impact of clinical information on diagnostic accuracy

This section aims to uncover whether the presence or absence of clinical information affected the overall accuracy of detecting primary infarctions by group and modality. Please refer to chapter 2 for further details regarding how measures of accuracy and confidence between and within participants were derived.

Table 8.1 represents the percentage of correct location scores for primary abnormalities by group, modality and whether clinical information was withheld or made available.

Primary accuracy (% correct)	Modality	Novice	Trainee	Expert	Neuro	Average
Information withheld	CT	70.8	87.5	94.8	88.5	84.7
	MR	61.7	80.0	99.0	90.6	81.5
Information given	CT	72.5	86.7	87.5	89.6	83.6
	MR	63.3	84.2	93.8	89.6	81.7
Average Total	CT	71.7	87.1	91.1	89.1	84.1
	MR	62.5	82.1	96.4	90.1	81.6

From the above table, on average novice readers detected more abnormal areas per case in CT when information was present whereas trainee performance was enhanced in CT when information was not present (see table 8.1). Radiology and neurology reader performance was improved in MRI when information was withheld. For radiology readers, having the information present in the both modality assessments appeared to worsen their performance.

When the results in table 8.1 are compared with raw accuracy values in studies four (CT chapter), five (MRI chapter) and seven (radiology versus neurology chapter), the results highlight the same patterns within the data, although the cross comparisons highlight that novice performance decreases much more than the group average in CT when information is withheld (1.6% decrease when information is withheld, compared with 0.1% increase than the average when information is present). As already considered, radiology accuracy is enhanced in CT when information is withheld (4.6% increase) and reduced when information is given (2.7% decrease).

Of interest In MRI, trainee accuracy is increased compared with the group average in study 5 when information is present (2.8%), but falls 1.4% when it is withheld. For the experts in MRI, once again accuracy falls when information is present (2.8%) but increases when it is withheld for radiologists (2.4%), and accuracy decreases for the neurologists marginally when information is present (0.2%) and improves marginally when it is withheld (0.8%). Although the above trends were

observed in descriptive statistics, and the reduction in radiology performance in MRI when information was presented was of borderline significance with a small effect size (*df* 1, Sig.055, F=3.74, Eta squared=0.1), and as the differences were often marginal, there were no official statistical differences between or within groups in CT or MR in overall accuracy.

Table 8.2 represents the mean confidence scores for primary abnormalities by group, modality and whether clinical information was withheld or made available.

Mean confidence	Modality	Novice	Trainee	Expert	Neuro	Total
Information withheld	CT	2.83	3.14	3.21	3.10	3.06
	MR	2.83	3.16	3.25	3.23	3.10
Information given	CT	2.94	3.26	3.27	3.10	3.16
	MR	3.01	3.27	3.33	3.21	3.20
Total	CT	2.88	3.20	3.15	3.16	3.11
	MR	2.92	3.21	3.29	3.21	3.15

For mean confidence ratings by group, table 8.2 demonstrates that novices were most confident when information was present, when viewing MR images. When information was not present, there were no differences between modalities in novice performance. The same trend emerged for radiology consultant readers, yet they performed marginally better in the MRI study when information was withheld. Whilst trainees were also most confident when rating MR cases when clinical information was present, there was no difference between CT and MR ratings when information was withheld. Neurology consultant readers performed better in the CT condition when information was withheld yet their performance between modalities was consistent. For all participant groups, MR images with information present appeared to marginally enhance confidence ratings. Although the aforementioned trends were noted from descriptive statistics, there were no significant differences either within or between groups in CT or MRI and it cannot be ascertained whether the descriptive statistics were down to individual preference for a particular modality over the other, or whether the impact of clinical information truly influenced reader performance. As only marginal differences were uncovered between modalities and between groups, the impact of clinical information on each stroke type was not considered.

8.2.2 Clinical Information and Diagnostic Decision-making (i.e. TP/FP/TN/FN) between Modalities

The following tables highlight the spread of true and false positive ratings and true and false negative ratings per group and modality to gain a more in-depth insight into which decisions the presence or absence of clinical information might be influencing. The figures were derived by totalling all individual decisions for the primary infarction (or lack thereof) for each case, each group, and each modality dependent upon whether information was withheld or given. It is important to remember that there were 6 control and 12 abnormal cases per experiment overall and therefore the original case spread was not equal. This figure was also halved in the present study owing to half cases having information present and the other half not having information presented. In addition, JAFROC scores i.e. to weight correct true positive and true negative scores against false positive and false positive decisions and further plot overall performance in this study was not calculated owing to the lack of a rating score per decision but rather a rating for the case overall. For more information regarding this aspect of the design please refer to chapter 2 (section 2.4.3) and methodology revisions in section 9.7 of the final chapter.

Table 8.3 Represents the spread of total decisions by reader group when information was either given or withheld in CT.

CT	Information?	TP	TN	FP	FN
Novice **(n=10)**	Given	65	19	26	16
	Withheld	71	14	25	14
SpR **(n=10)**	Given	73	19	26	7
	Withheld	75	21	17	7
Radiol **(n=8)**	Given	70	18	2	6
	Withheld	68	16	12	0
Neurol **(n=8)**	Given	66	21	11	4
	Withheld	61	22	6	8

Table 8.3 demonstrates that, in CT, novice and trainee readers report the most true positives when information is withheld, whereas radiology and neurology readers report the most when information is given. True negative decisions are reported the most when novices and experts are given the information and the opposite for trainees and neurologists. Novices, trainees and neurologists report the most false positives when information is given. Whilst neurologists report the most false negatives when information is withheld, radiologists report the most when information is given.

Table 8.4 Represents the spread of decisions by reader group when information was either given or withheld in MRI.

MRI	Information?	TP	TN	FP	FN
Novice **(n=10)**	Given	61	16	27	18
	Withheld	57	18	24	23
SpR **(n=10)**	Given	75	26	24	4
	Withheld	72	24	28	8
Radiol **(n=8)**	Given	69	21	4	2
	Withheld	71	24	0	1
Neurol **(n=8)**	Given	59	16	11	2
	Withheld	65	21	9	6

In MRI, radiologists and neurologists report the most true positives and true negatives when information does not accompany the images (see table 8.4). However, novices, radiologists and neurologists also report the most false positives when information does accompany the image. The false positive ration is increased for novices, radiologists and neurologists when the information is present and false negative reports are increased for novices, trainees and neurologists when information is withheld.

8.2.3 Eye movements and Reading in CT

The following hot-spot image 8.1 provides an example of the clinical information provided and the eye movements laid over the image.

Figure 8.1 Demonstrates the clinical history provided in a chronic case with an observers eye movement recordings.

The forthcoming tables consider the ten words that participants fixated on for the longest average duration and were revisited the most by each group, from both the CT and MRI studies. As the current literature is divided regarding word fixation (oculomotor model) and fixation duration underpinning attentional processes (processing model), both fixation count and duration were explored in the current study.

Table 8.5 represents the top ten words fixated upon and their respective fixation count per reader group in CT.

	Novice		Trainee		Radiology		Neurology	
Ranking	Word	Fix'n count	Word	Fix'n count	Word	Fix'n count	Word	Fix'n count
1	Hemaniopia	24	Hypoalbuminia	45	Ex-smoker	18	Resolving	22
2	Maintained	22	Clexnae	29	Inflammatory	17	TACS	19
3	Hypoalbuminia	19	CCF	27	Maintained	15	Hypoalbuminia	19
4	Resolving	18	Lymphoma	25	Myocardial	14	Family	19
5	Inflammatory	16	Carotid	22	Airway	12	Currently	17
6	Please	16	Unresponsive	19	Spontaneously	12	Inflammatory	16
7	Myocardial	13	Hypertension	19	TACS	12	IV	16
8	11:30am	12	Fistula	19	Hypoalbuminia	12	Behavioural	16
9	Airway	12	Cerebellar	18	CVD	11	Az (Alzheimers)	15
10	Dysphasia. Ex-smoker. Clexnae.	12	Abducens	17	Decreased. Trauma.	11	CN	14

Table 8.5 demonstrates that all readers fixated frequently on the word 'hypoalbuminia', three of the four groups fixated frequently on the words 'hemaniopia' and 'inflammatory', two of the four groups focussed on the words 'ex-smoker', 'resolving', 'maintained', 'myocardial', 'TACS' and 'airway'. Only the neurologists focussed in on the words 'family', 'behavioural', 'IV', 'CVD' and 'Alzheimers'.

Table 8.6 represents the top ten words fixated upon for the longest duration and their respective fixation count per reader group in CT. (N.B. fixation durations are represented in milliseconds)

	Novice		Trainee		Radiology		Neurology	
Ranking	Word	Fix'n duration	Word	Fix'n duration	Word	Fix'n duration	Word	Fix'n duration
1	Hemaniopia	1085	Hypoalbuminia	2233	Ex-smoker	937	TACS	1842
2	Maintained	1056	Superficial	1742	Trauma	793	Currently	1240
3	Hypoalbuminia	898	Clexnae	1683	Maintained	745	Resolving	1001
4	Resolving	817	CCF	1423	Spontaneously	618	CN	929
5	Dysphasia	670	Lymphoma	1383	Myocardial	590	IV	901
6	Myocardial	654	Carotid	1222	Hypoalbuminia	590	Az (Alzheimers)	845
7	Clexnae	642	Fistula	1172	Lymphoma	562	Family	786
8	Unresponsive	622	Hypertension	987	Inflammatory	514	Inflammatory	741
9	Hemiparesis	606	CN	967	Smoke	506	Behavioural	729
10	Hypertension	608	Abducens	929	Decreased	499	Hypoalbuminia	726

Table 8.6 demonstrates the fixation durations of each group in CT, whilst most of the words are repeated from table 1 results, the words that are fixated for a long period of time but not necessarily re-fixated are 'unresponsive', 'hemiparesis' and 'hypertension' by novice, the acronym 'CN' and word 'superficial' by trainees, 'Lymphoma' and 'smoke' by radiologists and there were no differences between the number of times a word was visited and the fixation durations by neurology readers. Indicating a consistency between the numbers of times an interesting word was visited and the fixation durations that accompany it between groups in CT. The word 'hypoalbuminia' was fixated on the most by trainee readers who fixated on the word 45 times with an average fixation time of over two seconds.

8.2.2 Eye movements and Reading in MRI

The following tables also compare the top ten words by each reader group and fixation count and duration within the MRI study.

Table 8.7 represents the top ten words fixated upon and their respective fixation count per reader group in MRI.

	Novice		Trainee		Radiology		Neurology	
Ranking	Word	Fix'n count	Word	Fix'n count	Word	Fix'n count	Word	Fix'n count
1	Dysmetria	16	Continuously	21	Initially	16	Hemiparesis	28
2	Psoriasis	15	Provoked	20	Represented	13	Fluctuating	14
3	Paresthesia	14	Shoulder	19	Previously	13	Comprehension	13
4	Photophobia	13	Paresthesia	19	Mini-stroke	12	Hemaniopia	13
5	Pulsations	12	Valsalva	18	Discharged	10	Continuously	12
6	Nystagmus	12	Disturbance	16	Valsalva	10	Paresthesia	11
7	Cerebrovascular	11	Ischemic	14	Attack	8	Disturbance	11
8	Hemaniopia	11	Comprehension	13	Ischemic	8	Represented	10
9	Fungating	11	Hemiparesis	13	Fully	8	Burning	10
10	Early. Fundoscopy. Infarction. Including. Continuously.	10	Forwards	12	Morning	8	Became	10

Table 8.7 demonstrates that three of the four groups fixated frequently on the words 'paresthesia' and 'continuously'. Two of the groups focussed on the words 'valsalva', 'ischemic', 'comprehension', 'hemiparesis', 'presented' and 'disturbance'. Only novice readers focussed on the following words; 'dysmetria', 'psoriasis', 'photophobia', 'pulsations', 'nystagmus', 'cerebrovascular', 'fungating', 'early', 'infarction', and 'including'.

Table 8.8 represents the top ten words fixated upon for the longest duration and their respective fixation count per reader group in MRI. (N.B. fixation durations are represented in milliseconds)

	Novice		Trainee		Radiology		Neurology	
Ranking	Word	Fix'n duration	Word	Fix'n duration	Word	Fix'n duration	Word	Fix'n duration
1	Psoriasis	709	Valsalva	972	Initially	470	Hemiparesis	1734
2	Nystagmus	555	Continuously	896	Represented	470	Comprehensive	630
3	Cerebrovascular	550	Provoked	885	Attack	451	Fluctuating	610
4	Dysmetria	534	Paresthesia	774	Valsalva	418	Hemaniopia	518
5	Pulsations	530	Shoulder	718	TIA	403	Paresthesia	495
6	Hemaniopia	526	Disturbance	612	Previously	391	Valsalva	450
7	Fungating	520	Spread	586	Mini-stroke	375	Disturbance	424
8	Fundoscopy	502	Ischemic	582	Fully	335	Continuously	422
9	Photophobia	494	Fundoscopy	527	Ischemic	319	Drooping	399
10	Continuously	435	Reduction. While.	518	Discharged	311	Fundoscopy	383

Table 8.8 demonstrates the fixation durations of each group in MR, whilst most of the words are repeated from table 3 results, the words that are fixated for a long period of time but not necessarily re-fixated are 'spread', 'fundoscopy', 'reduction' and 'while' by trainee readers, 'TIA' by radiologists and 'valsalva', 'drooping' and 'fundoscopy' by 'neurologists'. The word 'hemiparesis' was the most fixated upon word by neurologists who fixated on the word 28 times, for an average of 1733.8 milliseconds in MRI. In all tables above (0.41-0.44), trainees had the most fixations and fixation durations followed by neurologists, novices and then experts, respectively.

8.3 Discussion

8.3.1 Influence of Clinical Information on Accuracy and Confidence

The presence of information within this study had mixed results. Primarily, there was no significant increase or decrease in performance with the presence or absence of clinical information. Secondly, when information was present, novice, trainee and radiology confidence scores were enhanced,

although once again, this finding did not reach statistical significance. As confidence was enhanced but accuracy was not, it seems that having the additional information present may provide a false sense of security, which does not necessarily improve performance and might actually worsen it. Whilst clinical information marginally enhanced novice performance in descriptive statistical analyses in both modalities, it only enhanced trainee performance in the MR condition and reduced performance in the CT condition. Providing clinical information worsened the performance of radiologists in both modalities and neurology scores were indifferent to modality and presence or absence of information.

Schrieber (1963) stated that the presentation of clinical information increased the number of true positive ratings. For the more inexperienced readers (i.e. novices and trainees), information present in enhanced true positive ratings and decreased the number of true positives in CT. The opposite was true for more experienced readers; consultants made more true positive ratings in CT than MRI when the information was present. This trend appears to indicate that experienced readers are more reliant upon the information when clarity of image features is reduced, whereas inexperienced readers rely on information when clarity is enhanced and there are many more features to consider and rule out. This observed trend is in agreement with Loy and Irwig (2004) who suggested that some readers are more reliant on clinical information than others.

As overall accuracy was improved but was not statistically significant, the present study results are in agreement with Good *et al.,* (1990) and Tudor *et al.* (1997). Therefore, the present study is not in agreement with those of Doubilet (1981) who found a significant increase in performance of four radiologists when information was present and Norman, Brooks, Coblentz and Babcock (1992) who stated that both experts and novices were influenced by clinical information.

In reference to the literature surrounding false positive ratings, the presence of clinical information did appear to increase the number of false positives made by novices, trainees and neurologists in both the CT and MRI studies, which was in agreement with the study conducted by Tudor (1997). McNeil (1983) also stated that whilst the number of false positives increased with the presence of clinical information, false negatives decreased; whilst this appeared true for all participants reading MR images, it was only marginally true for neurologists reading CT scans. Overall, the results in the present study seem in agreement with Bonke *et al.* (1989), as clinical information did not statistically bias either reader group in either direction.

Leslie *et al.,* (2000) alluded to a reliance on clinical information within multidimensional imaging rather than 2D imaging; this assumption was not supported in the present study as there

were no significant differences between information present and withheld on resultant performance. However, there was a trend towards a difference dependent upon image modality being viewed. In addition, if more image sequences and image plains were made available, e.g. sagittal, axial, coronal, perhaps information might have been more influential and considered more important to anchor clinical decision-making. Future information studies should consider how information impacts upon performance within and between multi-slice, multi-plain and multi-modal studies of observer performance.

8.3.2 Reading, Eye Movements and Clinical Information

Whilst some word repeat fixations could be accounted for by either being completely unknown to novice readers or relatively uncommon for many of the medical readers (i.e. 'Hypoalbuminia', 'hemaniopia', 'CCF' (Carotid Cavernous Fistula), 'valsalva', 'abducens' and 'clexnae'), those who are aware of their respective meaning could be paying attention to these words for one of two reasons; 1) they are infrequent words and/ or 2) they carry important clinical meaning for the patient.

For novice readers, the main reason for attending to these words is more likely to be the 'novel' word explanation, especially considering novices also fixated on additional words that are not highly associated with stroke incidence e.g. psoriasis, nystagmus, fungating. In addition, whilst the novice words differed considerably between image modalities, the clinical reader groups remained consistent between studies, suggesting these words are of particular importance to both the reader group and clinical decision-making in stroke. Overall, this result appears in agreement with research by Rayner *et al.,* (1996), that readers are more likely to refixate on low-frequency words than on high-frequency ones. This finding appears to also be in-line with the processing model as promoted by Starr (2001). At the outset, it was anticipated that novice readers would be more interested in 'patient-centred' factors (e.g. 'patient was found lying on the floor') rather than medical terminology or clinical results (e.g. Glasgow Coma Scale scores); however, the most frequent fixated words did not appear to confirm this hypothesis as novice readers paid more attention to more complex and novel words in general, rather than those that pertained to patient status.

Patterns in the most frequently viewed words (and for the longest durations) highlight that trainees appear to focus in on novel acronyms such as 'CCF' and their adjacent explanation e.g. 'CCF' (Cavernous Carotid Fistula) within the text. Trainees also paid attention to existing disease or contributory factors and recent changes to patient status e.g. 'Lymphoma', 'spread', 'superficial', 'provoked', 'disturbance', as well as other novel or complex words considered by other readers groups such as 'clexnae', 'hypoalbuminia', and 'valsalva'. Whether trainees focus on these words

owing to their infrequency or because they do, or do not understand the words' semantic meaning, cannot be established. It is more likely, however, that the more experienced readers understand the meaning behind some of the infrequent words.

In the results summary, radiologists not only appear to focus in on stroke indicators and words that allude to a stroke diagnosis more than other readers (e.g. 'mini-stroke', 'myocardial', 'ischemic', 'attack', 'CVD' & 'TIA'), but also focus much attention on contra-indications away from an obvious possibility of stroke (e.g. 'trauma' and 'lymphoma'). Neurologists also appear to pay much attention to indications and contra-indications, but they also focus on associated neurological conditions, such as Alzheimers Disease, much more than other reader groups. Of all the groups, neurologists focussed on the words 'family' and 'discharge' more than other readers; perhaps neurologists consider not only the possibility of confirming or refuting a queried diagnosis of stroke but also consider discharge potential and family considerations, which is considered a routine part of their role as care manager.

The only references to time, which were of high interest to readers, were '11:30 am'; attended to mostly by novice readers and the word 'morning' for radiologists. As many cardiovascular events more frequently take place in the morning due to alterations in blood pressure (Willich *et al*., 1992; Stergiou *et al*., 2002), this reference could be an important clinical indicator for radiologists. Words that were considered more closely were those that alluded to a change in patient status over time i.e. 'previously', 'continuously', 'currently', 'spontaneously', 'maintained', 'decreasing' and/ or 'resolving' and were frequently attended to by all readers, irrespective of experience. Incidentally, these are common words but much attention was still paid to them and therefore context is paramount. All readers paid attention to the lifestyle indicators; particularly 'ex-smoker' but radiology readers paid the most attention to these incidences within the clinical histories. It is well documented that smoking is the most common contributory factor for cardiovascular disease, including stroke, and could be a reason why novices also paid much attention to this word.

It was anticipated that very high interest would be paid to anatomical regions in the body, which would have further implications for image interpretation i.e. 'left-sided weakness' would indicate obstructed blood flow within the opposing hemisphere, however, this results was not found within the ten selected words by novice, radiology or neurology readers. Trainee readers paid more attention to affected anatomical regions within the clinical history such as; 'shoulder', 'paraesthesis', 'hemiparesis', and 'abducens', than other reader groups and might be an indication of problem-solving through anatomical features, a relative lack of experience or a desire to consider every aspect

of information about the patient than other readers. This hypothesis sits with other research stating that trainees generate more clinical hypotheses than other readers, but also do not refute hypotheses as quickly as their more experienced peers (Garlatti and Sharples, 1998).

It is well known within the reading research literature that word length and frequency factors influence reading and the associated fixation frequencies and durations. Frequent words were indeed skimmed over in the present study, unless they were references to a change in patient status, whereas less frequent, and thus more novel, words are more likely to be examined (Starr, 2001). In the clinical context the infrequent words are often those that describe unique clinical indications pertaining to each patient. Whilst it is expected that infrequent and/ or longer words are fixated for longer, these words also carry differing clinical consequences for the patient and comparisons were drawn between each group with differing results dependent upon experience and specialty.

8.3.3 Limitations and modifications

In the present study, the clinical information which accompanied each case was sourced directly from patient records. To retrieve the clinical information in this manner was considered the most ecologically valid approach to apply to this series of image assessments, despite being under laboratory testing conditions regardless. However, the accuracy of the clinical information itself was not divided into 'poor', 'limited', 'ambiguous' or 'thorough' categories and at present, no standardisation procedures appear to be in place regarding what constitutes best practice communication. To some readers, the quality of clinical information provided might be considered incomplete or misleading. In an ECG study by Hatala, Norman & Brooks (1999), when accurate information preceded image inspection, the true positive ratio was enhanced but when misleading information was presented, accuracy was reduced by up to 25% for cardiology trainees, 19% for medical students and 5% for cardiologists compared with the control condition. Even when information from the referring clinician was of high quality, its presence appeared to influence all levels of experience, although as expected, trainee observers more than the experts. Future studies should consider how best to maintain a consistent quality of information between patient cases to ensure readers receive high, low or average quality, and even misleading information, dependent upon study design and research questions.

The results in this chapter suggest that there is no influence of clinical information on image interpretation in the groups studied and thus, information did not bias image interpretation. Although the number of observers in this study supersedes the majority of studies previously reported within the area, there is a possibility that this work is still underpowered for this type of

enquiry. It is recommended that future work should modify the number of cases, image sequences and/ or participants to establish whether information, genuinely do not affect image assessment and whether these results were limited to this study design and group of observers.

In terms of the reading research within this study, it is recognised that this is an incredibly small insight into reading within the clinical context from only a small sample of patient information. Reading research has been conducted for many decades and the present study aimed to offer a limited insight into reading within the medical imaging realm. Future studies investigating the reading of clinical information should increase the number of patient cases and it may be advisable to use a head-mounted system to enhance eye movement recordings further. Further studies could also pin-point where the reader examines the word itself i.e. at the start or end of the word, which could be compared with existing research dedicated to the investigation of reading itself.

8.3.4 Implications and Recommendations

In future, rather than cross-comparing separate measures of performance such as confidence ratings, lesion detection percentages and the spread of true positive versus false positive ratings per group, it might be more efficient and consistent to cross-compare ROC curves and AUC values between studies. This approach would facilitate observer performance comparisons irrespective of anatomical region (e.g. chest, breast or head), disease presentation (e.g. nodules, metastases or infarct), modality choice (X-ray, CT or MRI) and single slice versus 3D imaging. As previously discussed, future studies should not only continue to investigate the impact of clinical information across multiple disease manifestations, modalities and multi-slice studies but should also assess the accuracy of the clinical information itself, which may require a new set of guidelines to ascertain what constitutes good patient information in the first instance.

A final recommendation to overcome the potential bias of clinical information, whether good, bad or indifferent, was postulated by Thorne Griscom (2002), who suggested that whatever the situation; there are many advantages of always examining the images prior to the clinical histories and in clinical practice, radiologists may already do this. Although this approach requires discipline when faced with time pressures and a degree of imagination to consider potentially rare conditions or diseases, the flip-side of this approach was posited to be increased mental acuity, fewer false-positive decisions and a way of encouraged the reader to scan all of the image without being influenced by prior hypotheses, by the reader or other clinicians. However, whilst Loy and Irwig (2004) agreed with this recommendation, they also suggested that some perceptual advantages of

reading the clinical information first, might be lost if Thorne-Girscom's advice was followed, however the exact nature of these perceptual advantages could not be explained.

8.4 Conclusions

This chapter set out to examine two main research questions; whether clinical information as a whole, has any impact upon diagnostic accuracy and which clinical words or categories of words are attended to most by different observer groups. At present, although there is research surrounding the impact of information on performance, the results to-date have been mixed and although much research exists regarding the process of reading in many contexts, studies could not be found relating to the reading of clinical information itself, and therefore the present study was aimed at offering a preliminary insight into reading in the clinical realm.

Whilst there were differing trends between groups regarding both these lines of enquiry, no statistically significant differences could be drawn. Although, the descriptive results in this study indicate observed trends regarding the impact of clinical information, they are indeed very small. This finding may be particular to the present study or it may be indicative of this area of enquiry. Further research should continue to investigate this area, including the reading process itself, with larger samples of readers and cases to uncover whether findings within this study remain consistent.

8.5 Study Reflections

Whether clinical information does or does not bias image interpretation in the clinical setting does have important implications for clinical practice and/or service delivery. It is important to know whether there is robust evidence to suggest that clinical information should be read prior or post image assessment, to ascertain which approach might enhance or decrease accuracy, if at all. Results of this kind might even transcend to teleradiology practice i.e. for those reading images at a distance, is knowledge of patient status and clinical histories required or superfluous? If patient information is found to be unnecessary, what implications might this have for radiology as part of multidisciplinary team practice and/ or could all radiology interpretation be outsourced? Would the removal of clinical information reduce affiliated costs and expert time, if proven to be unnecessary? Unfortunately the current study cannot make any inferences beyond the scope of the present results, but future studies could be designed to consider these questions and the important implications these results might have for clinical practice.

Key Findings, Recommendations and Implications

9.1 The Gold-standard in stroke Detection and Error Prevalence

When linking back to the original aims and objectives, the present series of studies uncovered that radiologists do misinterpret CT and MR image findings in stroke presentation, however, the most disagreements occur in subtle findings such as small vessel disease and focal abnormality detection. As stroke detection experience increased, fewer errors were made, with the most experienced consultants disagreeing on very few features. Of the radiologists who did make errors, stroke detection was not their primary role within the radiology department although they all had a special interest in neuroradiology. Of these errors, most were a result of feature recognition, predominantly normal cortical fissures or sulci were misinterpreted as stroke.

The characteristics of expert readers, includes a quick time to 'hit' the abnormality, a quick recognition time and thorough but efficient appraisal of the surrounding cortical tissue. As experience increased, fewer fixations were needed to make a confident decision. Experts who demonstrated a 'gold-standard' of visual search also spent additional time, particularly in CT, searching for secondary abnormalities. Viewing time was prioritised around the middle slice, which contained the most additional subtle findings and indications of cerebrovascular disease such as small vessel changes, which tend to cluster around the corners of the ventricles. Whether experts already knew additional abnormalities normally cluster around these areas or whether they recognised the abnormalities within the scan, cannot be determined as a result of visual search assessment within the present study. Although if experts image interpretations are heavily influenced by prior experience and anatomical regions, this findings does agree with those of Garlatti and Sharples, who hypothesised that experts have a mental 'schema' for disease location and appearance (1998).

9.2 Novice, trainees and expert performance and visual search

In terms of how performance and visual search differs along the expertise acquisition continuum, novices and trainees are obviously more likely to make more mistakes than experts and a number of large differences existed between the groups in terms of visual search; trainees spent much longer viewing the images overall in both modalities than novices and experts. In general, trainees and

novices took longer to reach the area of interest and also spent longer appraising the abnormal area, the image slice when it first appears and the case overall, irrespective of modality. Errors were characteristic of both search and decision errors, as they appeared more concerned with opting in suspicious regions rather than risk missing something important.

Trainee readers frequently cross compared hemispheres, indicating training had taken place compared with novices who either over, or under appraised the images. Novices were also a less 'homogenous' group overall, both professionally and in terms of visual search behaviour. Novices had to construct their own interpretations of what was normal from abnormal in a very short time frame, whereas trainees exhibited uncertainty owing to a lack of caseload experience. The additional clarity in MR over CT also appeared to increase the number of features clearly seen, but also require systematic elimination, which increases trainee viewing and decision-making times considerably.

9.3 Visual Search and Decision-making between Modalities

Visual search behaviour differed substantially dependent upon the type of decision being made. As previously discussed, true positive and true negative decisions were marked by quite clear trends within and between groups throughout the image stack and false negative and false positive decisions were marked with inconsistencies within and between groups, even if it was hard to tease out the exact location of the false positive decisions retrospectively. Overall there was much more variability between the range of scores and search measures when lack of experience was evident, experts were indeed quicker, more accurate and more consistent as a group, even if some radiologists read more CT/MR scans of the brain than others.

When true positive decisions were made readers quickly saw the abnormality and examined the features for approximately 1.4 seconds in CT and 1.3 seconds in MRI, with recognition times increasing as experience decreased. True negative decisions and visual search patterns were also consistent between readers and modalities, spending much less time over a true negative than false positive decisions. The eye movements of readers when making false positives decisions were 'bipolar' in nature; either readers seldom dwelt upon suspicious areas for between 100 and 200 milliseconds, or they revisited the area a number of times, which amount to between 1.5 and 3 seconds duration. This finding is not consistent with the 1000 millisecond rule applied by Manning and this is most likely to be a result of more than one image being available to the observer. False negative decisions in both modalities suggest the AOI was simply missed or gazed over and not recognised as these decisions were accompanied by 1 second of gaze duration or less. It appears

therefore, that the optimum time threshold for making a true positive or true negative decision is more than 200 milliseconds but less than 1.5 seconds in CT and 1.3 seconds in MRI; the over 200 milliseconds rule appears to hold true to both modalities even if recognition is quickened up after this minimum gaze duration.

Visual search also differed considerably between stroke types and was also linked to stroke size and degree of 'spread'. Ruling out an abnormality and detecting small abnormalities were the most challenging tasks for all readers. More inexperienced readers struggled with acute cases in CT and subacute strokes in MR imaging as the boundaries were less defined compared with the 'block' of high signal in CT. Therefore, whilst enhanced clarity is preferred for subtle changes, enhanced clarity is not necessarily conducive to enhanced performance in the less experienced readers, it seems a more experienced eye is required to filter out what is normal from abnormal, reducing false positive decisions. When the reader is experienced, MRI also quickens reading time.

As discussed in chapter 7 (CT versus MRI) previous studies into modality preference have not rated one modality over the other, yet in the present series of studies experts were quicker and more accurate in MR than CT. Whilst MRI was preferable for acute case detection (mean time 1.1 second) and chronic case detection (mean time 1.1 second), subacute detection was quicker in CT (mean time 0.9 seconds). Therefore, MRI appears the preferred modality for early stroke detection, despite CT being the modality of choice due to accessibility and financial resources. If performance of early parenchymal changes is enhanced and quickened in MR, perhaps MRI should be the emergency imaging priority. Early imaging with MR might reduce the overall cost of repeat imaging and consultant reading time whereas CT could be used to image the infarct if the damage 'spreads'. As many papers favourably reviewed the accuracy of DW imaging, a future study comparing eye-movements between conventional MR and DWI is recommended.

In terms of secondary infarct detection, experts spent much more time searching for and appraising secondary abnormalities than novices and trainees; however 40% of secondary abnormalities were not reported by consultants, indicating consultants are also not immune from making errors (Fitzgerald, 2001). It is likely that experts were already more aware of the presence of incidental findings such as small vessel changes in neuroradiological reading tasks than the secondary search for chest nodules or breast calcifications or metastases for example, but abnormalities were still missed. Inexperienced readers were more likely to incorrectly rule out the presence of additional abnormalities, which is expected due to a lack of knowledge and experience within this area. Of those readers who did correctly identify these abnormalities, the region of

interest was gazed at between 400 and 1.4 seconds overall, which is slightly longer than primary infarcts (0.2-1.3 seconds).

Satisfaction of search was also evident for readers, contributing to overall errors between groups and modalities. A higher proportion of primary abnormalities were not seen when the secondary was looked at first by novices in CT and MRI, particularly if the secondary was larger i.e. subacute stroke type. Trainees were more influenced by satisfaction of search when the primary was seen first in CT but not in MRI and of the radiologists who saw the primary in CT, half neglected to see the secondary. In MRI, experts those who saw the primary, were more likely to see the secondary than if they had seen the secondary first. Indicating, enhanced clarity facilitates secondary searches in the experienced readers, but not the inexperienced readers.

9.4 Comparing Consultants

When level of experience was controlled for in the neurology versus radiology study, it was surprising how well neurologists performed, although there were still differences between the groups. For instance, whilst both groups were most confident in MR, neurologists made more errors in CT than MR cases, indicating the enhanced clarity provides a false sense of security in less experienced readers and whilst neurology specificity remained the same between modalities, radiology specificity was 20% better in MRI. Radiologists spent much longer appraising the surrounding tissue compared with neurologists, indicating once again a search for secondaries. Overall, neurology visual search behaviours appeared similar to those of experienced trainees. Satisfaction of search was less likely for radiologists and neurologists in MRI, even though the abnormalities were looked at for longer in CT.

Whether this group of neurologists might be particularly well-versed in image interpretation is unknown. Future studies might do better to match pairs on level of experience and select readers from a number of centres. In addition, some of the radiologists in the present study only read a few hundred CT/MR head scans a year and thus might actually under represent the performance of neuroradiologists as a whole. As both groups of readers performed well in the present series of studies, future studies might examine the detection of only subtle abnormalities within a series of normal cases, which might emulate the breast screening model.

9.5 The Influence of Clinical Information

Overall, the presence of clinical information did not bias readers in one way or another. Although no significant differences emerged there were trends between readers and modalities; experienced readers appeared more reliant upon the information when the clarity of the image was decreased, whereas inexperienced readers rely on information more when clarity was enhanced. The most interesting, and borderline significance, finding was that radiology performance decreased when information was given in both modalities and during testing some reported that the information was 'distracting'. The present study did not consider the quality of the information itself and although not explicitly considered here, future work could also consider the quality of the radiology report to ensure neurologists understand and can make the best use of the image findings for patient intervention (Plumb *et al*., 2009).

Although the overall presence of clinical information did not significantly bias one reader group in either direction in both reading tasks, it was interesting to examine which words were paid most attention to and how this differed between groups. This type of study has not been explored before and demonstrates clearly that radiologists and neurologists focus in on clear stroke indicators and contra-indicators as well as any change in patient status. Neurologists also focus in on words that describe care planning. Trainees paid particular attention to words that were stroke indicators, although they may equally have been interpreting their meaning. Novices focussed on complex or novel words that sometimes were relevant and sometimes not, to overall stroke likelihood.

9.6 Overall Thesis Limitations

The main limitation of eye movement technology is the degree of drift, which sometimes infiltrates the reliability of the system and cannot be controlled once the experiment has commenced. In certain instances, it is clear from the output data that drift has occurred and the raw data should be eliminated from the study. In some instances, the degree of drift is very subtle and unbeknown to the researcher and does not adequately reflect the search patterns of the observer. As eye movement technology advances, it is hoped the degree of drift will decline.

The second limitation of eye movement technology, which may have impacted upon this series of studies, considers the region of parafoveal vision. The degree of visual acuity of the parafovea is estimated to extend up to 5 degrees on either side of a foveal fixation, yet it is unknown how this 'global' or peripheral region impacts upon what is interpreted within the medical image and cannot be monitored using eye movement methodologies. It is unknown how much of what can be

seen in peripheral vision influences where the observer examines subsequently. For instance, an infarct might have been seen but not gazed at directly, influencing the reliability of the system output.

In reading research, it is considered likely that the parafovea examines the upcoming words i.e. those to the right of the fovea, in parallel to the current word being interpreted and only when this process is interrupted by an abstract word, which does not 'fit' does the reader fixate for a longer duration on the phrase (Starr, 2001; Yang & McConkie, 2001). The same might be true of the visual scanning process – the parafovea draws in the increased attention from the fovea of features that do not 'fit' in with the observers normal schema of the brain and thus, it is likely that many observers captured the infarct within this region of vision prior to direct fixation. This global vision phenomenon does impact upon how features are perceived and the fixation data which is experimentally captured and used as evidence to back up empirical hypotheses. It may be that some observers either did not fixate directly upon the infarct, as it may have been large and unambiguous, or they could extract enough information from skirting around the boundaries of the infarct, as seen by the consultant readers in study 1. Therefore, it is with caution that all eye movement data is used to answer important clinical questions, used as evidence to redesign future training programmes, or even as an external measure of validity or performance.

When considering the abnormality reporting, infrequently a few general reporting errors were made by participants i.e. circling the wrong hemisphere on the reporting sheet or highlighting an area which bore no relevance to clinical features or their eye movement data results. These perceived errors could be down to a number of factors; a) the location was recalled retrospectively rather than during the case analysis, b) consultant time pressure resulted in a desire to finish the task and continue on with their working day, or c) a limitation of the study reporting design i.e. having to report a 3D feature on a 2D reporting sheet. If the latter hypothesis is true, the reporting could have been more clearly defined if participants were able to mark the area within the image itself. Unfortunately due to design constraints this was not achievable within this CT or the following MRI study at the time of completion. Future reporting of a multislice image could include axial, sagittal and coronal representations which allow the observer to stipulate where the abnormality appeared and ended within the 'stack' i.e. by stating the slice number where it appeared and the slice number where it ended would allow pinpointing of the entire 'mass'.

Other design limitations included the inability of readers to scroll up and down image stacks, and use pan and zoom functions at will, which is possible within normal clinical practice. For experimental design reasons this could not be permitted here, however, a contribution to knowledge

has been granted by allowing exploration of multiple slices of the same patient in terms of time in each slice dependent upon stroke type, modality and reader type. Future studies should allow reporting on the image, multi-slice evaluation with an increased number of images and free-scrolling between the images, up and down the image 'stack', if possible.

In addition to image and software limitations, bias may have been inherent at the reader level; reader motivation may have been high, low or average dependent upon either a desire to do well in the exercise, or a desire to complete the exercise and continue on with their working day. If reader motivation is high, they may perform better than their average working day or reader motivation may have been low as readers were aware that patient care was not being affected by their performance within the task. Therefore, the results of the aforementioned studies are confined to the population studied on the particular day they were tested. Performance can also vary within participants and their performance on this task may be better or worse when tested on another day and/ or another time of day, for example. Therefore, care must be taken to generalise the results from one population based at a single test centre, to the entire radiology and neurology population. However, much care was taken to ensure independent variables were counterbalanced within the study, experimental cases were matched (as far as possible), and the same readers rated all sets of cases from both modalities to minimise confounding variables within the study parameters.

9.7 Methodology Revisions – ROC Analysis

ROC analysis was applied to all primary confidence ratings per case. Despite ROC being considered a robust and reliable method of analysis in medical imaging, there were limitations of this analytical approach in the current series of studies. Primarily, location information for all reader markings, whether true or false, was available for inclusion in the analysis, and therefore, a wealth of information was excluded from the ROC framework. However, whilst each lesion localisation was available, each reader 'mark' did not have an independent confidence rating assigned to it. A confidence rating was only assigned to the primary abnormality per case, which at the time of design, was considered the most important. The consequences of only rating the 'case' and not each lesion per image meant that free-response methods such as FROC, AFROC and JAFROC could not be reliably performed upon the data (Chakraborty, 2006). In addition, despite having the location and rating data for the primary abnormalities, LROC parameters dictate that the clinical image should either have one abnormality per image or none at all (Metz, 2006); these experimental cases frequently had secondary abnormalities, small vessel changes or focal abnormalities present within the image being studied. This early design decision meant that a wealth of location information could

not be included into the ROC paradigm and had to be considered separately within the percentage accuracy data and category spread of TN/TP/FN/FP decisions per observer group.

In summary, free-response studies allow the scoring of more than one lesion per image and also take an overall measure of performance by considering the number of true positive marks off-set by the number of incorrect marks, which ensures performance does not appear over-inflated in Binormal ROC curves. For instance a novice reader might appear to have 90% accuracy as they have only 'missed' 10% of cases; however, within the images that the reader scored correctly, there might also have been 6 other areas which were marked as suspicious, meaning that their accuracy percentage did not consider the considerable number of guesses within the same case. For the experienced reader, free-response studies offer a good approximation to the clinical task, which enhances the ecological validity of the experimental condition and study. As a cautionary note, location data can provide greater statistical power than conventional ROC but the result depends entirely upon the amount of location error that is permitted by the research team. In the present series of studies, the degree of location 'tolerance' was discussed at the outset within the scoring criteria.

In future visual search studies, the following amendments would be made to the overall study design;

- A more discrete rating scale i.e. based on 6 categories rather than a four-point rating scale would be adopted to collect more in-depth data regarding subtle differences in perception with reference to infarct prevalence or absence.
- Every suspicious region, which is denoted by a location point, will be accompanied by a confidence rating.
- Location reporting will consist of not only a score on a 2D representation, but also a marking on the digital image i.e. mouse click to the centre of the abnormality and/ or the ability to use trace function to highlight the boundaries of the lesion where the infarct shape might be irregular.
- Allow the SHOW MARK/ HIDE MARK function to ensure the radiologist can trace the marks they have made to avoid repeat visits and will also ensure the FP ratings can be correctly assigned to the correct image slice in future studies.

The above scoring will enable the results to be analysed using JAFROC software, which is currently regarded as the most comprehensive method of analysis for free-response studies, which also has a very close approximation to the clinical task (Chakraborty, 2006).

9.8 Future Training Recommendations

As errors were evident in both reading tasks, radiology trainees and some consultants would benefit from targeted training to increase experience of what constitutes normal from abnormal features in head and neck imaging, and to not only increase infarct detection, but also address misclassification errors that were evident within each reading task. Further training, if not already in place, should increase knowledge of neuroanatomy, cerebrovascular pathways and stroke type characteristics within each modality. Training should also address how to spot focal abnormalities and small vessel changes, and provide insights into how and when these features might become a problem for the patients' cognitive function. As previously eluded in chapter 8 (consultant comparisons), neurology and radiology disciplines could collaborate more closely in the medical school curriculum to strengthen links between disease epidemiology and disease manifestation/ presentation in medical images. In addition, previous work has demonstrated that fewer clinical errors are made when people work in teams (DoH, 2000; Fitzgerald, 2001).

A future training package akin to the voluntary self-assessment scheme in mammographic screening known as PERFORMS (Gale, 2003; Scott & Gale, 2006) could be designed and introduced into neuroradiology. The mammographic screening scheme runs biannually and provides anonymous data to individuals and radiology centres across the UK, allowing radiologists to keep track of their performance and training needs as well as gaining an insight into errors in breast screening throughout the UK, combining knowledge rather than working in solitude. A similar model could be applied to neuroradiology, which not only gathers data and provides feedback on stroke detection and performance but also other conditions such as tumour, M.S. and/ or Alzheimers, for example.

9.9 Thesis Conclusions

The repercussions of patients not receiving timely or appropriate stroke treatment include increasing likelihood of disability and death. This series of studies aimed to explore whether errors were evident in the detection and identification of stroke and how expertise appears to develop in one aspect of neuroradiology. Errors were evident and visual search behaviours differed on the basis of decisions made between reader groups, stroke types and image modalities examined. Although this series of studies originated from one centre, it is hoped the findings therein are a significant, and novel, contribution to knowledge in a particularly neglected and mixed area of research and although findings are not conclusive, offer up another respectable perspective on observer performance within radiology.

Research within this exciting area draws from psychology, cognitive science, radiology, statistics, physics, and ergonomics, allowing insights into errors, visual search behaviour and cognitive processing. Although this series of studies is only a miniscule part of the overall picture, collectively, research findings with a clear medical focus have the ability to filter down through clinical governance and alter daily practice. There is much scope for improvement within both research, training and clinical practice and it is hoped the present series of studies, sparks further interest and investigation into the visual and cognitive interpretations of neuroradiological images might, one day, improve training, performance and/ or patient outcomes.

Adams, H.P., Brott, T.G., Furlan, A.G., Gomez, C.R., Grotta, J., Helgason, C.M., Kwiatkowski, T., Lyden, P.D., Marler, J.R., Torner, J., Feinberg, W., Mayberg, M., & Thies, W., 1996. Guidelines for Thrombolytic Therapy for Acute stroke: A Supplement to the Guidelines for the Management of Patients with Acute Ischemic stroke. *Circulation*; 94:1167-1174.

Asplund, K., Stegmayr, B., & Peltonen, M., 1998. From the Twentieth to the Twenty-First Century: A Public Health Perspective on stroke. *In: Ginsberg MD, Bogousslavsky J, eds. Cerebrovascular Disease Pathophysiology, Diagnosis, and Management; Vol 2. Malden, Mass: Blackwell Science;* Chapter 64.

American Heart Association, 2007. Heart Disease and stroke Statistics—2007 Update: A Report From the American Heart Association Statistics Committee and stroke Statistics Subcommittee. *Circulation, American Heart Association*. Dallas, TX.

Association of British Neurologists, 1996. Neurology in the United Kingdom: Numbers of Clinical Neurologists and Trainees. *ABN;* London.

Berbaum, K.S., Franken, E.A. Jr., & Dorfman, D.D. *et al*., 1990. Satisfaction of visual search in diagnostic radiology. *Investigative Radiology;* 25: 133-49

Berbaum, K.S., El-Khoury, G.Y., Franken, E.A., Kathol, M., Montgomery, W.J., & Hesson, W., 1988. Impact of Clinical History on Fracture Detection with Radiography. *Musculoskeletal Radiology;* 168:507-511.

Birkelo, C.C., Chamberlain, W.E., & Phelps, P.S., 1947. Tuberculosis case finding. A comparison of the effectiveness of various roentgenographic and photofluorographic methods. *JAMA;* 133, 359-66.

Bonke, B., Koudstaal, P.J., Dijkstra, G., Van Hilligersberg, R., Van Knippenberg, F.C.E., Duivenvoorden, H.J., & Kappelle, L.J., 1989. Detection of lacunar infarction in brain CT-scans. *Neuroradiology;* 31:170-173.

Brennan, P.C., McEntee, M., Evanoff, M., Phillips, P., O'Connor, W. T., & Manning, D.J., 2007. Ambient Lighting: Effect of Illumination on Soft-Copy Viewing of Radiographs of the Wrist. *Musculoskeletal Imaging, AJR;* 188:W177-W180.

Brenner, R.J., Lucey, L.L., Smith, J.J., & Saunders, R., 1998. Radiology and medical malpractice claims: a report of the practice standards claims survey of the Physician Insurers Association of America and the American College of Radiology. *American Journal of Roentgenology;* 171:19-22.

British Society of Neuroradiologists, 2003. Effective guidelines for safe and effective practice, London.

Carlson, R.N., 2001. Physiology of Behaviour. Seventh edition, London.

Chan, J. H. M., Tsui, E. Y. K., Luk, S. H., Fung, A. S. L., Yuen, M. K., Szeto, M. L., Cheung, Y. K., & Wong, K. P. C. 2001. Diffusion-weighted MR imaging of the liver: distinguishing hepatic abscess from cystic or necrotic tumor. *Abdominal imaging;* 26: 161-165.

Chakraborty, D.P., 2007. FROC curves using a model of visual search. *Proc. SPIE.* Vol.6515.

Chakraborty, D.P. & Berbaum, K.S., 2004. Methodologies for Observer Studies Involving Detection and Localization: Modelling, Analysis and Validation. *Medical Physics;* 31(8): 2313.

Chakraborty, D.P., 2006. ROC Curves Predicted by a Model of Visual Search. *Physics in Medicine and Biology;* 51(14): 3463.

Department of Health, 2006. Asset 2 – Action on stroke Services: an Evaluation Toolkit for commissioners. London.

Department of Health, 2000. An Organisation with a Memory. London.

Department of Health, 2005. The National service Framework for Long-term Conditions. London.

Department of Health, 2006. Mending Hearts and Brains. London.

Department of Health, 2007. National stroke Strategy, London.

Department of Health, 2008. Implementing the National stroke Strategy – an imaging guide. London.

DeVries, H., 1943. The quantum character of light and its bearing upon threshold of vision, the differential sensitivity and visual acuity of the eye. *Physica;* 10:553-64.

Donovan, T., & Manning, D.J. 2006. Successful reporting by non-medical practitioners such as radiographers, will always be task-specific and limited in scope. *Radiography;* 12:7-12.

Doubilet, P., & Herman, P.G., 1981. Interpretation of Radiographs: Effect of Clinical History. *AJR;* 137:1055-1058.

Dorfman, D.D., Berbaum, K.S., & Metz, C.E., 1992. Receiver Operating Characteristic analysis. Generalization to the population of readers and patients with the jacknife method. *Invest Radiol;* 27: 723-31.

Duchowski, A.T., 2007. Eye Tracking Methodology: Theory and Practice. Second Edition, London.

Ericsson, K.A., Charness, N., Feltovich, P.J. & Hoffman, P.R. 2006.The Cambridge Handbook of Expertise and Expert Performance. *Cambridge University Press, New York, USA.*

Fawcett, T., 2006. An introduction to ROC analysis. *Pattern Recognition Letter;* 27:861-874.

Feigin, V.L., Lawes, C.M.M., Bennett, D.A. & Anderson, C.S. 2003. Stroke epidemiology: a review of population based studies of incidence, prevalence, and case fatality in the late 20th century. *The Lancet;* 2:43-53.

Fitzgerald, R., 2001. Error in Radiology. *Clinical Radiology;* 56:938-946.

Gale, A.G., 2003. PERFORMS - a self assessment scheme for radiologists in breast screening. *Seminars in Breast Disease;* 6(3):148-152.

Gale, A.G., Johnson, F., and Worthington, B.S., 1979. Psychology and Radiology. In Oborne, D.J., Gruneberg, M.M. and Eiser, J.R.S. (Eds.). *Research in Psychology and Medicine;*1, *Physical Aspects,* Academic Press, London.

Garlatti, S., & Sharples, M., 1998. The use of a computerized brain atlas to support knowledge-based training in radiology. *Artificial intelligence in Medicine;* 13:181-205.

Good, B.C., Cooperstein, L.A., & DeMarino, G.B., 1990. Does knowledge of the clinical history affect the accuracy of chest radiograph interpretation? *American Journal of Roentgenology;* 154:709-712.

Gonzalez, R.G., Schaefer, P.W., Buonanno, F.S., Schwamm, L.H., Budzik, R.F., Rordorf, G., Wang, B., Sorensen, A.G. & Koroshetz, W.J., 1999. Diffusion-weighted MR Imaging: Diagnostic Accuracy in Patients Imaged within 6 Hours of stroke Symptom Onset. *Radiology;* 210:155-162.

Gunderman, R.B., Siddiqui, A.R., Heitkamp, D.E., & Kipfer, H.D., 2003. The vital role of radiology in the medical school curriculum. *AJR;* 180:1239-1242.

Grotta, J.C., Chiu, D., Lu, M., Patel, S., Levine, S.R., Tilley, B.C., Brott, T.G., Haley, C., Lyden, P.D., Kothari, R., Frankel, M., Lewandowski, C.A., Libman, R., Kwaitkowski, T., Broderick, J. P., Marler, J.R., Corrigan, J., Huff, S., Mitsias, P., Talati, S., & Tanne, D., 1999. Agreement and Variability in the Interpretation of Early CT Changes in stroke Patients Qualifying for Intravenous rtPA Therapy. *American Heart Association;* 30:1528-1533.

Hatala, R., Norman, G.R., & Brooks, L.R., 1999. Impact of a Clinical Scenario on Accuracy of Electrocardiogram Interpretation. *JGIM;* 14:126-129.

Haughton, V.M., Rimm, A.A., Sobocinski, K.A., Papke, R.A., Daniels, D.L., Williams, A.L., Lynch, R., & Levine, R., 1986. A Blinded Clinical Comparison of MR Imaging and CT in Neuroradiology. *Neuroradiology;* 160:751-755.

Herzog, H., 2007. Methods and applications of positron-based medical imaging. Radiation *Physics and Chemistry*; 76:337–342.

Hobby, J.L., Tom, B.D.M., Todd, C., Bearcroft, P.W.P., & Dixon, A.K., 2000. Communication of doubt and certainty in radiological reports. *British Journal of Radiology;* 73:999-1001.

Hoeffner, E.G., Case, I., Jain, R., Gujar, S.K., Shah, G.V., Carlos, R.C., Thompson, B.G., Harrigan, M.R., Mukherji., S.K. 2004. Cerebral Perfusion CT: Technique and Clinical Applications. RSNA; 231 (3): 632-644.

Ikeda, D.M., Hylton, N.M., Kinkel, K., Hochman, M.G., Kuhl, C.K., Kaiser, W.A., Weinreb, J.C., Smazal, S.F., Degani, H., Viehweg, P., Barclay, J., & Schnall, M.D., 2001. Development, Standardization, and Testing of a Lexicon for Reporting Contrast-Enhanced Breast Magnetic Resonance Imaging Studies. *Journal of Megnetic Resonance Imaging;* 13:889-895.

Indredavik, B., Bakke, F., & Slordahl, S.A., 1999. stroke Unit Treatment: 10-Year Follow-Up. *stroke;* 30:1524-7.

Kanwisher, N., McDermott, J., & Chun, M.M., 1997. The Fusiform Face Area: A Module in Human Extrastriate Cortex Specialized for Face Perception. *The Journal of Neuroscience;* June; 17(11):4302–4311

Kloska, S.P., Nabavi, D.G., Gaus, C., Nam, E., Klotz, E., Bernd Ringelstein, E., & Heindel, W., 2004. Acute stroke Assessment with CT: Do We Need Multimodal Evaluation? *Radiology;* 233: 79-86.

Krupinski, E.A., Nodine, C.F., & Kundel, H.L., 1993. A perceptually based method for enhancing pulmonary nodule recognition. *Investigative Radiology;* 28:289-294.

Kundel, H.L., 1978. Visual Scanning, Pattern Recognition and Decision-making in Pulmonary Nodule Detection. *Investigative Radiology.*

Krupinski, E.A., 2000. The Importance of Perception Research in Medical Imaging. *Radiation Medicine;* 18: 329-334.

Krupinski, E.A., Williams, M.B., Andriole, K., Strauss, K.J., Applegate, K., Wyatt, M., Bjork, S., & Seibert, A., 2007. Digital Radiological Image Quality: Image Processing and Display. *American College of Radiology;* 4: 389-400.

Kundel, H.L., 2006. History of Research in Medical Image Perception. *Journal of the American College of Radiology;* June; 3:402-408.

Lansberg, M.G., Gregory, W.A., Beaulieu, C., & Marks, M.P., 2000. Comparison of diffusion-weighted MRI and CT in acute stroke. *American Academy of Neurology;* 54:1557-1561.

Leslie, A. Jones, A.J. & Goddard, P.R., 2000. The influence of clinical information on the reporting of CT by radiologists. *The British Journal of Radiology;* 73:1052-1055.

Loy, T.C. & Irwig, L., 2004. Accuracy of Diagnostic Tests Read With and Without Clinical Information: A Systematic Review. *American Medical Association;* 292(13):1602-1609.

Lusted, L.B., 1960. Logical analysis in roentgen diagnosis. *Radiology;* 74, 178-93.

Manning, D.J., Ethell, S.C., & Donovan, T., 2004. Detection or decision errors? Missed lung cancer from the posteroanterior chest radiograph. *The British Journal of Radiology*; *March;* 77: 231-235.

Manning, D.J., Gale, G.A., & Krupinski, E.A., 2005. Perception Research in Medical Imaging. *British Journal of Radiology;* 78:683-685.

Manning, D.J., Barker-Mill, S.C., Donovan, T., & Crawford, T., 2005. Time-Dependent Observer errors in pulmonary nodule detection. *British Journal of Radiology;* 78: 1-5.

Manning, D.J., Ethell, S.C., Donovan, T., & Crawford, T., 2006. How do Radiologists do it? The influence of experience and training on searching for chest nodules. *Radiography;* 12: 134-142.

Maxwell, S.E., & Delaney, H.D., 1990. Designing Experiments and Analyzing Data. Wadsworth Publishing Company. Belmont, California.

McCarthy, E., & Brennan, P.C., 2003. Viewing conditions for diagnostic images in three major Dublin hospitals: a comparison with WHO and CEC recommendations. *British Journal of Radiology;* 76: 94-97.

McEntee, M.F., & Martin, B., 2010. The varying effects of ambient lighting on low contrast detection tasks. *Proc SPIE. International Society for Optical Engineering;* 7627-22.

McEntee, M.F., & Gafoor, S., 2009. Ambient Temperature is an Important Consideration in the Radiology Reading Room. *Proc MIPS* (XIII) Santa Barbara California.

McCarthy, E., & Brennan, P.C., 2003. Viewing conditions for diagnostic images in three major Dublin hospitals: a comparison with WHO and CEC recommendations. *British Journal of Radiology;* 76: 94-97.

McConkie, G.W., 1988. Eye-movement control during reading: I. The location of initial eye fixations during words. *Visual Research;* 28:1107-1108.

McNeil. B.J., Hanley, J.A., Harris-Funkenstein, H., & Wallman, J., 1983. Paired Receiver Operating Characteristic Curves and the Effect of History on Radiographic Interpretation: CT of the Head as a Case Study. *Radiology;* 149: 75-77.

Mello-Thoms, C., 2002. What attracts the eye to the location of missed and reported breast cancers? *ACM;* 58113(1): 111-117.

Metz, C.E., 1978. Basic principles of ROC analysis. *Sem. Nuc. Med;* 8: 283-298.

Metz, C.E., 1986. ROC Methodology in radiographic imaging. *Investigative Radiology;* 21(9) 720-733.

Metz, C., 2006. Receiver Operating Characteristic Analysis: A Tool for the Quantitative Evaluation of Observer Performance and Imaging Systems. *American Journal of Radiology;* 3(6): 413-422.

Mitchell, J.B., Ballard, D.J., Whisnant, J.P., Ammering, C.J., Samsa, G.P., & Matchar, D.B., 1996. What role do neurologists play in determining the costs and outcomes of stroke patients? *stroke;* 27:1937-1943.

Mohr, J.P., Biller, J., Hilal, S.K., Yuh, W.T.C., Tatemichi, T.K., Hedges, S.D., Tali, E., Nguyen, H., Mun, I., Adams, H.P., Grimsman, K., & Marler, J.R., 1995. Magnetic Resonance versus Computed Tomographic Imaging in Acute stroke. *stroke;* 26:807-812.

Mullins, M.E., Lev, M.H., Schellingerhout, D., Koroshetz, W.J., & Gilberto-Gonzalez, R., 2002. Influence of Availability of Clinical History on Detection of Early stroke Using Unenhanced CT and Diffusion-Weighted MR Imaging. *AJR;* 179:223-228.[a]

Mullins, M.E., Schaefer, P.W., Sorensen, A.G., Halpern, E.F., Ay, H., He, J., Koroshetz, W.J., & Gonzalez, R.G., 2002. CT and Conventional and Diffusion-weighted MR Imaging in Acute stroke: Study in 691 Patients at Presentation to the Emergency Department. *Neuroradiology;* 224:353-360.[b]

NHS Modernisation Agency, 2005. Action on Neurology: Improving Neurology Services – a practical guide. London.

NHS Purchasing and Supply Agency, 2007. Diagnostic imaging update: Imaging developments at the Radiological Society of North America (RSNA) conference. Centre for Evidence-based Purchasing, London.

Nodine, C.F., & Kundel, H.L., 1990. A visual dwell algorithm can aid search and recognition of missed lung nodules in chest radiographs. In Borogan D, editor. *Visual search. Taylor Francis.* London: 399-405.

Nodine, C.F., & Krupinski, E.A., 1998. Perceptual skill, radiology expertise, and visual test performance with NINA and WALDO. *Academic Radiology*; 5: 603–612.

Norman, G.R., Brooks, L.R., Coblentz, C.L., & Babcock, C. J., 1992. The correlation of feature identification and category judgments in diagnostic radiology. *Mem Cognit*; 20:344-355.

Norman, G.R., Brooks, L.R., Coblentz, C.L., & Babcock, C. J., 1992. Expertise in visual diagnosis: a review of the literature. *Academic Medicine;*67(10):78-83.

Obuchowski, N.A., 2000. Sample Size Tables for Receiver Operating Characteristic Studies. *AJR;* 175:603-608.

O'Regan, J.K., 1992. Optimal viewing position in words and the strategy-tactics theory of eye movements in reading. In Eye movements and Visual Cognition: Scene Perception and Reading by Rayner, K: 333-354, Springer-Verlag.

Osborn, A.G., Blazer, S.I., Salzman, K.L., Katzman, G.L., Provenzale, J., Castillo, M., Hedlund, G.L., Illner, A., Harnsberger, H.R., Cooper, J.A., Jones, B.V., & Hamilton, B.E., 2005. Diagnostic Imaging: Brain. Amirsys, USA.

Phillips, P.W., Manning, D.J., Donovan, T., Crawford, T., & Higham, S., 2005. A Software Framework for Diagnostic Medical Image Perception with Feedback, and a Novel Perception Visualisation Technique. *Proc. SPIE*; 5749: 572-580.

Phillips, P.W., Manning, D.J., Crawford, T., Burling, D., Tam, C.H., Taylor, A., 2008. Searching in Axial and 3D CT Visualisations. *Proc. SPIE*; 6917.

Plumb, A.A.O., Grieve, F.M., & Khan, S.H., 2009. Survey of hospital clinicians' preferences regarding the format of radiology reports. *Clinical Radiology;* 64: 386-394.

Rayner, K., 1996. Eye movement control in reading: a comparison of two types of models. *Journal of Experimental Psychology in Human Perception and Performance;* 22:1188-1200.

Rivett, G.C. ,1998. From Cradle to Grave: fifty years of the NHS. London, King's Fund.

Rogers, E., 1995. VIA-RAD: a blackboard-based system for diagnostic radiology. *Artificial Intelligence in Medicine;* 7, 343-60.

Rose, A., 1948. The sensitivity performance of the human eye on an absolute scale. *J Opt Soc Am;* 38: 196-208.

Royal College of Physicians, 2007. Clinical Effectiveness and Evaluation Unit. London.

Royal College of Radiologists, 2004. Special Interest Training Curricula. London: The Royal College of Radiologists.

Royal College of Radiologists, 2007. Structured Training Curriculum for Clinical Radiology. London: the Royal College of Radiologists.

Sakas, G., 2002. Trends in medical imaging: from 2D to 3D. *Computers & Graphics;* 26: 577-587.

Samuel, S., Kundel, H.L., Nodine, C.F., & Toto, L.C., 1995. Mechanism of satisfaction of search: eye position recordings in the reading of chest radiographs. *Radiology;* 1: 242-9.

Schreiber, M.H., 1963. The clinical history as far in Roentgenogram interpretation. *Journal of American Medical Association;* 185:137-139.

Schriger, D.L., Kalafut, M., Starkman, S., Krueger, M., & Saver, J.L., 1998. Cranial Computed Tomography Interpretation in Acute stroke. *Journal of American Medical Association;* 279 (16): 1293-1297.

Schwartz, S. 2010. Visual Perception: A Clinical Orientation, Fourth Edition. McGraw-Hill Companies, Inc. USA.

Scott, H., & Gale, A.G., 2006. Breast screening: PERFORMS identifies key training needs. *The British Journal of Radiology;* 79:S127-S133.

Scott, H., Gale, A.G., & Hill, S., 2008. How are False Negative Cases perceived by Mammographers? Which Abnormalities are misinterpreted and which go undetected? *Proc. SPIE*; 6917: 691-713.

Seifert, V., 2003. Intraoperative MRI in neurosurgery: Technical overkill or the future of brain surgery? *Neurology India;* 51:329-332.

Spector, R.H. Visual Fields: Chapter 116 in "Clinical Methods: The History, Physical, and Laboratory Examinations. 3rd edition. Walker, H.K., Hall, W.D. & Hurst, J.W. *Boston:* Butterworths*;* 1990.

Starr, M.S., & Rayner, K., 2001. Eye movements during reading: some current controversies. *Trends in Cognitive Sciences;* 5 (4):156-163.

Stergiou, G.S., Vemmos, K.N., Pliarchopoulou, K.N., Synteos, A.G., Roussias, L.G., & Mountokalakis, T.D., 2002. Parallel Morning and Evening Surge in stroke Onset, Blood Pressure, and Physical Activity. *stroke;* 33:1480-1486.

Sureshbabu, W., 2005. PET/CT Imaging Artifacts. *Journal of Nuclear Medicine Technology;* 33(3):156-161

Swets, J.A., Pickett, R.M., & Whitehead, S.F., 1979. Assessment of diagnostic technologies. *Science;* 205: 753-9.

Swensson, R.G., 1996. Unified measurement of observer performance in detecting and localizing target objects on images. *Medical Physics;* 23: 1709-1725.

Thomas, E.L. & Lansdown, E.L., 1963. Visual Search Patterns of Radiologists in Training. *Radiology; Aug;* 81, 288-91.

Thorne-Griscom, N., 2002. A Suggestion: Look at the Images First, Before You Read the History. *RSNA;* 223:9-10.

Tobii Technology Whitepaper. 2010 Tobii Eye Tracking. An introduction to eye tracking and Tobii Eye Trackers. *Tobii Technology AB.*

Tudor, G.R., Finlay, D., & Taub, N., 1997. An assessment of inter-observer agreement and accuracy when reporting plain radiographs. *Clinical Radiology;* 52: 235-238.

Van Everdingen, K.J., van der Grond, J., Kappelle, L.J., Ramos, L.M.P., & Mali, W.P.T.M., 1998. Diffusion-Weighted Magnetic Resonance Imaging in Acute stroke. *American Heart Association;* 29: 1783-1790.

Weiner, M., Schuff, N., Mueller, S., Zhan, W., Zhang, Y., Miller, B. & Chui, H., 2008. Multimodality of neurodegenerative diseases. *Proc. SPIE* 6916-06:110.

Willich, S.N., Goldberg, R.J., Maclure, M., Perriello, L., & Muller, J.E., 1992. Increased onset of sudden cardiac death in the first three hours after awakening. *American Journal of Cardiology;* 70:65-68.

Wintermark, M., Meuli, R., Browaeys, P., REichhart, M., Bogousslavsky, J., Schnyder, P., & Michel, P., 2007. Comparison of CT perfursion and angiography and MRI in selecting stroke patients for acute treatment. *Neurology;* 68:694-697.

World Health Organisation, 2004. The atlas of heart disease and stroke.

Yang, S.N., & McConkie, G.W., 2001. Eye movements during reading: a theory of saccade initiation times. *Vision Research;* 41:3567-3585.

Appendix

Pilot Study 1: Radiology image perception and observer performance: how does expertise and clinical information alter interpretation? stroke detection explored through eye-tracking.

Observer Rating Sheet

Is the abnormality?

1 – Definitely absent

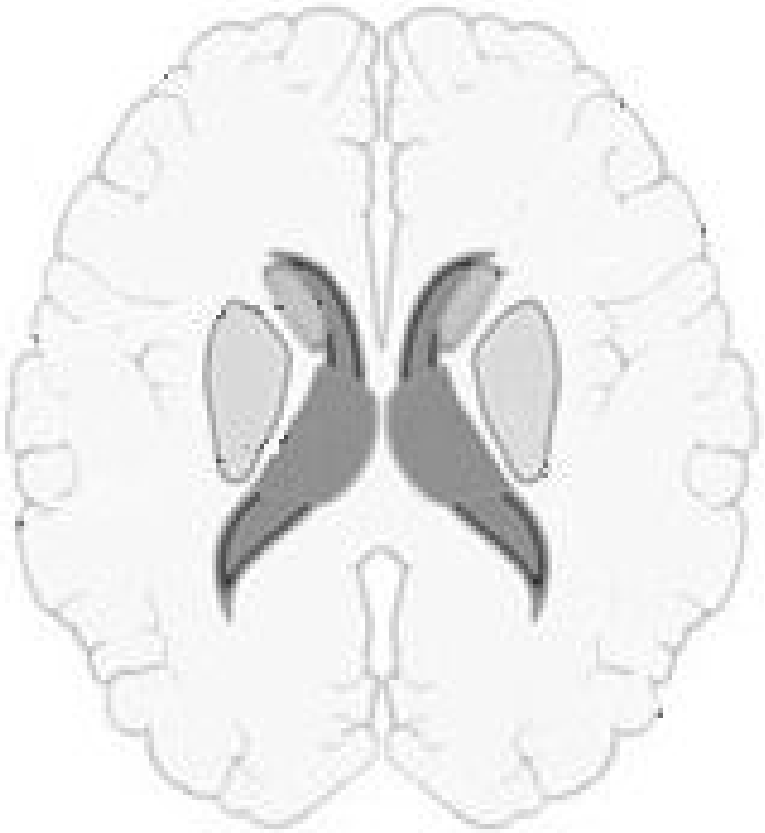
2 – Probably absent

3 – Unsure

4 – Probably present

5 – Definitely present

Please circle on the brain atlas below where you saw the abnormality (if at all):



Pilot Patient Clinical information

Patient 1: Acute stroke 1

The images you are about to see come from a 45 year-old woman. She noticed the sudden onset of right body weakness and trouble speaking prior to hospital admission.

Patient 2: Acute stroke 2 – Alexia without Agraphia

The images you are about to see come from a 63 year-old woman. She had a history of high blood pressure and non-insulin dependent diabetes mellitus, and complained of an inability to read. She also had the occasional problem 'finding the right words'. She could write a grammatically complete sentence, but was then unable to read it back.

Patient 3: Acute stroke 3 – Fluent Aphasia (CT)

These images are from an 86 year-old man. He had a history of cardiac problems (atrial fibrillation) and diabetes mellitus. He also had the sudden onset of fluent aphasia; the lack of ability to perceive the pitch, rhythm, and motional tone of speech.

Patient 4: Acute stroke 4 – Speech Arrest

These images are from a 63 year-old, right-handed male. He had a history of Diabetes and High blood pressure; both were being treated with medication. He had a 5 minute spell of tingling in the right cheek, followed by an inability to speak. He was able to communicate with hand gestures, and his speech returned during transport to the hospital, less than 60 minutes after the episode's onset.

Patient 5: Subacute stroke 1

A 48 year-old right handed man developed sudden difficulty speaking while at work. His past medical history was significant for high blood pressure, but he took no medication. On initial exam, his blood pressure was 168/106 with regular heart rate of 100. He was alert and attentive but he had some difficulty with complex grammatical phrases.

Patient 6: Subacute stroke 2 – Loss of Sensation

These images are from a 65 year-old right-handed man. He had a history of cardiac problems (atrial fibrillation) and previous stroke; he had suffered a very mild paralysis of the left side of his body. He suddenly experienced tingling in the left hand and arm. On examination he failed to explore the left half of space (hemispatial neglect).

Patient 7: Normal Case

These images are from a 76 year-old woman. Prior to admission at the hospital was in good health. On admission her heart rate was 125/ 80 with a heart rate of 65.

Patient 8: Chronic stroke

These images are from a 71 year old woman. She had a history of cardiac weakening (Ischemic Cardiomyopathy) and renal insufficiency. She developed the sudden onset of paralysis in the left side of her body and failed to explore the left side of space. She became mute soon after admission.

Observer Rating Sheet: Studies 1, 2, 3, 4 and 5

Is the **main abnormality?** Please circle your response (1-4) below:

1 – Definitely absent

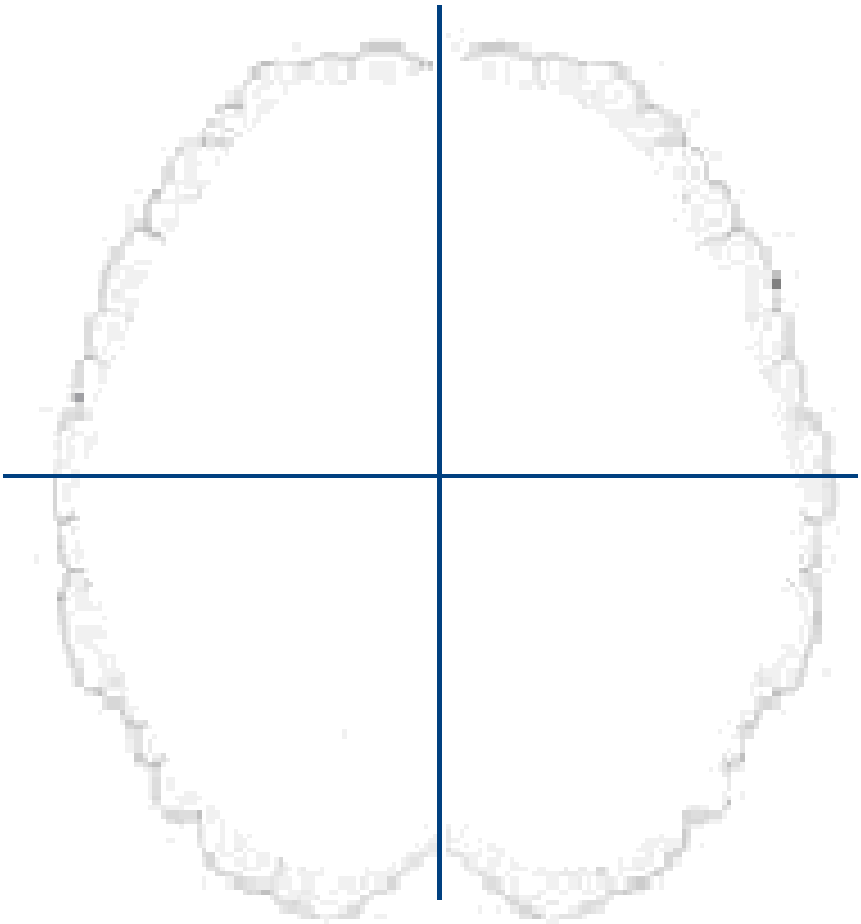
2 – Probably absent

3 – Probably present

4 – Definitely present

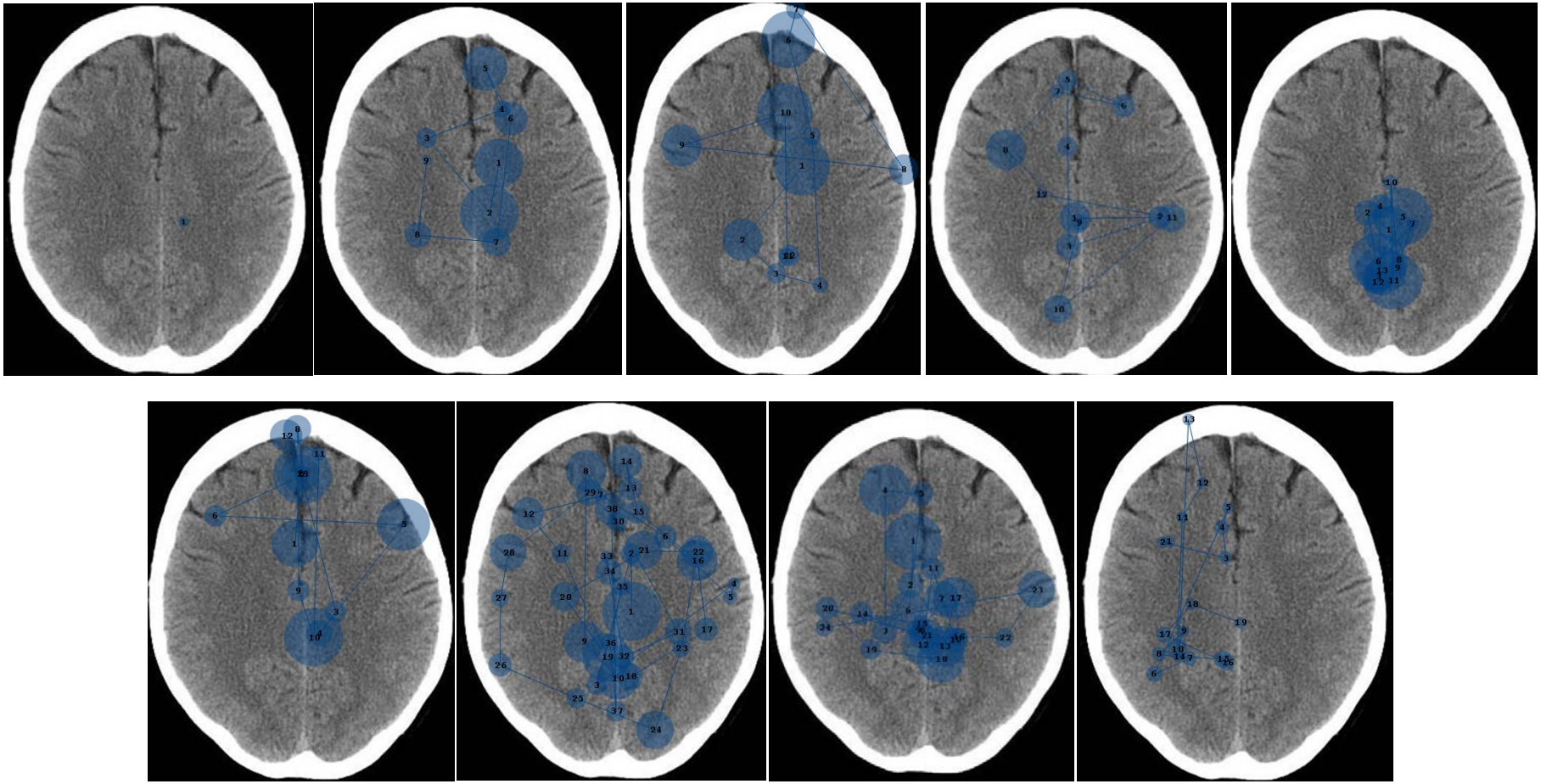
Please circle on the brain representation below where you saw the main abnormality.

Please also identify any **small vessel changes (SVC)** and/or any **focal abnormality (FA),** if at all:



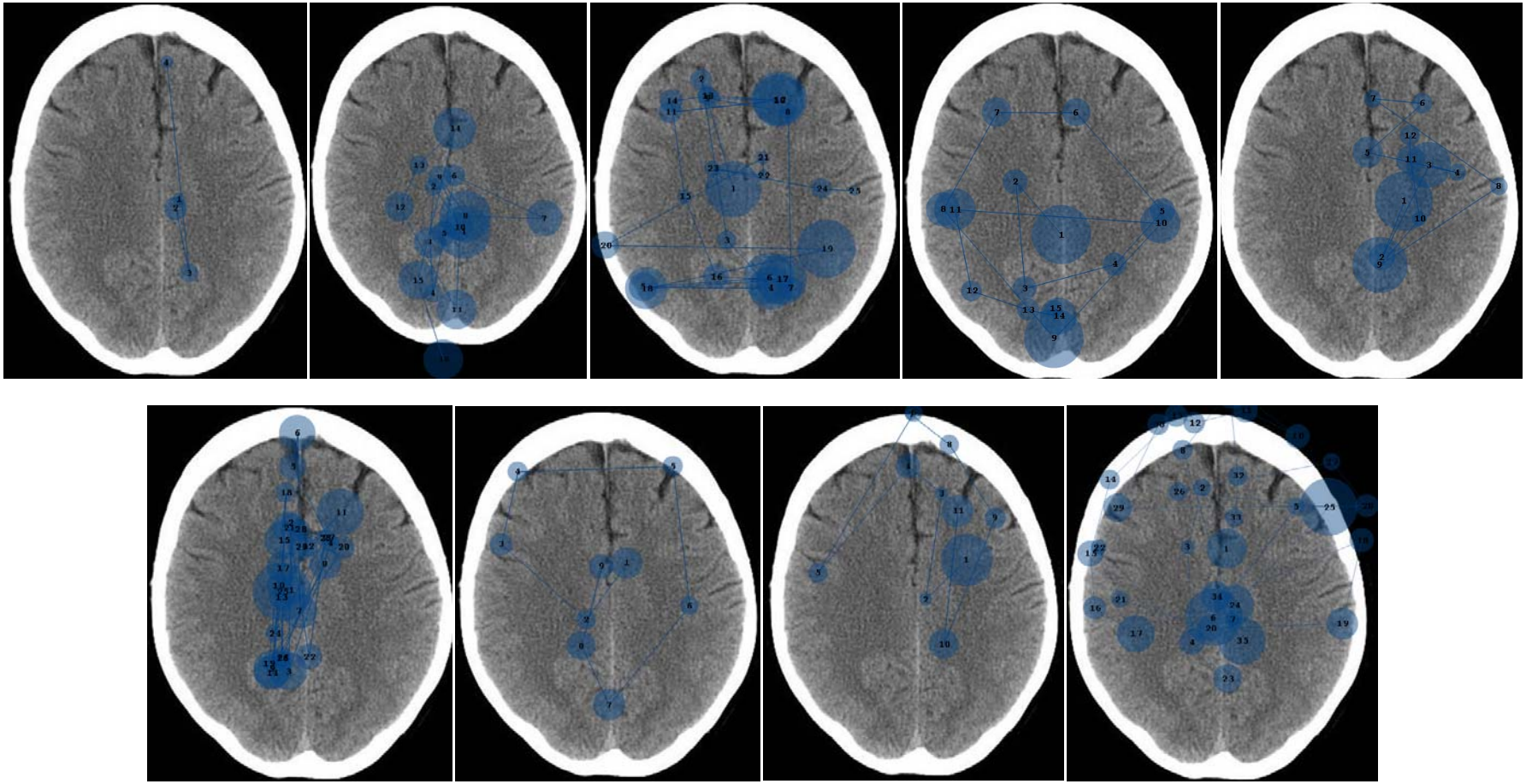
Appendix

CT gaze-tracker images of novice readers (case NBH)

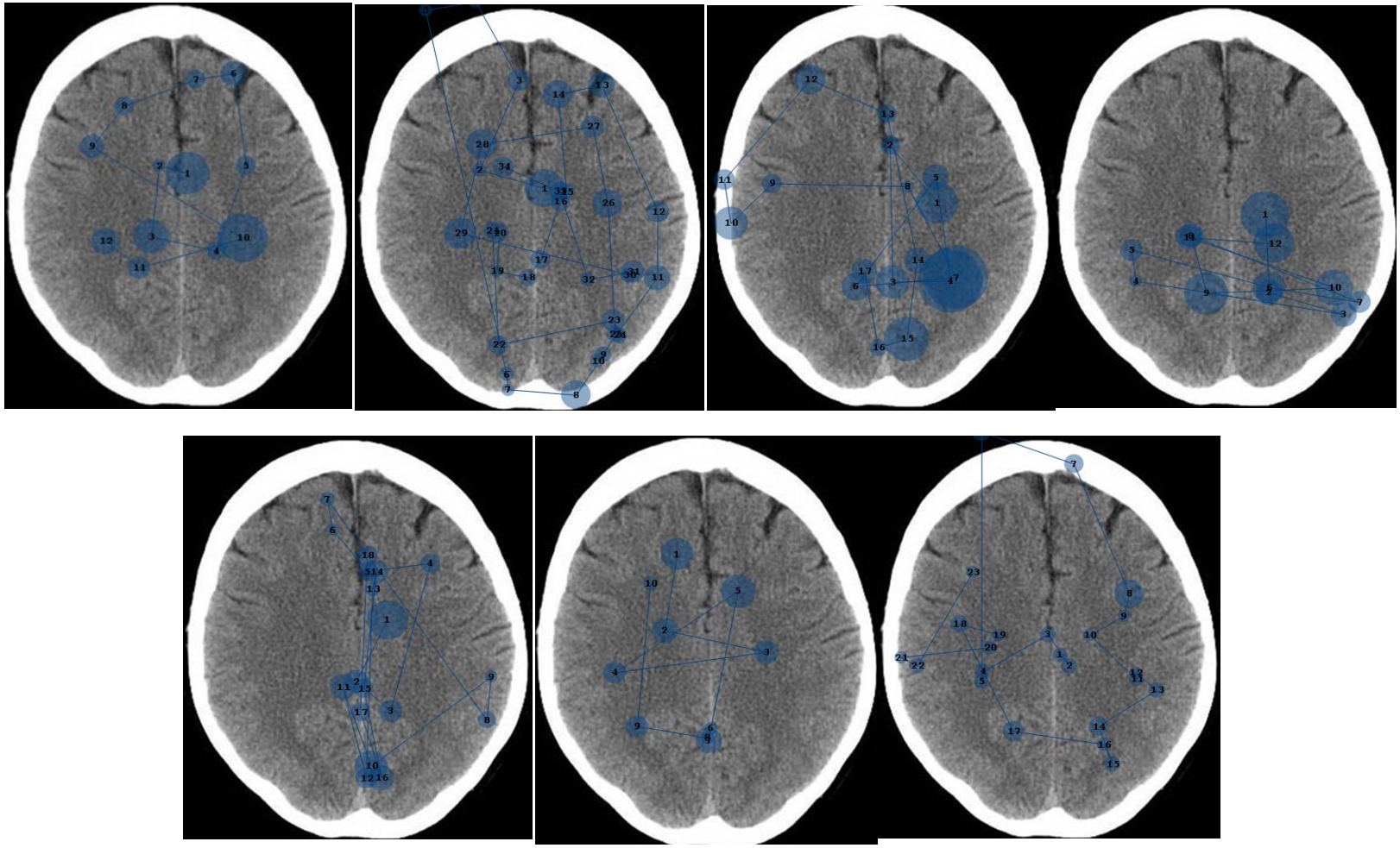


Appendix

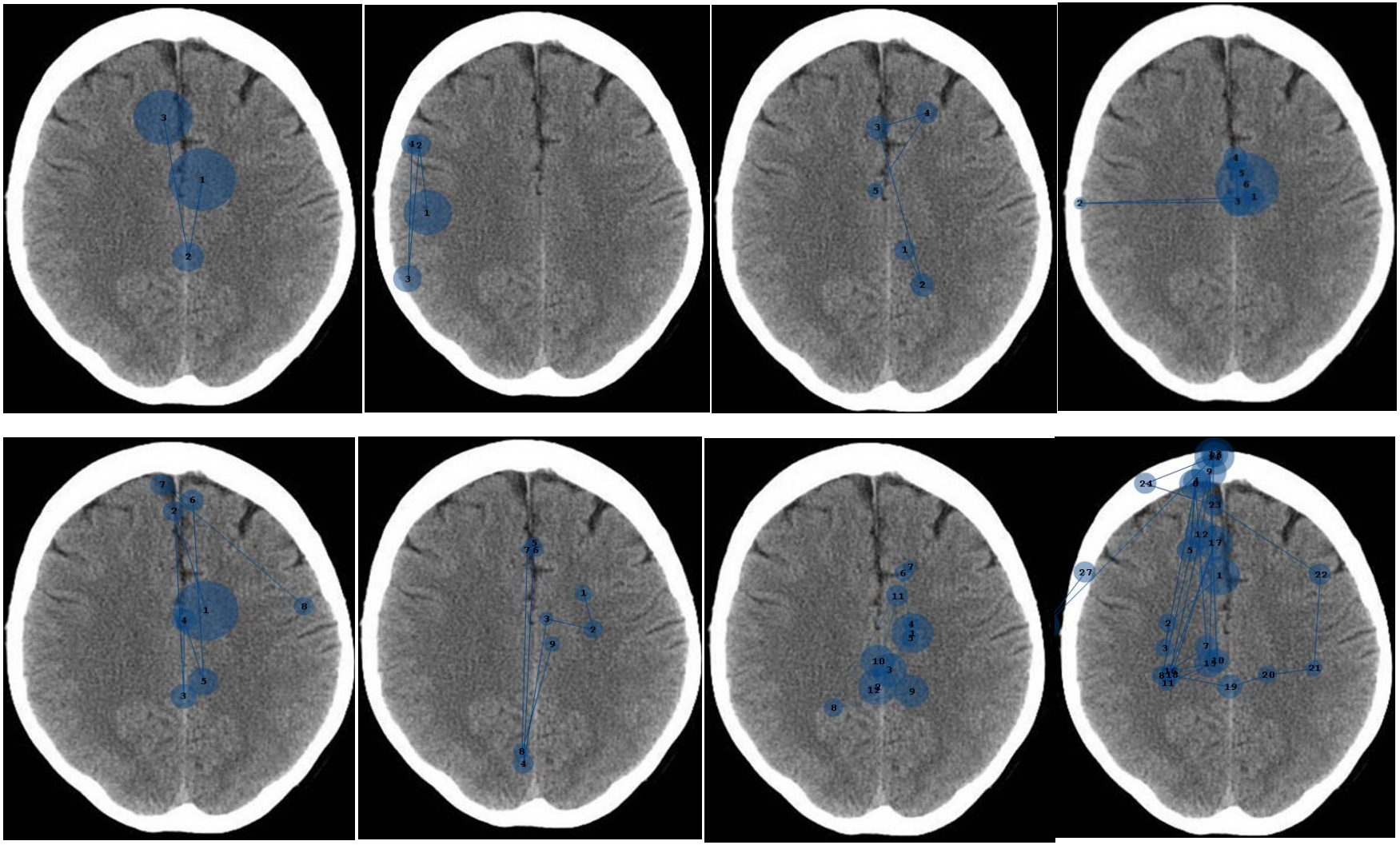
CT gaze-tracker images of trainee readers (case NBH)



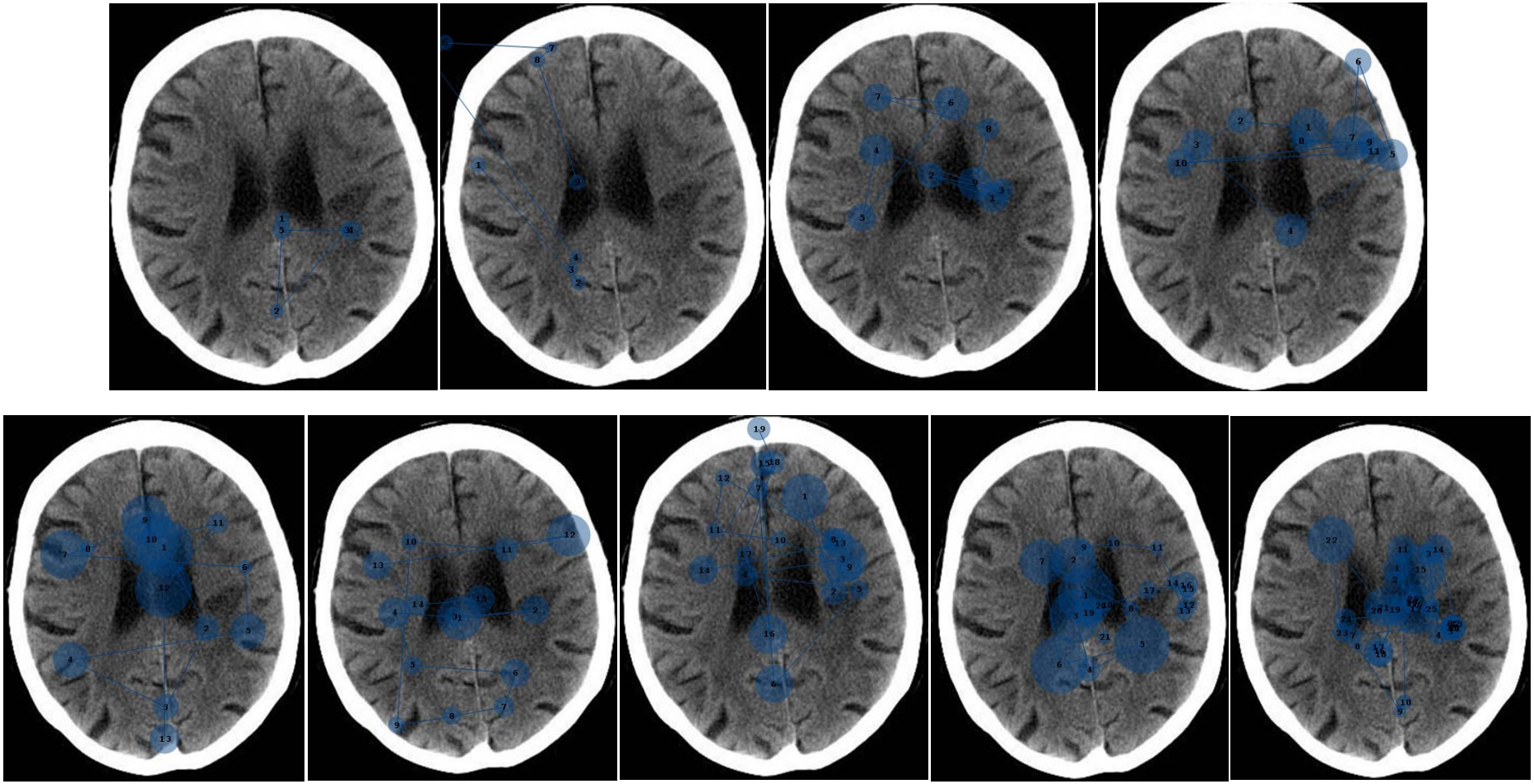
CT gaze-tracker images of expert readers (case NBH)



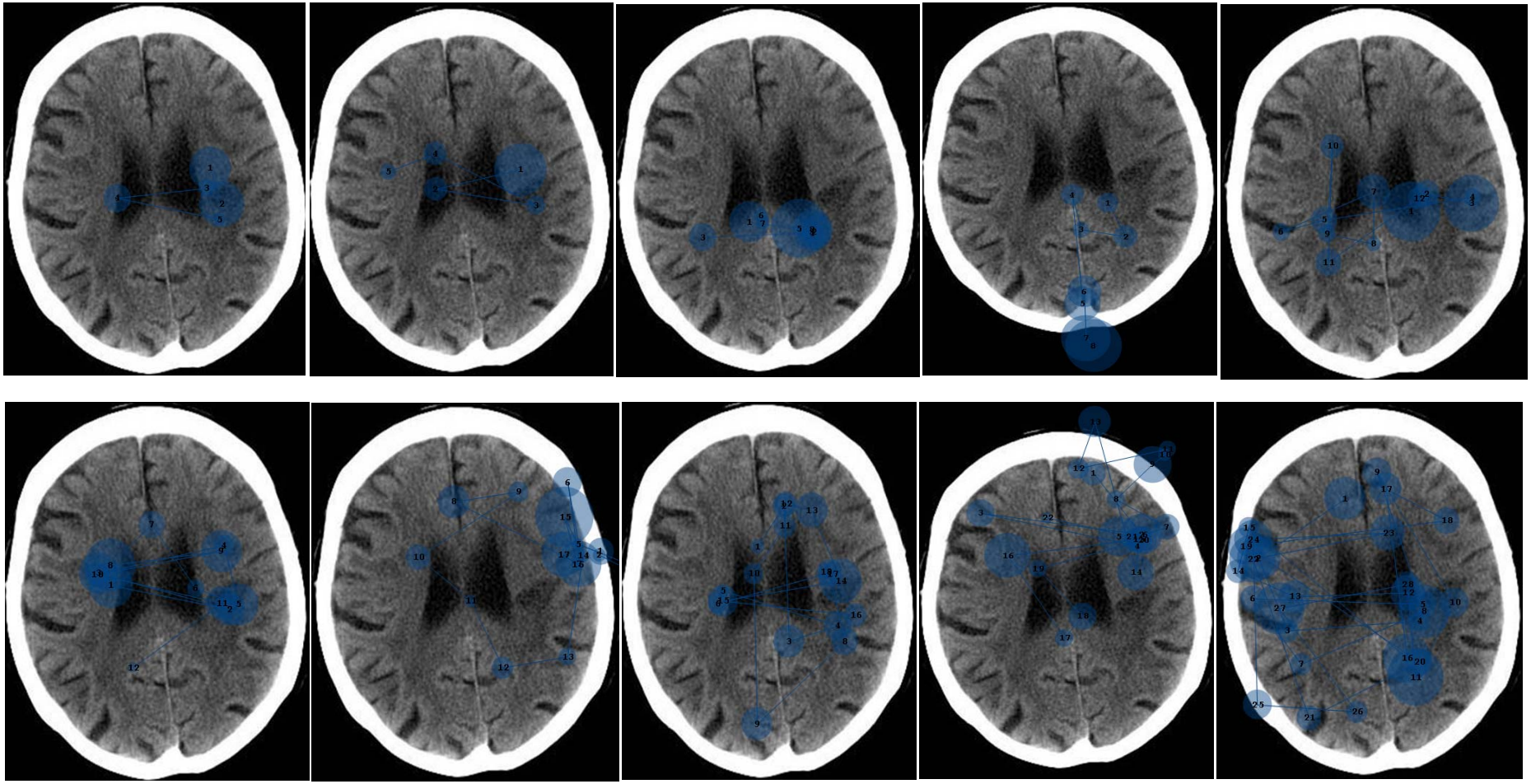
CT gaze-tracker images of neurology readers (case NBH)



CT gaze-tracker images of novice readers (case AAB)

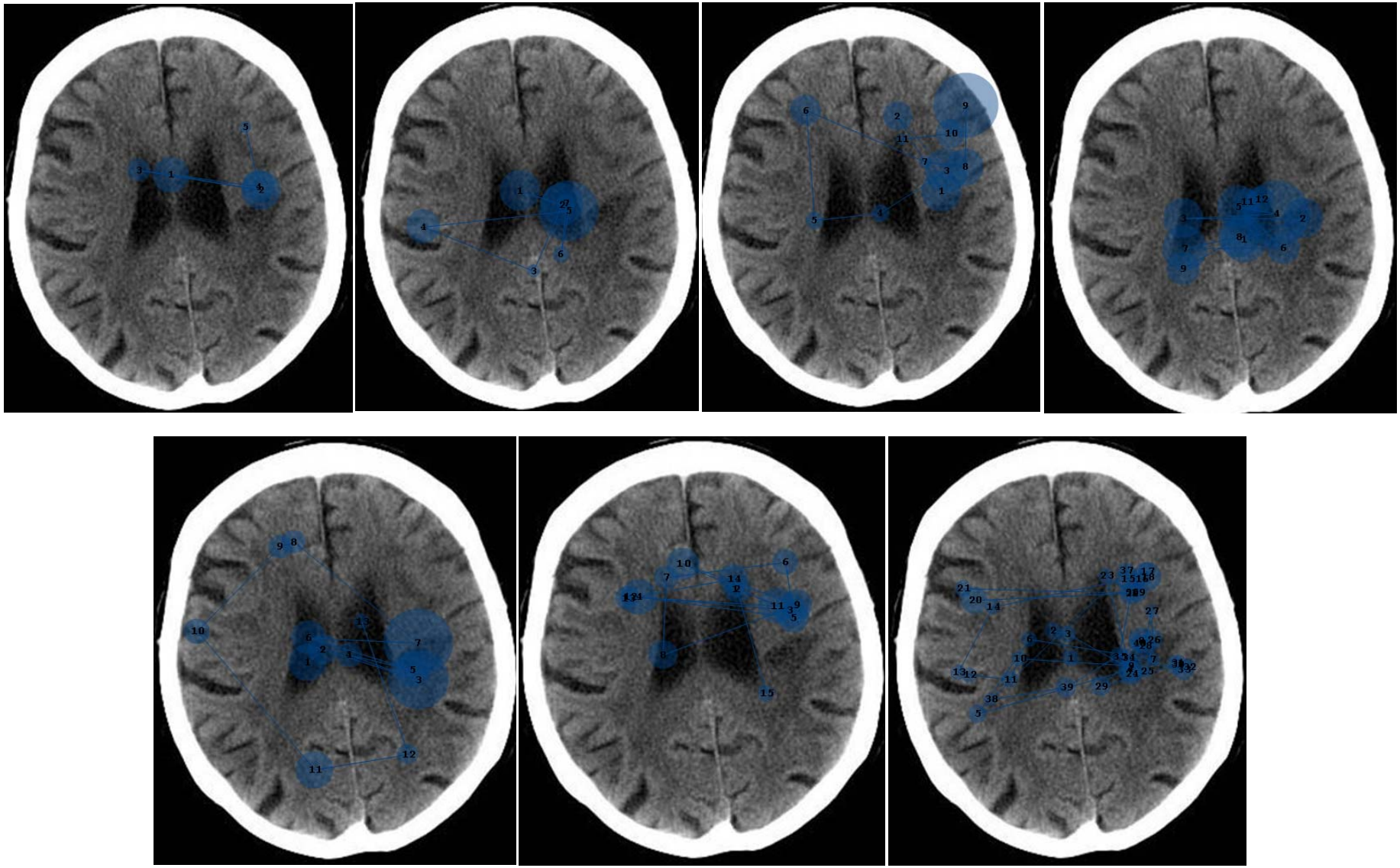


CT gaze-tracker images of trainee readers (case AAB)

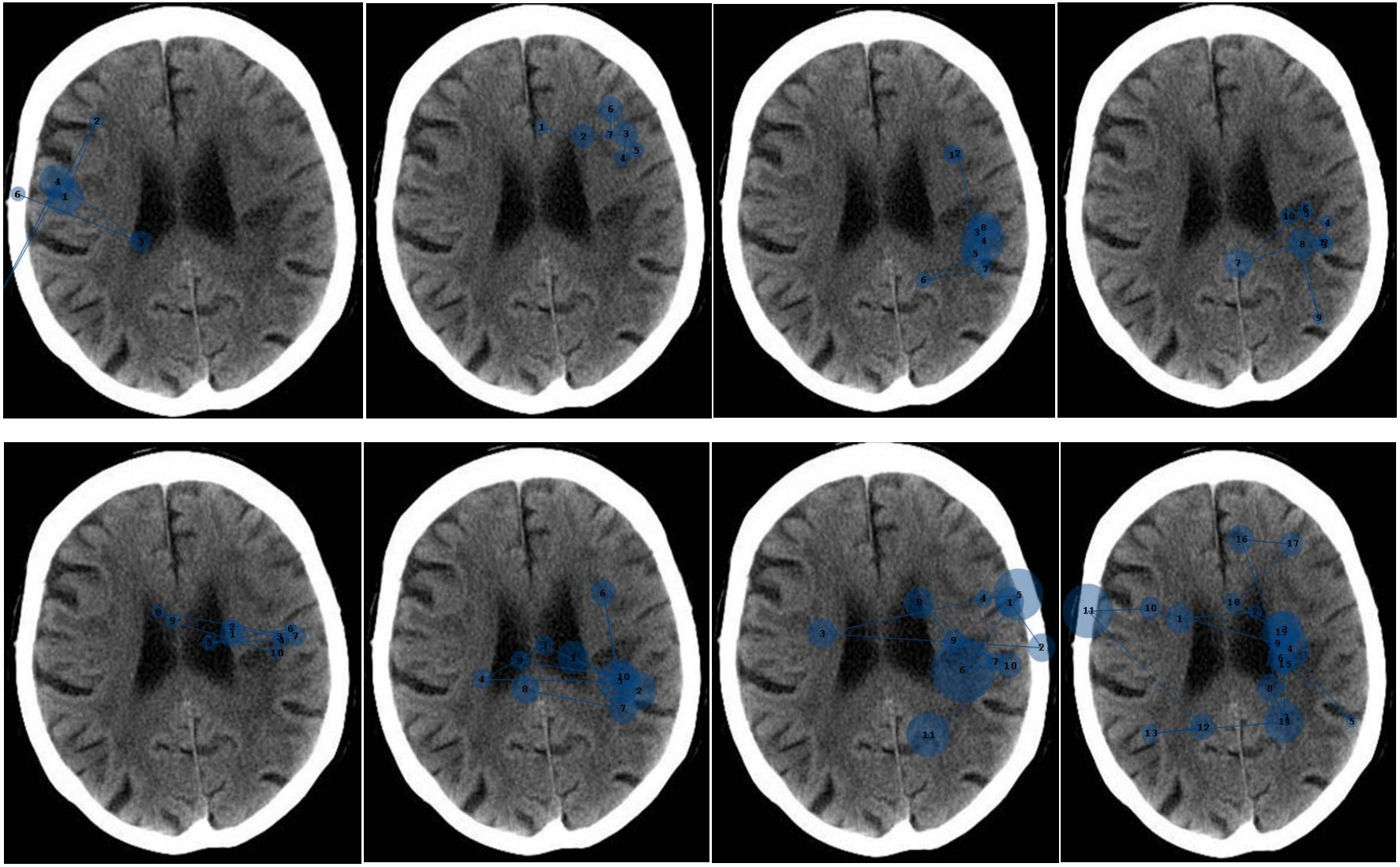


Appendix

CT gaze-tracker images of expert readers (case AAB)

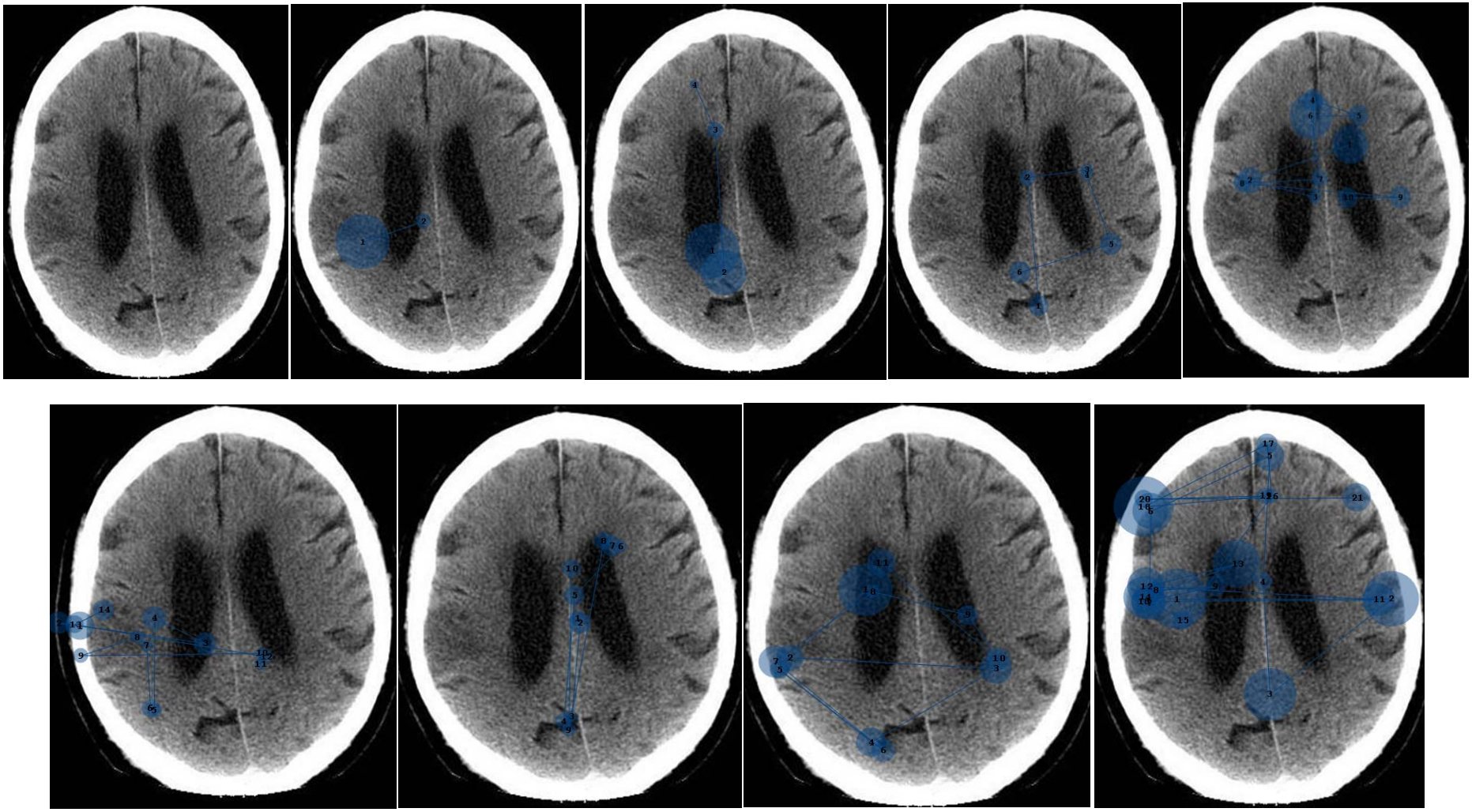


CT gaze-tracker images of nuerology readers (case AAB)

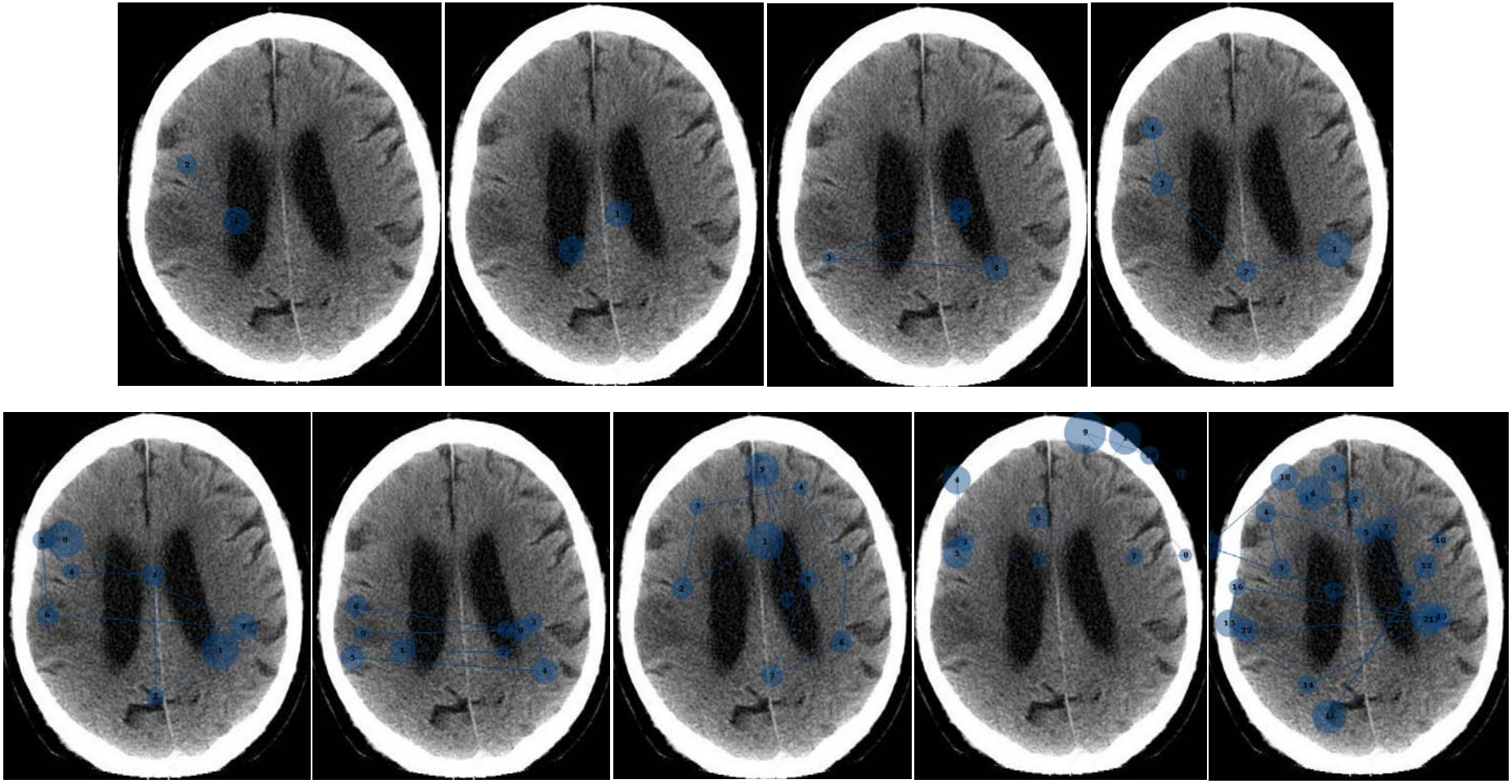


Appendix

CT gaze-tracker images of novice readers (case SDP)

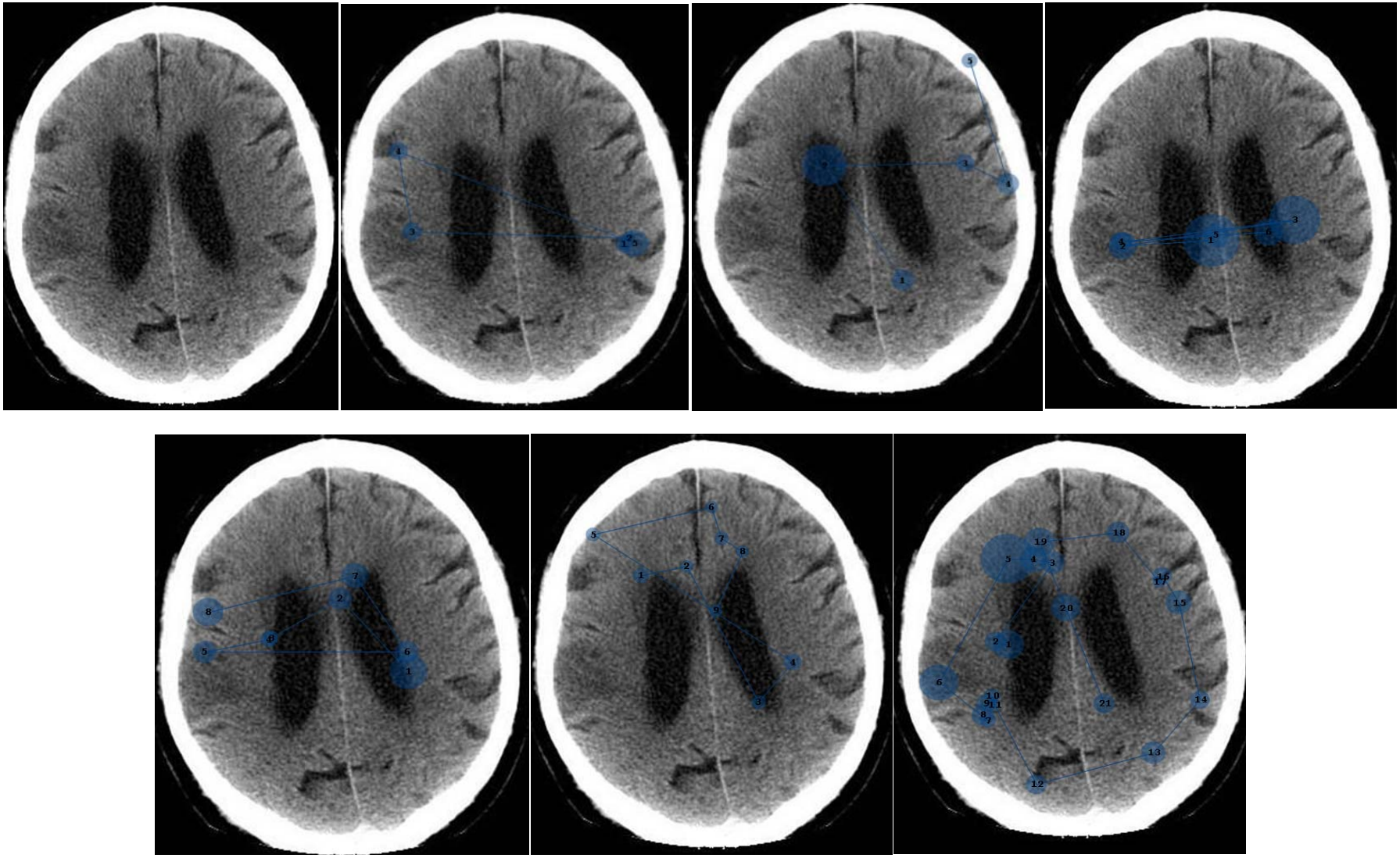


CT gaze-tracker images of trainee readers (case SDP)



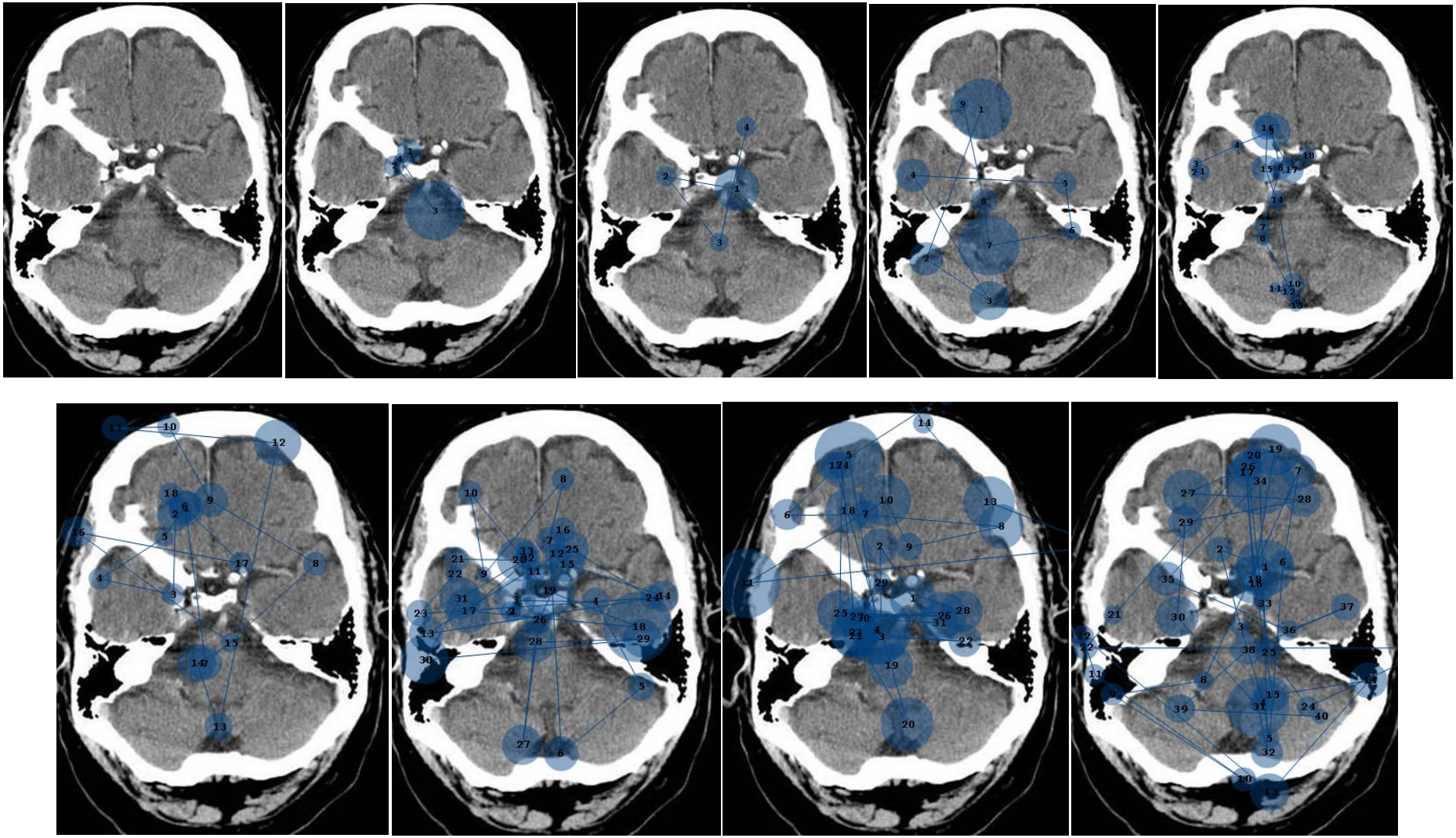
Appendix

CT gaze-tracker images of expert readers (case SDP)



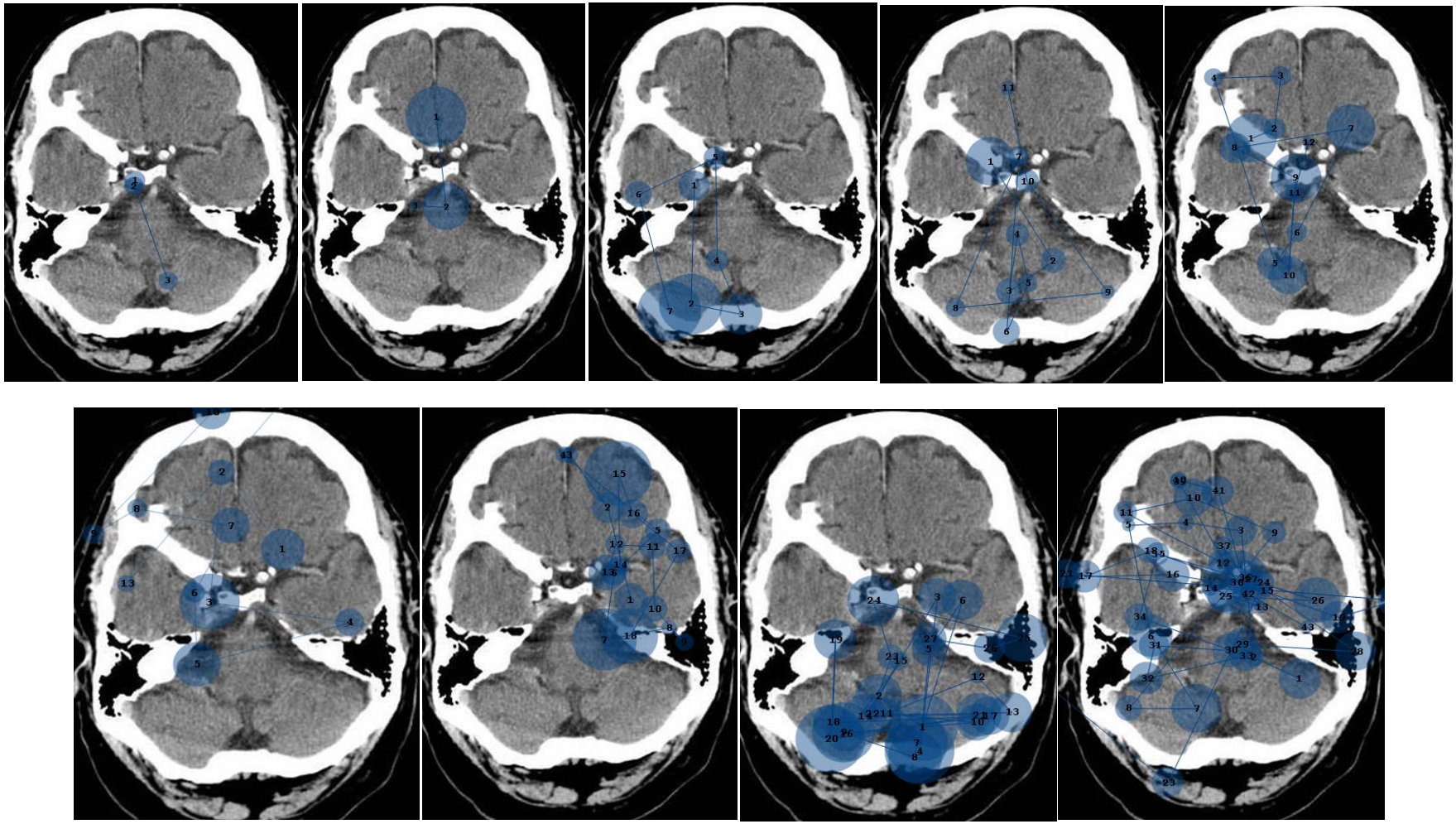
Appendix

CT gaze-tracker images of novice readers (case CRW)

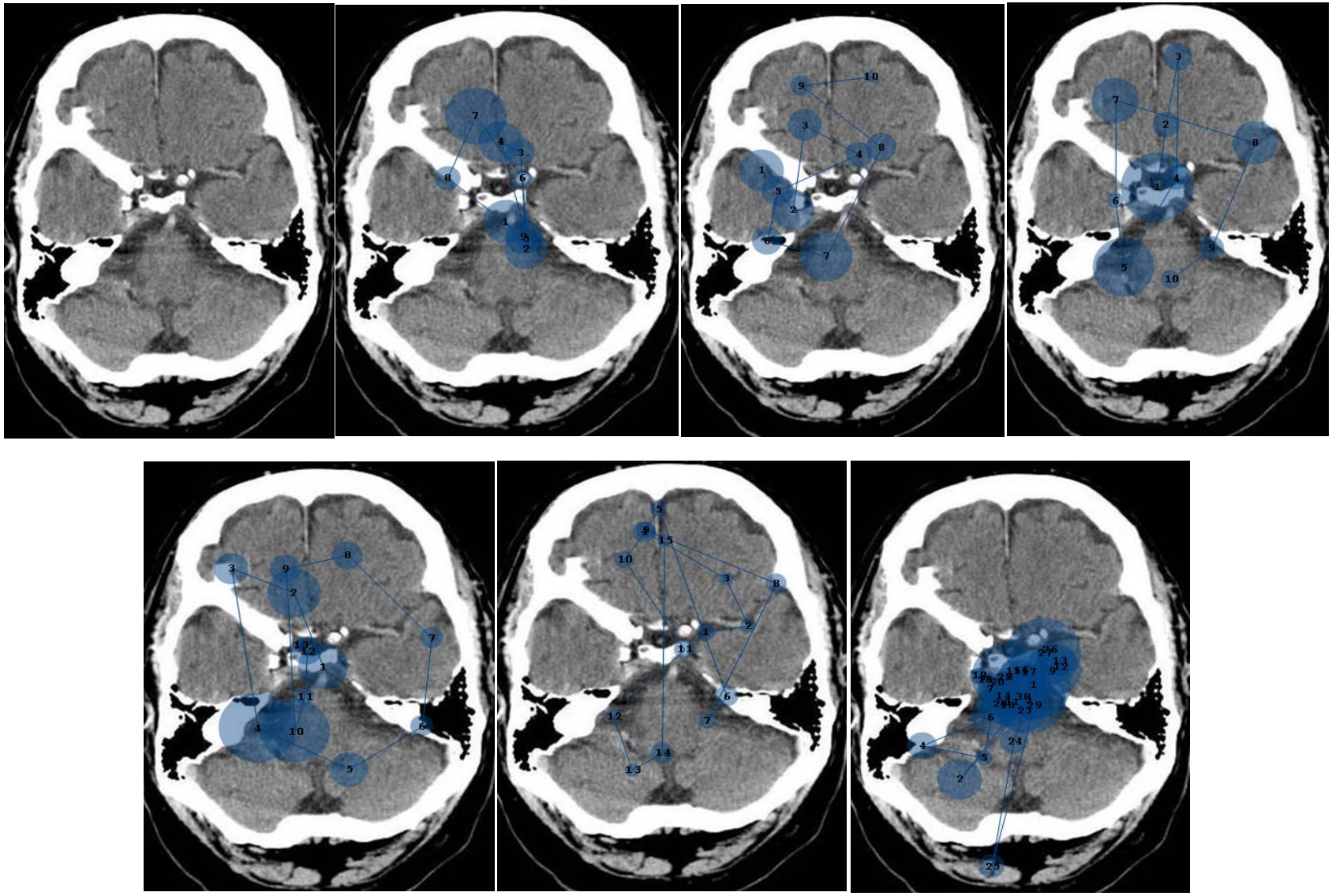


Appendix

CT gaze-tracker images of trainee readers (case CRW)

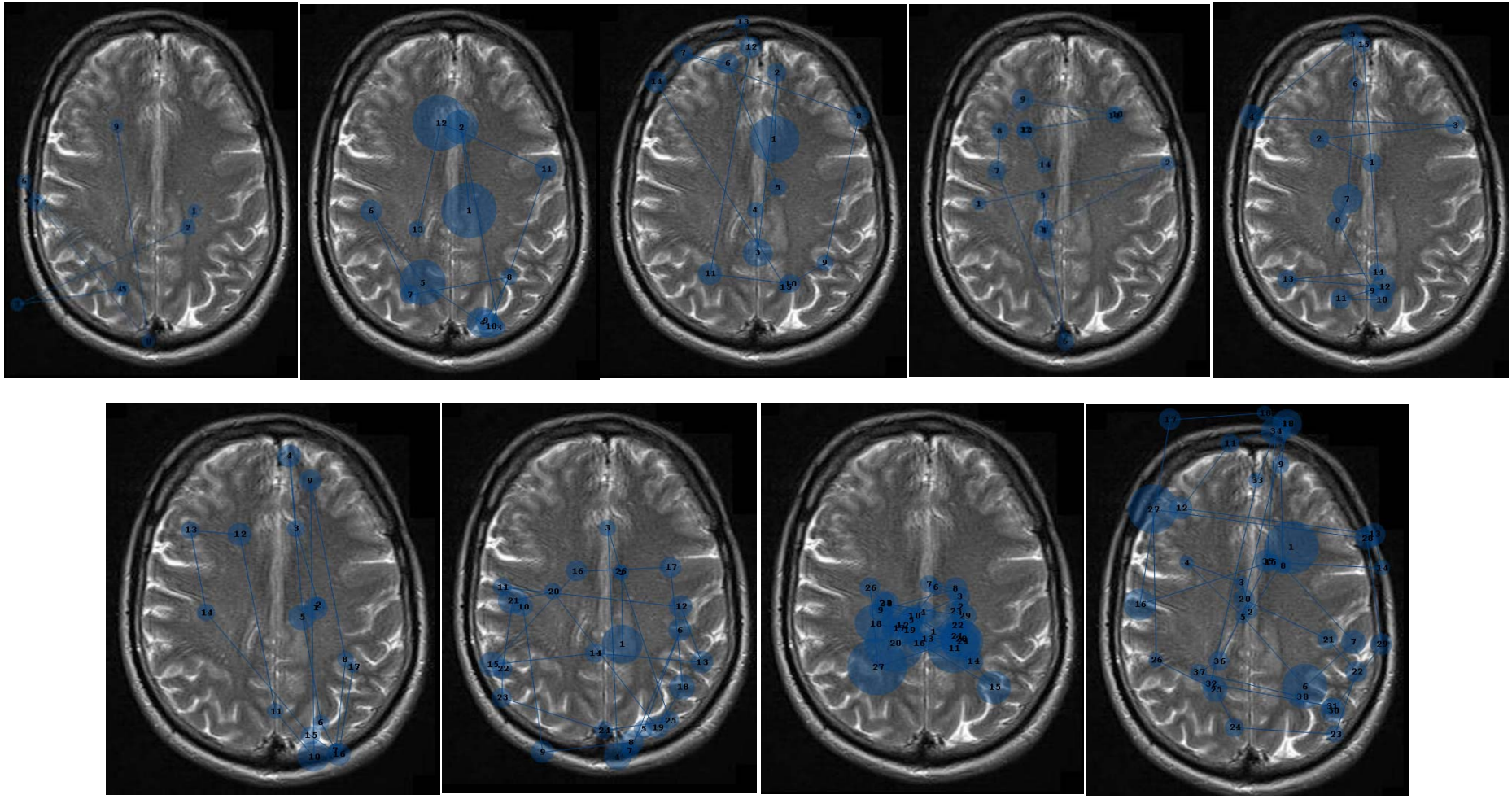


CT gaze-tracker images of expert readers (case CRW)

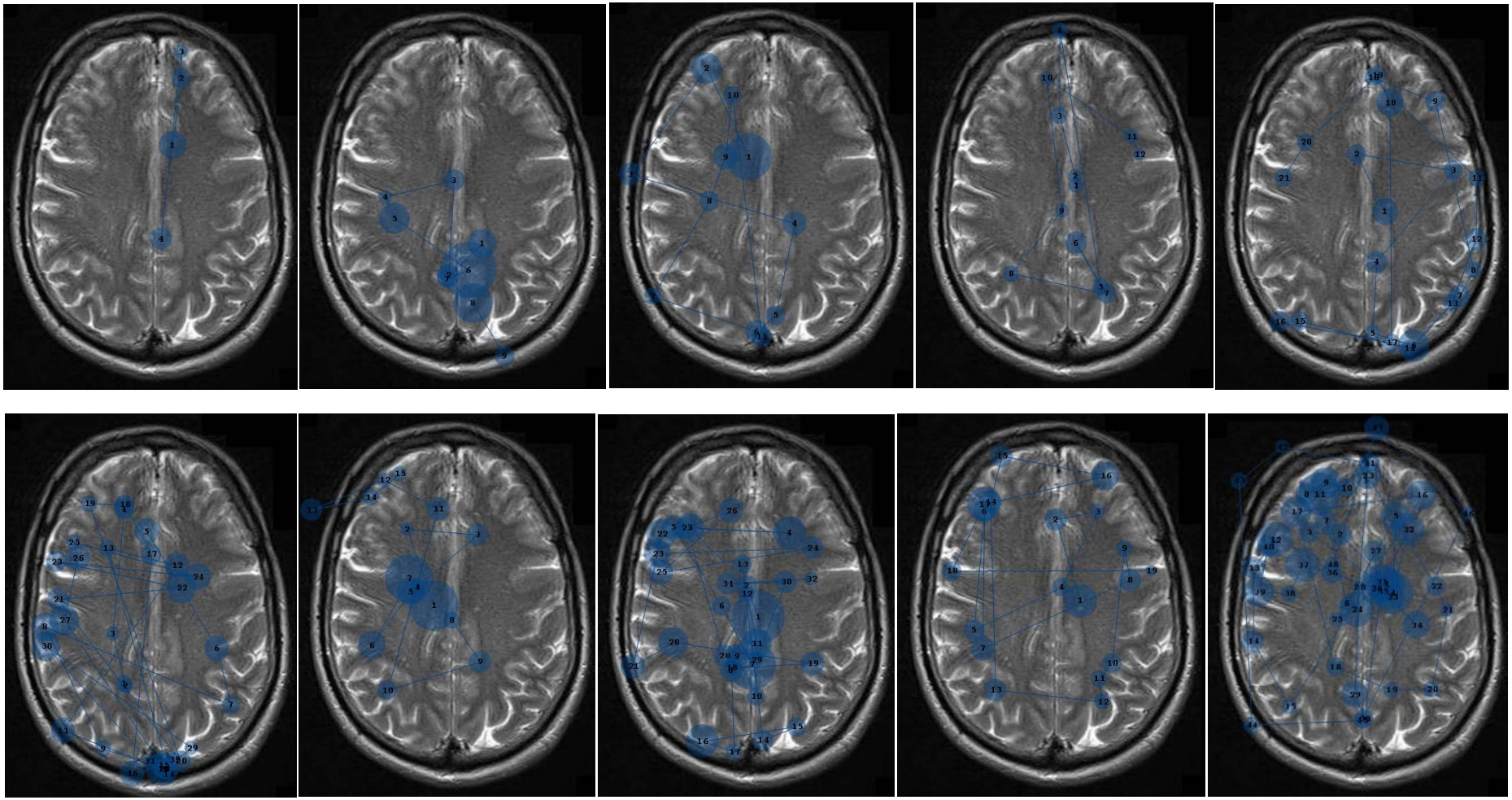


Appendix

MRI gaze-tracker images of novice readers (case NHG)

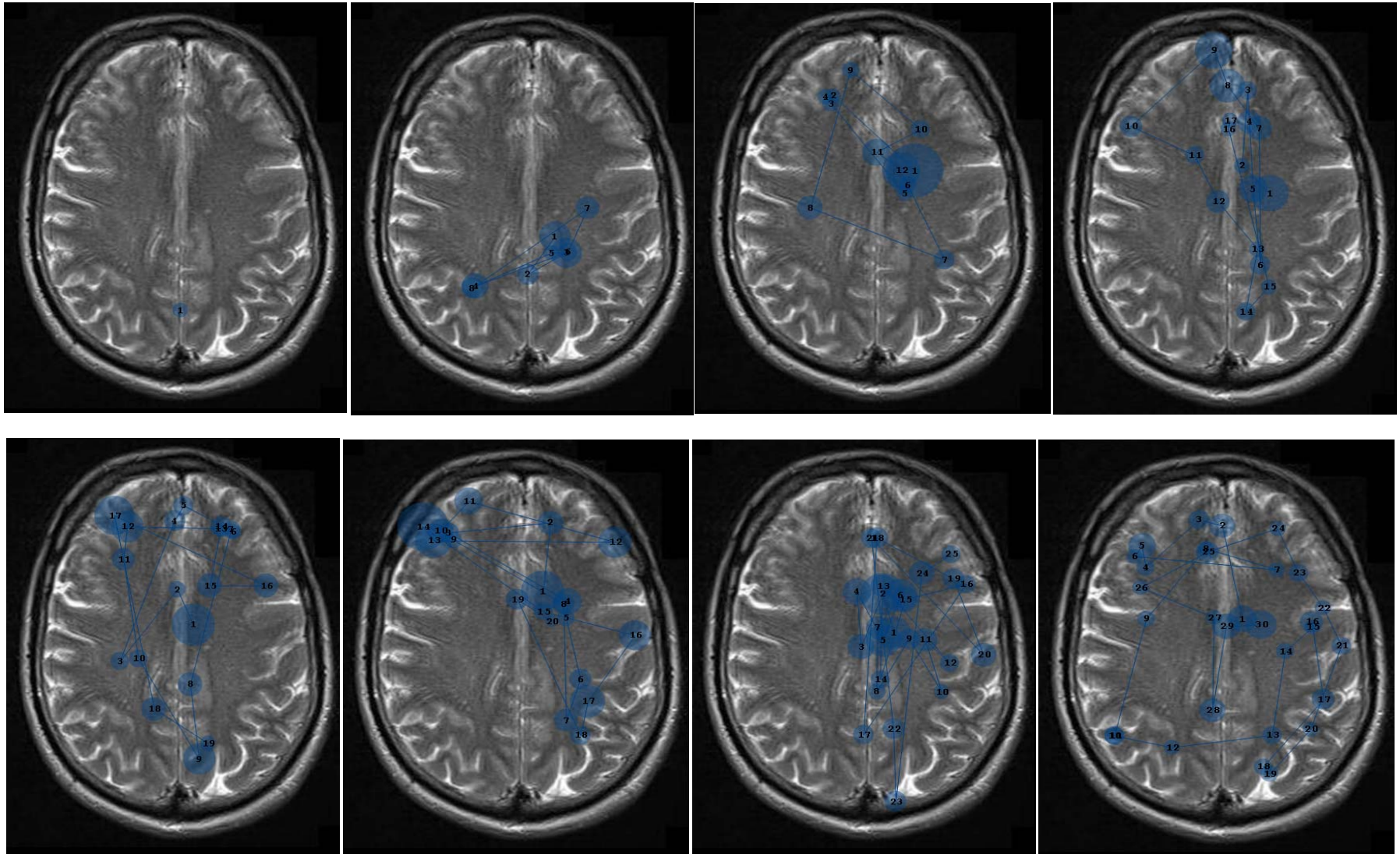


MRI gaze-tracker images of trainee readers (case NHG)

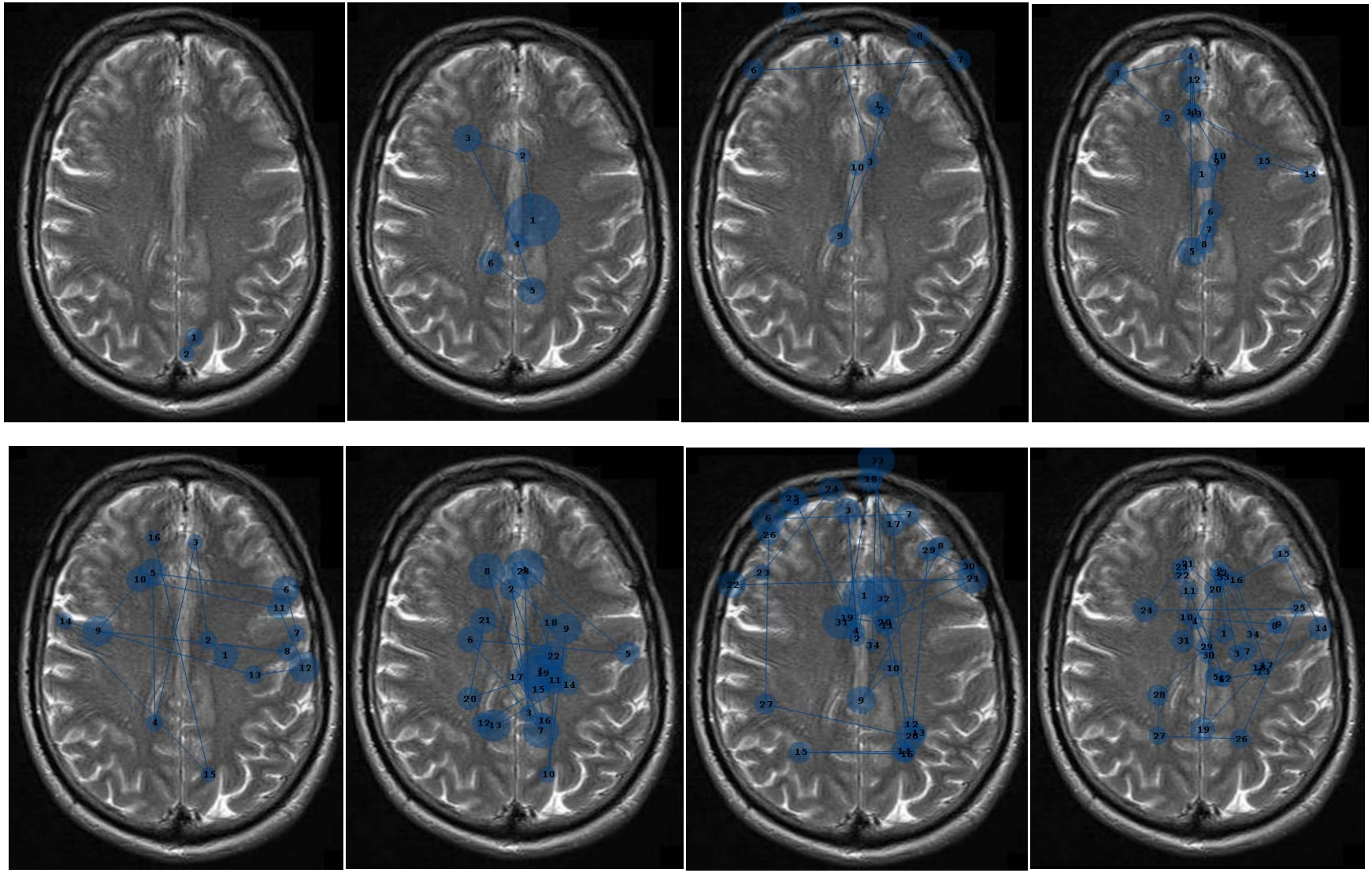


Appendix

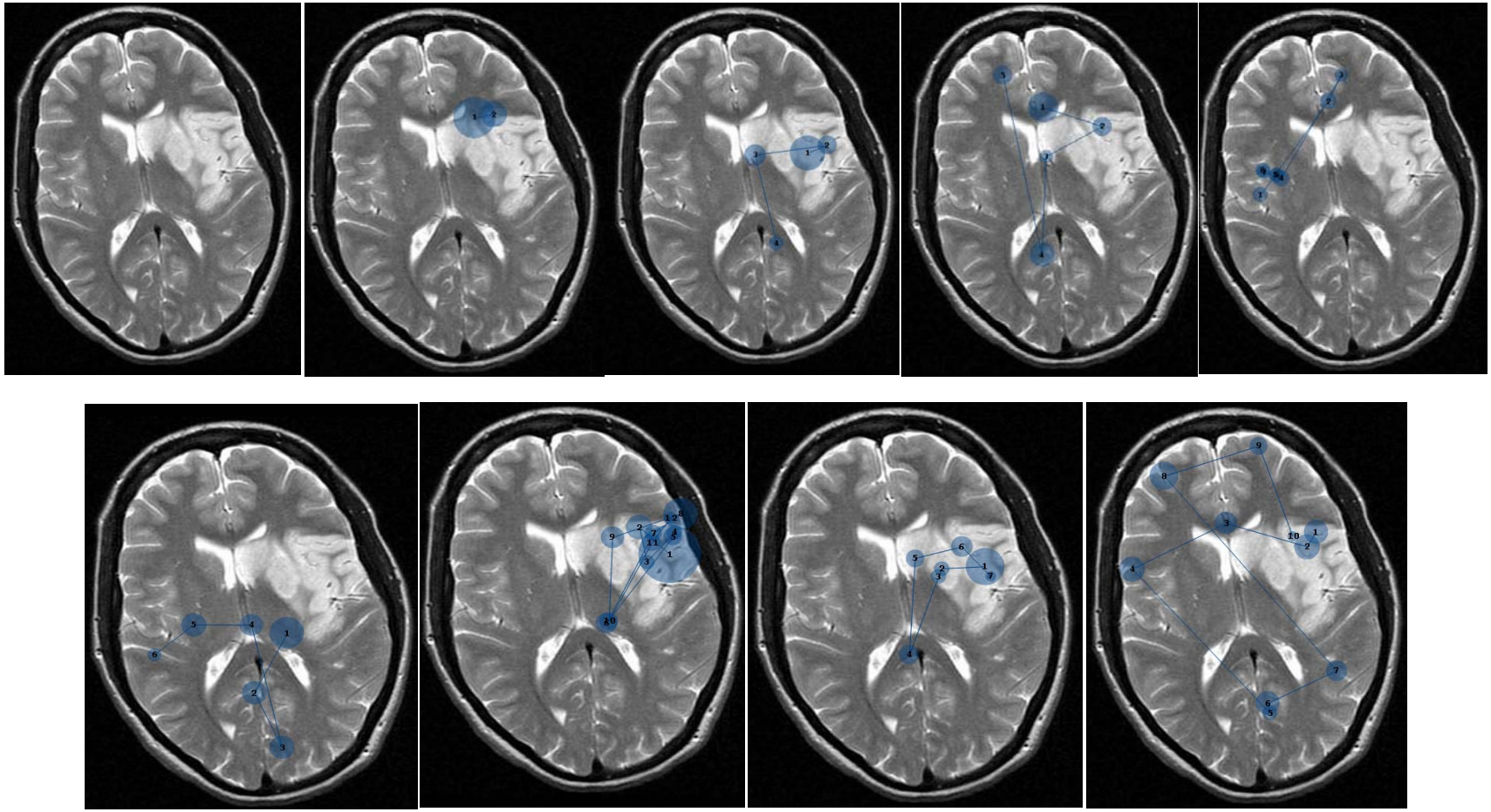
MRI gaze-tracker images of expert readers (case NHG)



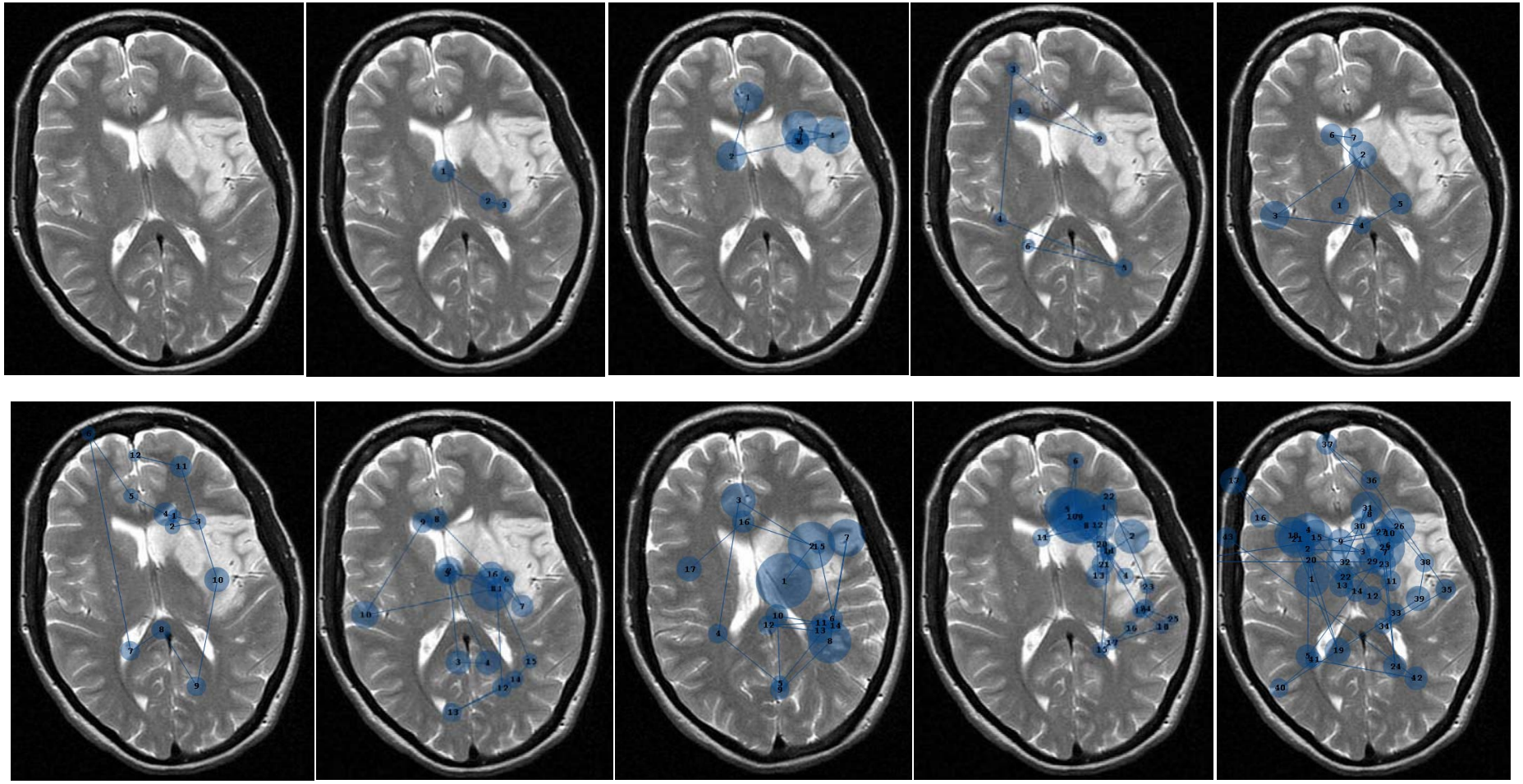
MRI gaze-tracker images of neurology readers (case NHG)



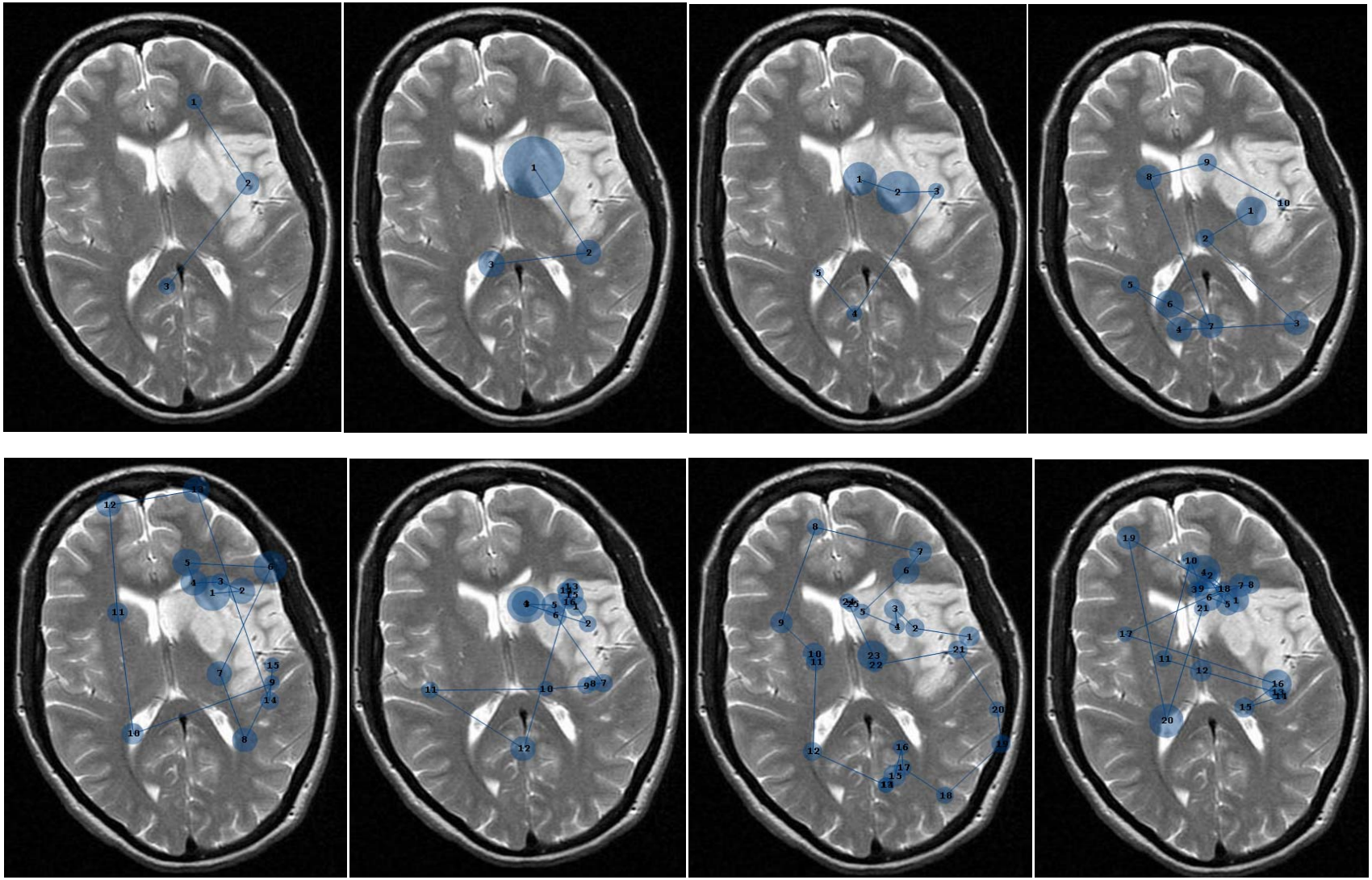
MRI gaze-tracker images of novice readers (case ACW)



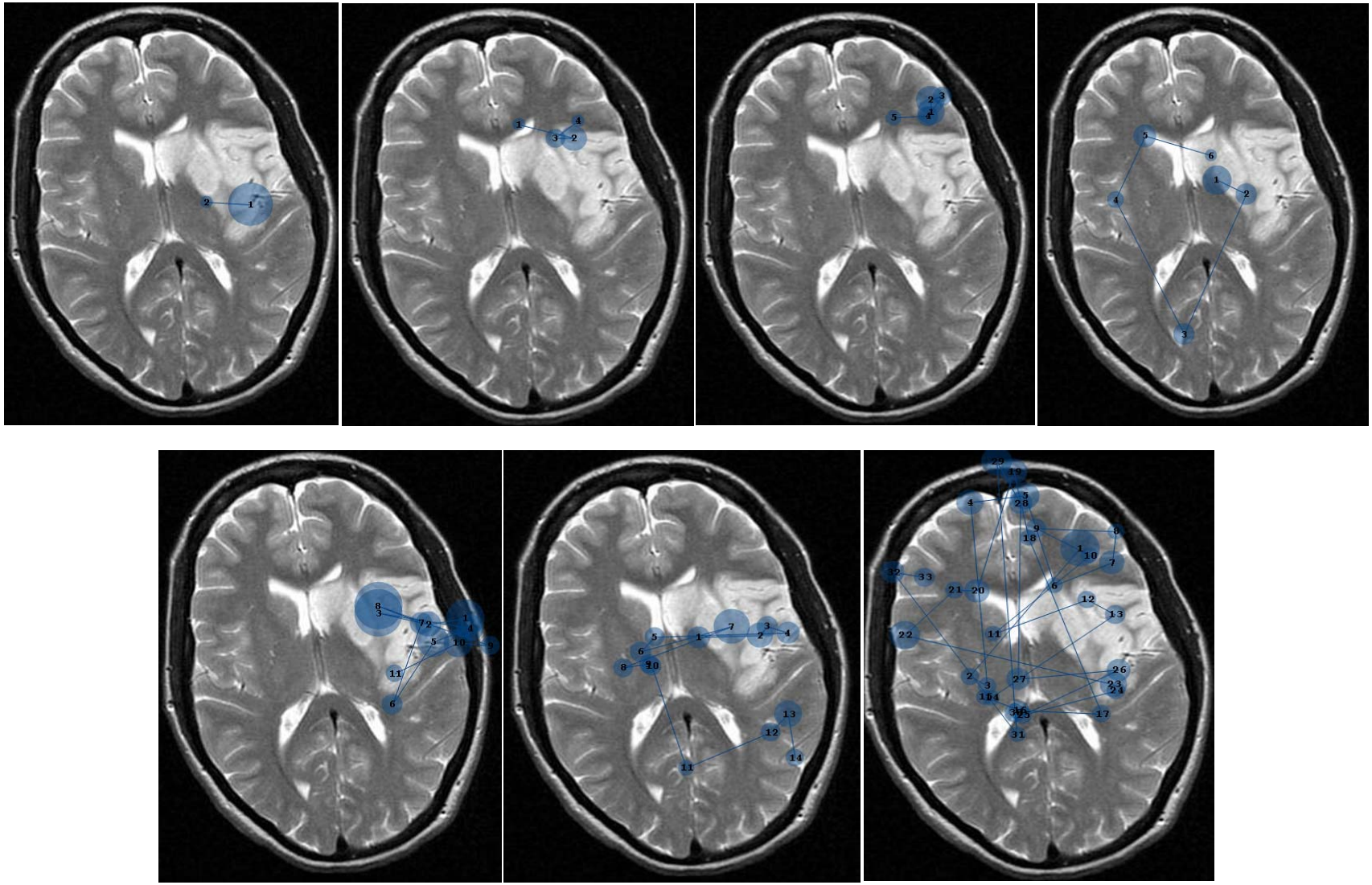
MRI gaze-tracker images of trainee readers (case ACW)



MRI gaze-tracker images of expert readers (case ACW)

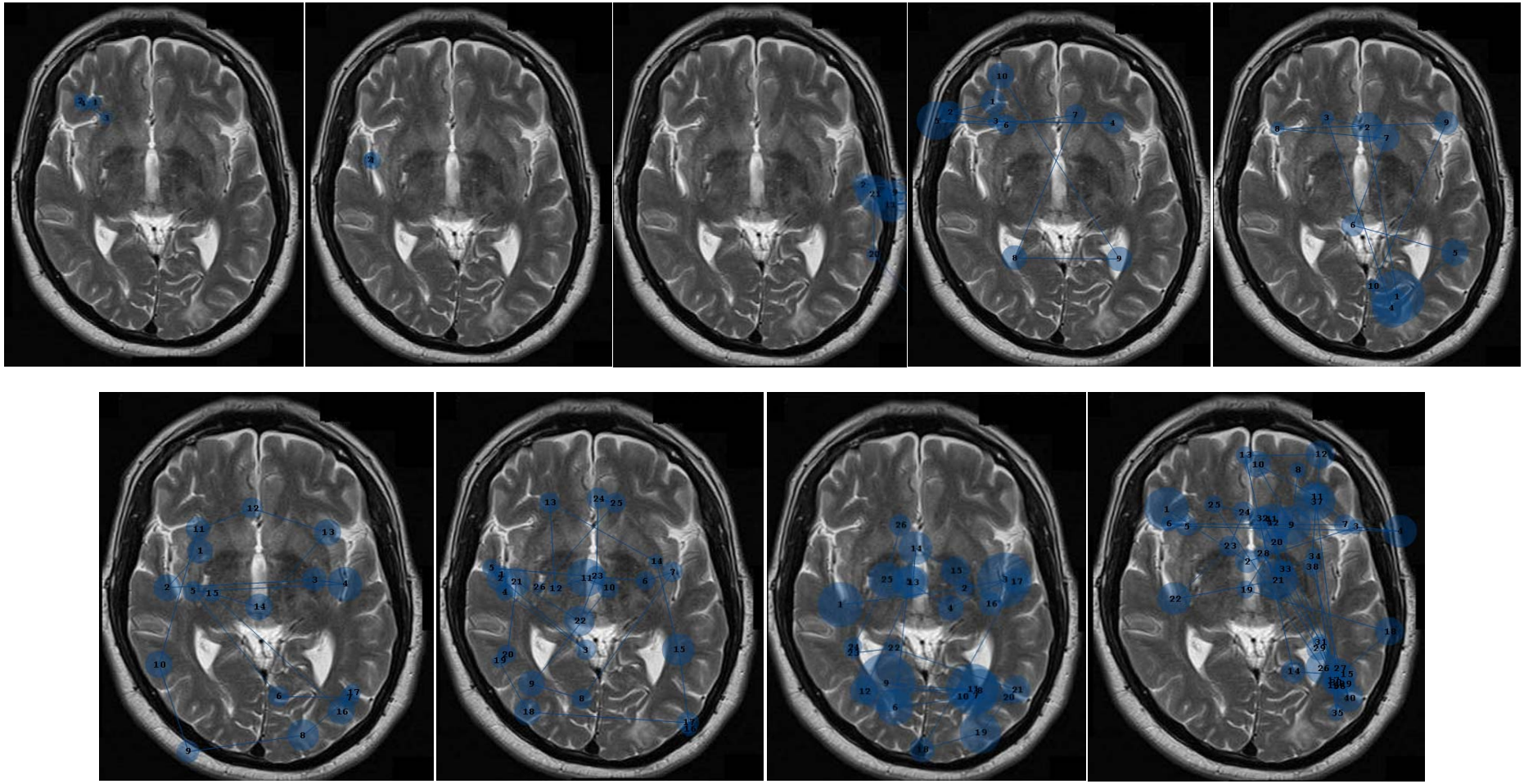


MRI gaze-tracker images of neurology readers (case ACW)

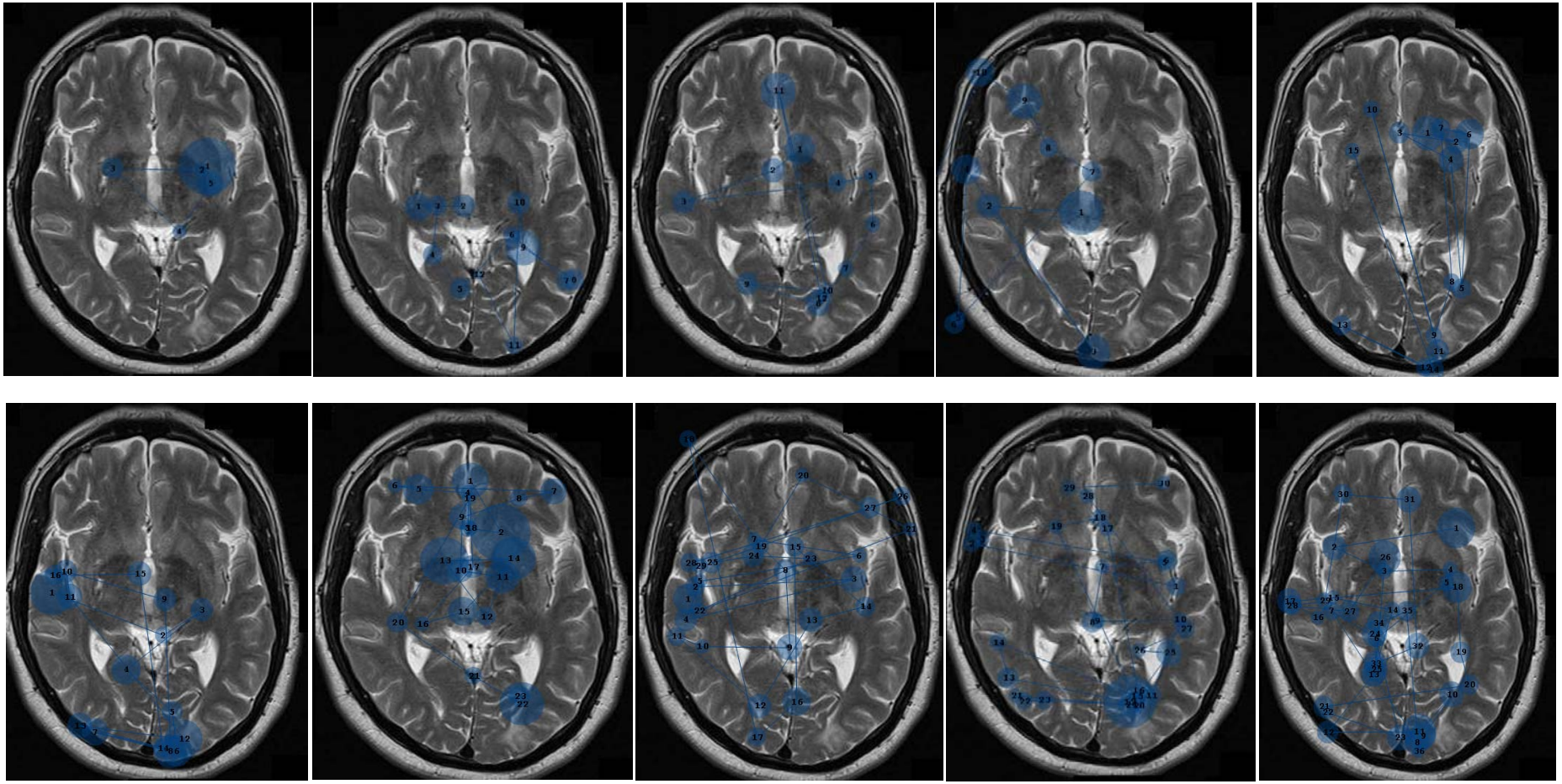


Appendix

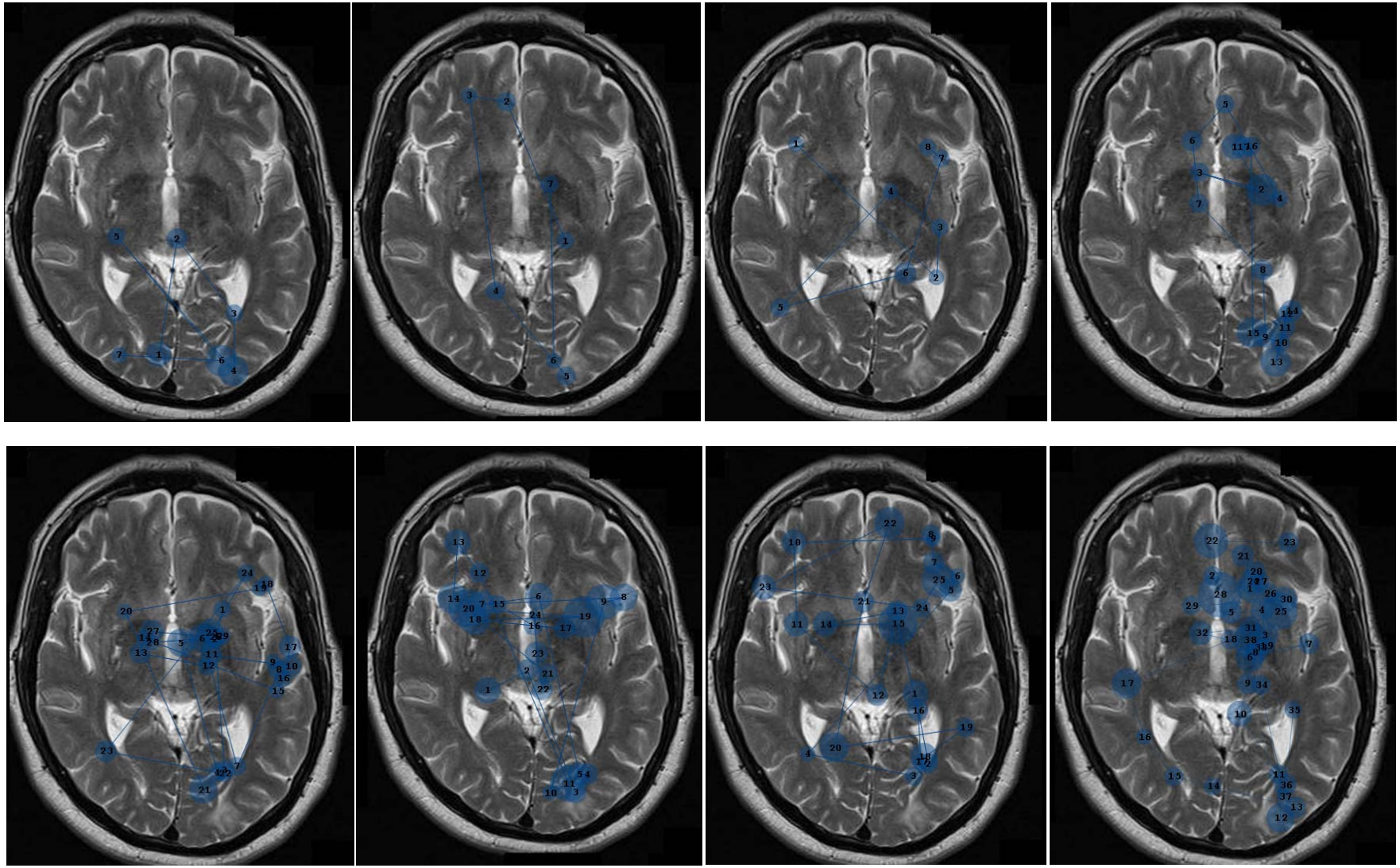
MRI gaze-tracker images of novice readers (case SAC)



MRI gaze-tracker images of trainee readers (case SAC)

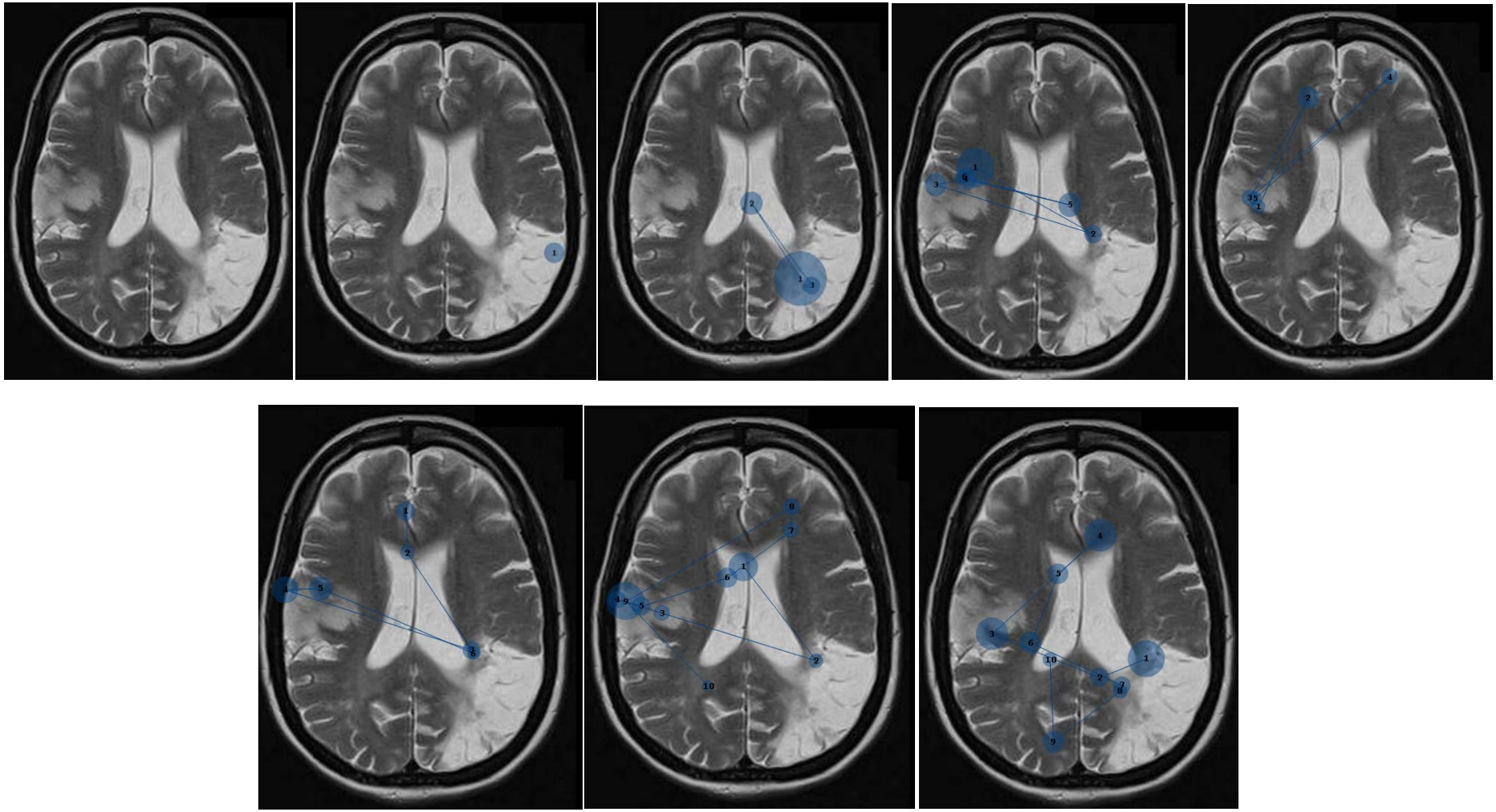


MRI gaze-tracker images of expert readers (case SAC)

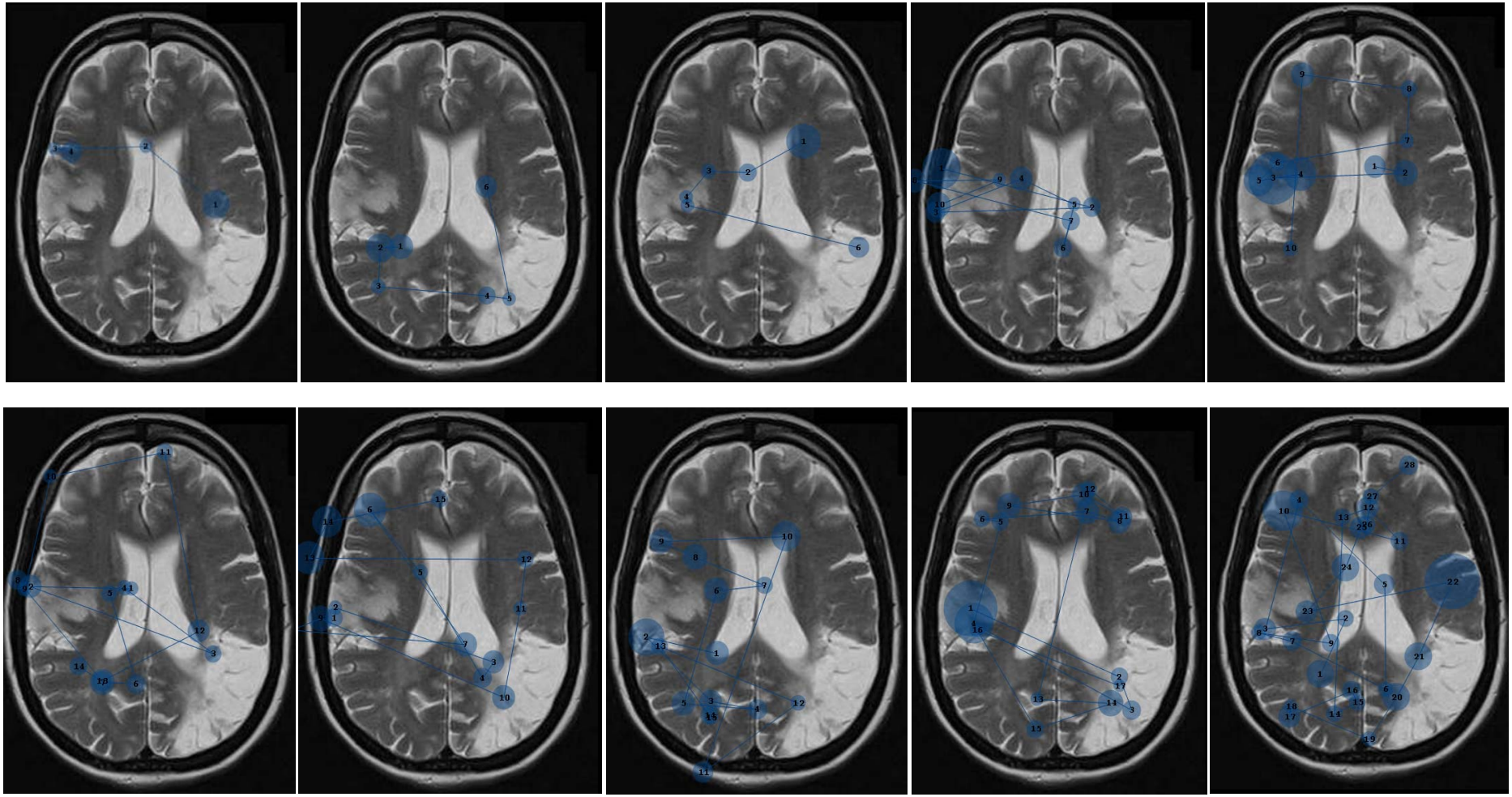


Appendix

MRI gaze-tracker images of novice readers (case CSS)



MRI gaze-tracker images of trainee readers (case CSS)



MRI gaze-tracker images of expert readers (case CSS)