Examining Neuroradiologists' Task-Specific Diagnostic Confidence in Reporting Acute and Chronic Stroke for Retrospectively Accelerated 0.5 T MR Images

by

Michelle Christina Pryde

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

at

Dalhousie University
Halifax, Nova Scotia
July 2022

Table of Contents

List of Tables

List of Figures

Abstract

Accelerated MRI is key for emergency medicine situations like acute ischemic stroke (AIS). Though acceleration alters image quality, neuroradiologists may remain able to report stroke pathology. This thesis examines trade-offs between acceleration factor (R) and neuroradiologists' task-specific diagnostic confidence in reporting AIS and chronic stroke, and assesses correlations between confidence and image quality metrics (IQMs).

18 participants were scanned using 0.5 Tesla MRI. Image data were retrospectively undersampled (R = 1-7X) and reconstructed via compressed sensing. Diagnostic confidence was ranked (1-5 Likert scale) and was correlated to IQMs via non-linear regression modelling.

Neuroradiologists' confidence remained high at R = 7X ($p > 0.05$) for AIS but decreased at R = 3X ($p < 0.05$) for chronic stroke. No IQMs correlated with confidence for AIS but all correlated to various degrees with confidence for chronic stroke, suggesting IQM performance does not necessarily indicate an image's usefulness for a specific diagnostic task.

List of Abbreviations and Symbols Used

AC: Agreement Coefficient (Gwet's)

AC1: unweighted Agreement Coefficient (Gwet's)

AC2: quadratic weighted Agreement Coefficient (Gwet's)

ADC: apparent diffusion coefficient

AHA: American Heart Association

AIS: acute ischemic stroke

ASA: American Stroke Association

CI: confidence interval

CS: compressed sensing

CSF: cerebral spinal fluid

CT: computed tomography

CTA: computed tomography angiography

CTP: computed tomography perfusion

DICOM: Digital Imaging and Communications in Medicine

DWI: diffusion weighted imaging

ED: emergency department

FDA: Food and Drug Administration

FLAIR: fluid attenuated inversion recovery

FOV: field of view

FSE: fast spin-echo

FSIM: Feature SIMilarity

FT: Fourier transform

GRAPPA: GeneRalized Autocalibrating Partial Parallel Acquisition

IQM: image quality metric

IQR: interquartile range

MR: magnetic resonance

MRA: magnetic resonance angiography

MRI: magnetic resonance imaging

MSE: mean squared error

NaN: not a number

NCCT: noncontrast CT

NEX: number of excitations

NQM: noise quality measure

NSH: Nova Scotia Health

p: polynomial decay rate

Pa: observed agreement

Pe: expected agreement by chance (*also*, chance agreement)

PFI: partial Fourier imaging

PI: parallel imaging

POC: point of care (*also,* point-of-care)

Q1: first quartile

Q2: second quartile (*also*, median)

Q3: third quartile

R: undersampling factor (*also*, acceleration factor)

RMSE: root mean square error

SENSE: SENSitivity Encoding

SNR: signal-to-noise ratio

SOC: standard of care (*also,* standard-of-care)

SROCC: Spearman rank order correlation coefficient

SSIM: Structural SIMilarity

SSR: sum of squares residuals

SWI: susceptibility weighted imaging

T: Tesla

T1: longitudinal relaxation time

T2: transverse relaxation time

T2*: effective transverse relaxation time

TE: echo time

TI: inversion time

TOF: time-of-flight

TR: repetition time

UCLA: University of California, Los Angeles

VIF: visual information fidelity

κ: Cohen's kappa

κ_{QW}: quadratic weighted Cohen's kappa

κ_{UW}: unweighted Cohen's kappa

Acknowledgements

I would like to provide a sincere and warm thank you to the following for their expertise and support throughout my degree: my supervisor, Dr. Steven Beyea; my unofficial (official, in my opinion) co-supervisor, Dr. James Rioux; my committee members: Steven, James, and Dr. Tim Bardouille; my lab mates, Sarah Reeve, Taylor Bouchie, Ally Klassen, and Robert Weaver; my collaborators, including Dr. Adela Elena Cora, Dr. Matthias H. Schmidt, Dr. David Volders, Mohammed Abdolell, and Allister Mason; the BIOTIC team, especially Beverly Lieuwen and Dr. Chris Bowen; Dalhousie University's School of Biomedical Engineering (BME), including faculty, staff, and fellow peers and alumni of the BME program, as well as the Faculties of Graduate Studies (FGS), Engineering, and Medicine, and the Department of Diagnostic Radiology; Nova Scotia Health and the Dalhousie Student Health and Wellness Clinic; and last, but most definitely not least, my wonderful partner, family, and friends (and their dogs ☺)! You all played an integral part, and I cannot express my appreciation enough. Thank you!

I would also like to acknowledge the provided funding sources for the research work of this thesis, as follows: grants from Research Nova Scotia, Natural Sciences and Engineering Research Council of Canada (NSERC) Discovery program, and Brain Repair Centre Knowledge Translation program; scholarships from NSERC Canada Graduate Scholarships – Master's program, Exxon Mobil Canada Ltd. Post-Graduate Scholarship, and A.S. Mowat Prize; prizes from BME, as well as the Image-Guided Therapeutics and Diagnostics Symposium, and helpful cost-offsetting stipends from FGS, as well as the International Society for Magnetic Resonance in Medicine.

Finally, special thanks to Dr. Geoff Maksym for chairing my thesis defence and to my examining committee, Steven, James, Tim, and Kim, for making my defence a challenging, yet fun, experience!

Chapter 1: Introduction

This chapter includes a description of the stroke pathology of interest in this thesis, with particular focus on diagnosis and treatment of acute ischemic stroke (AIS). As the current standard-of-care (SOC) diagnostic imaging modalities for rapid diagnosis of AIS, trade-offs between computed tomography (CT) and conventional clinical magnetic resonance imaging (MRI) systems (those at 1.5 Tesla (T) or 3 T main magnetic field strengths) will be explored, segueing into the motivation for this research work via a 0.5 T MRI system and setting the framework for the main objectives and hypotheses of this thesis.

1.1. Stroke

Stroke is defined as brain tissue deprived of vascular perfusion that can therefore infarct (i.e. die due to obstruction of local circulation) due to the tissue being starved of oxygen and nutrients [1] and can be both acute and chronic. There are two overarching types of acute stroke [2]: ischemic stroke, which is due to a blood clot, and hemorrhagic stroke, which is due to a bleed.

For the purposes of this thesis, chronic stroke may be defined as chronic ischemic lesion burden, which may otherwise be known as white matter disease. Chronic stroke tends to be more diffuse compared to acute stroke and does not occur precipitously (i.e. suddenly, progressing rapidly). Chronic stroke may occur as a result of acute stroke or other risk factors, such as aging. With aging, small blood vessels that feed the brain's white matter can occlude (i.e. close) or sclerose (i.e. harden/thicken), resulting in a

narrowing of their lumen (i.e. space where blood flows through vessel) over time, which can also starve surrounding brain tissue from nutrients and oxygen.

Both acute and chronic stroke are clinically important. However, acute stroke are significantly more time-sensitive because they have the potential to be treated to minimize their lasting damage or potential for causation of death. For chronic stroke, assessing age and other risk factors is important to their treatment regime, but they do not carry the same burden of time-sensitivity. Since chronic stroke is not typically considered a medical emergency, it will therefore be secondary in importance to this introductory discussion.

1.1.1. Acute Ischemic Stroke

Of the two types of acute stroke, this thesis work focuses on AIS. AIS is more common than hemorrhagic stroke, of which the latter occurs in less than approximately one quarter of stroke cases [1] [2]. AIS tends to be quite focal within brain tissue, occurs precipitously, and can range in severity [2]. Severity is not to be confused with seriousness, however; all AIS are serious and require immediate medical attention.

There are two types of AIS, as follows: (1) embolic and (2) thrombotic (Fig. 1). With embolic AIS, a blood clot formed elsewhere in the body travels up through the bloodstream to the brain [3]. If the blood vessel in the brain has a decreased cross-sectional luminal area, the clot may lodge (i.e. embolism), resulting in restricted vascular perfusion. On the other hand, thrombotic AIS is due to arterial blockages resultant from atherosclerosis, particularly in the major arteries of the neck that provide blood to the brain [3]. Atherosclerosis is the build-up of plaques, or fatty deposits, along the inner

wall of the vessel [4], which cause the vessel wall to occlude. If these plaques rupture, this can trigger the process of clot formation [4] (i.e. thrombosis), also resulting in restricted vascular perfusion.

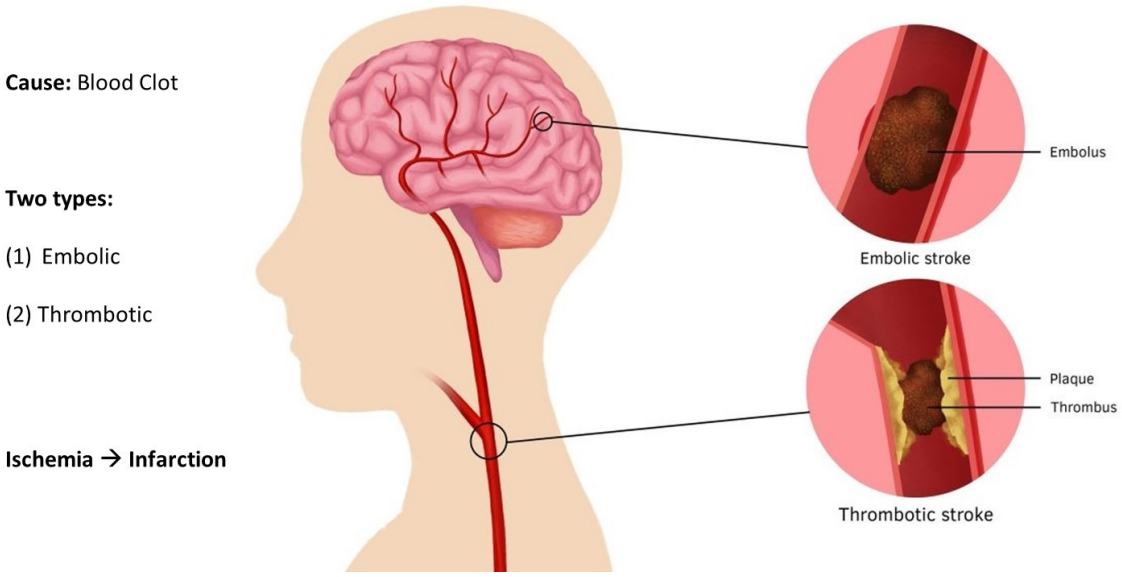


Fig. 1. Acute ischemic stroke illustration from website [4] with added text on left.

The result of either an embolic or thrombotic AIS is therefore ischemia at the embolic or thrombotic site, known as the infarct core, while brain tissue surrounding the infarct core may be hypoperfused and, as a result, not yet infarcted [1]. Infarction ensues once brain tissue is starved from vital oxygen and nutrients, thus hypoperfusion may lead to infarction. However, hypoperfused brain tissue, known as the *ischemic penumbra*, is salvageable due to collateral blood flow [1], whereas infarcted brain tissue is irreversibly damaged [1]. Therefore, rapid diagnosis and treatment of acute stroke is time-critical to achieving better patient outcomes by preventing infarction and salvaging hypoperfused tissue.

1.1.2. A Medical Emergency

The global impact of stroke is such that it is the second leading cause of mortality [5] and one of the major leading causes of acquired disability [6]. In Canada, stroke is the second leading cause of dementia in the form of vascular dementia [7] and the third leading cause of mortality [8]. Roughly over three quarters of stroke are AIS [9], and, according to the American Heart Association (AHA), the prevalence of stroke was 101.5 million people, globally, in 2019 [9]; this means that at least as many as 76 million people had AIS in 2019, alone.

"Time is brain" is an adage, coined by Gomez [10], which points to the time-sensitivity of AIS such that delays in diagnosis and treatment lead to destruction of neurons, synapses, and myelinated fibers [11] in the brain. For example, Kunz *et al.* [12] concluded that every 10-minute delay in treatment yielded an average 39 days of life with disability, which gives context to the degree of acuteness associated with AIS, elucidating it as a medical emergency.

To minimize their negative impact, medical emergencies like AIS require rapid diagnosis in order for treatment procedures to be identified and administered as early as possible. This thesis therefore focuses, primarily, on the opportunity to diagnosis AIS rapidly in order to limit the negative impact of AIS. Given the high global prevalence of occurrence of AIS out of all stroke, research into the field of rapid diagnosis of AIS is warranted: the more rapid the diagnosis of AIS, the faster the patient can be triaged to receive the appropriate treatment, which is critical to both limiting and reversing the

extent of damage due to brain tissue infarct [1], thereby greatly improving patient outcomes.

1.2. Diagnosis and Treatment of Acute Ischemic Stroke in Emergency Medicine

In this section (*Section 1.2.*), existing clinical workflows to diagnose AIS will be addressed, focusing on rapid diagnostic imaging SOCs. Relevant magnetic resonance (MR) stroke imaging protocols will then be discussed in greater detail to aid in the justification, later on in this thesis, as to why certain imaging protocols were the focus of this research work. A brief discussion of AIS treatment follows.

1.2.1. Existing Clinical Workflows

Existing clinical workflows for the diagnosis and treatment of AIS involve a coordinated multi-step collaboration between the many different healthcare professionals involved. From the time paramedics are called and/or the patient enters the doors of the emergency department (ED), the medical centre's stroke team [13] is notified, and their stroke protocol is activated. The suspected stroke patient will be triaged as emergent [13] and will immediately proceed to assessment.

As part of the standard stroke protocol, a patient with suspected stroke is not triaged to receive the appropriate treatment without a diagnosis of AIS, which is largely in part provided by diagnostic brain imaging modalities, such as CT and/or MRI, since these modalities may be used to report stroke pathology. Due to the emergent nature of AIS, it is critical that diagnostic imaging be rapid in order to proceed, as quickly as possible, to the necessary treatment.

1.2.2. Imaging Standards-of-Care for Rapid Diagnosis

According to the most up-to-date recommendations for emergency evaluation of patients with AIS from Canadian Stroke Best Practices by J. M. Boulanger *et al.* [13],

> All patients with suspected acute stroke should undergo brain imaging with non-contrast CT or MRI [13].

The most up-to-date guidelines for emergency evaluation of patients with AIS from the AHA/American Stroke Association (ASA) by Powers *et al.* [14] state,

> All patients with suspected acute stroke should receive emergency brain imaging evaluation on first arrival to a hospital before initiating any specific therapy to treat AIS [14].

Interestingly, [14] was updated from the previous AHA/ASA guidelines only a year prior which stated,

> All patients admitted to hospital with suspected acute stroke should receive brain imaging evaluation on arrival to hospital. In most cases, noncontrast CT (NCCT) will provide the necessary information to make decisions about acute management [15].

Although [15] doesn't make an up-front mention to use CT *or MRI* like the Canadian guidelines [13], it does reference select patients where MRI may be part of the protocol for diagnosis and treatment of AIS – but otherwise the guidelines [15] are not highly recommending of MRI and are therefore restrictive of its suggested use in this context. The updated AHA/ASA guidelines state more generally, recommending the use of emergency brain imaging [14] in the same section that used to specifically recommend noncontrast CT (NCCT) [15]. The change stems from the fact that the new guidelines pointedly recommend, in an equally non-restrictive way, the use of CT or MRI for emergency evaluation. The updated recommendations from [15] to [14] within the span of one year speaks to the fact that this is a rapidly evolving field as active investigations are ongoing. Despite these recommendations and guidelines, however, in most United

States and Canadian EDs, the reality is such that the current modality serving as the SOC for first-line imaging of indications and subsequent rapid diagnosis of AIS is CT. In fact, [13] acknowledges,

> In most Canadian centers, a CT approach may be more practical and more readily available than an MR approach. Choice of imaging modality should be based on most immediate availability and local resources [13].

A typical clinical workflow in the context of CT as first-line imaging includes rapid assessment of the suspected stroke patient by various standard means, ordering an emergent CT scan of the brain, and ensuring immediate access to the CT scanner [13]. The patient is transported to the CT scanner, which is often located at the point of care (POC) in the ED. The CT protocol commonly used for imaging of acute stroke first acquires images via NCCT, which are primarily sensitive to detection of hemorrhagic stroke and, to a lesser extent [16], sensitive to detection of AIS. Although hemorrhagic stroke is less common than AIS, hemorrhagic stroke is ruled out first via NCCT [1] in order to dictate the appropriate treatment. NCCT may also help to rule out stroke mimics [1], examples of which include infections or brain tumours. Depending on the clinical situation and/or the medical centre, some CT protocols for imaging of acute stroke will acquire images additional to NCCT, such as CT angiography (CTA) and/or CT perfusion (CTP) [1].

The standard CT protocol for imaging of acute stroke at Nova Scotia Health (NSH) [17], for example, includes NCCT, CTA, and CTP. Typically, if NCCT yields positivity for hemorrhagic stroke then the patient may or may not be imaged further via CTA prior to proceeding immediately to the appropriate treatment. Alternatively, if NCCT yields

positivity for AIS then the patient would be imaged further via CTA and/or CTP prior to proceeding immediately to the appropriate treatment. Assuming hemorrhagic stroke has been excluded, if NCCT is negative for AIS, but AIS is still strongly suspected based on clinical presentation, then the patient may be imaged further via CTA and/or CTP. In many cases, however, despite AIS being strongly suspected, the acquired CT images (whether NCCT, CTA and/or CTP) may remain negative for AIS; if this is the case then subsequent diagnostic imaging, using conventional clinical MRI systems, is usually necessitated.

Generally, a patient whose CT or MRI scan yields positivity for AIS will be admitted, post-treatment, as an inpatient to the stroke unit for monitoring [14]. Additionally, a patient whose CT images yielded negativity for AIS, but who remains a suspected AIS patient and is awaiting an MRI scan, will also typically be admitted as an inpatient to the stroke unit for monitoring. In the latter case, however, there are challenges associated with scheduling conventional clinical MRI scanners. Scheduling challenges, due to well-known waitlists for accessing MRI exams, commonly cause suspected AIS patients to experience delays in accessing a clinical MRI scan and, consequently, to experience delays in accessing treatment [13] in the case their clinical MRI scan yields positivity for AIS. Further to this point, longer waits to receive clinical MRI scans yield higher rates of inpatient bed occupancy, which reduces capacity for other patients who require inpatient admission to the stroke unit and increases associated monetary costs [71], [72]. However, all of this assumes that the medical center is not only equipped with an MRI machine, but that there is timely access to it. As a result, if MRI is not available then

patients with suspected AIS may be treated without undergoing MRI subsequent to CT. Therefore, some stroke protocols and/or medical centres will proceed to treatment of a suspected AIS without further imaging solely based on clinical presentation, despite the CT scan remaining negative for AIS [14]. However, studies show that imaging provides the necessary information to radiologists, such that the patient receives a better diagnosis and treatment, as well as an improved prognosis [13], [73].

Outside of North America, however, some EDs have deemed MRI, instead of CT, to be the primary imaging modality and thus SOC for rapid diagnosis of AIS. For example, as early as 2009, the French National Authority for Health's clinical practice guidelines for early management of stroke [18] state,

> MRI is the most effective examination for early diagnosis of signs of recent ischemia and it can also show intracranial bleeding. It should be the preferred course of action**. If MRI is possible as a first-line examination, it should be accessible as an emergency procedure and short protocols should be used including the following sequences: diffusion, FLAIR, gradient echo (grade B).** If emergency access to MRI is not possible, a cerebral CT scan should be carried out. This examination does not consistently show signs of recent ischemia, but it can be used to view intracranial haemorrhaging [18, p. 10, emphasis added].

Additionally, as referenced above, some North American EDs *are* following suit, using MRI as the imaging modality of choice defined in their acute stroke protocols. For example, University of California, Los Angeles (UCLA) Health [19] states:

> At UCLA, all acute patients should go to MRI first, if available within 15 minutes of arrival and no contraindication to MR, ordering "Interventional Stroke MRI Protocol". If MRI not available within 15 minutes, all acute patients should instead go to multimodal CT, ordering "Reduced contrast dose Stroke CT protocol" [19, Sec. I.D.].

It is interesting to note that the protocols from both France and UCLA Health make reference to the preference of MRI over CT with the caveat that the MRI be available. Furthermore, the Canadian guidelines [13] reaffirm that "decisions regarding MRI scanning should be based on MRI access, availability and timing of appointments" [13],

all of which can be limiting factors to choosing what they call an “MR approach” [13].

Despite the general consistency, however, in identifying MRI as more sensitive to CT in diagnosing AIS, the fact that CT is more commonly used as first-line imaging speaks to not only the current lack of dedicated emergency medicine POC MRI systems, but also to the aforementioned scheduling challenges currently associated with booking clinical stroke exams on conventional clinical systems.

Therefore, based on culminating the aforementioned guidelines, MRI complements the use of CT in cases where NCCT has not only ruled out hemorrhagic stroke but also yielded negativity for AIS despite AIS still being suspected due to clinical presentation, and where an MRI system is immediately available after imaging via NCCT. The rationale for *why* MRI might compliment CT in these cases is that, because hemorrhagic stroke is less common than AIS, a higher proportion of stroke cases will involve identifying the presence or absence of AIS once hemorrhagic stroke has been ruled out, and MRI is superior to CT in diagnostic sensitivity for AIS. Further, personal communication with neuroradiologists at NSH reaffirmed that if an AIS diagnosis can be confirmed via MRI, it is preferred, rather than proceeding to treatment for AIS solely based on clinical presentation and an AIS-negative NCCT in the absence of information provided by MRI.

1.2.3. Magnetic Resonance Stroke Imaging Protocols

Across MRI stroke protocols, there may be multiple and various different sequences with their parameters nuanced to the medical centre’s MRI system(s) and to their health care teams’ needs to detect and diagnosis AIS safely, timely, and accurately. However, it is standard for MRI stroke protocols to have the following sequences, each with

different contrasts in order to provide unique information: some form of gradient echo sequence, such as perfusion weighted imaging (PWI) or susceptibility weighted imaging (SWI); time of flight (TOF) MR angiography (MRA); transverse relaxation time (T2) fluid attenuated inversion recovery (FLAIR)**;** and diffusion weighted imaging (DWI) with its corresponding apparent diffusion coefficient (ADC) map [1]. Some medical centres may also use complementary sequences, examples of which may include additional T2-weighted images [1] [20] and/or longitudinal relaxation (T1)-weighed images [20]. For example, the MRI protocol for imaging of acute stroke at NSH includes axial SWI, oblique TOF MRA, axial T2 FLAIR, axial DWI/ADC, as well as axial T2 fast spin-echo (FSE) and sagittal T1 FLAIR. Specifically, axial T2 FSE is complementary to axial T2 FLAIR and sagittal T1 FLAIR is then complimentary to the T2-weighted images due to the different contrast (i.e. T1 versus T2) and orientation (i.e. sagittal versus axial). These differences help a neuroradiologist confirm whether something visualized by one specific contrast is in fact stroke pathology and not some other type of pathology or artifact, respectively.

Similar to CT stroke protocols, it is important for MRI stroke protocols to not only be able to screen for, and rule out, hemorrhagic stroke but of course also be able to diagnose AIS. Table I outlines some of the MRI stroke protocol sequences, why they are used, the main weighting they rely on, (e.g. T1, T2, effective T2 (T2*)), and the general length of their echo time (TE) and repetition time (TR) pulse sequence parameters.

TABLE I: Example of some MRI stroke protocol (a) sequences, (b) why they are used, (c) the main weighting they rely on and (d) the general length of their TE and TR pulse sequence parameters.

(a) sequence	(b) use	(c) weighting	(d) TE & TR length
SWI	▪ first sequence in protocol ▫ typically gradient echo ▪ screens for hemorrhagic stroke ▫ capable of visualizing hemorrhage (i.e. bleeding) and veins [1] [21] ▫ more sensitive MRI sequence for this	▪ T2*-weighting	TE: long [21] TR: short [21]
TOF MRA	▪ detection of AIS ▫ shows signal intensity contrast between flowing and stationary tissue ◦ ideal for imaging of vascular flow (e.g. visualization of any occluded or sclerosed blood vessels) [1]	▪Not Applicable	TE: short [28] TR: moderate [28]
T2 FLAIR	▪ detection of AIS ▫ shows signal hyperintensity when there is an ischemic infarct [1] ▫ aids in the identification of the stage of AIS disease progression (since ischemia can lead to infarction) ▪ assessment of chronic stroke (as it is defined in Section 1.1.) [1]	▪ T2-weighting ▫ suppresses otherwise bright signal from CSF making it appear dark	TE: long [67] TR: long [67]

TE = echo time; TR = repetition time; SWI = susceptibility weighted imaging; TOF MRA = time-of-flight magnetic resonance angiography; T2 = transverse relaxation time; FLAIR = fluid attenuated inversion recovery; T2* = effective transverse relaxation time; CSF = cerebral spinal fluid

When pulse sequence parameters, TE and TR, are listed as *long* or *short*, these lengths are with respect to the relaxation time (e.g. T1/T2/T2*) of interest to generate the desired contrast; they are not with respect to one another. In fact, TE is always shorter than TR for most sequences (for any conventional definition of these times). TE lengths primarily affect T2- and T2*-weighting such that, in general, lengthening TE will increase T2 or T2*-weighting. On the other hand, TR lengths primarily affect T1-weighting such that, in general, shortening TR will increase T1-weighting. As an example, T2 FLAIR is a T2-weighted sequence and thus TE and TR are lengthened. However, TR is lengthened not only to minimize T1-weighting but also to allow for inversion recovery of signal required to suppress otherwise bright signal from cerebral spinal fluid (CSF).

DWI is an additional MRI stroke protocol sequence and is both T2-weighted and diffusion weighted. DWI obtains contrast via sensitization to the ADC through the use of diffusion gradients. The sequence parameters that control the strength and timing of the diffusion-encoding gradients can be combined into a single number known as a b-

value typically on the order of 1000 s mm^{-2}, with higher b-values introducing more diffusion weighting [29], in addition to a matching acquisition at b = 0 s mm^{-2}. In AIS, DWI shows signal hyperintensity when there is restricted diffusion resultant from blood vessel occlusion, and parametric ADC maps calculated from the DWI data can be used to quantify the level of restricted diffusion [1]. In the presence of restricted diffusion as a result of ischemia and/or infarction signal hyperintensity on DWI typically corresponds with signal hypointensity on ADC maps. Note that DWI is most useful for identifying early AIS disease progression since the images can visualize ischemic tissue that has not yet infarcted [1].

However, given that signal hyperintensity is seen with either ischemia and/or infarction on DWI, T2 FLAIR images can be used in conjunction to aid in the distinction between the two and thus aid in the identification of AIS disease progression. For example, if DWI shows signal hyperintensity (i.e. bright spot) in one anatomic region of the brain, while that same anatomic region shows normal signal intensity on the T2 FLAIR image, then AIS is likely in early disease progression and may represent ischemic tissue that has not yet infarcted. DWI signal hyperintensity and corresponding normal T2 FLAIR is known as the DWI-T2 FLAIR mismatch [1]. On the other hand, DWI and corresponding T2 FLAIR signal hyperintensity may indicate infarction. The disease progression of AIS is important in the determination of the appropriate treatment for the patient, since infarcted tissue is considered to be irreversibly damaged but ischemic tissue may be salvaged upon treatment.

1.2.4. Treatment

Diagnosing hemorrhagic stroke versus AIS is critical in determining the action plan for treatment and whether or not a thrombolytic therapy (also known as fibrinolytic therapy), such as Tissue Plasminogen Activator (or tPA) should be administered. tPA acts to dissolve blood clots, thereby restoring blood flow through the originally clotted vessel. In the case of hemorrhagic stroke, a neurosurgical approach to treatment is taken since administering tPA would clearly further bleeding in the brain and could therefore cause patient death. In the case of AIS, however, clot dissolution and subsequent restoration of blood flow can be life-saving. In the case of treating AIS, tPA should be administered as soon as possible [13], with the caveat that it cannot be administered past the upper limit of a critical time window. Once the upper limit of this time window has been surpassed, a neurosurgical approach, such as endovascular therapy (EVT), is required (again, with the caveat that this therapy cannot be performed past the upper limit of a separate critical time window) [13].

The duration of these critical time windows are not listed because they may depend on the geographic location, the chosen therapy, as well as the clinical context and judgements made by the physicians involved. The existence of these critical time windows alludes to a major pitfall associated with delays in accessing a standard clinical MRI stroke exam due to scheduling challenges on conventional MRI systems and a lack of dedicated POC MRI systems: by the time the suspected AIS patient receives their clinical MRI exam, these critical time window(s) may have passed. The time criticality of

diagnosing and treating AIS is reinforced by studies, such as [72]-[75], which have identified additional important benefits to MRI in emergency stroke management.

1.3. Computed Tomography versus Magnetic Resonance Imaging

In this section (*Section 1.3.*), a basic understanding of image acquisition via CT and conventional clinical MRI systems will be provided solely for the purposes of understanding one of the major trade-offs for using these imaging modalities in the diagnosis of AIS in emergency medicine: image acquisition speed. Additional trade-offs for the use of CT versus conventional clinical MRI systems in emergency medicine will be explored, including a discussion of where conventional clinical MRIs have attempted to address these trade-offs, but where pitfalls in their use remain, and thus where the use of the 0.5 T MRI system becomes impactful. Finally, in Section 1.4, an introduction to compressed sensing (CS) as a method for prospectively accelerating MR image acquisition, in addition to assessing image quality via subjective and objective methods, will be provided.

1.3.1. An Introduction to System Basics: Image Acquisition

Considering the imperativeness of diagnosing and treating AIS rapidly, it is important to understand how CT and MRI acquire images since image acquisition is an important part to consider, among others, in the overall scan-to-treatment time.

A CT imaging system consists of the CT machine and a motorized table upon which the patient lies supine (Fig. 2) [30]. The CT machine consists of an x-ray source and x-ray detectors, which rotate around the patient [30].

Fig. 2. Example of a computed tomography imaging system from [23].

CT imaging systems operate on the basis of acquiring computed digital images of transverse slices of the anatomy of interest [31], whereby axial CT scans are commonly used for brain imaging [30]. For example, one transverse slice constitutes a single scan, or two-dimensional brain image [31]; thus, in order to obtain a 3D volume of 2D images representative of the entire brain [30], multiple CT scans must be acquired. However, acquiring multiple scans for CT image acquisition is rapid. Given the circular rotation of the CT machine, the x-ray source produces a fan-shaped x-ray beam, whereby multiple rows of x-ray detectors detect the x-rays passing through the patient and allow for concurrent imaging of multiple slices of the body [30].

To compare, a conventional clinical MRI system (Fig. 3) consists of the bore within which the patient lies supine, and the surrounding concentric layers of equipment [32] housed within the MRI machine.

Fig. 3. Example of a conventional clinical magnetic resonance imaging system from [22].

There are many components that make up the hardware of an MRI system, but there are three components of main importance to prospectively acquiring images. For example, as illustrated in Fig. 4, the layer furthest from the patient is the main magnet (in blue), the middle layer contains the gradient coils (in purple), and the layer closest to the patient contains the radiofrequency (RF) transmitter and receiver [32] (in yellow and green, respectively).

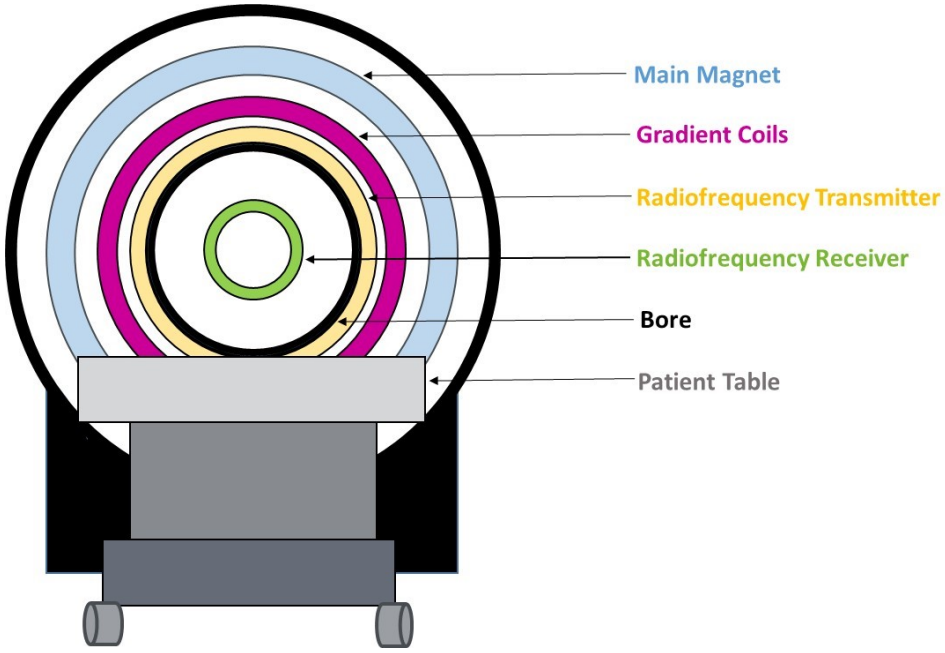


Fig. 4. Cross-sectional illustration of a magnetic resonance imaging system showing the important pieces of hardware required for prospective image acquisition.

The main magnetic field of an MRI system, B_0, is generated by the main magnet [32]. Note that for conventional clinical MRI systems, then, B_0 = 1.5 T or 3 T. As for the gradient coils, there are three orthogonal linear gradient directions (e.g. one coil for each orthogonal direction [32] in an (x, y, z) coordinate system). There are different implementations of imaging that use these three orthogonal gradient directions for different strategies, but a common implementation strategy is such that (x, y, z) corresponds to frequency-encoding, phase-encoding, and slice-selection directions, respectively. For example, slice-selection with the z gradient would yield axial anatomical image slices with frequency- and phase-encoding in the remaining two dimensions. The RF transmitter and receiver is responsible for transmitting [32] the RF pulse at the chosen bandwidth and for receiving the induced RF signals from within the excited slice [32]. For brain imaging, the RF transmitter housed within the MRI machine transmits the RF pulse, while the additional RF receiver placed snug around the patient's head [32] receives the induced RF signals.

MRI systems operate on the basis of acquiring digitized Fourier components of the image, called k-space data. Raw k-space data are typically acquired in 2D via a combination of frequency encoding in one dimension (e.g. kx) and phase encoding in the second dimension (e.g. ky) at the location of the anatomical slice selected via the frequency bandwidth of the RF pulse. During phase-encoding, the position in k-space is selected along the phase-encoding direction, while during frequency-encoding, an entire line of k-space data points may be collected in the frequency-encoding direction (i.e. multiple data points, separated by microseconds, are digitized along the kx

direction at the selected ky position). This data collection process is repeated until all ky positions are selected and all corresponding kx data points are digitized at each ky position.

The resolution and field of view (FOV), which are specified prior to k-space data acquisition, dictate the total number of ky positions to select and the corresponding total number of kx data points to digitize at each ky position necessary for the 2D Cartesian grid of k-space to be considered fully acquired. Depending on whether or not all k-space data were acquired as dictated by the resolution and FOV, k-space data acquisition may be considered prospectively fully sampled or undersampled, respectively. The factor by which k-space is undersampled can be referred to as the undersampling factor, R (*also*, acceleration factor).

Computing an inverse discrete Fourier transform (FT) on prospectively fully sampled 2D k-space data would yield the corresponding 2D image free of any visible aliasing artifacts. However, computing an inverse discrete FT on prospectively undersampled 2D k-space data would yield the corresponding 2D image showing visible aliasing artifacts. Depending on the method of undersampling, however, certain reconstruction techniques, along with an inverse discrete FT, can be applied prospectively to reduce aliasing artifacts. Theoretically, this process would be repeated for all 2D slices, followed by an inverse discrete FT and/or image reconstruction in order to yield a set of 2D images that represent the anatomy in 3D. A helpful analogy for understanding sampling fully acquired k-space data is the Nyquist-Shannon sampling theorem [89], whereby insufficient sampling results in aliasing (e.g. undersampling k-space data yields an

aliased MR image) and sufficient sampling results in no aliasing (e.g. fully sampling k-space data yields a non-aliased MR image).

Considering, in many cases, each excitation of the MRI signal is followed by collection of an entire line of frequency encoding k-space data points at a single k-space position in the phase encoding direction, phase encoding would be a limiting factor to MR image acquisition speed. As such, considering CT image acquisition is rapid via concurrent imaging of multiple slices in comparison to MR image acquisition, which is slice-by-slice, (and, for each slice, phase-encoding is time-cumbersome), CT bodes well for time-critical scenarios. In fact, image acquisition speed in conjunction with the understanding of what AIS is – and why rapid diagnosis is critical – delineates largely why CT is primarily the current first-line imaging SOC across the United States and Canada for diagnosing AIS in emergency medicine. However, despite considering image acquisition speed as being a barrier to implementing MRI in the ED for diagnosis of AIS, while for CT it is considered a pro, both imaging modalities have additional trade-offs for their use, specifically in the context of emergency medicine.

1.3.2. Trade-Offs for Use in Emergency Medicine

Additional to CT's rapid image acquisition [24], CT machines are widely accessible, available, and inexpensive. However, CT imaging does have drawbacks. Not only does CT bombard the patient with harmful ionizing radiation [30] in the form of x-rays, CT images yield poor soft tissue contrast specifically in the case of visualizing AIS [24], [16]. On the other hand, MRI systems are not widely accessible or available, and they are cost-prohibitive, along with their comparably slow image acquisition [24]. However,

beneficially, MRI does not expose the patient to harmful radiation [27], and MR images yield superior soft tissue contrast [24], [16].

Despite the drawbacks of slow image acquisition speed, lower accessibility and availability, and higher financial costs, considering both that brain tissue is a soft tissue and MRI is a non-ionizing imaging modality, it would otherwise appear logical to use MRI for indications of AIS in the ED. In fact, research has been performed on attempting to make acute MRI protocols for use in emergency medicine given the widespread knowledge that MRI is superior to CT, specifically in regards to its ability to visualize soft tissue. For example, [76] [80] [81], among many other studies, utilize standard clinical MRI systems for AIS research in emergency medicine.

	CT		MRI	
Image Acquisition Speed	RAPID	✓	SLOW	✗
Modality	Ionizing	✗	Non-Ionizing	✓
Soft Tissue Contrast	Poor	✗	Superior	✓
Accessibility	✓		✗	
Availability	✓		✗	
Cost	$		$$$	

Fig. 5. Main pros and cons for the use of computed tomography versus magnetic resonance imaging in emergency medicine. Information gathered from [16], [24], [27].

These studies show that MRI is useful in diagnosing AIS, but the barriers to its use in the ED remain (Fig. 5), such as low accessibility and availability, and high monetary costs [27], whereby CT seems to outperform MRI in these areas, thus still justifying the use of CT as SOC in most of the United States and Canada for imaging of AIS in the ED. The opportunity for this thesis to contribute to these bodies of research, therefore, is based

on exploration of novel MRI systems and acquisitions to overcome some of these barriers to the use of MRI in AIS diagnosis, such as image acquisition speed.

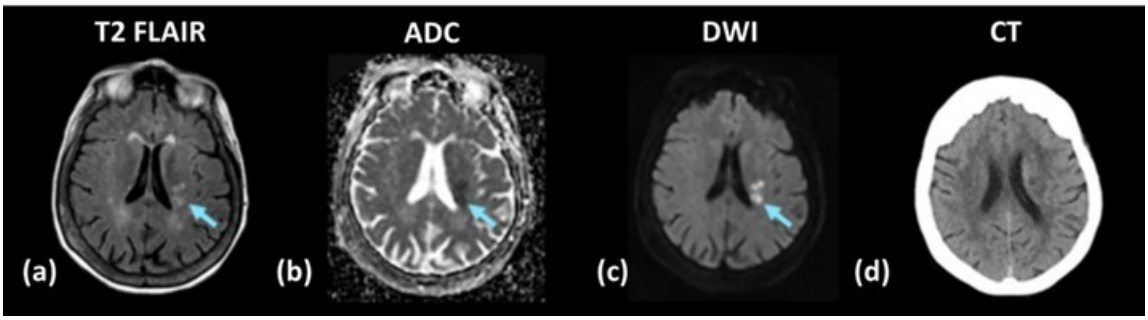


Fig. 6. Non-accelerated axial 0.5 Tesla magnetic resonance imaging brain scans, (a) T2 FLAIR, (b) ADC, and (c) DWI, versus a (d) computed tomography brain scan for a patient with suspected stroke. T2: transverse relaxation time; FLAIR: fluid attenuated inversion recovery; ADC: apparent diffusion coefficient; DWI: diffusion weighted imaging.

Take Fig. 6, for example, which shows a CT scan versus fully acquired 0.5 T MRI scans for a patient with suspected stroke. The blue arrows show that the stroke pathology MRI visualized, CT completely missed; this exemplifies why it is critical to investigate accelerated protocols for stroke at low-field so that one day an MRI stroke protocol can be implemented in emergency medicine that is *as fast as clinically useful*.

1.3.3. Low-Field Magnetic Resonance Imaging

This thesis investigates the use of an FDA and Health-Canada approved head-only POC 0.5 T MRI system (*Synaptive Medical, Toronto, Ontario, Canada*) because it was designed and engineered specifically in response to the aforementioned barriers inherent to preventing conventional clinical MRI systems for use in diagnosing AIS [25]. In order to mimic POC use, this 0.5 T MRI system was installed in the Biomedical Translational Imaging Centre (BIOTIC) (Fig. 7), which is located across the hallway from the emergency medicine interventional suites at the Queen Elizabeth II Health Sciences

Centre (*Halifax, Nova Scotia, Canada*) – approximately the same distance to the hospital's current CT imaging suite.

Fig. 7. Photograph taken by Dr. Steven Beyea of Synaptive Medical's head-only point-of-care 0.5 Tesla magnetic resonance imaging system located in the Biomedical Translational Imaging Centre (BIOTIC) lab at the QEII Health Sciences Centre in Halifax, Nova Scotia, Canada.

The 0.5 T system has several advantages in terms of accessibility compared to standard clinical scanners. First, the 0.5 T POC MRI system weighs approximately 1100 kg [25], which is much lighter than standard clinical MRI systems that can weigh anywhere from 4,000 – 10,000 kg. Second, the 0.5 T POC MRI system can be installed through a loading dock and elevator [25] versus installation through rigging and building modifications [25] necessary with installation of the heavier and larger clinical systems. Third, the 0.5 T POC MRI system does not require venting of cryogens [25], such as helium, which is beneficial for the following two reasons: (1) helium is an expensive limited resource, and (2) eliminating the requirement to set up a ventilation system ultimately translates into the aforementioned advantage of less building modifications.

Lastly, with a low-field magnet of 0.5 T, the fringe field is more compact [25] because of both the reduced physical size of the main magnet and the reduced strength of the

main magnetic field than that of its 1.5 T and 3 T counterparts; therefore, stray magnetic fields extend shorter distances from the scanner. This is important because larger fringe fields require larger hospital square footage for installation – and hospital square footage is scarce and in high demand [25]. In addition, the lower the magnetic field strength, the cheaper the magnet, in general; therefore, in comparison to magnets at 1.5 T or 3 T in standard clinical MRI systems, the 0.5 T magnet is less expensive.

	MRI			
	Clinical (1.5 T or 3 T)		**POC Head-Only (0.5 T)**	
Image Acquisition Speed	SLOW	✗	SLOW	✗
Modality	Non-Ionizing	✓	Non-Ionizing	✓
Soft Tissue Contrast	Superior	✓	Superior	✓
Weight	>> 1134 kg	✗	< 1134 kg	✓
Installation	Rigging and building modification	✗	Installed through loading dock and elevator	✓ ✓
Cryogens	Venting system required	✗	Closed conduction cooling	✓
Fringe Field	> 1.03 m	✗	< 1.03 m	✓
Cost	$$$		$	

Fig. 8. Example of pros and cons for emergency medicine use of conventional clinical magnetic resonance imaging (MRI) systems versus the head-only point-of-care (POC) 0.5 Tesla (T) MRI system. Information gathered from [25].

Additional aspects of the 0.5 T MRI system that are not outlined in Fig. 8 are the safety considerations that render the 0.5 T system more accessible for patient safety in the ED as compared to conventional clinical scanners. To explain, the three hardware components of main importance for prospective image acquisition – main magnet, gradient coils, and RF transmitter and receiver – each have an associated safety risk that must be considered. For example, if the patient has an implant device composed of

magnetic material, the main magnet may yield a ballistic effect on the implant. Further, the gradient coils may cause the patient to experience peripheral nerve stimulation, deafening acoustic noise, and/or, should the patient have metal implants, heating around those implants, all due to rapid switching of gradient coils. Lastly, as a result of thermal energy absorption due to RF power deposition, the RF transmitter may yield skin burns, general body heating, and/or heating around implants. Although these safety considerations are present regardless of whether a conventional clinical MRI scanner or the 0.5 T MRI system are being used, the design of the 0.5 T MRI system reduces the extent of the risks when compared to conventional clinical scanners.

For example, given that the 0.5 T MRI system (B_0 = 0.5 T) is at either 1/3 or 1/6 of the magnetic field strength of 1.5 T or 3 T MRI scanners respectively, the ballistic effects are reduced. Additional examples include the head-only design and SAR constraints of this low-field scanner, which essentially prevent peripheral nerve stimulation and general body tissue heating, respectively. A major safety concern for MRI in the ED with this system is therefore the presence of implant devices in the above-shoulder region; although the system's head-only design mitigates risk caused by implants in the below-shoulder region, the risk of implants in the above-shoulder region remains (albeit, to a lesser extent due to the design of the system). These examples yield an improved MRI safety profile when considering the use of MRI in an emergency medicine context, and thus for the diagnosis of AIS in the ED, which justifies the use of this system, from a safety standpoint, for this research work. Lastly, it is worth mentioning that, although there appears to be a larger number of safety considerations when imaging with MRI

systems as compared to CT imaging, the major safety consideration being radiation exposure with CT imaging is fundamental to being able to collect scans since the imaging modality relies upon the use of x-rays. As it relates to safety considerations, the benefit of MRI is that, regardless of whether using a conventional clinical scanner or the 0.5 T system, sequencing parameters can be selected, and other measures taken, such that safety risks are mitigated.

Nonetheless, despite the 0.5 T MRI system addressing the drawbacks of both MRI in comparison to CT and standard clinical MRI in comparison to POC MRI, the barrier to MRI utilization in the ED remains: MR image acquisition speed is too slow. Hence, the problem of accelerating MR image acquisition remains unsolved, which does not bode well when time is of the essence – and where minutes matter – as required to achieve positive patient outcomes in cases of AIS. In fact, a study by Kidwell *et al.* in 2000 concluded that feasibility of emergent use of MRI in AIS patients exists, but rapid MRI imaging must be available [74].

1.4. Image Quality versus Acquisition Speed

When attempting to accelerate MR image acquisition, it is imperative to measure image quality. This section (*Section 1.4.*) briefly outlines the trade-off between acquisition speed and image quality, introduces how this trade-off may be addressed, and explains how to assess the extent to which image quality might remain clinically important.

1.4.1. Compressed Sensing: A Method for Prospectively Accelerating MR Image Acquisition

CS is one imaging technique, among many, that can be implemented to accelerate MR image acquisition. The nuts and bolts of CS, as well as other reconstruction techniques, including techniques to accelerate image acquisition, will be discussed in further detail in Section 2.1.

In the meantime, CS can be understood as the ability to compress an image, similar to that which is done in JPEG compression. However, different from JPEG compression, where all data is first collected and then redundant information is discarded, CS reconstructs images whereby k-space data were either not fully acquired (in the case of prospective MR image acquisition) or fully acquired but undersampled (in the case of retrospective image acquisition) – in either case, k-space data is pseudo-randomly missing and the image acquisition is considered to be accelerated. Despite the acquisition being accelerated, however, the trade-off is such that missing k-space data comes at the cost of reduced image quality, especially when reconstructed using only an FT. Advantageously, there are alternate reconstruction methods, in particular CS (*Sections 2.1.4. and 2.1.5.*), that can be implemented to mitigate this trade-off, but it nonetheless remains necessary to measure the quality of the resultant images.

1.4.2. An Introduction to Objective versus Subjective Image Quality

Image quality (Fig. 9) can refer to either the objective or subjective quality of an image. Objective image quality is calculated by a computer using image quality metrics (IQMs), which can be thought of as unique mathematical formulas that output numbers, each quantifying something specific based on the images input to the calculation.

Subjective image quality, on the other hand, is based on human perception. Of course, in a clinical context, these humans are radiologists and their perceptions are based on expertise in performing specific diagnostic tasks in order to make clinical decisions.

One might wonder why IQMs are important when, in a clinical setting, what matters is that radiologists can perform the necessary diagnostic task and confidently make an accurate clinical decision from the images. In this sense, subjective image quality ultimately matters most. Therefore, when it comes to research, such as assessing large MR image datasets, generating new MRI protocols, or doing MRI sequence parameterization, radiologists need to assess the associated images, first, in order to allow for clinical translation. However, a significant portion of time and energy must be dedicated by a radiologist in order to assess these images.

IQMs therefore become important when correlated with radiologists' diagnostic confidence scores, which could allow for *a priori* assessment in order to streamline which images need to be shown to radiologists and which images can be discarded for future studies, ultimately utilizing less of radiologists' valuable time going forward while ensuring their efforts can be directed where they will have the most impact. To this point, it is critical to keep in mind that just because a certain IQM performs the best based on its computed score of an image does not automatically extrapolate to the image being the best, clinically – despite this currently being a habitual way of thinking in the MRI field.

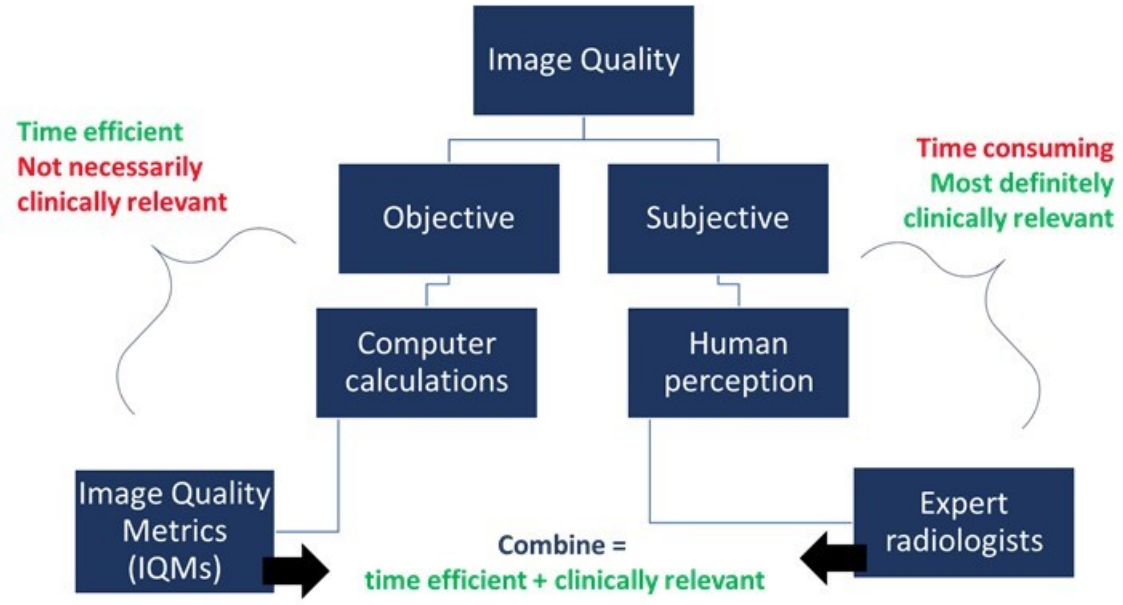


Fig. 9. An outline of image quality in the context of medical MR images.

IQMs have been used in studies seeking to understand if a correlation exists between objective and subjective image quality, whereby objective image quality is computed via IQMs and subjective image quality is measured via raters' scoring subjectively perceived image quality. In MRI literature, these studies appear have been implemented in a progressive manner, as follows: IQMs computed on natural images with non-expert raters scoring subjectively perceived *overall* image quality [68]; IQMs computed on non-pathological medical MR images with non-expert raters scoring subjectively perceived *overall* image quality [69]; and IQMs computed on non-pathological medical MR images with expert radiologist raters scoring subjectively perceived *overall diagnostic* image quality [42]. The latter of these studies yielded IQM studies more clinically relevant since a radiologist's expert opinion of overall diagnostic image quality ultimately matters most in a clinical context.

Specifically, Mason *et al.* [42], investigated 10 full-reference IQMs, concluding that 3 IQMs – the noise quality measure (NQM), the visual information fidelity (VIF) criterion, and the Feature SIMilarity (FSIM) index – correlated more closely with radiologists' scores of *overall diagnostic* image quality for images absent of pathology than the commonly used IQMs root mean square error (RMSE) and the Structural SIMilarity (SSIM) index (Fig. 10) [42].

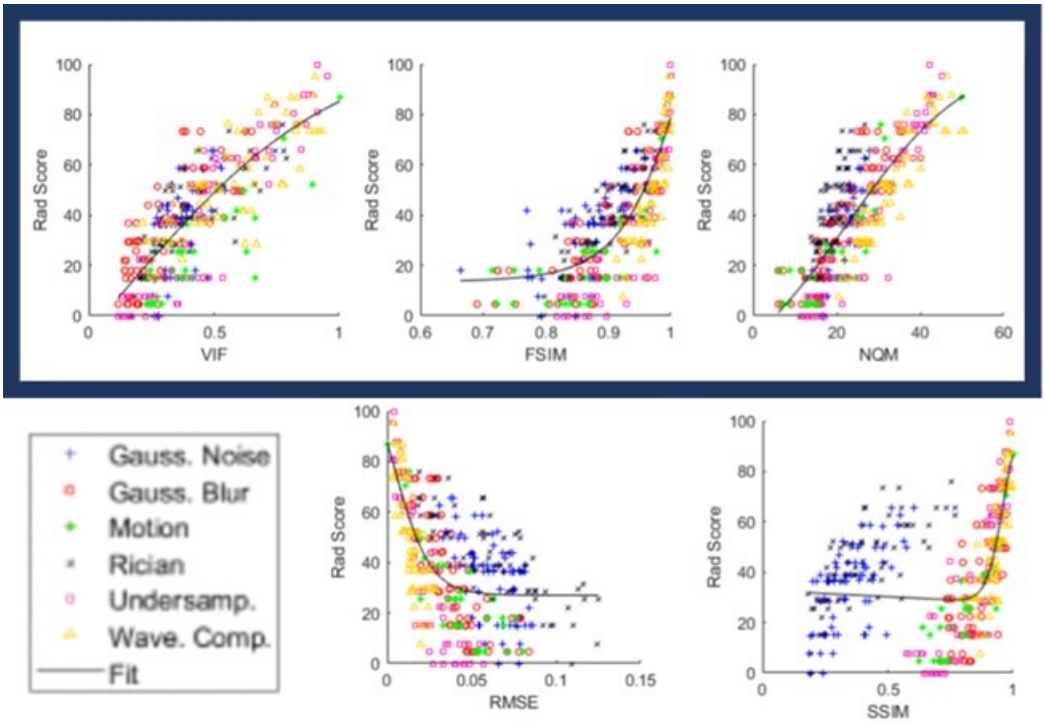


Fig. 10. Example of a study [42] where image quality metrics (IQMs) were computed on non-pathological medical MR images with expert radiologist raters scoring subjectively perceived overall diagnostic image quality. Examples of the correlation graphs were cropped from [42, Fig. 3.], where noise quality measure (NQM), visual information fidelity (VIF), and Feature SIMilarity (FSIM) correlated more closely with radiologists' scores of overall diagnostic image quality for images absent of pathology than the commonly used IQMs root mean square error (RMSE) and Structural SIMilarity (SSIM).

However, overall diagnostic quality might not necessarily indicate an image's usefulness for a specific diagnostic task. As such, the research work in this thesis

explores the IQMs that Mason *et al.* [42] identified as correlating with *overall diagnostic* image quality and assesses how they perform when computed on pathological images where expert neuroradiologists' score their ability to perform specific diagnostic tasks as they would in a clinical setting. This next progression can therefore be defined as follows: IQMs computed on stroke pathological medical MR images with expert neuroradiologist raters scoring subjectively perceived *task-specific diagnostic* image quality, rather than *overall diagnostic* image quality as it were in the recent study by Mason *et al.* [42].

The concept of As Low As Reasonably Achievable (ALARA) can aid in linking the prominent trade-offs between CT and MRI with diagnostic image quality in the ED, which are radiation dose and image acquisition time, respectively. Image quality in CT is defined by radiation dose, whereby a higher radiation dose yields an image of better diagnostic quality observed by radiologists. For CT imaging, ALARA therefore defines the lowest possible radiation dose required to yield a diagnostically acceptable image – and no dose more to make the image better than necessary, so as to protect the patient from undue harm attributable to unnecessary levels of x-ray radiation exposure. On the other hand, with MRI, it is assumed that a longer image acquisition theoretically yields an image of better diagnostic quality. Therefore, for MRI, ALARA could define the lowest possible image acquisition time required to yield a diagnostically acceptable image – and no time longer to make the image better than necessary, so as to protect the patient from undue harm attributed to unnecessary delays in diagnosis and receiving appropriate treatment.

Hence, the concept of ALARA is a guide to prevent diminishing returns to the patient associated with higher radiation doses and longer image acquisition times for CT and MRI, respectively, when there is no clear improvement in a radiologist's ability to achieve the same diagnosis – only an unnecessary improvement in image quality at the expense of the patient's outcome.

1.5. Motivation of Thesis Work

The motivation behind the research work in this thesis is the clinically relevant integration of accelerated MRI scans – specifically, acceleration of the image acquisition – to the diagnosis of AIS in emergency medicine in the ED. A decrease in MR image acquisition time may achieve the following: (1) a more positive patient experience; (2) an increased patient throughput, thereby reducing associated burdens on the healthcare system; and (3) the ability for neuroradiologists to rapidly and confidently diagnose AIS from MRI scans in the ED.

Specifically, this thesis examines the research problem of determining neuroradiologists' task-specific diagnostic confidence in reporting the presence or absence of AIS, as well as assigning Fazekas scoring for chronic stroke, using MR images prospectively acquired via the head-only point-of-care 0.5 T MRI system. Images were fully-sampled at acquisition and were then retrospectively accelerated via CS. Note that a Fazekas scale is used to rank the extent of chronic ischemic lesion burden.

Neuroradiologists' task-specific diagnostic confidence is assessed because it is a requirement that resultant images remain diagnostically useful when trying to solve the problem of speeding up image acquisition. Further to subjective image quality, as

defined by neuroradiologists' task-specific diagnostic confidence, objective image quality is an important metric to assess in terms of acting as a surrogate measure for neuroradiologists' task-specific diagnostic confidence, specifically as it pertains to the use of diagnostic images. Hence, this thesis examines the correlation of various IQMs with neuroradiologists' task-specific diagnostic confidence.

1.6. Thesis Objectives and Hypotheses

The main objectives to addressing the research problems in this thesis are outlined in this section (*Section 1.6.*). Methods-based objectives will be discussed first, considering these were required to be completed before the two hypotheses could be tested.

1.6.1. Methods-Based Objectives

Objective 1 is to retrospectively accelerate MR image acquisition through k-space data undersampling and CS reconstruction. Objective 2 is to compute objective image quality scores of the undersampled images output from Objective 1 via full-reference IQMs.

1.6.2. Hypothesis-Based Objectives and Associated Hypotheses

Objective 3 is to assess neuroradiologists' diagnostic confidence in performing specific stroke-related diagnostic tasks using the images output from Objective 1. The specific stroke-related diagnostic tasks were as follows: (1) identifying the presence/absence of acute stroke (hereafter referred to as the *acute stroke diagnostic task*) and (2) assigning a Fazekas score for chronic stroke (hereafter referred to as the *chronic stroke diagnostic task*). The first hypothesis of this thesis, Hypothesis A, is associated with Objective 3, and is as follows: Neuroradiologist raters' diagnostic

confidence scores corresponding to performing the acute stroke diagnostic task will be less sensitive to increasing R than neuroradiologist raters' diagnostic confidence scores corresponding to the chronic stroke diagnostic task.

Objective 4 is to assess the relationship between IQM scores and radiologists' confidence scores. The second hypothesis of this thesis, Hypothesis B, is associated with Objective 4, and is as follows: the IQMs FSIM, NQM and VIF will perform better than RMSE and SSIM for both the acute and chronic stroke diagnostic tasks.

Chapter 2: Theory and Background

In Chapter 1, the concept of fully acquiring a 2D grid of k-space data were introduced, noting that the image reconstruction was the implementation of an FT to yield the corresponding non-aliased anatomical 2D image. It was then stated that k-space data acquisition was time-consuming due to the fact that multiple phase-encoding lines need to be collected for each slice, which is then repeated for each of the multiple slices within the anatomy of interest. As such, the number of phase-encoding lines collected is directly proportional to image acquisition time, hence acquiring the phase encoding lines contribute to a slow MR image acquisition [38], [54]. Advantageously, there are well-tested image acquisition and reconstruction techniques that can be implemented to accelerate MR image acquisition.

2.1. Accelerated Image Acquisition/Reconstruction Techniques

Accelerated image acquisition has traditionally been achieved through reducing the number of excitations (NEX) [38], and/or through undersampling k-space data through partial Fourier imaging (PFI) [40], parallel imaging (PI) [40], CS, or CS-SENSitivity Encoding (SENSE) [77]. An illustrative example of how image acquisition time can be cut in half will be provided for each accelerated image acquisition/reconstruction technique.

2.1.1. Reducing Number of Excitations

MR images can be acquired at various NEX. The natural number NEX value that is set during sequence parameterization of the MRI scan dictates how many times each slice is sampled. Therefore, reducing NEX proportionally accelerates image acquisition, and this

is based on *a priori* knowledge that the number of k-space phase-encode lines collected is proportional to image acquisition speed [41]. For example, if NEX = 2 is reduced to NEX = 1, image acquisition would take half of the total time. However, signal averaging theory [65] tells us that, upon acquiring fewer averages of the same signal, signal-to-noise ratio (SNR) will decrease.

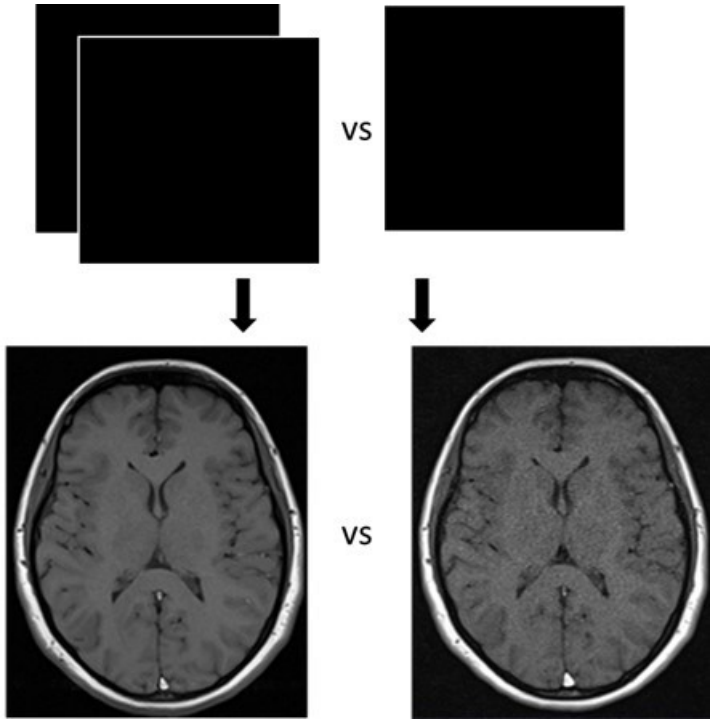


Fig. 11. Illustrative example of a magnetic resonance (MR) image resultant from a Fourier transform reconstruction of reduced number of excitations (NEX) k-space data. Each black square represents the same 2D slice of fully acquired k-space data: fully acquiring the same 2D slice of k-space data at NEX = 2 (top left) versus fully acquiring the same 2D slice of k-space data at NEX = 1 (top right) will yield a loss of signal-to-noise ratio, as shown in the example MR images (bottom, left to right). From NEX = 2 to NEX = 1, image acquisition time is cut in half. Example MR images are from [70].

As such, implementing image reconstruction on a reduced NEX k-space data acquisition via an FT would yield a loss of SNR in the corresponding 2D image acquired at NEX = 1 as compared to that which was acquired at NEX = 2, as exemplified in Fig. 11.

Lastly, it is worth noting that, if reduced to NEX = 1, the remaining option to accelerate image acquisition even further is to start undersampling k-space data.

2.1.2. Partial Fourier Imaging

PFI accelerates image acquisition by undersampling k-space data on the basis of exploiting the inherent conjugate symmetry of the Fourier domain.

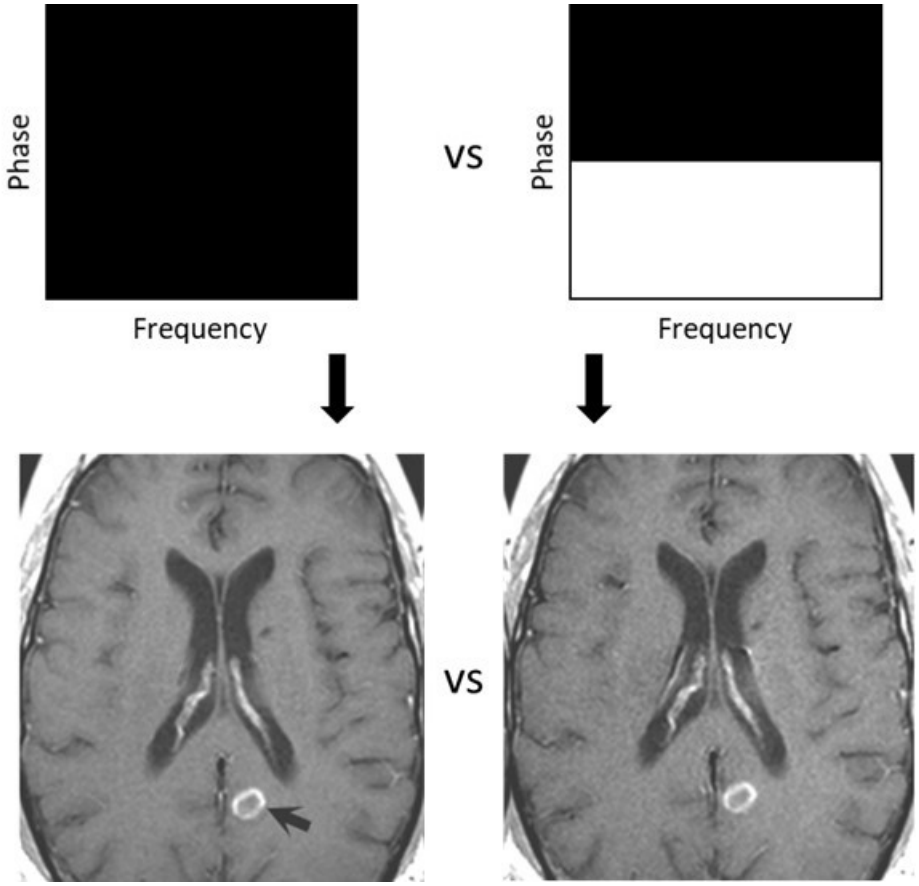


Fig. 12. Generalized illustrative example of undersampling k-space data through partial Fourier imaging (PFI). Top row: black corresponds to acquired k-space data; white corresponds to k-space data not acquired that would be zero-filled or estimated using conjugate symmetry during image reconstruction. Fully acquired 2D slice of k-space data (top left) versus acquiring only half of k-space (top right) represents an undersampling factor of 2. Image acquisition would be cut in half compared to if k-space data were fully acquired but yields a loss of signal-to-noise ratio in images from left to right (bottom). Bottom two images are cropped out from [66, Figs. 78.1A and 78.1B], respectively.

Fig. 12 is an illustrative example for undersampling k-space data by a factor of two through PFI. The missing k-space data can be zero-filled, yielding a lower SNR [40], or estimated during image reconstruction using conjugate symmetry to reduce artifacts resultant from the abrupt transition between sampled and unsampled regions as would be the case with zero-filling. However, even if k-space data are estimated, the SNR is not recovered since the estimated data are not statistically independent from the acquired data. Whether zero-filled or estimated, applying image reconstruction via an FT to the k-space data would yield an anatomical image with lower SNR, compared to if k-space data were fully acquired.

2.1.3. Parallel Imaging

PI [82] [83] can accelerate image acquisition through regular, or periodic, undersampling of k-space phase-encoding lines. PI operates by concurrently acquiring the same line of k-space data from multiple coils – but this requires specific software and multi-coil hardware design – and higher acceleration factors can actually yield a lower SNR [43]. An associated constraint with PI, however, is such that *a priori* knowledge of coil sensitivity is required in order to reconstruct the image [77].

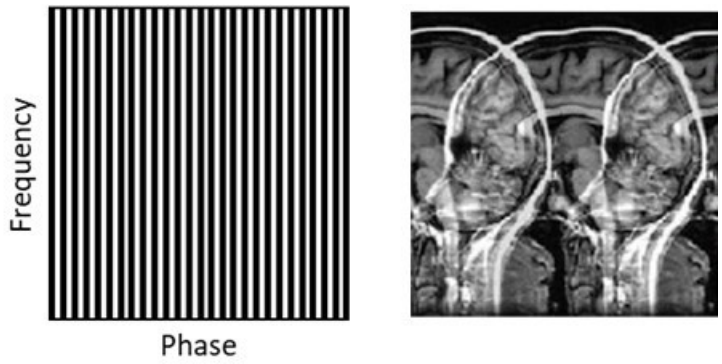


Fig. 13. Illustrative example of regularly, or periodically, undersampling k-space phase-encoding lines through parallel imaging. Left: black lines correspond to acquired phase-encoding lines; white lines correspond to phase-encoding lines not acquired. Acquiring every second phase encoding line represents an undersampling factor of 2, and image acquisition would be cut in half compared to if all lines of k-space were acquired. Right: example of coherent aliasing artifacts in an image (cropped from [37, Fig. 2]) resultant from the Fourier transform reconstruction of regularly, or periodically, undersampled k-space data.

Fig. 13, on the left, shows a regular, or periodic, 2D Cartesian pattern for undersampling k-space phase encoding lines by a factor of two; during prospective image acquisition, this pattern is called a sampling trajectory. If this sampling trajectory of k-space were to be acquired, reconstruction via an FT would yield coherent aliasing artifacts in the resultant image, as exemplified in Fig. 13 on the right, due to insufficient sampling of phase-encoding lines. Once in image space, the coherent aliasing can be removed with an appropriate reconstruction method; this occurs with the SENSE approach to PI. Alternative to image-space methods, like SENSE, there are also k-space methods, like GeneRalized Autocalibrating Partial Parallel Acquisition (GRAPPA).

2.1.4. Compressed Sensing

In 2007, Lustig *et al.* [26] amalgamated the constructs of a signal processing technique, called *compressed sensing (CS)* [44], with MRI in order to accelerate image

acquisition. CS operates on the basis of three fundamental principles: (1) sparsity in some domain [26], (2) pseudo-random undersampling of k-space data to yield incoherent (noise-like) aliasing artifacts [26], and (3) optimization-based nonlinear reconstruction that enforces both sparsity and data consistency [26].

CS [Eqn. 1] [53] aims to reconstruct an undersampled image in an iterative way based on the trade-off between data consistency and image sparsity.

$$minimize\ \|F_u m - y\|_2^2 + \lambda\|\Psi m\|_1$$

[Eqn. 1]

The data consistency term, $\|F_u m - y\|_2^2$, computes the least-squares difference [53] between undersampled k-space data, y, actually acquired and k-space data resultant from the reconstruction, $F_u m$. The image sparsity term, $\|\Psi m\|_1$, computes the ℓ_1 norm sparsity of the reconstructed image, m, within the domain based on which sparsifying transform, Ψ, was selected [53]. The sparsifying transform enforces sparsity in the corresponding transform domain, while thresholding is applied to the data in the sparsified domain to make the data ever sparser. Based on the parameters chosen, CS aims to reconstruct an image that is consistent with the phase-encoding lines of k-space *actually* acquired and that is sparse in the transform domain. The regularization parameter, λ, weighs this trade-off between data consistency and image sparsity [53], thus controlling how much of an effect sparsifying and thresholding are allowed to have on k-space data in an iterative way until both data consistency (i.e. minimizing the least squares difference between the reference and reconstructed undersampled image) and

transform sparsity (i.e. “noise” removal via thresholding) are enforced in an optimized way.

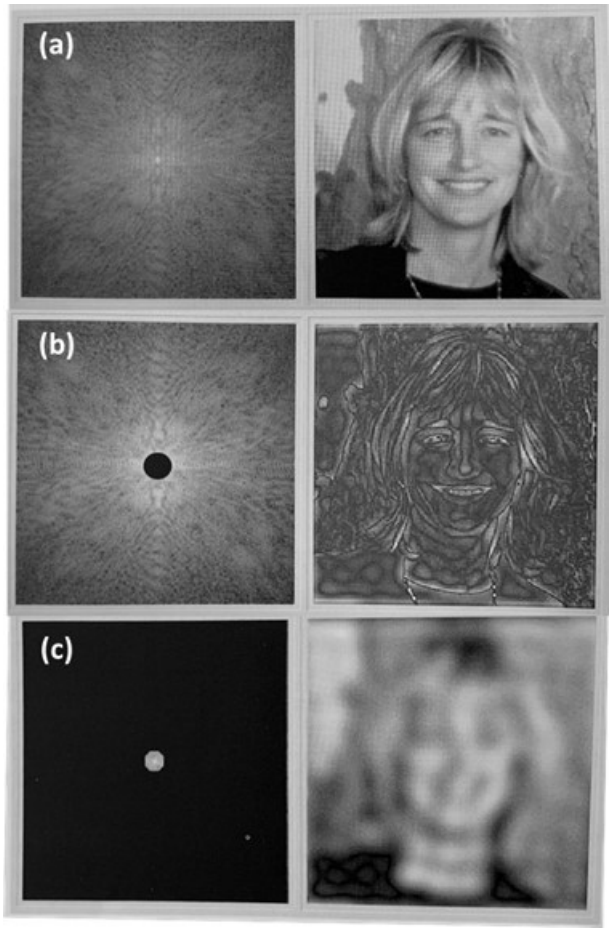


Fig. 14. A representative example of the type and location of information stored in k-space data (combined [54, Figs. 3.30-3.32]). k-Space data are shown on the left with the corresponding image (upon Fourier transform reconstruction) on the right. (a) Fully acquired k-space data yield a non-aliased image [54, Fig. 3.30]. (b) Fully sampling high-frequency signal resolution information contained at the edges of k-space data yields an image with edge resolution, but where very little information exists about the features within edge boundaries [54, Fig. 3.31]. (c) Fully sampling low-frequency signal information contained at the centre of k-space data yields an image with very little information about edge resolution but with overall image contrast [54, Fig. 3.32].

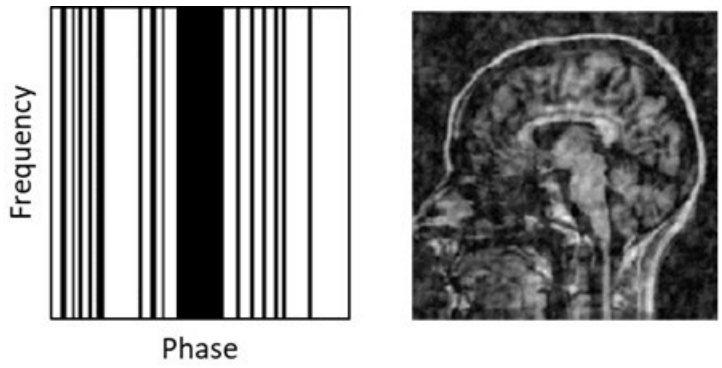


Fig. 15. Illustrative example of pseudo-randomly undersampling k-space phase-encoding lines through compressed sensing. Left: black lines correspond to acquired phase-encoding lines; white lines correspond to phase-encoding lines not acquired. Fully acquiring the centre of k-space and randomly undersampling the rest of k-space, amounting to a total undersampling factor of 2, would cut image acquisition in half compared to if all lines of k-space were acquired. Right: example of incoherent (noise-like) aliasing artifacts in an image (cropped from [37, Fig. 2]) resultant from the Fourier transform reconstruction of pseudo-randomly undersampled k-space data.

Undersampling through CS operates on the premise of *a priori* knowledge about the nature of k-space data and where particular information is located. To explain, the centre of k-space data is where the low frequency signal (i.e. contrast) information is located [54], while the edges of k-space data are where the high-frequency signal (i.e. resolution) information is located [54], as shown in a representative example in Fig. 14.

Specifically, undersampling through CS is performed in a pseudo-random fashion. Pseudo-random undersampling consists of fully-sampling the central $1/n^{th}$ of k-space data (i.e. fully sampling low frequency signal information), yet randomly undersampling the rest of k-space data (i.e. high frequency resolution information). The outer regions of k-space always contain far less signal power than the center of k-space, and hence are more suitable for undersampling without loss of important information. As such, k-space data is randomly sampled, but not truly randomly sampled due to fully-sampling

the centre; hence utilizing the term *pseudo*-random. Fig. 15, on the left, shows an illustrative example of a pseudo-random 2D Cartesian pattern for undersampling k-space phase encoding lines by a factor of 2 (e.g. the number of phase-encoding lines fully sampled at the centre of k-space, plus the number of phase encoding lines randomly sampled at the edges of k-space, would total to half the number of total phase-encoding lines required for k-space to be considered fully acquired).

If this sampling trajectory of k-space (Fig. 15 left) were to be acquired, image reconstruction via an inverse discrete FT would yield incoherent (noise-like) aliasing artifacts in the resultant image, as exemplified on the right in Fig. 15, due to insufficient sampling of phase-encoding lines at the edges of k-space. Therefore, rather than an inverse discrete FT, a chosen CS image reconstruction technique is implemented to remove these incoherent aliasing artifacts (and to be considered as a CS reconstruction, the three fundamental principles of CS previously mentioned at the beginning of Section 2.1.4. must be satisfied).

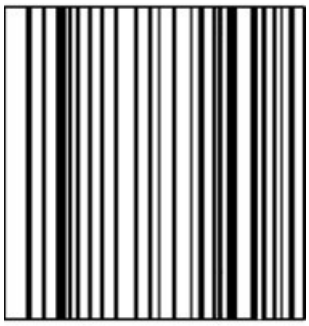


Fig. 16. Illustrative example of irregularly, or non-periodically (truly random) undersampling k-space phase-encoding lines. Black lines correspond to acquired phase-encoding lines. White lines correspond to phase-encoding lines not acquired. Amounting to a total undersampling factor of 2, image acquisition would be cut in half compared to if all lines of k-space were acquired.

Fig. 16 shows irregular, or non-periodic (truly random) sampling of k-space phase-encoding lines. It is worth mentioning that if this sampling trajectory of k-space were to be acquired then reconstruction via an FT would also yield incoherent aliasing artifacts in the resultant image due to insufficient sampling of phase-encoding lines, as is similarly described for pseudo-random undersampling. Although a CS image reconstruction takes advantage of these noise-like artifacts to iteratively *denoise* the reconstructed image, it is imperative that the centre of k-space be fully sampled due to the fact that the centre of k-space contains most of the signal power in the Fourier domain, while the rest of k-space is sparser.

In contrast to reducing NEX, and PFI and PI techniques, CS may recover SNR – despite sampling less of k-space data – due to its inherent denoising reconstruction [26] relative to a simple FT. One of the drawbacks of CS, however, is that image reconstruction is time-consuming [45]. However, access to graphics processing units (GPUs) has increased significantly in recent years, and this has led to improvements in CS reconstruction times. Additionally, increased access to GPUs has also enabled much of the work in artificial intelligence (AI), and there has been extensive literature in recent years [46] – [52] on applications of AI to hasten CS reconstruction. Nonetheless, it is important to note that the acceleration of image reconstruction will not be explored in the research work of this thesis.

2.1.5. Compressed Sensing-Sensitivity Encoding

In order to accelerate image acquisition further than that which is capable by CS and PI, independently, CS and PI can be combined to create, for example, CS-SENSE [77].

Since CS and PI have different requirements that fundamentally do not interfere with one another, it is possible to combine CS and PI as a single reconstruction method and therefore take advantage of both at once. The research work in this thesis investigates image acquisition acceleration from the context of CS, including coil sensitivity information.

2.2. Retrospective Application of Compressed Sensing in the Literature

A 2D Cartesian grid of k-space can not only be prospectively sampled (*Section 1.3.1., pg. 18-19*), but it can also be retrospectively sampled. For example, prospectively fully acquired k-space data may then be retrospectively fully sampled or undersampled. Computing an inverse discrete FT on prospectively fully acquired, and then retrospectively fully sampled, 2D k-space data would yield the corresponding 2D image free of any visible aliasing artifacts. However, computing an inverse discrete FT on prospectively fully acquired, and then retrospectively undersampled, 2D k-space data would yield the corresponding 2D image showing visible aliasing artifacts. As long as the method of 2D undersampling creates incoherent aliasing in the sparse domain, image reconstruction via CS can be applied retrospectively to reduce aliasing artifacts, yielding a set of 2D images that represent the anatomy in 3D. Relevant to the research work of this thesis, CS can be used, retrospectively, to accelerate image acquisition speed and therefore determine the relative acquisition time [Eqn. 2]

$$Relative\ Acquistion\ Time = \frac{image\ acquisition\ time}{R},$$

[Eqn. 2]

where R is the acceleration factor, as dictated by the pseudo-random undersampling pattern.

The literature extensively investigates standard CS in body MRI applications [55] – [59], whereby the anatomies of interest or patient populations (e.g. pediatric) being imaged yield severe motion. For example, at the time of the literature review of this thesis, out of the 483 articles listed for a Web of Science search (*keywords: compressed sensing MRI accelerate*), the main applications were cardiac and lung imaging, likely due to inherent severe motion with respiration, with only a handful of clinically-relevant neurological applications to standard CS. Examples of the latter include pediatric use cases [60], routine/standard brain imaging protocols [33], [34], [78] and healthy volunteer brain imaging [35]. Specifically, [33] attempted up to R = 4X, while [35] investigated up to R = 5X, but only R = 2X and R = 3X, respectively, were deemed by neuroradiologists to yield acceptable images. Although neuroradiologists' confidence were included in these studies in order to ensure the provision of clinical relevance to images acquired at various R, the reason for investigating image acquisition time was not geared for the intended use case of acute stroke, or even emergency medicine for that matter, but simply for routine/standard brain imaging. Not to mention that the R deemed to be acceptable, clinically, remained limited to R = 2-3X. In this sense, it may not be worth adding the longer reconstruction time required by CS when PI could instead be implemented to likely yield similar results.

Even further attempts to accelerate using CS-SENSE by [78] remained low at R = 4X being deemed to be acceptable, while [77] stated that R = 6X images were of poor

quality due to being visually blurry and showing undersampling artifacts, and that reconstructed images at R > 6X failed to be acceptable.

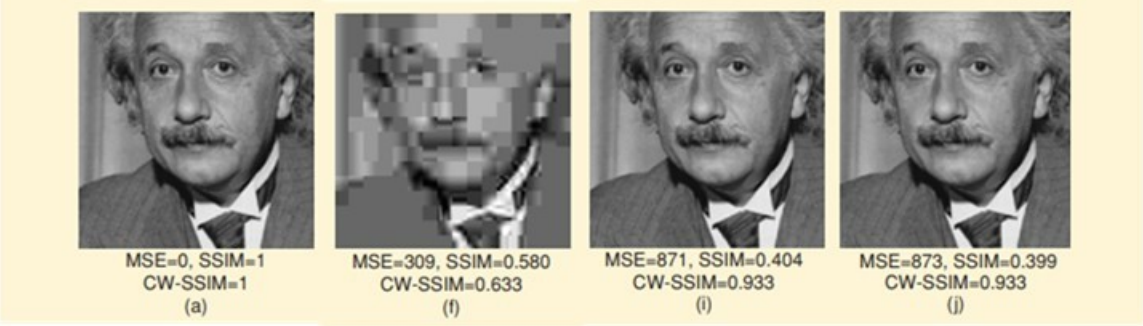
Further, only [34] applied to stroke, whereby acceptable image quality was deemed by neuroradiologists at an average reduction in image acquisition time by approximately one third. Overall, there is evidently a gap in the literature as it relates to accelerated image acquisition via CS for application to MRI in emergency medicine, specifically in the context of AIS. This, in addition to the imperative need to yield clinical relevance deemed by neuroradiologists, justifies the purpose of this research work and its application of retrospective CS to MRI of AIS.

2.3. Image Quality Metrics

Image quality metrics (IQMs) are visual signal analysis methods that yield objective quantification of image quality. There are various categories of IQMs depending on the available image information prior to the computation: no-reference, reduced reference and full-reference [61]. While the other two categories of IQMs are being developed in the literature, full-reference IQMs constitute the more conventional existing approach. Full reference IQMs are computed using both the available reference image and the altered image derived from the reference image. The research work of this thesis uses full-reference IQMs, where *reference* images refer to the image derived from fully acquired, fully sampled k-space data, while the *altered* (typically referred to as *undersampled* or *accelerated*, throughout) images refer to the image derived from fully acquired, undersampled k-space data. The following full-reference IQMs will be used in this research work: root mean square error (RMSE), the Structural SIMilarity (SSIM)

index [61], the Feature SIMilarity (FSIM) index [39], noise quality measure (NQM) [62], and the visual information fidelity (VIF) criterion [63].

Fig. 17 is an illustration of different ways in which an image's quality may be altered, showing various image distortion/degradation techniques, and demonstrates how IQMs respond differently to different alterations. Although the types of image alterations in Fig. 17 are not used in the research work of this thesis, the example may help provide intuition as to why there are many different ways of trying to quantify image quality, and why a metric that works in some cases may fail in others.



[FIG2] Comparison of image fidelity measures for "Einstein" image altered with different types of distortions. (a) Reference image. (f) JPEG compression. (i) Spatial shift (to the right). (j) Spatial shift (to the left).

Fig. 17. Example of various types of image distortions/degradations from [64, Fig. 2].

Fig. 17 demonstrates how some IQMs are more sensitive to particular types of image alterations and/or correlate better with subjective human perception of image quality. For example, mean squared error (MSE) equals zero for the reference image [64, Fig. 2 (a)] in Fig. 17, representing the metric's best possible objective image quality, while MSE = 873 and MSE = 309 for [64, Fig. 2 (j)] and [64, Fig. 2 (f)] (Fig. 17), respectively. According to MSE, the higher the score, the worse the objective image quality. However, this is likely not congruent with subjective human perception of image quality when comparing images [64, Fig. 2 (a), (j), (f)] (Fig. 17), since [64, Fig. 2 (j)] is not only visibly

higher quality than [64, Fig. 2 (f)] but relatively visibly identical in quality to [64, Fig. 2 (a)], despite the higher MSE score of [64, Fig. 2 (j)] otherwise indicating "worse" quality. This example opens a small window of insight into the idea that high objective IQM performance may not necessarily always indicate the quintessential image, especially in the context of medical MR images. High objective image quality has no bearing if the image is not useful for its specific diagnostic purpose, while the opposite may also be true in that, despite low objective image quality, an image might still be useful for its purpose. Although common in MRI literature, the idea that a higher IQM performance yields a better diagnostic image is extrapolatory thinking, and it should be challenged.

2.3.1. Root Mean Square Error

RMSE [Eqn. 3] compares the image quality between the reference and altered images, both in the spatial domain, by calculating the average pixel-by-pixel (i.e. voxel-by-voxel for MR images) difference via the following formula:

$$RMSE\ score = \sqrt{\frac{\sum(Reference\ Image - Altered\ Image)^2}{N}}$$

[Eqn. 3]

In theory, mathematically, RMSE scores could range from [0, +∞), but, in practice, the maximum of the range will be bound by the images input to the calculation. For RMSE, decreasing scores indicate better objective image quality (a score of zero indicates the best RMSE image quality score).

2.3.2. The Structural Similarity Index

The Structural SIMilarity (SSIM) index was developed based on the assumption that the human visual system can extract structural information from an image, whereby changes in the extracted structural information are perceived as image distortions [61]. SSIM quantifies similarities in luminance, contrast, and structural information between reference and altered images. SSIM scores range from zero to one, whereby scores decreasing from one to zero indicate best to worst SSIM image quality scores, respectively. See [61] for more detailed information about SSIM.

2.3.3. The Feature Similarity Index

The Feature SIMilarity (FSIM) index compares the image quality between the reference and altered images via a complex function. FSIM relies on low-level details and operates on the basis that "visually discernable features coincide with points where the Fourier waves at different frequencies have congruent phases" [39]. This means that, based on physiological and psychophysical evidence, the human visual system is able to visually discern features that are at points of high phase congruency [39]. FSIM also considers changes in pixel intensity as per the gradient amplitude in order to take into account the effects that contrast imparts on the human visual system [39]. FSIM scores range from zero to one, whereby scores decreasing from one to zero indicate best to worst FSIM image quality scores, respectively. See [39] for more detailed information about FSIM.

2.3.4. The Noise Quality Measure

The noise quality measure (NQM) was developed to quantify the impact of the degradation type, called *additive noise*, on the human visual system [62]. NQM accounts for how contrast sensitivity (i.e. the ability of the human visual system to discern between an object and its background) varies with the following: (1) the distance from which the image is viewed, (2) the image dimension, and (3) the spatial frequency content [62]. NQM also accounts for variations in how the human visual system perceives contrast, such as "local luminance mean" [62], "contrast interaction between spatial frequencies" [62], and "contrast masking effects" [62]. NQM scores range from zero to positive infinity, whereby increasing scores indicate better objective image quality (a score of zero indicates the worst NQM image quality score). See [62] for more detailed information about NQM.

2.3.5. The Visual Information Fidelity Criterion

The visual information fidelity (VIF) criterion was developed to quantify image information lost from the reference image to the altered image [63]. VIF computations include the use of a human visual system model to quantify the extent to which information perceived by the human visual system from the altered image is identical to that which is perceived from the reference image [63]. VIF scores range from zero to one, whereby scores decreasing from one to zero indicate best to worst VIF image quality scores, respectively. See [63] for more detailed information about VIF.

Chapter 3: Methods

3.1. Data Acquisition

3.1.1. Nova Scotia Health Research Ethics Board Approval

The NSH REB-approved protocols submitted for the purposes of this thesis were: (1) 0.5 T Technique Development: "Development and Optimization of Point-of-Care Magnetic Resonance Imaging," NSH-REB ROMEO #: 1025269 (abbreviated as *TD*); and (2) CT Negative Stroke Protocol: "Use of 0.5 T MRI in CT Negative [AIS]," NSH-REB ROMEO #: 1026395 (abbreviated as *CTneg*).

3.1.2. Participant Recruitment and Demographics

Informed consent was obtained from all potential participants of this research work by a Research Coordinator, and MRI screening was performed by an MRI Technologist prior to all research scans. Post-scan, the participant was asked by the Research Coordinator to complete an exit questionnaire. Once the participant completed the exit questionnaire, their participation in the study was complete. Eligible patients were recruited on the basis of an attending neurology physician identifying them as potential participants of this research work.

Under the TD protocol, recruited patients were those diagnosed with AIS on initial POC CT scans during their ED clinical care. A total of 15 patients were recruited under the TD protocol who were de-identified as TD 001-015. The following 12/15 patients (5F, 7M, average age 64y) completed participation and were therefore included in the research work of this thesis: TD 002, TD 004-005, and TD 007-015. The following 3/15 patients were excluded from the research work of this thesis: TD 001, 003, and

006. TD 001 withdrew their consent to participate in the study due to claustrophobia, k-space data were incompletely saved for TD 003, and TD 006 was recruited 9 days post-stroke. The latter initiated the process by which a neuroradiologist checked all cases of patients with diagnosed or suspected AIS to ensure these patients had received their clinical CT scan within 72 hours (later changed to 120 hours) prior to recruitment. If outside the specified window (72 or 120 hours), patients were not eligible for the study and were therefore not recruited to participate. Further NSH REB-approval was granted under the TD protocol for the recruitment of volunteers as healthy controls. There were a total of 2 participants recruited as healthy participants who were de-identified as STRHC 001-002. 2/2 participants (2F, 0M, average age 27.5y) completed participation and were therefore included in the research work of this thesis.

Under the CTneg protocol, recruited patients were those with suspected AIS but with unconfirmed AIS on initial POC CT scans during their ED clinical care. As of May 7, 2021, there were a total of 5 patients recruited under the CTneg protocol who were de-identified as CTneg 001-005. The following 4/5 patients (3F, 1M, average age at acute symptoms 60.5y) completed participation and were therefore included in the research work of this thesis: CTneg 001-002 and CTneg 004-005. CTneg 003 was excluded from the research work of this thesis due to exhaustion, which prevented the patient from completing the scanning process.

In summary, data were acquired from a total of 18 participants (10F, 8M, average age 59y). 16/18 participants were diagnosed and/or suspected AIS patients (8F, 8M, average

age 63y). 2/18 participants were volunteers as healthy controls (2F, 0M, average age 27.5y).

3.1.3. Participant Imaging Information and Scanning Parameters

Patients recruited under the TD protocol were scanned by the responsible personnel (e.g. MRI Technologist) on the POC 0.5 T MRI system once they had completed their SOC during their emergency room visit and were in stable condition as admitted inpatients. All healthy control participants recruited under the TD protocol were scanned by the responsible personnel (e.g. BIOTIC Research Scientist) on the POC 0.5 T MRI system during a time at their own convenience since they were not receiving clinical care.

Patients recruited under the CTneg protocol were both admitted as inpatients and still in the process of receiving their clinical care, but deemed stable by the treating physician, when they were scanned on the POC 0.5 T MRI system. The patient received their 1.5 T or 3 T clinical MRI scan whenever they were able to be scheduled in; therefore, depending on availability, some patients received their clinical MRI scan prior to their 0.5 T research MRI scan, while others received their research scan before their clinical scan. Nonetheless, all patients recruited under the CTneg protocol were scanned on the 0.5 T MRI system during Wednesday or Friday research imaging hours within a maximum of 72 hours (which was later changed to 120 hours) after receiving their clinical CT scan.

The TD protocol set the sequences and parameter specifications for imaging of participants on the POC 0.5 T MRI system, whereby the appropriate personnel (e.g. clinicians, BIOTIC Research Scientists, etc.) built from the clinical stroke protocol used on

the 1.5 T MRI systems at the QEII Health Sciences Centre. The following series of sequences were included, listing scan times in parentheses: axial SWI (69 s), axial T2 FLAIR (266 s), axial T2 FSE (255 s), oblique TOF MRA (458 s), sagittal T1 FLAIR (264 s), and DWI/ADC (97 s). Occasionally scans were repeated, if necessary. The order of scans was not always the same patient to patient.

TABLE II: Parameters for axial (a) T2 FLAIR and (b) DWI/ADC sequences from the TD protocol.

	TR (ms)	**TE (ms)**	**TI (ms)**	**resolution (mm x mm)**		**matrix size**		**FOV (mm x mm)**	**NEX**	**b-value (s mm^{-2})**	**PI**	**# of slices**
				acquired at	*interpolated to*	*acquired at*	*after interpolation*					
(a) Axial T2 FLAIR	5893	86	1904	1.0 x 1.2	0.5 x 0.5	250 x 216	512 x 512	250 x 250	3	---	---	28
(b) Axial DWI/ADC	3945	83	---	2.0 x 2.0	0.9375 x 0.9375	120 x 120	256 x 256	240 x 240	---	0 & 1000	2x	28

T2 = transverse relaxation time; FLAIR = fluid attenuated inversion recovery; DWI = diffusion weighted imaging; ADC = apparent diffusion coefficient; TD = technique development; TR = repetition time; TE = echo time; TI = inversion time; FOV = field of view; NEX = number of excitations; PI = parallel imaging.

Parameters for relevant sequences to this thesis are listed in Table II.

3.2. Data Processing: Retrospective Compressed Sensing Data Pipeline

A data pipeline was coded in MATLAB 2021b to implement retrospective MR image acquisition acceleration via CS, to compute IQM scores, and to generate Digital Imaging and Communications in Medicine (DICOM) images. Raw fully acquired (at NEX = 3) k-space data from the prospective axial T2 FLAIR scans of all 18 study participants were saved and input to the data pipeline. Axial DWI images and their corresponding ADC maps were not processed via the data pipeline and therefore remained non-accelerated; the images used were those acquired directly from the POC 0.5 T MRI system via Synaptive Medical's reconstruction pipeline.

Access to the internal workings of Synaptive Medical's reconstruction pipeline software was unavailable during the timeframe of this thesis work. Therefore, the coded

pipeline was implemented such that its non-altered, fully acquired (retrospectively fully sampled) reference images (R = 1X) were as similar as possible, qualitatively, to the images output from Synaptive Medical's reconstruction pipeline (i.e. images acquired directly from the 0.5 T MRI system from prospective scans).

3.2.1. Undersampling and Compressed Sensing Reconstruction Implementation

All k-space data input to the pipeline had the structure [500 frequency encoding lines x 216 phase encoding lines x 28 slices x 3 averages (NEX) x 16 channels]. This means that each 2D k-space grid was 500 x 216, corresponding to (500 x 216) pixels in image space. The acquired image space FOV, based on k-space data acquisition parameters, was (500 x 200) mm, with pixel width (i.e. image resolution) of (1 x 0.926) mm.

Pre-processing was implemented prior to retrospective undersampling and CS reconstruction, which involved combining k-space averages to yield a data structure of [500 frequency encoding lines x 216 phase encoding lines x 28 slices x 16 channels], and zero-padding to match Synaptive's reconstruction and FOV. Specifically, the ideal image space FOV, based on DICOM dimensions of (512 x 512) pixels, was (250 x 250) mm, and the ideal pixel width, based on DICOM dimensions and desired image space FOV, was (0.488 x 0.488) mm. Therefore, k-space data were zero-padded to yield a data structure of [1024 frequency encoding lines x 410 phase encoding lines x 28 slices x 16 channels] with (0.488 x 0.488) mm pixel width, based on the acquired image space FOV.

Pre-processed fully-sampled k-space data corresponding to each participant's T2 FLAIR scan were then undersampled from R = 1X to each R = 2-7X via a linear 2D

Cartesian undersampling pattern (i.e. sampling mask). k-Space data at each R = 1-7X, corresponding to each participant's T2 FLAIR scan, were reconstructed via CS (Ψ = ℓ_1-wavelet, λ = 0.01) using the Berkeley Advanced Reconstruction Toolbox (BART) [36], yielding 18 reference, and 108 undersampled, T2 FLAIR image datasets. Images reconstructed from fully sampled k-space data (i.e. R = 1X) were the *reference* images, while images reconstructed from undersampled k-space data (i.e. R = 2-7X) were the *undersampled* images.

Post-processing was implemented in image space after retrospective undersampling and CS reconstruction, which involved cropping from 1024 pixels to 512 pixels (in the dimension corresponding to frequency encoding in k-space) and zero padding from 410 pixels to 512 pixels (in the dimension corresponding to phase encoding in k-space) to maintain the (0.488 x 0.488) mm pixel width.

Based on recommendations in the literature [1], and input from the clinical team at NSH, the T2 FLAIR images and DWI with corresponding ADC maps were the focus since they are the images primarily used by neuroradiologists to report AIS and chronic stroke. In terms of relevance to a rapid MRI stroke protocol, the T2 FLAIR acquisition was the one chosen to accelerate because it was the much longer acquisition at 266 s. Not only was the DWI acquisition shorter at 97 s, and therefore its acceleration seemed less impactful, the capability to save a useable form of the fully sampled k-space data from prospective DWI scans on the 0.5 T MRI system did not exist. Further to these points, there was limited access to fully sampled k-space data from the other sequences acquired under the TD protocol (i.e. they were less amendable to acceleration), not to

mention the fact that one of the other sequences was already considered rapid relative to the other sequences in the protocol.

All sampling masks were generated in MATLAB via genPDF (part of the SparseMRI software package [79]) with parameters set such that the central eighth of k-space data were fully-sampled with a polynomial decay rate (p) of seven. Although there are many pseudo-random undersampling patterns to choose from to implement CS reconstruction, a 2D Cartesian pseudo-random linear undersampling pattern is a standard example [37]. As defined within genPDF, p = 7 was chosen based on what was both allowable and consistent given that both the central eighth of k-space data were always fully-sampled while the entire grid of k-space data were to be undersampled at various R. The choice to fully-sample the central eighth of k-space was empirical. Further, despite it being more common for image reconstruction to be implemented in 2D given a slice selection sequence and therefore 2D undersampling of k-space data, the BART-based CS image reconstruction was implemented in 3D to allow for normalized pixel intensity across all slices, rather than on a per slice basis.

In terms of the rationale behind choosing the reconstruction method, CS is a well-known and well-tested MR image reconstruction method that denoises images determined from undersampled data relative to a simple FT. BART is a MATLAB-based computational MRI toolbox that can be used to implement PI- and CS-based image reconstructions [36] and is accessible as a Github-downloadable, free, and open-source toolbox. Within BART, ESPIRiT calibration was used to generate coil sensitivity information, and the 'pics' command was used to implement CS reconstruction, which

included the coil sensitivity information. Ψ and λ were chosen to be ℓ1-wavelet and 0.01, respectively, because these are relatively standard examples of CS parameters in the literature [55]. While it is acknowledged that the chosen parameters may not be optimal in this specific scenario, given that there are no known explicitly similar examples to reference within the literature, CS parameter tuning is out of the scope of this research work.

Pseudocode

1 **For** each participant's T2 FLAIR image dataset
2 Load raw fully sampled k-space data acquired at NEX = 3
3 **For** each R = 2-7X
4 Set λ
5 Set coil sensitivity mapping parameters
6 Set reconstruction parameters
7 Sum raw fully sampled k-space data across NEX dimension
8 Zero pad each slice to DICOM-equivalent dimensions, retaining acquired FOV
9 FT slice dimension from spatial to frequency domain
10 Load linear 2D Cartesian sampling mask
11 Zero pad to equivalent dimensions as above (line 8)
12 Undersample k-space data (line 9 * line 11 outputs)
13 Generate coil-sensitivity maps from fully-sampled data output from line 9
14 Generate coil-sensitivity maps from undersampled data output from line 12
15 Reconstruct reference images (input line 9 & line 13 outputs)
16 Reconstruct undersampled images (input line 12 & line 14 outputs)
17 **For** reconstructed reference image slices output from line 15
18 Crop, then zero pad, retaining DICOM-equivalent dimensions & acquired FOV
19 Draw ROI around head to generate binary mask
20 Apply binary mask to zero fill non-ROI pixels (line 18 * line 19 outputs)
21 **For** reconstructed undersampled image slices output from line 16
22 Crop, then zero pad to equivalent dimensions as above (line 18)
23 Apply binary mask to zero fill non-ROI pixels (line 19 * line 22 outputs)

3.2.2. Two-Dimensional Image Quality Metric Computation Implementation

All reconstructed reference and undersampled T2 FLAIR image datasets were prepared for the 2D IQM computations. Part of this preparation was normalizing all

image datasets. As a result, RMSE was normalized to the intensity of the reference image, and therefore ranged from zero to one, indicating best to worst RMSE image quality scores, respectively. RMSE, SSIM, FSIM, NQM, and VIF were then computed for each undersampled image slice (R = 2-7X) based on the corresponding reference image slice (R = 1X).

The choice to use RMSE, SSIM, FSIM, NQM, and VIF comes from previous literature by Mason *et al.* [42], which demonstrated correlations between radiologists' scores of overall diagnostic image quality versus IQM scores for "clinically normal MR image[s]" [42] that were subject to numerous types and levels of degradations. The clinically normal images were from patients, but were selected so long as the 2D slice to be used in the study did not show evidence of pathology [42]. In this study [42], VIF, FSIM, and NQM were shown to correlate best with radiologists' opinions of overall diagnostic image quality, while RMSE and SSIM, although standard IQMs used extensively in MRI literature, did not correlate best with radiologists' opinions. Despite the fact that RMSE and SSIM have been used in the literature, aiming to guide the development of accelerated MRI protocols, they are typically not made relevant to radiologists' opinions, which diminishes the clinical focus of MRI research. The research work in this thesis is therefore focused on assessing correlations between IQMs and neuroradiologists' *task-specific* diagnostic confidence.

Section 3.2.1. Pseudocode continued…

```
23      Determine maximum dynamic range out of reference and undersampled
          images datasets
24      Normalize maximum dynamic range of reference and undersampled image
          datasets to 2^15
```

25 Eliminate all slices corresponding to 2D array of zeros
26 Compute 2D IQMs (RMSE, SSIM, FSIM, NQM, VIF) via GetMetric function
27 Export results to spreadsheet file.

3.2.3. DICOM Image Library Generation

DICOM images were generated and saved. Each patient dataset consisted of the following 9 images: the reference T2 FLAIR image (R = 1X) and its corresponding undersampled images (R = 2-7X), the reference DWI image and its corresponding ADC map.

Section 3.2.2. Pseudocode continued…

28 Cast all reference and undersampled images to 16-bit unsigned integers
29 **For** all reference image slices
30 Obtain DICOM header info
31 Read DICOM data info
32 Generate new DICOM series instance but keep current study instance
33 Generate new DICOM series number and series description
34 Generate new DICOM window width and window center
35 **For** all undersampled image slices
36 Repeat lines 30-34
37 Save DICOM reference and undersampled images into separate image libraries

3.2.4. Image Dataset De-Identification and Randomization

All image datasets were de-identified and randomized to generate the study schedule (*Appendix A.1.*). Images from the same participant were always separated by images from at least two other participants, and image degradation types and levels were dispersed randomly throughout the six-week timeframe.

3.3. Task-Specific Diagnostic Confidence Study with Neuroradiologist Raters

This study was implemented over the course of six weeks (*Appendix A.1.*) with three board-certified neuroradiologists as participating raters. Each image dataset series consisted of a scanned participant's T2 FLAIR image dataset (non-accelerated or

accelerated), DWI image dataset and corresponding ADC map dataset (both non-accelerated). Therefore, for each image dataset series, a T2 FLAIR image dataset, alongside the corresponding DWI image dataset and ADC maps dataset, was shown to neuroradiologists. With 18 total scanned participants, each having a T2 FLAIR reference image dataset (R = 1X) and a T2 FLAIR image dataset at each of the six undersampling factors (R =2-7X), the study included a total of 126 image dataset series shown to each neuroradiologist.

As such, every week for the six-week duration of the study, each neuroradiologist received a new Questionnaire (*Appendix A.2.*) based on the corresponding image dataset series schedule (*Appendix A.1.*) and, for each image dataset series, was individually asked to perform two stroke-specific diagnostic tasks (acute and chronic), as they would in a clinical setting, ranking their diagnostic confidence on a 1-5 Likert scale (0% confidence, 5 = 100% confidence).

Specifically, the acute stroke diagnostic task that neuroradiologist raters were asked to perform was, “Using the DWI and T2 FLAIR images, would you report the presence of an acute stroke?” The answers available for neuroradiologist raters to select were either *YES* or *NO*, where *YES* indicated the presence of an acute stroke and *NO* indicated the absence of an acute stroke. Neuroradiologist raters were then asked to rank their diagnostic confidence in reporting the presence or absence of an acute stroke. The answers available to select were as follows: 0 % confident = 1; 25 % confident = 2; 50 % confident = 3; 75 % confident = 4; and 100 % confident = 5.

The chronic stroke diagnostic task that neuroradiologist raters were asked to perform was, “What Fazekas score (0-3) would you report for identification of chronic ischemic lesion burden?” The answers available for neuroradiologist raters to select were as follows: absent (*= 0*); punctate foci, or “caps” or pencil-thin lining (*= 1*); beginning confluence, or smooth “halo” (*= 2*); or large confluent areas, or irregular periventricular signal extending into deep WM (*= 3*). Neuroradiologist raters were then asked to rank their diagnostic confidence in reporting this Fazekas score. The answers available to select were as follows: 0 % confident = 1; 25 % confident = 2; 50 % confident = 3; 75 % confident = 4; and 100 % confident = 5.

Further, when possible, neuroradiologist rater(s) located the slice(s) best visualizing acute and chronic stroke pathology. If more than one slice was located, the IQM scores for each slice across the range of located slices were averaged. If only a single slice was located, or two adjacent slices were located, adjacent slices on either side were included and therefore the IQM scores for 3 or 4 slices, respectively, were averaged. In the absence of rater assistance and/or pathology, 3-4 non-zero-IQM-score central slices were selected, and IQM scores for these slices were averaged.

Best attempts were made to mimic what is performed in the clinical setting, whereby each image dataset series (the T2 FLAIR image dataset and corresponding DWI image dataset and ADC map dataset) was shown to neuroradiologists in a side-by-side view panel within the software they use in the clinical setting. Additionally, the questionnaire (*Appendix A.2.*) was developed with the help of expert neuroradiologists as a best

attempt to ensure they were clinically-pointed and relevant to the diagnosis of acute and chronic stroke.

Note that images that were part of an additional study, external to this thesis, were included for both efficient utilization of raters' time, as well as to enlarge the dataset associated with the research work of this thesis to mitigate rater recognition bias. An additional tactic to address potential recognition bias was ensuring the inclusion of the following within the datasets: healthy (non-stroke) participants, stroke-positive patients on first-line CT imaging, stroke-negative patients on first-line CT imaging, and of the latter, stroke-positive and stroke-negative patients on MRI.

Calibration

Pre-study, all neuroradiologists completed a single calibration questionnaire following the same implementation as the study described above. The calibration questionnaire was identical to that shown in Appendix A.2., but consisted only of the corresponding combinations of Patient ID and Acceleration Factor image datasets: TD 009-010 at R = 1X, CTneg 002 at R = 3X, CTneg 001 at R = 5X, and both TD 002 and STRHC 002 at R = 7X. The calibration questionnaire served as a representative example of image datasets shown to neuroradiologist raters prior to the main implementation of the study to ensure their perceptions of the Likert scale were aligned with one another, and thus "calibrated" in their ratings going forward. In addition, neuroradiologists were sent a calibration email on a weekly basis (*Appendix A.3.*) to mitigate deviations from the initial calibration and to remind them that a high Likert score should be reported to indicate findings that they view to be either a true positive or a true negative, while a

low Likert score should be reported to indicate findings that they view to be either a false negative or a false positive.

Verification

Post-study, all neuroradiologists completed a single verification questionnaire following the same implementation as the study described above. The verification questionnaire was identical to that shown in Appendix A.2., but included only the first two questions (acute stroke diagnostic task and associated diagnostic confidence) and consisted of only the corresponding combinations of Patient ID and Acceleration Factor image datasets: TD 002 and CTneg 002 at R = 4X; TD 009, CTneg 001, and STRHC 001 at R = 5X; and TD 007, CTneg 004, and STRHC 001 at R = 6X. The verification questionnaire served as a representative example of image datasets shown to neuroradiologist raters after the main implementation of the study to "verify" that a range of confidence ratings remained consistent and also to address any potential "errors" observed in the weekly questionnaires, such as typos, inaccurate stroke/no stroke results, empty questionnaire fields, etc.

3.4. Data Analysis

To determine how to test Hypotheses A and B (*Section. 1.6.2.*) with statistical significance, the following algorithm was followed for data resultant for both the acute and chronic stroke diagnostic tasks: if neuroradiologists were in fair agreement then their confidence scores corresponding to each R were to be pooled at each R. Fair agreement was deemed as the inter-rater reliability measurement > 0.20, where the inter-rater reliability measurement was either Cohen's kappa (κ) [90] or Gwet's

Agreement Coefficient (AC) [91]. Note that κ is computed using the quantities observed agreement (Pa) and expected agreement by chance (Pe), whereas Gwet's AC adjusts for agreement by chance.

If κ or Gwet's AC ≤ 0.20 then neuroradiologists were to be reported as individual raters at each R. Depending on these results, it was then to be determined whether pooled or individual rater data were Gaussian via kurtosis calculations. Gaussianity was defined as a kurtosis value between 2.0 and 4.0 (inclusive). If the distribution of data were Gaussian, as defined by a kurtosis between 2.0 and 4.0 (inclusive), mean ± SD at each R value would be reported (per rater, if poor inter-rater reliability), plotting bar graphs of the mean including corresponding error bars of the SD, at each R value. Results were then to be plotted as confidence scores versus acceleration factor, and linear and quadratic trends were to be fit to the data to assess statistical significance of trends. However, if the distribution of data were non-Gaussian, as defined by a kurtosis < 2.0 or > 4.0, then boxplots were generated to analyze the distribution of diagnostic confidence scores at each R value, reporting the first quartile (Q1), the second quartile (Q2, also median), the third quartile (Q3), the interquartile range (IQR), the minimum and maximum values of the range, and quantile skewness. Note that quantile skewness is a measure of whether the distribution is asymmetric about the median, while indicating the directionality of the skewness (e.g. positive value indicates right skewness, negative value indicates left skewness).

Since performing the acute stroke diagnostic task involved binary decision-making (i.e. deciding whether stroke was present or absent), neuroradiologist raters were either

accurate or inaccurate in their performance of the acute stroke diagnostic task. An accurate response was either: (i) if stroke was present and the neuroradiologist rater selected *YES*, or (ii) if stroke was absent and the neuroradiologist rate selected *NO*. An inaccurate response was either: (i) if stroke was present and the neuroradiologist selected *NO*, or if stroke was absent and the neuroradiologist rater selected *YES*. As such, the neuroradiologist raters' accuracy in performing the acute stroke diagnostic task could be calculated; however, since the associated diagnostic confidence in performing the acute stroke diagnostic task is subjective, accuracy in diagnostic confidence scoring cannot be calculated.

Neuroradiologist raters' accuracy in performing the chronic stroke diagnostic task could not be calculated because diagnosis of chronic stroke is undefined in the sense that the Fazekas score is determined based on the subjective (albeit, expert) opinion of the neuroradiologist rater – unlike the acute stroke diagnostic task, which involves binary decision-making and thus allows neuroradiologist raters' accuracy in performing the task to be calculated. When it comes to Fazekas scoring, there is no dichotomy, but rather a sliding scale of intensity of chronic ischemic lesion burden. Similar to the acute stroke diagnostic task, however, the associated diagnostic confidence in performing the chronic stroke diagnostic task is subjective and therefore accuracy in diagnostic confidence scoring cannot be calculated.

Results were compared to those of the *Calibration* (*Section 3.3, pg. 64*) datasets and were highlighted as either *Accurate with Calibration Score, Consistent with Calibration Score,* or *Inconsistent with Calibration Score* (*Appendix A.4.*). Results were also

compared to those of the *Verification* (*Section 3.3., pg. 65*) datasets and were highlighted as *now Accurate, remained Accurate,* r*emained Consistent*, or *Inconsistent with Verification Score* (*Appendix A.4.*).

3.4.1. Hypothesis A Testing

Inter-Rater Reliability Computations

The following inter-rater reliability measurements were computed in RStudio (2021, Version 1.4.1106) corresponding to both the acute and chronic stroke diagnostic task confidence scores for all 126 images (7 R values x 18 participants) between each of the three neuroradiologist raters: pairwise unweighted Cohen's kappa (κ_{UW}) ± 95 % confidence interval (CI), pairwise quadratic weighted Cohen's kappa (κ_{QW} ± 95 % CI), pairwise Gwet's unweighted Agreement Coefficient (AC1) ± 95 % CI and pairwise Gwet's quadratic weighted Agreement Coefficient (AC2) ± 95 % CI.

Statistical Analysis

Gaussianity of confidence score data were determined via kurtosis computations performed in MATLAB. Boxplots were generated corresponding to neuroradiologist raters' pooled Likert scores at each R for the acute and chronic stroke diagnostic task confidence scores; and Q1, median (Q2), Q3, IQR, range, and quantile skewness were reported. Quantile skewness was calculated via [(Q3 – Q2) – (Q2 – Q1)] / (Q3 – Q1).

Wilcoxon signed-rank (exact) p-values were computed, and clinically significant p-values were reported based on post-hoc computed Bonferroni corrected Wilcoxon signed-rank (exact) p-values. Based on the corrected Wilcoxon signed rank test results for comparisons between R = 1X and R > 1X for both the acute and chronic stroke

diagnostic tasks, Hypothesis A (*Section 1.6.2.*) could be concluded. All Wilcoxon signed-rank statistical tests were performed in RStudio.

3.4.2. Hypothesis B Testing

Hypothesis B (*Section 1.6.2.*) was measured using regression analysis and statistical testing based on the methods used by Mason *et al.* [42] and previous studies [68], which were adapted to suit the nature of the data resultant from the research work of this thesis.

Regression Analysis

Neuroradiologist raters' diagnostic confidence scores for undersampled images were evaluated as pooled and averaged raw Likert scores from 1-5 for the acute stroke diagnostic task and as pooled and averaged raw Likert scores that were rescaled from 0-100 for the chronic stroke diagnostic task. Note that pooling and averaging of scores was allowable by the results of the inter-rater reliability computations from Hypothesis A testing. Whether pooled and averaged data were plotted as raw Likert scores or rescaled Likert scores depended on the interquartile range values corresponding to neuroradiologists' diagnostic confidence scores across R = 2-7X determined from Hypothesis A testing.

For all T2 FLAIR undersampled images, neuroradiologist raters' diagnostic confidence scores were separately plotted for each task versus objective IQM scores. The plotted data were fit to a constrained logistic function for a non-linear regression model in MATLAB using lsqcurvefit. Corresponding sum of squares residuals (SSR) and Spearman rank order correlation coefficient (SROCC) values were computed in MATLAB. SSR and

SROCC represent the logistic model's goodness-of-fit to the plotted data and the correlation of the plotted data, respectively.

Statistical Testing

Wilcoxon signed-rank (exact) tests were computed on absolute residuals, and clinically significant p-values were reported based on post-hoc computed Bonferroni corrected Wilcoxon signed-rank (exact) p-values. Based on the corrected Wilcoxon signed rank test results for comparisons between IQMs for both the acute and chronic stroke diagnostic tasks, Hypothesis B (*Section 1.6.2.*) could be concluded. All Wilcoxon signed-rank statistical tests were performed in RStudio.

Chapter 4: Results and Discussion

In this chapter, results and discussion will be provided in aggregate, transitioning between each result and its accompanying discussion. The only results and discussion that will not be provided in aggregate are those corresponding to the retrospectively accelerated MR images output from the data pipeline; the results will be presented in the following section (*Section 4.1.*), but the discussion will follow, separately, in Section 4.3.

4.1. Neuroradiologist Raw Scoring

In this section (*Section 4.1.*), neuroradiologists' raw scoring data will be presented in order to portray raw agreement results and therefore aid in the examination of inter-rater reliability. Hence, standard measures of computed inter-rater reliability will be presented next as part of Hypothesis A (*Section 1.6.2.*). *Calibration* and *Verification* results (*Appendix A.4.*) will also be commented on.

A total of 126 image datasets (18 participants * 7 R values) were scored by each neuroradiologist rater, yielding a total of 378 scores (54 scores = 18 participants * 3 neuroradiologist raters, per R) for each of the following: (a) acute stroke: (i) absence/presence scores and (ii) corresponding diagnostic confidence scores; (b) chronic stroke: (i) Fazekas scores and (ii) corresponding diagnostic confidence scores.

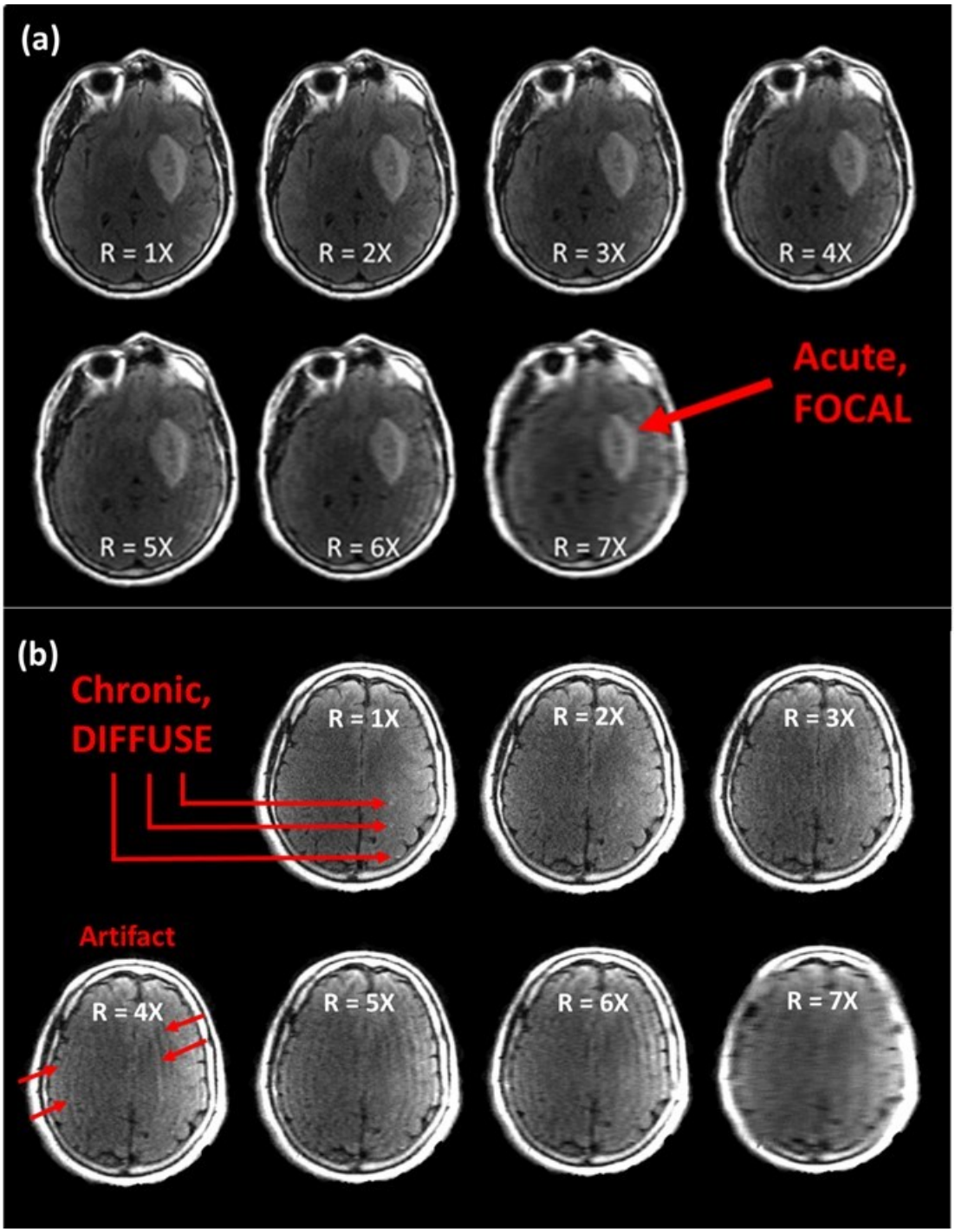


Fig. 18. Example T2 fluid attenuated inversion recovery (FLAIR) images output from the undersampling and compressed sensing (CS) reconstruction pipeline from a single recruited acute ischemic stroke (AIS) patient. Images show (a) acute stroke lesions, corresponding to an identical anatomical slice and (b) chronic stroke lesions, corresponding to an identical anatomical slice, both at R = 1X (fully-sampled reference images) and R = 2-7X (undersampled images). Undersampling resulted in variations in image quality.

Fig. 18 shows an example of 2D slices selected from an accelerated MR image dataset output from the data pipeline corresponding to a single recruited AIS patient. The same 2D slice is shown for R = 1-7X for AIS, while a different 2D slice (but from the same image dataset) is shown for R = 1-7X for chronic stroke. These slices, in particular, were chosen in a best attempt to illustrate the impact of accelerating image acquisition via CS reconstruction on image quality in the presence of AIS (Fig. 18 a) and chronic stroke (Fig. 18 b) pathology. Presenting these images early in this chapter serves as a visual guide to aid in the understanding of what might impact neuroradiologist raters' task-specific diagnostic confidence scores based on the image alterations observed, laying the foundation for some of the conclusions that are drawn.

4.1.1. Acute Stroke Diagnostic Task

The acute stroke diagnostic task that neuroradiologist raters were asked to perform was, "Using the DWI and T2 FLAIR images, would you report the presence of an acute stroke?" where accuracy in performing the task could be calculated. In this section (*Section 4.1.1.*), raw pairwise inter-rater agreement between Raters 1, 2, and 3 in performing the acute stroke diagnostic task will be outlined.

Results

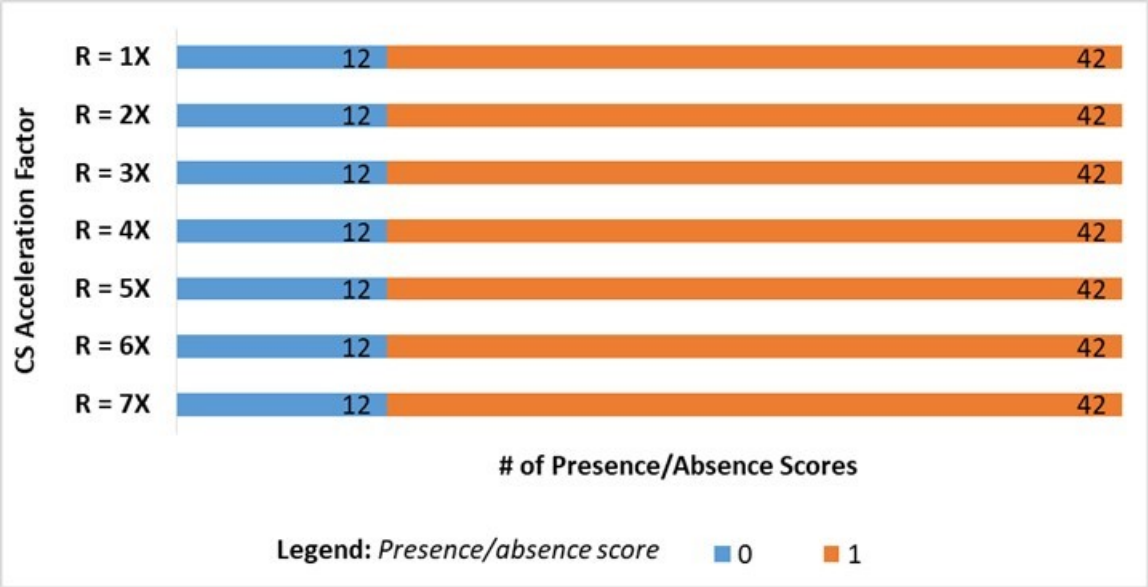


Fig. 19. Neuroradiologist raters' binary scores in reporting the presence or absence of acute stroke: 0 = absence of stroke, 1 = presence of stroke. Each row tallies to 54 scores (18 participant image datasets * 3 neuroradiologist raters, per R).

Fig. 19 illustrates the tally of neuroradiologist raters' presence or absence scores at each R value, whereby rows are lined up by the undersampling factor, R, and columns are lined up by the presence/absence scores.

18/18 (100 %) *Calibration* (*Section 3.3., pg. 64*) dataset scores were highlighted as *Accurate with Calibration Score* (*Appendix A.4.*). 4/24 (17 %) *Verification* (*Section 3.3., pg. 65*) dataset scores were highlighted as *now Accurate*, while 20/24 (83 %) were highlighted as *remained Accurate* (*Appendix A.4.*).

Discussion

Raters 1, 2 and 3 all yielded 100 % accuracy in performing the acute stroke diagnostic task, which also corresponded to 100 % accuracy to the *Calibration* results. That is, all neuroradiologist raters selected *YES* when an acute stroke was present or selected *NO*

when an acute stroke was absent. Raw inter-rater agreement in performing the acute stroke diagnostic task was therefore 100 % (126/126 images) between: (a) Rater 1 and Rater 2, (b) Rater 1 and Rater 3, and (c) Rater 2 and Rater 3, yielding an average raw inter-rater agreement across all raters of 100 %.

Since accuracy in performing the acute stroke diagnostic task could be calculated, scores were compared using the *accuracy* rather than *consistency*. The fact that there were 4/24 scores highlighted from the *Verification* results as *now Accurate* is indicative of a limitation corresponding to the acute stroke diagnostic task results since the 4 *now Accurate* scores replaced the original scores for the corresponding datasets and thus yielded the 100 % accuracy in performing the acute stroke diagnostic task. Nonetheless, the 4 original scores were deemed to be errors, and, when consulted post-*Verification*, all neuroradiologist raters confirmed that an error had been made on their part, whether it be a simple oversight (e.g. empty questionnaire fields) or typographical error (e.g. inaccurate stroke/no stroke results, typos, etc.). The combination of all *Verification* study results yielding accurate scores (without neuroradiologist raters' knowledge of why specific datasets were chosen for the V*erification*) along with neuroradiologist raters' confirmation post-*Verification* that they had initially made errors, rendered it reasonable to replace the scores and move forward with analyses.

4.1.2. Diagnostic Confidence in Acute Stroke Task

Neuroradiologist raters were asked to rank their diagnostic confidence in performing the acute stroke diagnostic task on a Likert scale from 1-5 (with 1 being no confidence and 5 being 100 % confident). In this section (*Section 4.1.2.*), raw pairwise inter-rater

agreement between Raters 1, 2, and 3 in their associated diagnostic confidence in performing the acute stroke diagnostic task will be outlined.

Results

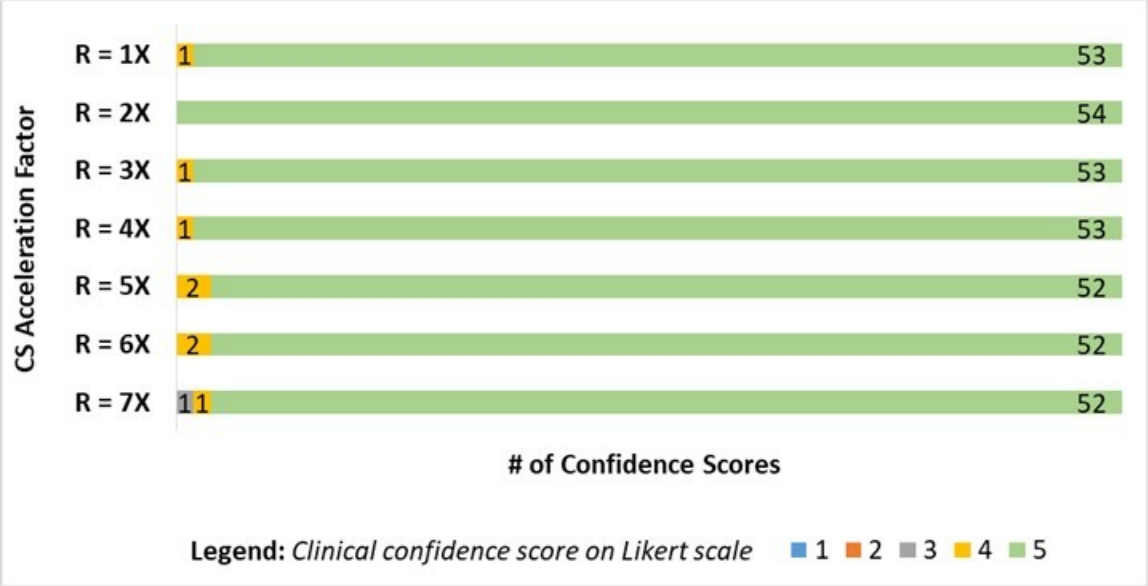


Fig. 20. Neuroradiologist raters' diagnostic confidence scores in reporting the presence or absence of acute stroke on a 5-point Likert scale: 1 = 0 % confident, 2 = 25 % confident, 3 = 50 % confident, 4 = 75 % confident, 5 = 100 % confident. Each row tallies to 54 scores (18 participant image datasets * 3 neuroradiologist raters, per R).

Fig. 20 illustrates the tally of neuroradiologist raters' diagnostic confidence scores at each R value, whereby rows are lined up by the undersampling factor, R, and columns are lined up by the diagnostic confidence scores.

16/18 (89 %) *Calibration* (*Section 3.3., pg. 64*) dataset scores were highlighted as *Consistent with Calibration Score* (*Appendix A.4.*), while 2/18 (11 %) were highlighted as *Inconsistent with Calibration Score*. Both inconsistent scores were from Rater 3, such that, in both cases, their original score was a Likert scores of 4, but their *Calibration* score was a 5.

23/24 (96 %) *Verification* (*Section 3.3., pg. 65*) dataset scores were highlighted as *remained Consistent*, while 1/24 (4 %) was highlighted as *Inconsistent with Verification Score* (*Appendix A.4.*). The one inconsistent score was from Rater 3, such that their original score was a Likert score of 5, but their *Verification* score was a 4.

Discussion

Despite Raters 1, 2 and 3 all yielding 100 % accuracy in performing the acute stroke diagnostic task, their associated diagnostic confidence in performing the acute stroke diagnostic task was not always rated a Likert score of 5 (i.e. 100 % confident). Fig. 20 shows that, for the acute stroke diagnostic task, neuroradiologist raters' raw diagnostic confidence scores are mainly Likert scores of 5 for the R = 1-7X dataset, with a minimal distribution of Likert scores across 3-4, and no Likert scores across 1-2. Diagnostic confidence scores corresponding to performing the acute stroke diagnostic task from Fig. 20 can be broken down, as follows: (a) both Rater 1 and Rater 2 had 100 % diagnostic confidence for 126/126 (100 %) images; and (b) Rater 3 had 100 % diagnostic confidence for 117/126 (93 %) images, 75 % diagnostic confidence for 8/126 (6 %) images, and 50 % diagnostic confidence for 1/126 (1 %) images. Raw pairwise inter-rater agreement is therefore as follows: 100 % (126/126 images) between Rater 1 and Rater 2, 93 % (117/126 images) between Rater 1 and Rater 3, and 93 % (117/126 images) between Rater 2 and Rater 3, yielding an average raw inter-rater agreement across all raters of 95 %. The 5 % disagreement is solely from disparities in Rater 3's scoring since Rater 1 and 2 are in perfect raw agreement, which may indicate that Rater 3 could be more conservative in their clinical decision making process.

Since accuracy could not be calculated for diagnostic confidence in performing the acute stroke diagnostic task, *Calibration* and *Verification* scores were compared using *consistency* rather than *accuracy*. Despite the fact that none of the *Calibration* or *Verification* scores replaced the original scores for the corresponding datasets (due to the fact that accuracy could not be determined), it is at least worth noting that inconsistencies existed and therefore may pose as a limitation to this study. It is interesting, nonetheless, that all inconsistences arose from Rater 3, who was suspected to be more conservative in their clinical decision making process. It is also interesting to note that none of the inconsistencies in diagnostic confidence scores corresponded to any of the original 4/24 results in performing the acute stroke diagnostic task that were changed to the *Verification* study results. This seems to further solidify that neuroradiologist raters did in fact simply make oversight or typographical errors in performing the acute stroke diagnostic task for these 4/24 datasets.

4.1.3. Chronic Stroke Diagnostic Task

The chronic stroke diagnostic task that neuroradiologist raters were asked to perform was, “What Fazekas score (0-3) would you report for identification of chronic ischemic lesion burden?” where accuracy could not be calculated considering the Fazekas scale is a subjective scale and there is no gold standard for judging accuracy. In this section (*Section 4.1.3.*), raw pairwise inter-rater agreement between Raters 1, 2, and 3 in performing the chronic stroke diagnostic task will be illustrated.

Results

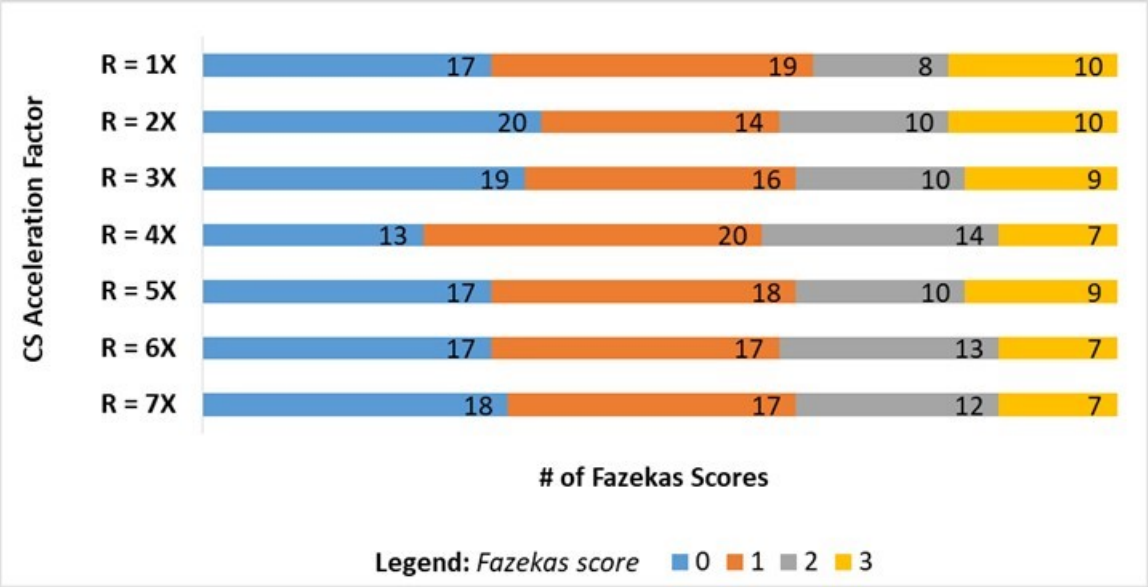


Fig. 21. Neuroradiologist raters' Fazekas scores (i.e. reporting identification of chronic ischemic lesion burden) on the following Fazekas scale: 0 = absent; 1 = punctate foci, or "caps" or pencil-thin lining; 2 = beginning confluence, or smooth "halo"; 3 = large confluent areas, or irregular periventricular signal extending into deep white matter. Each row tallies to 54 scores (18 participant image datasets * 3 neuroradiologist raters, per R).

Fig. 21 illustrates the tally of neuroradiologist raters' Fazekas scores at each R value, whereby rows are lined up by the undersampling factor, R, and columns are lined up by the Fazekas scores (0-3).

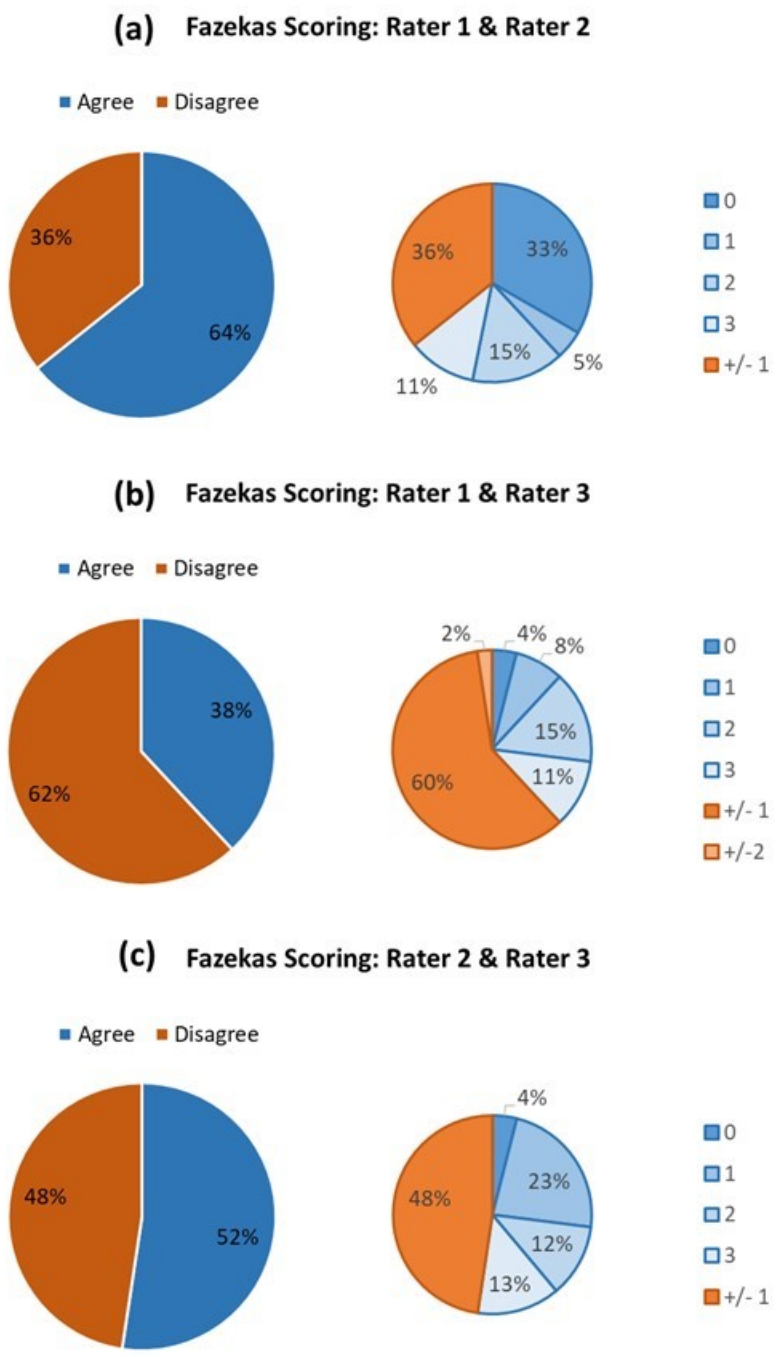


Fig. 22. Raw pairwise inter-rater agreement in the identification of chronic ischemic lesion burden on the Fazekas scale between neuroradiologist raters (a) 1 and 2, (b) 1 and 3, and (c) 2 and 3. Blue represents agreement and orange represents disagreement. Left (a-c): breakdown of raw overall agreement (%) and disagreement (%). Right (a-c): corresponding breakdown into agreeing Fazekas scores and +/- difference in disagreeing Fazekas scores. All percentages are out of 126 total scores (18 participant image datasets * 7 R values) and sum to 100 % per chart.

The pie charts on the left in Fig. 22 illustrate raw overall agreement and disagreement between pairwise raters in Fazekas scoring, shown as percentages of the total number of image datasets. The pie charts on the right in Fig. 22 illustrate the breakdown of the raw overall agreement between pairwise raters in Fazekas scoring, while raw overall disagreement is further broken down into the Fazekas score difference (+/-) between pairwise raters.

8/18 (44 %) *Calibration* (*Section 3.3., pg. 64*) dataset scores were highlighted as *Consistent with Calibration Score* (*Appendix A.4.*), while 10/18 (56 %) were highlighted as *Inconsistent with Calibration Score*. All three neuroradiologist raters were responsible for inconsistent scores.

Discussion

Fig. 21 shows that, for the chronic stroke diagnostic task, neuroradiologist raters' raw Fazekas scores are distributed across all Fazekas scores for the R = 1-7X dataset, with the majority of all scores being 0's or 1's. Taking the results from Fig. 21, the raw pairwise inter-rater agreement between Raters 1, 2, and 3 in performing the chronic stroke diagnostic task can be better visualized by breaking down results into pie charts, as shown in Fig. 22 (a-c).

Average raw inter-rater agreement across all raters is (64 + 38 + 52) % / 3 = 51 % (Fig. 22). Between all 3 raters, Rater 3 overall yielded the least number of Fazekas scores of 0 and the most number of Fazekas scores of 3, which may further indicate that Rater 3 could be more conservative in their clinical decision making process. On the other hand, all pairwise inter-rater disagreements are only different by +/- 1 Fazekas score. The only

exception to this is for Rater 1 and Rater 3, whereby some images in disagreement were different by +/- 2, but this was only the case for 3/126 images.

It is important to note that, clinically, the Fazekas scale might not even be mentioned. However, chronic ischemic lesion burden will be reported, as seen, and the description given may correspond to a particular Fazekas score. The Fazekas scale was therefore used for the purposes of this thesis to create a sense of order, rather than trying to categorize various clinical descriptions of chronic ischemic lesion burden. Nonetheless, in the clinical context, chronic ischemic lesion burden will be subjectively assessed by an expert physician (e.g. neuroradiologist) as either pathologically relevant or non-pathological. For the former, referring the patient to neurology for follow-up care would be necessitated, while a neurology referral may not be necessary for the latter, especially if the chronic ischemic lesion burden is due to the natural aging process, for example.

Since accuracy could not be calculated for performing the chronic stroke diagnostic task and there were no associated obvious typographical errors to correct, a *Verification* study was not performed. Further, *Calibration* scores were compared using *consistency* rather than *accuracy*. Despite the fact that none of the *Calibration* scores replaced the original scores for the corresponding datasets (due to the fact that accuracy could not be determined), it is worth noting that inconsistencies existed and therefore may pose as a limitation to this study. All inconsistences were only ever different by +/- 1 Fazekas score, except one, which was different by +2. This seems to further solidify the fact that Fazekas scoring is highly subjective and accuracy is therefore too difficult to calculate.

4.1.4. Diagnostic Confidence in Chronic Stroke Task

Neuroradiologist raters were asked to rank their diagnostic confidence in performing the chronic stroke diagnostic task on a Likert scale from 1-5 (with 1 being no confidence and 5 being 100 % confident). In this section (*Section 4.1.4.*), raw pairwise inter-rater agreement between Raters 1, 2, and 3 in their associated diagnostic confidence in performing the chronic stroke diagnostic task will be illustrated.

Results

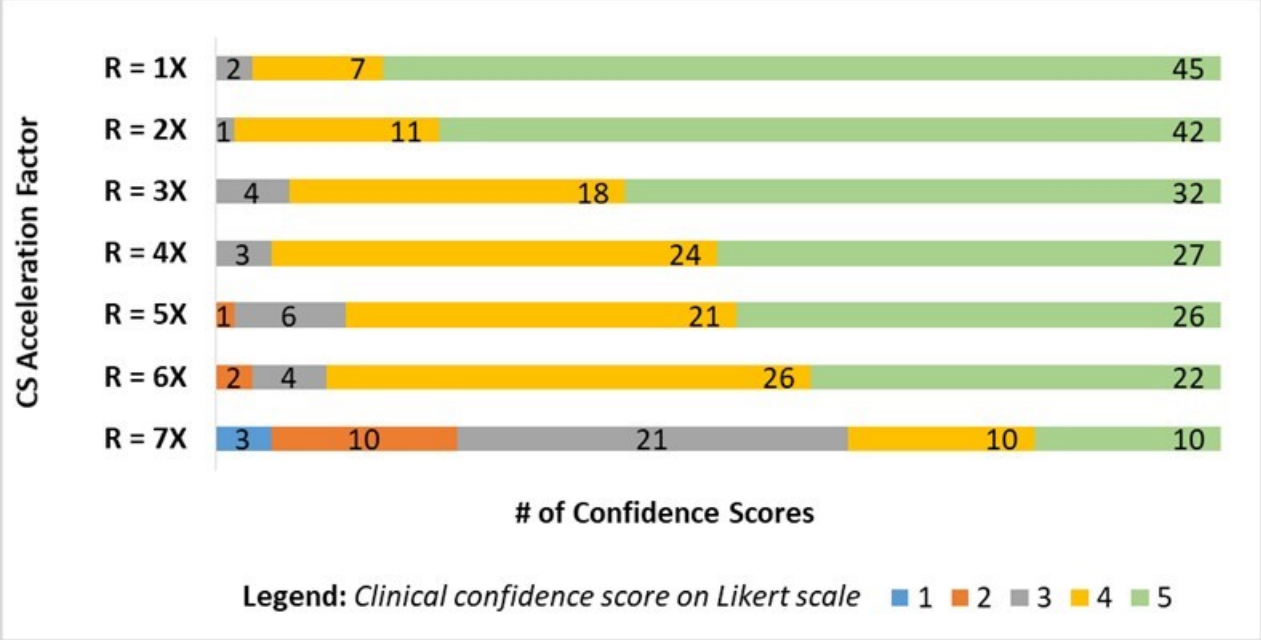


Fig. 23. Neuroradiologist raters' diagnostic confidence scores in reporting Fazekas scores (i.e. identification of chronic ischemic lesion burden) on a 5-point Likert scale: 1 = 0 % confident, 2 = 25 % confident, 3 = 50 % confident, 4 = 75 % confident, 5 = 100 % confident.

Fig. 23 illustrates the tally of neuroradiologist raters' diagnostic confidence scores associated with performing the chronic stroke diagnostic task at each R value, whereby rows are lined up by the undersampling factor, R, and columns are lined up by the diagnostic confidence scores.

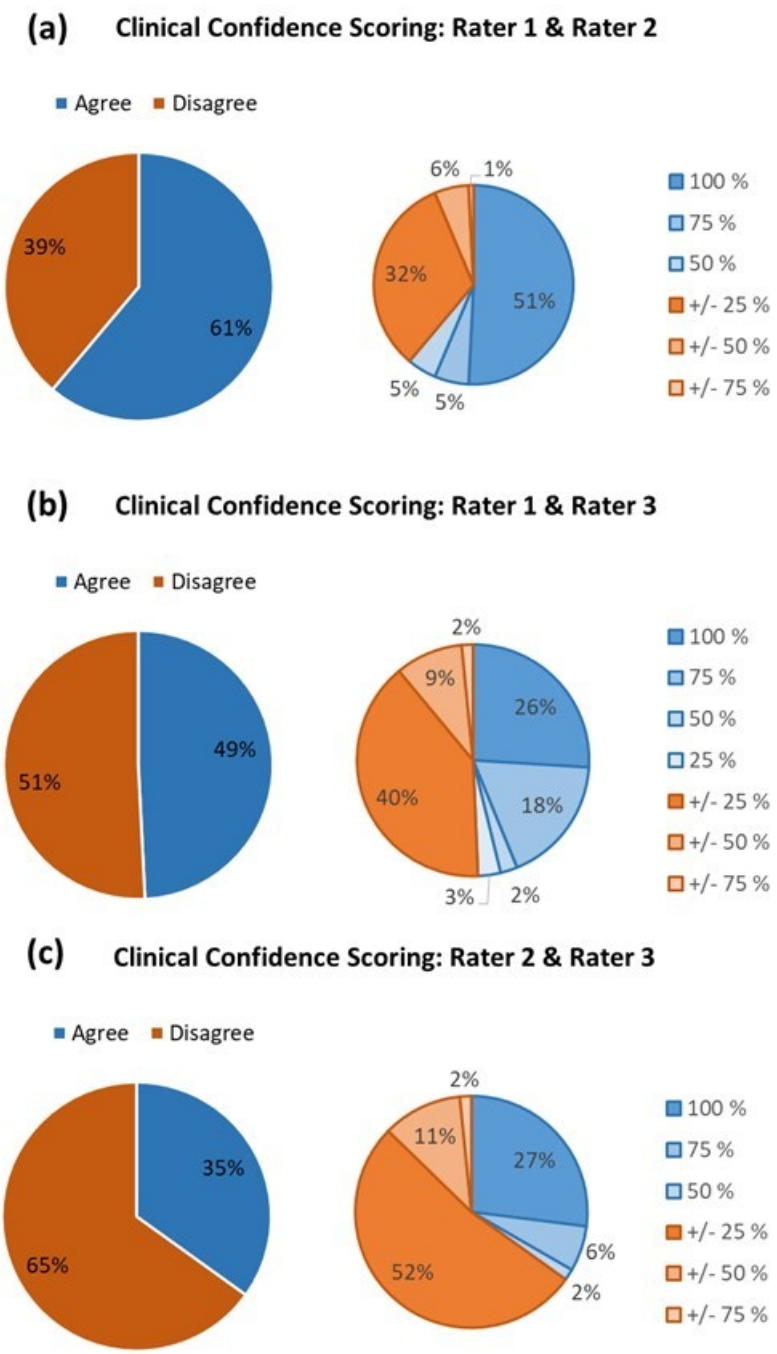


Fig. 24. Raw pairwise inter-rater agreement in diagnostic confidence associated with the identification of chronic ischemic lesion burden between neuroradiologist raters (a) 1 and 2, (b) 1 and 3, and (c) 2 and 3. Blue represents agreement and orange represents disagreement. Left (a-c): breakdown of raw overall agreement (%) and disagreement (%). Right (a-c): corresponding breakdown into agreeing diagnostic confidence scores and +/- difference in disagreeing diagnostic confidence scores. All percentages are out of 126 total scores (18 participant image datasets * 7 R values) and sum to 100 % per chart.

The pie charts on the left in Fig. 24 illustrate raw overall agreement and disagreement between pairwise raters in diagnostic confidence scores, shown as percentages of the total number of image datasets. The pie charts on the right in Fig. 24 illustrate the breakdown of the raw overall agreement between pairwise raters in diagnostic confidence scores, while raw overall disagreement is further broken down into the diagnostic confidence score difference (+/-) between pairwise raters.

10/18 (44 %) *Calibration* (*Section 3.3., pg. 64*) dataset scores were highlighted as *Consistent with Calibration Score* (*Appendix A.4.*), while 8/18 (56 %) were highlighted as *Inconsistent with Calibration Score*. All three neuroradiologist raters were responsible for inconsistent scores.

Discussion

Fig. 23 shows that, for the chronic stroke diagnostic task, neuroradiologist raters' raw diagnostic confidence scores are distributed across all Likert scores, but the distribution tends towards lower diagnostic confidence scores as the dataset's R value increases from 1-7X. Taking the results from Fig. 23, the raw pairwise inter-rater agreement between Raters 1, 2, and 3 in performing the chronic stroke diagnostic task can be better visualized by breaking down results into pie charts, as shown in Fig. 24 (a-c).

Average raw inter-rater agreement across all raters is (61 + 49 + 35) % / 3 = 48 % (Fig. 24). Rater 1 and Rater 3 both yielded diagnostic confidence scores across the entire range [0 %, 100 %]. Rater 2 solely rated diagnostic confidence scores in the range [50 %, 100 %]. Despite Rater 1 and Rater 3 being the only two raters to yield diagnostic confidence scores in the range [0 %, 25 %], they only agreed on 25 % diagnostic

confidence for 4/126 images and never agreed on 0 % diagnostic confidence. Between all raters, Rater 3 yielded the most diagnostic confidence scores in the range [0 %, 25 %] and yielded the least diagnostic confidence scores in the range [50 %, 100 %], which may further indicate that Rater 3 could be more conservative in their clinical decision making process. On the other hand, all pairwise inter-rater disagreements were different by either +/- 25 %, +/- 50 %, or +/- 75 % diagnostic confidence, but the majority were only different by +/- 25 % diagnostic confidence.

Since accuracy could not be calculated for diagnostic confidence in performing the chronic stroke diagnostic task and there were no associated obvious typographical errors to correct, a *Verification* study was not performed. Further, *Calibration* scores were compared using *consistency* rather than *accuracy*. Despite the fact that none of the *Calibration* scores replaced the original scores for the corresponding datasets (due to the fact that accuracy could not be determined), it is worth noting that inconsistencies existed and therefore may pose as a limitation to this study. All inconsistences were only ever different by +/- 1 Likert score, except two, which were different by +2. It is interesting to note that there were not only more inconsistencies in performing the chronic stroke diagnostic task than there were in the associated diagnostic confidence scores (10/18 versus 8/18), but that, out of the 10 Fazekas score inconsistences, only 4 of them also had an inconsistency in the corresponding diagnostic confidence score.

4.2. Hypothesis A Testing

This section (*Section 4.2.*) examines Hypothesis A, which fell under Objective 3 (*Section 1.6.2.*). Objective 3 was to assess neuroradiologists' diagnostic confidence in performing specific stroke-related diagnostic tasks; and Hypothesis A (*Section 1.6.2.*) was as follows: neuroradiologist raters' diagnostic confidence scores corresponding to performing the acute stroke diagnostic task will be less sensitive to increasing R than neuroradiologist raters' diagnostic confidence scores corresponding to the chronic stroke diagnostic task.

To test Hypothesis A, pairwise inter-rater reliability was first computed from raw diagnostic confidence scores between each of the three neuroradiologist raters to determine whether raters' scores could be pooled, or must be kept independent. Since the scores could be pooled and the data were non-Gaussian, boxplots were generated to analyze the distribution of diagnostic confidence scores at each R. Second, Wilcoxon signed-rank (exact) tests (significance level, $p = 0.05$) were performed, reporting post-hoc Bonferroni adjusted p-values.

4.2.1. Cohen's Kappa Inter-Rater Reliability Measurements

Cohen's kappa (κ) is a standard inter-rater reliability measurement reported in MRI literature. This section (*Section 4.2.1.*) computes both unweighted Cohen's kappa (κ_{UW}) and quadratic weighted Cohen's kappa (κ_{QW}) to measure inter-rater reliability across all three neuroradiologist raters' diagnostic confidence scores in performing the acute and chronic stroke diagnostic tasks, using the raw diagnostic confidence score data from Sections 4.1.2. and 4.1.3, respectively.

Results

TABLE III: Pairwise inter-rater reliability between each of the 3 neuroradiologist raters' diagnostic confidence scores in reporting (i) presence/absence of acute stroke and (ii) Fazekas scores for chronic stroke via κ_{UW} ± 95 % confidence interval (CI). NaN: not a number.

(i) Acute Stroke				(ii) Chronic Stroke			
	Rater 1	**Rater 2**	**Rater 3**		**Rater 1**	**Rater 2**	**Rater 3**
Rater 1	--	NaN ± NaN	0 ± 0	**Rater 1**	--	0.246 ± 0.137	0.252 ± 0.115
Rater 2	NaN ± NaN	--	0 ± 0	**Rater 2**	0.246 ± 0.137	--	0.055 ± 0.081
Rater 3	0 ± 0	0 ± 0	--	**Rater 3**	0.252 ± 0.115	0.055 ± 0.081	--

Pairwise κ_{UW} ± 95 % CI values in Table III (i) corresponding to the acute stroke diagnostic task indicate "equivalent to chance" agreement (κ = 0), as defined by [84], or indicate a computational error (κ = not a number (NaN)). All pairwise κ_{UW} ± 95 % CI values in Table III (ii) corresponding to the chronic stroke diagnostic task indicate "slight" (κ = 0.01–0.20) or "fair" (κ = 0.21– 0.40) agreement, as defined by [84].

TABLE IV: Pairwise inter-rater reliability between each of the 3 neuroradiologist raters' diagnostic confidence scores in reporting (i) presence/absence of acute stroke and (ii) Fazekas scores for chronic stroke via κ_{QW} ± 95 % confidence interval (CI). NaN: not a number.

(i) Acute Stroke				(ii) Chronic Stroke			
	Rater 1	**Rater 2**	**Rater 3**		**Rater 1**	**Rater 2**	**Rater 3**
Rater 1	--	NaN ± NaN	0 ± 0	**Rater 1**	--	0.508 ± 0.138	0.471 ± 0.145
Rater 2	NaN ± NaN	--	0 ± 0	**Rater 2**	0.508 ± 0.138	--	0.342 ± 0.129
Rater 3	0 ± 0	0 ± 0	--	**Rater 3**	0.471 ± 0.145	0.342 ± 0.129	--

Pairwise κ_{QW} ± 95 % CI values in Table IV (i) corresponding to the acute stroke diagnostic task yield the same results as pairwise κ_{UW} ± 95 % CI shown in Table III (i). All pairwise κ_{QW} ± 95 % CI values in Table IV (ii) corresponding to the chronic stroke diagnostic task indicate "fair" (κ = 0.21– 0.40) or "moderate" (κ = 0.41–0.60) agreement, as defined by [84].

Discussion

According to Hypothesis A testing, it was stated that if κ ≤ 0.20 then neuroradiologists were to be reported as individual raters, while if neuroradiologist

rater's were in fair agreement ($\kappa > 0.20$) [84] then their confidence scores corresponding to each R were to be pooled.

For the acute stroke diagnostic task, despite pairwise κ_{UW} and κ_{QW} values being ≤ 0.20 (or $\kappa = NaN$), the results shown in Fig. 20 clearly depicted high raw inter-rater agreement between each of the three neuroradiologist raters. In fact, although not technically a degenerate distribution, the results depict such high raw-inter rater agreement that the data approaches a degenerate distribution (e.g. dataset with only a single value, such as a Likert score of 5). Therefore, doubt was cast on all κ values corresponding to the acute stroke diagnostic task and their relevancy as appropriate measurements of computed inter-rater reliability. The results $\kappa = 0$ and $\kappa = NaN$ are attributed to a phenomenon known as the *kappa paradox* (*Appendix B*). The *kappa paradox* speaks to the instability of κ as a measure of inter-rater reliability in certain circumstances. For example, κ falls apart in cases of extreme inter-rater agreement.

For the chronic stroke diagnostic task, despite the fact that all pairwise $\kappa_{UW} \pm 95\ \%$ CI are not > 0.20, but all pairwise $\kappa_{QW} \pm 95\ \%$ CI are > 0.20, it was deemed allowable to pool neuroradiologist raters' diagnostic confidence scores at each R factor. The fact that all pairwise κ_{QW} values were greater than all corresponding κ_{UW} values is explained by the fact that when neuroradiologist raters were in disagreement, the majority of the time they were only in disagreement by +/- 1 Likert score, and κ_{QW} accounts for the degree of disagreement. Whereas κ_{UW} is binary in measuring "agreement" or "disagreement", κ_{QW} penalizes to a lesser extent smaller differences in disagreements, the latter of which bears more relevance from a clinical standpoint. Nonetheless, despite κ_{QW}, in theory,

seeming like an appropriate measurement of computed inter-rater reliability, the *kappa paradox* is still at play, deeming κ_{QW} an inappropriate measurement of inter-rater reliability. The *kappa paradox* (*Appendix B*) is remediated by calculating Gwet's AC1 and AC2 values.

4.2.2. Gwet's Agreement Coefficient Inter-Rater Reliability Measurements

Gwet's Agreement Coefficient (AC) is an inter reliability measurement uncommonly reported in MRI literature; however, since κ has inherent paradoxes which were at play, as noted above in Section 4.2.1., Gwet's AC values were computed. This section (*Section 4.2.2.*) computes both Gwet's unweighted AC (AC1) and Gwet's quadratic weighted AC (AC2) to measure inter-rater reliability across all three neuroradiologist raters' diagnostic confidence scores in performing the acute and chronic stroke diagnostic tasks, using the raw diagnostic confidence score data from Sections 4.1.2. and 4.1.3., respectively.

Results

TABLE V: Pairwise inter-rater reliability between each of the 3 neuroradiologist raters' diagnostic confidence scores in reporting (i) presence/absence of acute stroke and (ii) Fazekas scores for chronic stroke via Gwet's AC1 ± 95 % confidence interval (CI). AC1: unweighted Agreement Coefficient.

(i) Acute Stroke				**(ii) Chronic Stroke**			
	Rater 1	**Rater 2**	**Rater 3**		**Rater 1**	**Rater 2**	**Rater 3**
Rater 1	--	1.00 ± 0.00	0.927 ± 0.047	**Rater 1**	--	0.555 ± 0.103	0.394 ± 0.107
Rater 2	1.00 ± 0.00	--	0.927 ± 0.047	**Rater 2**	0.555 ± 0.103	--	0.233 ± 0.104
Rater 3	0.927 ± 0.047	0.927 ± 0.047	--	**Rater 3**	0.394 ± 0.107	0.233 ± 0.104	--

All pairwise Gwet's AC1 ± 95 % CI values in Table V (i) corresponding to the acute stroke diagnostic task indicate "almost perfect" or "perfect" agreement. All pairwise Gwet's AC1 ± 95 % CI values in Table V (ii) corresponding to the chronic stroke diagnostic task indicate "fair" or "moderate" agreement.

TABLE VI: Pairwise inter-rater reliability between each of the 3 neuroradiologist raters' diagnostic confidence scores in reporting (i) presence/absence of acute stroke and (ii) Fazekas scores for chronic stroke via Gwet's AC2 ± 95 % confidence interval (CI). AC2: quadratic weighted Agreement Coefficient.

(i) Acute Stroke				(ii) Chronic Stroke			
	Rater 1	**Rater 2**	**Rater 3**		**Rater 1**	**Rater 2**	**Rater 3**
Rater 1	--	1.00 ± 0.00	0.994 ± 0.006	**Rater 1**	--	0.927 ± 0.031	0.853 ± 0.050
Rater 2	1.00 ± 0.00	--	0.994 ± 0.006	**Rater 2**	0.927 ± 0.031	--	0.839 ± 0.048
Rater 3	0.994 ± 0.006	0.994 ± 0.006	--	**Rater 3**	0.853 ± 0.050	0.839 ± 0.048	--

All pairwise Gwet's AC2 ± 95 % CI values in Table VI (i) for the acute stroke diagnostic task indicate "almost perfect" or "perfect" agreement. All pairwise Gwet's AC2 ± 95 % CI values in Table VI (ii) for the chronic stroke diagnostic task indicate "almost perfect" agreement.

Discussion

Considering the high raw inter-rater reliability between all three neuroradiologist raters for the acute stroke diagnostic task, there is, of course, no change in "perfect" Gwet's AC1 values between raters when calculating Gwet's AC2 for the same raters (i.e. both are 1.00 ± 0.00). For "almost perfect" Gwet's AC1 values between raters, however, there is a change when calculating Gwet's AC2 for the same raters, but this change is such that Gwet's AC2 is only minimally higher. The slight increase from Gwet's AC1 to Gwet's AC2 is due to the fact that Gwet's AC1 does not account for the degree of disagreement, penalizing all degrees of disagreement the same. For example, Likert scores of 1 versus 5 are penalized the same as Likert scores of 4 versus 5. Since Gwet's AC2 accounts for the degree of disagreement, disagreements of a smaller degree are penalized to a lesser extent. For example, Likert scores of 1 versus 5 would be penalized more heavily than Likert scores of 4 versus 5.

For the chronic stroke diagnostic task, considering the moderate raw inter-rater reliability between all three neuroradiologist raters, Gwet's AC1 values were reflective of this. But, similar to the acute stroke diagnostic task, Gwet's AC2 values ("almost perfect") are higher than Gwet's AC1 values ("fair" or "moderate") for the same reasoning as explained above.

Overall, all Gwet's AC1 and AC2 values were more reflective of the data than κ_{UW} and κ_{QW}, respectively, for both the acute and chronic stroke diagnostic tasks, which is explained by the fact that Gwet's AC values address the inherent paradoxes of κ. For both the acute and chronic stroke diagnostic tasks, the increase in values from Gwet's AC1 to AC2 arises from the fact that the majority of disagreements were the difference of only +/- 1 Likert score, and therefore the penalization of disagreements for Gwet's AC2 computations was the lowest possible. The extent of increase from Gwet's AC1 values to Gwet's AC2 values was minimal, however, for the acute stroke diagnostic task since the raw inter-rater reliability was already high, whereas it was more substantial for the chronic stroke diagnostic task since the raw inter-rater reliability was only moderate, but, again, the majority of disagreements were the difference of only +/- 1 Likert score. Considering Gwet's AC2 weights disagreements proportionately, while Gwet's AC1 does not, Gwet's AC2 is arguably more useful than Gwet's AC1, given neuroradiologist raters' diagnostic confidence is ranked on a 5-point Likert scale.

To conclude, as outlined in Section 3.4, since neuroradiologists were in fair agreement (inter-rater reliability > 0.20 based on the arguably more appropriate measurement Gwet's AC2 rather than κ_{QW}), it was deemed allowable to pool: (1) all 378

(126 image datasets x 3 neuroradiologists) diagnostic confidence scores corresponding to the acute stroke diagnostic task, and (2) all 378 (126 image datasets x 3 neuroradiologists) diagnostic confidence scores corresponding to the chronic stroke diagnostic task, yielding 54 (378/7 R values) diagnostic confidence scores at each R value per task. The allowability to pool the data is relevant to completing the testing of Hypothesis A.

4.2.3. Acute Stroke Diagnostic Task Boxplots

The results presented in this section (*Section 4.2.3.*) help visualize the distribution of the pooled data prior to completing the testing of Hypothesis A via post-hoc Bonferroni adjusted Wilcoxon signed-rank (exact) test results for pairwise comparisons of neuroradiologist raters' pooled Likert scores between R = 1X and each R > 1X. Boxplots of neuroradiologist raters' pooled Likert scores are shown, reporting the Q1, median (i.e. Q2), Q3, IQR, minimum and maximum values of the range, and quantile skewness at each R value, corresponding to the acute stroke diagnostic task.

Results

TABLE VII: First quartile (Q1), second quartile (Q2, median), third quartile (Q3), interquartile range (IQR), range (minimum and maximum), and quantile skewness corresponding to neuroradiologist raters' pooled Likert scores at each acceleration factor (R) for the acute stroke diagnostic task. Quantile skewness: [(Q3 – Q2) – (Q2 – Q1)] / (Q3 – Q1).

R	Q1	Median (Q2)	Q3	IQR	Range		Quantile Skewness
					Min	Max	
1	5	5	5	0	4	5	NA
2	5	5	5	0	5	5	NA
3	5	5	5	0	4	5	NA
4	5	5	5	0	4	5	NA
5	5	5	5	0	4	5	NA
6	5	5	5	0	4	5	NA
7	5	5	5	0	3	5	NA

Q1, median (Q2), Q3, IQR, and range values corresponding to neuroradiologist raters' pooled Likert scores at each R for the acute stroke diagnostic task are shown in Table VII. The kurtosis value of the Likert scores pooled across R = 1-7X is 61.4.

The boxplots of neuroradiologist raters' pooled Likert scores at each R corresponding to the acute stroke diagnostic task are shown as follows in Fig. 25:

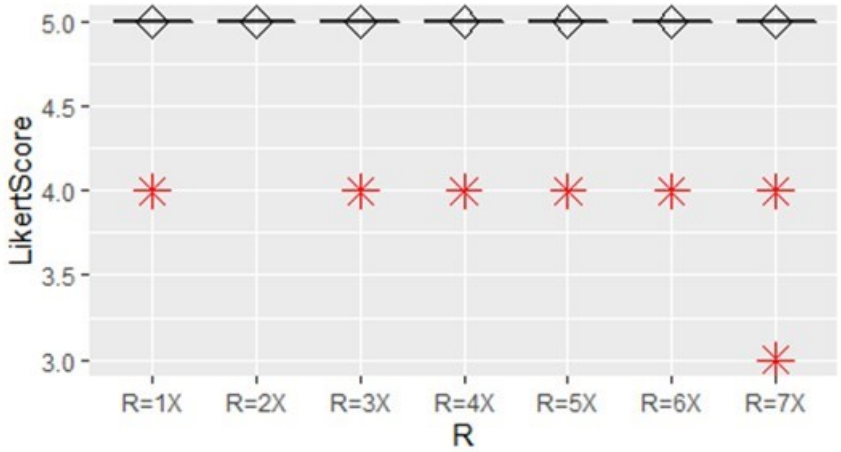


Fig. 25. Neuroradiologist raters' pooled diagnostic confidence scores in reporting presence/absence of acute stroke, plotted versus R, and represented as boxplots. Diamond represents median. Asterisks are diagnostic confidence score outliers.

Discussion

Note that Q1 – Q3, IQR, range, and quantile skewness are reported, rather than mean ± standard deviation (SD), because the spread of neuroradiologist raters' Likert scores pooled across R = 1-7X corresponding to the acute stroke diagnostic task deviates from normality, as defined by the computed kurtosis value of 61.4. In fact, there is technically no spread in the data, whereby Q1 = Q2 = Q3 = 5 and IQR = 0 for all R values from 1-7X, as shown by the values in Table VII. Further, Fig. 25 illustrates the values in Table VII, such that the median Likert score (Q2) and IQR remain at 5 and 0, respectively, as R increases from 1-7X. Considering the median remains at a Likert score of 5 and the

boxplots (Fig. 25) have no associated *box* (since IQR = 0 and Q1 = Q2 = Q3 = 5), it can be concluded that there is no skewness of neuroradiologist raters' Likert scores associated with the acute stroke diagnostic task. Therefore, although the data is neither a Gaussian distribution nor a degenerate distribution, but a non-Gaussian distribution, it could be suggested that the distribution of the data does approach degeneracy (e.g. *almost all* Likert scores are 5).

However, despite a lack of skewness, there are associated outliers. To explain, the variability, albeit minimal, is reflected by the range (Table VII). Although the maximum Likert score remains at 5 across the given range, the minimum value generally decreases as R increases, but from a maximum of 5 to only a minimum of 3. However, despite the slight variability in the minimum values of the range of Likert scores across R = 1-7X, the minimum values are actually considered outlier Likert scores, as illustrated in Fig. 25.

It is logical that all outliers are the minimum values of the range of Likert scores, rather than existing within boxplot quartiles, given the explanation that comes later, which discusses this result in relation to: (1) the nature of the images provided to neuroradiologist raters, (2) how these images are in fact used to perform the acute stroke diagnostic task, and (3) how (1) and (2) relate to the resultant Likert scores. However, it is worth noting, and elaborating on, the fact that there is a Likert score outlier at R = 1X, but no Likert score outlier at R = 2X. An outlier in the latter case would be expected rather than in the former case, considering the images at R = 1X were the fully-sampled, non-accelerated images, and R = 2X images were undersampled images. However, it is suspected that an outlier exists at R = 1X, but not at R = 2X, because the

CS reconstruction is inherently denoising. To explain further, the R = 1X images were not those output from Synaptive Medical's MR image data reconstruction pipeline because we did not have initial access to this pipeline. As such, the data reconstruction pipeline in this thesis was developed in such a way as to best replicate the images output from Synaptive Medical's reconstruction pipeline. As a result, the images output from the reconstruction pipeline developed in this thesis were unfiltered and not denoised (which, in contrast, they are in Synaptive Medical's reconstruction pipeline). Due to this fact, the CS reconstruction may have resulted in slightly better image quality in the R = 2X images, as compared to R = 1X images. However, this is merely speculation in order to rationalize the existence of an unexpectedly present versus expected-but-not-present outlier at R = 1X and R = 2X, respectively. Nonetheless, nothing of significance can be determined in terms of differences in overall image quality between R = 1X and R = 2X because the statistically meaningful median (Q2) and IQR values remain the same, at 5 and 0, respectively, for both R = 1X and R = 2X.

To conclude, since only 7/378 (less than 2 %) total pooled Likert scores across R = 1-7X were considered outlier Likert scores, and given that the median and IQR remained at 5 and 0, respectively, it can be concluded that neuroradiologists' diagnostic confidence in performing the acute stroke diagnostic task remained high across the given range of R from 1-7X. Specifically, it can be concluded that there was no change in neuroradiologists' diagnostic confidence in performing the acute stroke diagnostic task as images were undersampled.

It is important to note the limitation that the trend of Q1 = Q2 = Q3 = 5 and IQR = 0 across R = 1-7X should not be extrapolated to fit data beyond the given range of R = 1-7X. The reported data (Table VII) and boxplots (Fig. 25) only help to support Hypothesis A (*Section 1.6.2.*) by providing additional rigour to the results of the Wilcoxon signed-rank test. The boxplots, specifically, aid in the visualization of the statistical distribution of neuroradiologist raters' diagnostic confidence in performing the acute stroke diagnostic task.

4.2.4. Chronic Stroke Diagnostic Task Boxplots

The results presented in this section (*Section 4.2.4.*) help visualize the distribution of the pooled data prior to completing the testing of Hypothesis A via post-hoc Bonferroni adjusted Wilcoxon signed-rank (exact) test results for pairwise comparisons of neuroradiologist raters' pooled Likert scores between R = 1X and each R > 1X. Boxplots of neuroradiologist raters' pooled Likert scores are shown, and Q1, Q2 (median), Q3, IQR, range (minimum and maximum), and quantile skewness at each R value, corresponding to the chronic stroke diagnostic task, are reported.

Results

TABLE VIII: First quartile (Q1), second quartile (Q2, median), third quartile (Q3), interquartile range (IQR), range (minimum and maximum), and quantile skewness corresponding to neuroradiologist raters' pooled Likert scores at each acceleration factor (R) for the chronic stroke diagnostic task. Quantile skewness: [(Q3 – Q2) – (Q2 – Q1)] / (Q3 – Q1).

R	Q1	Median (Q2)	Q3	IQR	Range		Quantile Skewness
					Min	Max	
1	5	5	5	0	3	5	NA
2	5	5	5	0	3	5	NA
3	4	5	5	1	3	5	-1
4	4	4.5	5	1	3	5	0
5	4	4	5	1	2	5	1
6	4	4	5	1	2	5	1
7	3	3	4	1	1	5	1

Q1, median (Q2), Q3, IQR, range, and quantile skewness values corresponding to neuroradiologist raters' pooled Likert scores at each R for the chronic stroke diagnostic task are shown in Table VIII. The kurtosis value of the Likert scores pooled across R = 1-7X is 4.5.

The boxplots of neuroradiologist raters' pooled Likert scores at each R corresponding to the chronic stroke diagnostic task are shown as follows in Fig. 26:

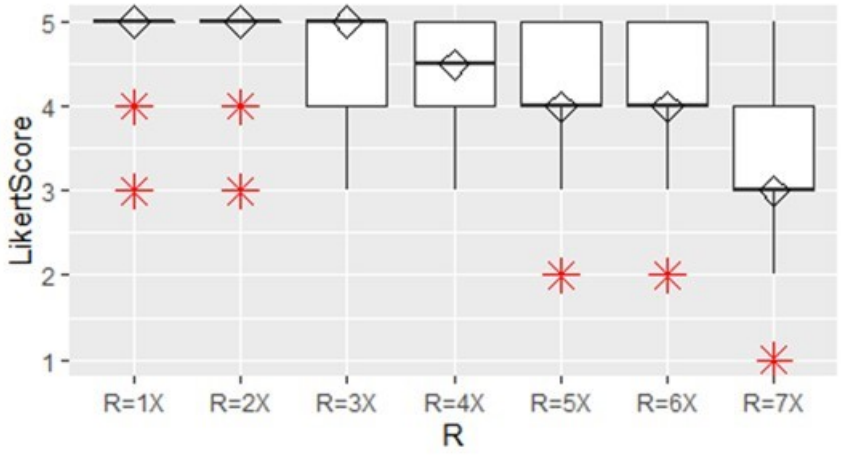


Fig. 26. Neuroradiologist raters' pooled diagnostic confidence scores in reporting Fazekas scores (i.e. reporting identification of chronic ischemic lesion burden), plotted versus R, and represented as boxplots. Diamond represents median. Asterisks are diagnostic confidence score outliers.

Discussion

Note that Q1 – Q3, IQR, range, and quantile skewness are reported, rather than mean ± SD, because the spread of neuroradiologist raters' Likert scores pooled across R = 1-7X corresponding to the chronic stroke diagnostic task deviates from normality, as defined by the computed kurtosis value of 4.5. In fact, for R values from 1-2X, there is technically no spread in the data, whereby Q1 = Q2 = Q3 = 5 and IQR = 0, as shown by the values in Table VIII. Further, Fig. 26 illustrates the values in Table VIII, such that the

median Likert score (Q2) and IQR remain at 5 and 0, respectively, as R increases from 1-2X. Considering the median remains at a Likert score of 5 and the boxplots (Fig. 26) have no associated *box* (since IQR = 0 and Q1 = Q2 = Q3 = 5), it can be concluded that there is no skewness of neuroradiologist raters' Likert scores at R = 1X and 2X associated with the chronic stroke diagnostic task.

However, despite a lack of skewness across R = 1-2X, there are associated outliers. Although the maximum Likert score remains at 5 across R = 1-2X, the minimum value remains at 3 and is considered an outlier Likert score, as illustrated by Fig. 26. Specifically, the boxplots at R = 1X and 2X (Fig. 26) show outliers existing at Likert scores of 3 and 4.

As R values increase from R = 3-7X, however, not only does the median decrease, but the skewness of the data begins to change, favouring Likert scores of a lower value. For example, the median Likert score decreases from 5 (R = 3X) to 4.5 (R = 4X) to 4 (R= 5-6X) to 3 (R = 7X), while – although IQR remains 1 (R = 3-7X) – the quantile skewness changes from -1 (R = 3X) to 0 (R = 4X) to 1 (R = 5-7X), indicating that the distribution of neuroradiologist raters' diagnostic confidence changes from having a negative skewness to being approximately symmetric to having a positive skewness.

Further, in addition to skewness across R = 3-7X, there is also variability and associated outliers. The variability, albeit minimal, is reflected by the range (Table VIII); although the maximum Likert score remains at 5 across R = 3-7X, the minimum Likert score decreases, as R increases, from 3 (R = 3-4X) to 2 (R = 5-6X) to 1 (R = 7X). Specifically, the minimum values 2 (R = 5-6X) and 1 (R = 7X) are considered outlier Likert

scores, while minimum non-outlier Likert scores (Q1 values) decrease from 3 (R = 3-6X) to 2 (R = 7X). The non-outlier decrease is logical, and the reasoning for this is effectively the same as that which will be discussed below. These results (Fig. 26, Table VIII), can be better visualized by breaking down the diagnostic confidence scores corresponding to performing the chronic stroke diagnostic task.

Results

The pie charts in Fig. 27 illustrate the breakdown of the proportion of Likert scores and outlier Likert scores shown as percentages of the total number of image datasets at each R. Since the Likert scores are discrete values, some information on the proportion of scores that exist within the quartile groups and/or below the upper and lower quartiles may be lost upon inspection of the boxplots. The pie charts extract this information and thus complement the boxplot interpretation to understand the distribution of the non-Gaussian data. The vertical and horizontal axes overlaying each pie chart are a visual representation analogous to the quartile groups in a box plot.

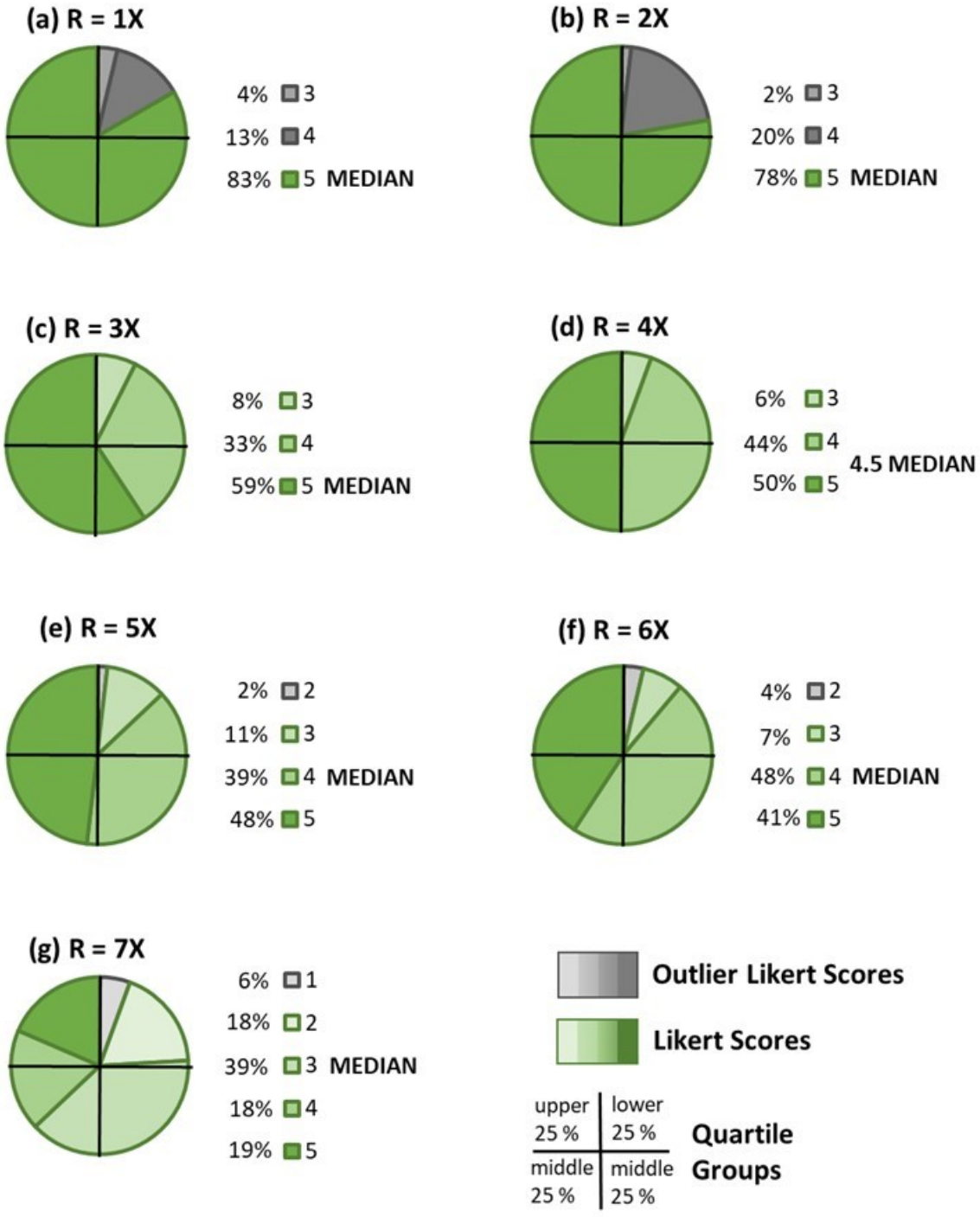


Fig. 27. Proportion of neuroradiologist raters' pooled diagnostic confidence scores associated with the identification of chronic ischemic lesion burden that are outliers at acceleration factors, R = 1-7X (a-g). Median Likert score is 5 at R = 1-3X (a-c), 4.5 at R = 4X (d), 4 at R = 5-6X (e-f), and 3 at R = 7X (g).

Discussion

At R = 1-2X, it is logical that there are fewer Likert score outliers at 3 compared to at 4, while it is also logical that the overall percentage of outlier Likert scores at R = 2X is higher compared to at R = 1X, considering R = 1X images are fully-sampled and R = 2X images are undersampled, but to the least extent compared to those at R = 3-7X. It is observed that out of all R values, R = 1-2X have the largest percentages of Likert scores at 5, representing the fact that neuroradiologists have higher diagnostic confidence performing the chronic stroke diagnostic task at R = 1-2X.

At R = 3-4X, there were no Likert score outliers, but the skewness of Likert scores does change. The percentages of scores at 5 decrease and at 4 increase from R = 3X to R = 4X, somewhat levelling out at R = 4X. The median at 4.5 suggests that 50% of the data is above Q2 at scores of 5, and 50% of the data is below Q2 at scores of 3 and 4. Fig. 27 (d) clearly depicts this, noting the left half of the pie chart only represents scores of 5, while the right half represents scores of 3 and 4, the majority of which are scores of 4.

Lastly, while the quantile skewness did not change, the percentage of Likert score outliers not only increased from R = 5X to R = 7X, but the value of the outlier score also dropped from 2 to 1 between R = 5-6X and R = 7X.

Overall, the decreasing median across R = 1-7X, in conjunction with increasing quantile skewness towards lower Likert score values indicate that neuroradiologists' diagnostic confidence in performing the chronic stroke diagnostic task decreases across the given range of R values. Specifically, it can be concluded that there was no change in neuroradiologists' diagnostic confidence in performing the chronic stroke diagnostic

task as images were undersampled from R = 1X to 2X, but that there was a change in neuroradiologists' diagnostic confidence in performing the chronic stroke diagnostic task as images were undersampled from R = 3-7X.

It is important to note the limitation that trends should not be extrapolated to fit data beyond the given range of R = 1-7X. The reported data (Table VIII) and boxplots (Fig. 26) only help to support Hypothesis A (*Section 1.6.2.*) by providing additional rigour to the results of the Wilcoxon signed-rank test, aiding in the visualization of not only the statistical distribution of neuroradiologist raters' diagnostic confidence, but also in the statistical distribution of associated skewness to their diagnostic confidence in performing the chronic stroke diagnostic task.

4.2.5. Wilcoxon Signed-Rank Tests

The results presented in this section (*Section 4.2.5.*) allow the completion of the testing of Hypothesis A via post-hoc Bonferroni adjusted Wilcoxon signed-rank (exact) p-values. To complete the testing of Hypothesis A, Table X lists the Wilcoxon signed-rank (exact) test results for pairwise comparisons of neuroradiologist raters' pooled Likert scores between R = 1X and each R > 1X.

Results

TABLE IX: Wilcoxon signed-rank (exact) test results (alternative hypothesis: greater) for pairwise comparisons of neuroradiologist raters' pooled diagnostic confidence scores corresponding to the (i) acute and (ii) chronic stroke diagnostic tasks, between (x) R = 1X and each (y) R > 1X. Clinically significant p-values are listed in bold text.

Pairwise Comparisons		Exact P-Values	
(x)	(y)	(i)	(ii)
R = 1X	R = 2X	0.8413	0.3328
R = 1X	R = 3X	0.5000	**3.449×10^{-3}**
R = 1X	R = 4X	Not Applicable	**1.319×10^{-4}**
R = 1X	R = 5X	0.2819	**3.814×10^{-5}**
R = 1X	R = 6X	0.1587	**3.940×10^{-6}**
R = 1X	R = 7X	0.1587	**3.845×10^{-9}**

TABLE X: Post-hoc Wilcoxon signed-rank (exact) test results (alternative hypothesis: greater) for pairwise comparisons of neuroradiologist raters' pooled diagnostic confidence scores corresponding to the (i) acute and (ii) chronic stroke diagnostic tasks, between (x) R = 1X and each (y) R > 1X. Clinically significant p-values are listed in bold text.

Pairwise Comparisons		Bonferroni Adjusted Exact P-Values	
(x)	(y)	(i)	(ii)
R = 1X	R = 2X	1.000	1.000
R = 1X	R = 3X	1.000	**2.069×10^{-2}**
R = 1X	R = 4X	Not Applicable	**7.914×10^{-4}**
R = 1X	R = 5X	1.000	**2.286×10^{-4}**
R = 1X	R = 6X	0.9522	**2.364×10^{-5}**
R = 1X	R = 7X	0.9522	**2.310×10^{-8}**

Wilcoxon, and Bonferroni adjusted Wilcoxon, signed-rank (exact) p-values are listed in Tables IX and X, respectively, for the greater alternative hypothesis test corresponding to the acute and chronic stroke diagnostic tasks. Clinically significant p-values ($p < 0.05$) are listed in bold text.

TABLE XI: Results of Wilcoxon signed-rank (exact) test (alternative hypothesis: greater) based on Bonferroni adjusted p-values from Table X, corresponding to the (i) acute and (ii) chronic stroke diagnostic tasks.

(y)	(x) R = 1X	
	(i)	(ii)
R = 2X	= 0	= 0
R = 3X	= 0	> 0
R = 4X	= 0	> 0
R = 5X	= 0	> 0
R = 6X	= 0	> 0
R = 7X	= 0	> 0

Based on the Bonferroni adjusted p-values in Table X (i-ii), the results of the Wilcoxon signed-rank (exact) test for the acute and chronic stroke diagnostic tasks are shown in Table XI (i-ii), respectively.

Discussion

Due to testing multiple comparisons, the Wilcoxon signed-rank (exact) p-values in Table IX (i-ii) were adjusted, post-hoc, via the Bonferroni adjustment, as shown in Table X (i-ii). Note that the p-value in Tables IX (i) and X (i) for the pairwise comparison between R = 1X and 4X is listed as *Not Applicable* given that the pooled Likert scores for the acute stroke diagnostic task at R = 1X and 4X are identical.

The Wilcoxon signed-rank test tests whether the median of the difference between x and y is symmetric about zero (= 0), or statistically significantly greater than zero (> 0) or less than zero (< 0). For Hypothesis A testing, x is defined as neuroradiologist raters' pooled diagnostic confidence scores corresponding to the acute and chronic stroke diagnostic tasks at R = 1X from Figs. 20 and 23, respectively, and y is defined as neuroradiologist raters' pooled diagnostic confidence scores corresponding to the acute

and chronic stroke diagnostic tasks at each R > 1X (R = 2-7X) from Figs. 20 and 23, respectively. The null hypothesis is such that the distribution of the difference between x - y is symmetric about 0, while the alternate hypotheses can be such that either (a) the distribution of x - y is > 0 (i.e. true location shift is statistically significantly *greater* than zero) or (b) the distribution of x - y < 0 (i.e. true location shift is statistically significantly *less* than zero).

For all pairwise comparisons between R = 1X and R > 1X for the acute stroke diagnostic task in Table X (i), and between R = 1X and R = 2X for the chronic stroke diagnostic task in Table X (ii), the null hypothesis was accepted at the significance level of $\alpha = 0.05$. Therefore, for the acute stroke diagnostic task, there was no statistical significance for any pairwise comparisons ($p > 0.05$). However, for the chronic stroke diagnostic task, statistical significance existed for pairwise comparisons between R = 1X and each R = 3-7X ($p < 0.05$), as shown in Table X (ii). As per Table XI (i-ii), '= 0' indicates that the medians of the difference between pooled Likert scores at R = 1X and each R > 1X for the acute stroke diagnostic task, and at R = 1X and R = 2X for the chronic stroke diagnostic task, were symmetric about zero, indicating that neuroradiologists' diagnostic confidence did not change across the respective acceleration factors for each diagnostic task. However, '> 0' in Table XI (i-ii) indicates that the medians of the difference between the pooled Likert scores at R = 1X and each R = 3-7X for the chronic stroke diagnostic task were statistically significantly greater than zero (i.e. true location shift is *greater* than zero), indicating that neuroradiologists' diagnostic confidence scores were greater at R = 1X than they were at each R = 3-7X for the chronic stroke diagnostic task.

The conclusion that neuroradiologists' diagnostic confidence did not change across R = 1-7X and R = 1-2X for the acute and chronic stroke diagnostic tasks, respectively, is therefore determined by the lack of statistical significance from the Wilcoxon signed-rank test results. This conclusion is supported by the boxplots in Figs. 25 and 26 across R = 1-7X and R = 1-2X, respectively, which illustrated both a lack of skewness in diagnostic confidence and a median that remained at the maximum Likert score of 5 (which corresponds to 100 % diagnostic confidence). Further, the conclusion that neuroradiologists' diagnostic confidence was greater at R = 1X than at R = 3-7X for the chronic stroke diagnostic task, is determined by the statistical significance from the Wilcoxon signed-rank test results and is supported by the other results. The boxplots in Fig. 26 across R = 3-7X illustrate skewness in diagnostic confidence and median Likert scores that both trend towards lower Likert scores and thus lower associated diagnostic confidence, while the boxplot in Fig. 26 at R = 1X illustrates a median Likert score at the maximum value of 5 in conjunction with no skewness in diagnostic confidence.

To conclude, Hypothesis A (*Section 1.6.2.*) was proven, such that neuroradiologist raters' confidence scores corresponding to the task of identifying the presence/absence of acute stroke (i.e. acute stroke diagnostic task) was less sensitive to increasing R than neuroradiologist raters' confidence scores corresponding to the task of assigning a Fazekas score for chronic stroke (i.e. chronic stroke diagnostic task).

4.3. Retrospectively Accelerated Magnetic Resonance Images

Results

See Fig. 18 in Section 4.1. for an example of typical retrospectively accelerated T2 FLAIR images output from the data pipeline from a recruited AIS patient.

Discussion

As expected, it can be qualitatively observed (Fig. 18) that as R increases – or relative acquisition time decreases – image quality decreases in the form of resolution loss. Quantitatively, since Hypothesis A (*Section 1.6.2.*) was proven, and based on the post-hoc Bonferroni corrected Wilcoxon signed-rank (exact) test results, a decrease in relative acquisition time from 266 seconds (R = 1X) to 38 seconds (R = 7X) did not change neuroradiologists' diagnostic confidence in performing the acute stroke diagnostic task.

A reduced acquisition time to this extent, in conjunction with neuroradiologists' remaining confident in performing the acute stroke diagnostic task, yields promise for application to diagnosing AIS in emergency medicine. Furthermore, although chronic stroke is not as time-critical of a scenario as acute stroke, the fact that neuroradiologists remain confident in performing the chronic stroke diagnostic task for images up to R = 2X indicates that valuable information on *both* acute and chronic stroke may be gained from scans with relative acquisition time of 133 s (R = 2X).

From a clinical standpoint, one possible reason that the diagnostic confidence in performing the acute stroke diagnostic task might not have been as sensitive to increasing R as the chronic stroke diagnostic task could be the fact that acute stroke

presents with more focal, high contrast image features, while chronic stroke presents with more diffuse, low contrast image features, as shown in Fig. 18 (a) and (b), respectively. Specifically, it can be qualitatively observed in Fig. 18 (b) that diffuse, low contrast image features are more likely to be masked by CS artifacts.

An additional possible reason why the diagnostic confidence in performing the acute stroke diagnostic task was less sensitive to increasing R than the chronic stroke diagnostic task could be the limitations of this study. To explain, for the acute stroke diagnostic task, although neuroradiologists rely on the T2 FLAIR and DWI/ADC map images to diagnose acute stroke, the process of *diagnosing* an acute stroke is such that the DWI/ADC map images will be reviewed, first, for purposes of sensitivity, and the T2 FLAIR images will be reviewed, second, for purposes of specificity. As a result, the T2 FLAIR image is not the image predominant to bearing the diagnosis; but, it still bears important weight in performing the diagnostic task. The specific limitation is therefore such that only the T2 FLAIR images were accelerated, yet the diagnostic task required the use of *both* the DWI/ADC maps and T2 FLAIR images to perform the diagnostic task, with the DWI/ADC maps being the predominant images used in the diagnosis. As such, one would expect neuroradiologists' diagnostic confidence to remain high considering they are predominantly using the non-accelerated images to make the diagnosis.

To attempt to reduce the extent of this specific limitation, it was explicitly stated in the questionnaire used in the study with neuroradiologist raters' that the acute stroke diagnostic task must involve the use of both the DWI and T2 FLAIR images, as follows: "Using the DWI and T2 FLAIR images, would you report the presence of an acute

stroke?" Knowing that neuroradiologists use both DWI/ADC map and T2 FLAIR images to diagnose acute stroke in a clinical setting, the reason both image contrasts were explicitly stated in the questionnaire was to combat the fact that the wording of the questionnaire might only trigger to a neuroradiologist rater the aspect of *specificity* in achieving a diagnosis and, as a result, they may only use the DWI images (considering the acute stroke diagnostic task was to *report the presence/absence of acute stroke*, and *presence/absence* may imply *specificity* only).

Nonetheless, regardless of the facts that: (1) the predominant image used in the diagnose of acute stroke is the DWI/ADC map images, and these images were not accelerated in this study, and (2) the T2 FLAIR is secondarily involved in the diagnosis of acute stroke, and these images were accelerated in this study, and (3) the neuroradiologist raters' diagnostic confidence was related to acceleration factors, the important takeaway is the fact that the overall acquisition time of a stroke protocol can be reduced – and this is what matters most in the case of diagnosing acute stroke in emergency medicine.

For the chronic stroke diagnostic task, on the other hand, not only are neuroradiologist raters' diagnostic confidence scores likely more sensitive to R because chronic stroke is more diffuse and the CS artifacts may mask chronic lesions, the primary image used to identify chronic lesions is the T2 FLAIR image; the DWI/ADC map images are not used at all in this identification. Since T2 FLAIR images were accelerated, one would expect that diagnostic confidence scores related to the chronic stroke diagnostic task would decrease to a greater extent, as compared to the diagnostic confidence

scores related to the chronic stroke diagnostic task where the T2 FLAIR images were not predominantly used.

To further explain why diffuse lesions characteristic of chronic stroke may be masked and the focal lesions characteristic of acute stroke may be more likely to prevail at higher accelerations goes back to the initial explanation of pseudo-random undersampling k-space data. The critical information in an AIS diagnostic MR image, for example, would be features within edge boundaries, such as overall image contrast, and therefore that which may allow one to distinguish pathological versus non-pathological tissue. The k-space data that improves the visibility of these lesions is important to retain. CS reconstructions require pseudo-random undersampling, and the use of this undersampling and reconstruction technique may then allow a greater possibility for a neuroradiologist to make these necessary distinctions. Fully-sampling the centre of k-space data provides the necessary overall image contrast, while randomly undersampling the edges of k-space data yield a loss of resolution which may be afforded, in some cases. The necessary information required to distinguish a high contrast, focal lesion may be more likely to prevail if overall image contrast remains due to fully-sampling the centre of k-space data, while already-hard-to-resolve diffuse low contrast lesions resultant from chronic stroke may not afford the resolution loss with the random undersampling of the edges of k-space data. In addition, because the diffuse lesions tend to be lower contrast, the information may be lost to the inherent denoising of the CS reconstruction, whereas focal high contrast lesions may be more

likely to remain above the thresholding level. Although worth attempting to rationalize these differences, more research would need to be done to move beyond speculation.

Finally, although the acute and chronic stroke diagnostic tasks do not yield the same results (i.e. statistical significance for the same pairwise R values), the results for the acute stroke diagnostic task hold priority. Chronic stroke is not typically diagnosed at the POC, and if identifiable, is typically considered a fringe benefit of receiving an MRI stroke exam in an emergency medicine setting (which, again, does not always happen). Should meaningful clinical findings be observed about chronic ischemic lesion burden on acute stroke MRI scans, chronic ischemic lesion burden will be reported, but a follow-up MRI exam and neurologist referral will typically be ordered so that the chronic ischemic lesion burden and its associated risk factors and potential treatments can be properly assessed and managed without interfering with the higher priority task of diagnosing and treating acute stroke.

Additional limitations for this research work were the process for selecting the three board-certified neuroradiologists, considering it was based on availability with limited resources, and the fact that Rater 3 reported gaining familiarity with repeating images. For the latter, Rater 3 reported that this familiarity may have artificially boosted their diagnostic confidence scores; however, Rater 3 in fact yielded scores indicating that they may be more conservative in their clinical decision making process, compared to Rater 1 and Rater 2. Nonetheless, it is worth noting that recognition bias is a limitation and should be addressed in future studies.

4.4. Magnetic Resonance Imaging Stroke Protocol Acquisition Times

Results

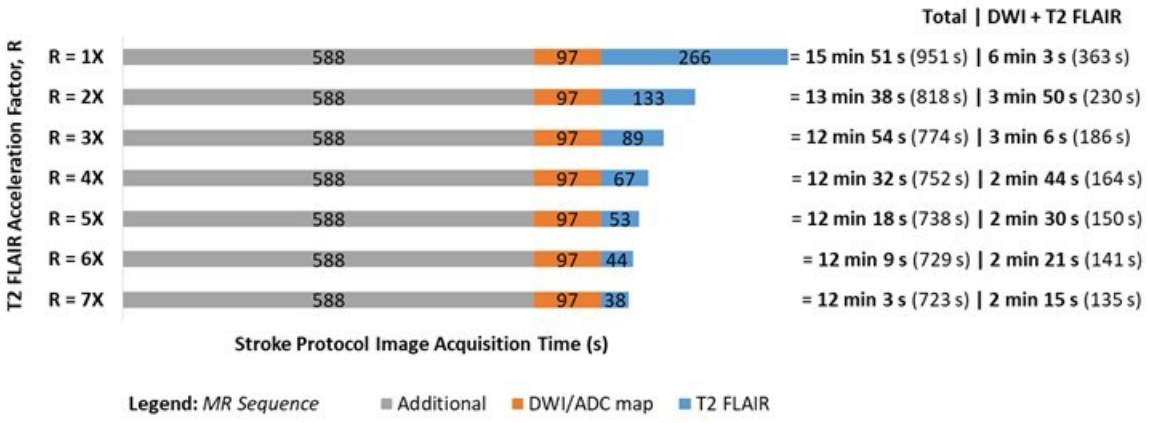


Fig. 28. Stroke protocol acquisition times corresponding to magnetic resonance imaging sequences: (1) axial DWI/ADC map (non-accelerated), (2) axial T2 FLAIR either non-accelerated (R = 1X) or accelerated (R = 2-7X), and (3) additional (non-accelerated). Total stroke protocol time | proportion of total stroke protocol time from DWI/ADC and T2 FLAIR sequence acquisition. Additional sequences: axial SWI, axial T2 FSE, and sagittal T1 FLAIR (did not include optional oblique TOF MRA). ADC: apparent diffusion coefficient; DWI: diffusion weighted imaging; FLAIR: fluid attenuated inversion recovery; FSE: fast spin-echo; R: undersampling (or acceleration) factor; SWI: susceptibility weighted imaging.

Fig. 28 illustrates the stroke protocol acquisition time, which is comprised of the non-accelerated prospective (1) axial DWI/ADC map sequence, (2) axial T2 FLAIR sequence, and (3) additional sequences used in the diagnosis of AIS at 97 s, 266 s, and 588 s respectively, yielding a total stroke protocol image acquisition time of 15 min 51 s. Specifically, the additional sequences being referred to Fig. 28 were axial SWI, axial T2 FSE, and sagittal T1 FLAIR but this research work focused on the DWI/ADC maps and T2 FLAIR images. Note that the stroke protocol includes an optional non-accelerated prospective oblique TOF MRA sequence, which is 7 min 38 (458 s); however, since it is optional, it was not included in Fig. 28 and will not be included in the discussions of

acquisition times, below. The acquisition times provided are computed by the scan computer based on the set sequence parameters in Table II (*Section 3.1.3.*), such as TR, and number of phase encoding lines, number of slices, etc., but do not take into account, or include, calibration scans, prescans, data transfer time, patient prep time, etc.

Discussion

Although only the T2 FLAIR sequence was accelerated in this research work, the impact of accelerating this sequence is such that the overall stroke protocol acquisition time is reduced. For example, as shown by Fig. 28, a retrospective acceleration factor of R = 7X implemented on the T2 FLAIR sequence brings the T2 FLAIR image acquisition down to 38 s, yielding a total stroke protocol image acquisition time of 12 min 3 s – this is a 3 min 48 s reduction in the overall stroke protocol acquisition time from the protocol that included the T2 FLAIR images at R = 1X (non-accelerated). Saving almost 4 minutes in acquisition time is clinically relevant because it brings the MRI stroke protocol closer to the CT stroke protocol in terms of the time the patient spends on the scanning table, and while CS reconstruction can end up lengthening the overall time-to-image-availability, there are ongoing investigations into methods, such as machine learning, to improve this aspect of the workflow.

Referring to the non-accelerated acquisition time (R = 1X), between the two main sequences whose images are primarily used in the diagnosis of AIS (DWI/ADC map and T2 FLAIR), it can be observed that the T2 FLAIR sequence constitutes a larger portion of the scan time than the DWI/ADC map sequence. Specifically, the non-accelerated T2

FLAIR sequence is 4 min 26 s, while the DWI/ADC map sequence is only 1 min 37 s, yielding a 6 min 3 sec acquisition time, in combination. Accelerating the T2 FLAIR sequence up to R = 7X renders this portion of the acquisition only 2 min 15 s, which is almost a 3-fold reduction in acquisition time.

4.5. Hypothesis B Testing

This section (*Section 4.5.*) examines Hypothesis B, which fell under Objective 4 (*Section 1.6.2.*). Objective 4 was to assess the relationship between IQM scores and neuroradiologists' confidence scores; and Hypothesis B (*Section 1.6.2.*) was as follows: the IQMs FSIM, NQM and VIF will perform better than RMSE and SSIM for both the acute and chronic stroke diagnostic tasks.

To test Hypothesis B, neuroradiologist raters' diagnostic confidence scores corresponding to the acute stroke diagnostic task for all T2 FLAIR undersampled images were plotted versus IQM scores. Neuroradiologist raters' diagnostic confidence scores corresponding to the chronic stroke diagnostic task for all T2 FLAIR undersampled images were also plotted versus IQM scores. Sum of squares residuals (SSR) values corresponding to the non-linear logistic regression model fit to objective IQM scores, with respect to neuroradiologist raters' subjective diagnostic confidence, were computed, whereby a smaller SSR represents a better model fit. Spearman rank order correlation coefficient (SROCC) values were computed on the correlation between subjective and objective data, whereby SROCC = +/- 1 indicates a perfect correlation, and SROCC = 0 indicates no correlation. Wilcoxon signed-rank (exact) tests were

performed to determine if FSIM, NQM, and VIF performed better than RMSE and SSIM for the acute and chronic stroke diagnostic tasks.

4.5.1. Non-Linear Regression Models

Results

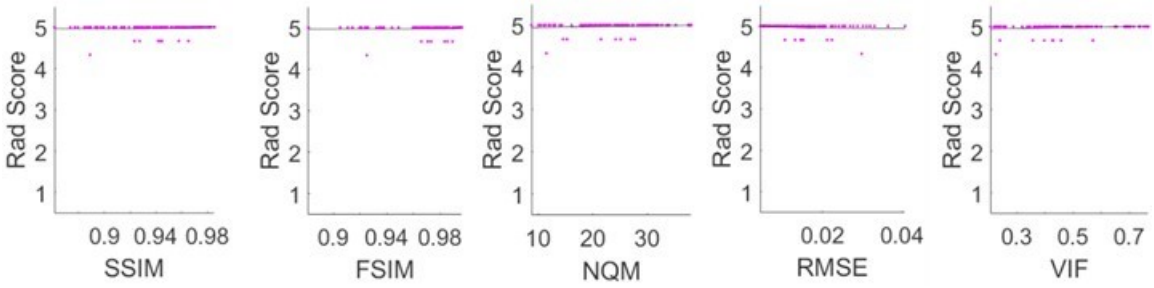


Fig. 29. Neuroradiologist raters' diagnostic confidence scores corresponding to the acute stroke diagnostic task versus image quality metric (IQM) scores computed for all undersampled T2 fluid attenuated inversion recovery (FLAIR) images. The plotted data were fit to a constrained logistic function for a non-linear regression model. T2: transverse relaxation time; SSIM: Structural SIMilarity; FSIM: Feature SIMilarity; NQM: noise quality measure; RMSE: root mean square error; VIF: visual information fidelity.

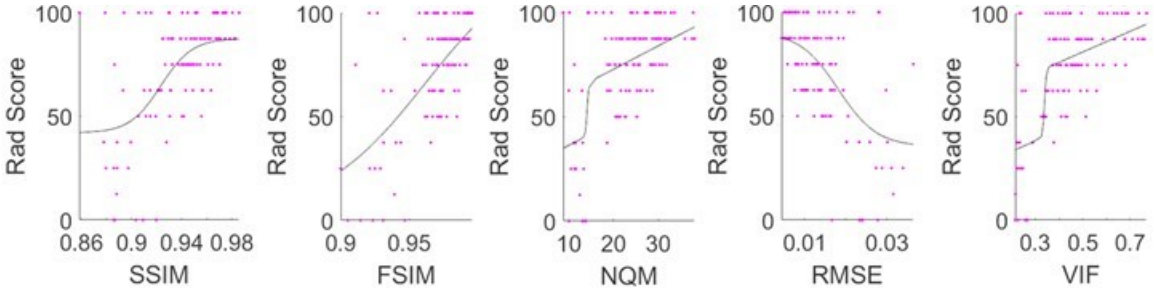


Fig. 30. Neuroradiologist raters' diagnostic confidence scores corresponding to the chronic stroke diagnostic task versus image quality metric (IQM) scores computed for all undersampled T2 fluid attenuated inversion recovery (FLAIR) images. The plotted data were fit to a constrained logistic function for a non-linear regression model. T2: transverse relaxation time; SSIM: Structural SIMilarity; FSIM: Feature SIMilarity; NQM: noise quality measure; RMSE: root mean square error; VIF: visual information fidelity.

Figs. 29 and 30 show neuroradiologist raters' diagnostic confidence scores corresponding to the acute and chronic stroke diagnostic tasks, respectively, plotted versus IQM scores computed for all T2 FLAIR undersampled images, and fit to a

constrained logistic function for a non-linear regression model. Note that a higher SSIM, FSIM, NQM, and VIF score indicates higher objective image quality, while an RMSE score of zero corresponds to the highest objective image quality.

Discussion

For Fig. 29, neuroradiologist raters' diagnostic confidence scores were evaluated as pooled and averaged raw scores (Likert scale 1-5) due to an IQR of zero (Table VII) corresponding to neuroradiologists' diagnostic confidence scores across the given range of acceleration factors, R = 2-7X, for the acute stroke diagnostic task. For Fig. 30, neuroradiologist raters' diagnostic confidence scores were also evaluated as pooled and averaged raw scores (Likert scores 1-5, however rescaled from 0-100) due to IQR values of zero or one (Table VIII) corresponding to neuroradiologists' diagnostic confidence scores across R = 2-7X for the chronic stroke diagnostic task. For both Figs. 29 and 30, the data were fit to a constrained logistic function for a non-linear regression model, where, since five IQMs were tested, there are five plots (per diagnostic task), each with a unique non-linear fit. Lastly, as previously mentioned in this thesis, the IQMs chosen were those identified to best correlate with radiologists' scores of overall diagnostic image quality (FSIM, NQM, VIF) [42], as well as IQMs most commonly used in MRI literature (RMSE, SSIM [85] [86]).

Note that there were a few reasons that the y-axis for Fig. 29 remains in the range of the Likert scale from 1-5 and the y-axis for Fig. 30 exists as Likert scores rescaled from 0-100. First, there are no statistically significant changes in neuroradiologist raters' confidence in performing the acute stroke diagnostic task, and therefore the logistic fit

only provides information that confirms this. When the y-axis of Fig. 29 is rescaled from 0-100, as shown in Fig. 31 (a), below, the result is misleading considering the zoomed in view (Fig. 31 b) of the logistic fit's attempt to fit the data, whereby it could be very easily mistaken that a trend exists when one, in fact, does not exist.

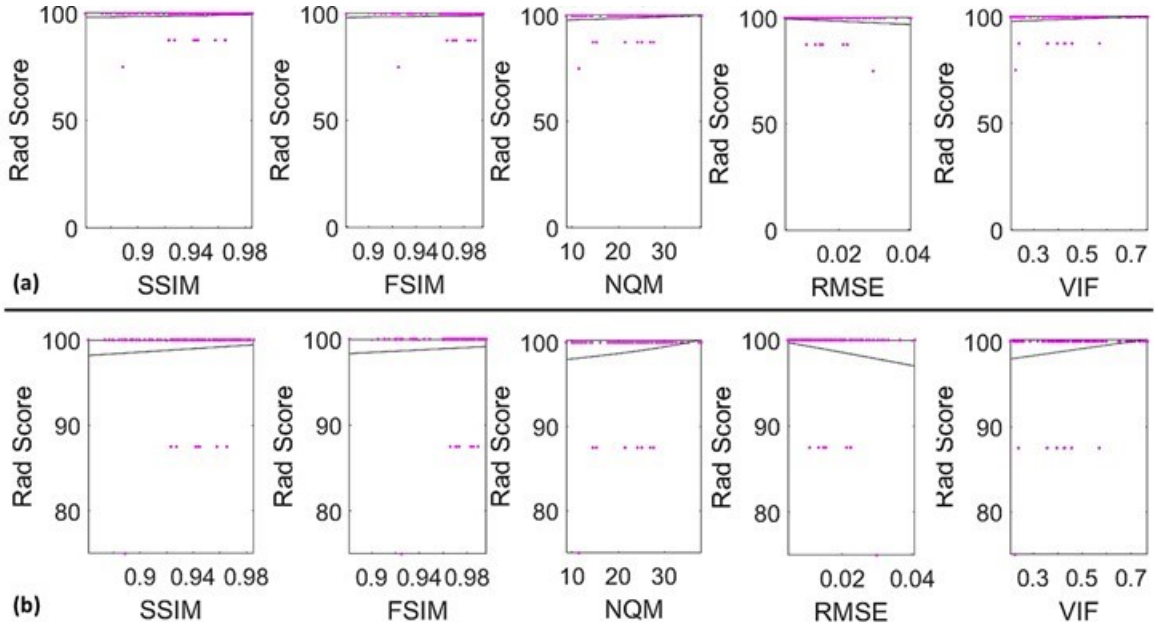


Fig. 31. Example of what Fig. 29 looks like with its (a) y-axis rescaled from 0-100 and (b) subsequently zoomed in.

Second, there are statistically significant changes in the neuroradiologist raters' confidence in performing the chronic stroke diagnostic task, and rescaling from 0-100 (Fig. 30) illustrates the logistic fit to the data more clearly. Third, the acute and chronic stroke diagnostic tasks, although being discussed with reference to one another, are not being directly compared in any statistically significant manner, so the fact that they are being plotted on axes that are scaled differently is irrelevant.

4.5.2. Sum of Squares Residuals, Spearman Rank Order Correlation Coefficient, and Kurtosis Values

Results

TABLE XII: (i) Sum of squares residuals (SSR) and (ii) kurtosis of the raw signed residuals, both corresponding to the non-linear logistic regression model fit to objective IQM scores, with respect to neuroradiologist raters' subjective diagnostic confidence scores associated with performing the (a) acute and (b) chronic stroke diagnostic tasks, with results reported for the (iii) Spearman rank order correlation coefficient (SROCC) on the correlation between subjective and objective data. A smaller SSR represents a better model fit, while SROCC = +/- 1 indicates perfect correlation, and SROCC = 0 indicates no correlation.

	(a) ACUTE					(b) CHRONIC				
	SSIM	**FSIM**	**NQM**	**RMSE**	**VIF**	**SSIM**	**FSIM**	**NQM**	**RMSE**	**VIF**
(i) SSR	1.13	1.13	1.12	1.10	1.11	4.78×10^4	4.12×10^4	4.28×10^4	4.74×10^4	3.58×10^4
(ii) Kurtosis	18.5	18.7	17.5	17.4	17.5	4.3	4.2	3.9	4.2	4.8
(iii) SROCC	0.05	0.01	0.10	-0.17	0.12	0.50	0.53	0.46	-0.48	0.55

Results for the corresponding SSR values; kurtosis values of the raw signed residuals; and SROCC values are shown in Table XII (a-b) (i-iii). To demonstrate the distribution of the raw signed residuals, Figs. 32 and 33 show the histograms associated with each IQM and diagnostic task.

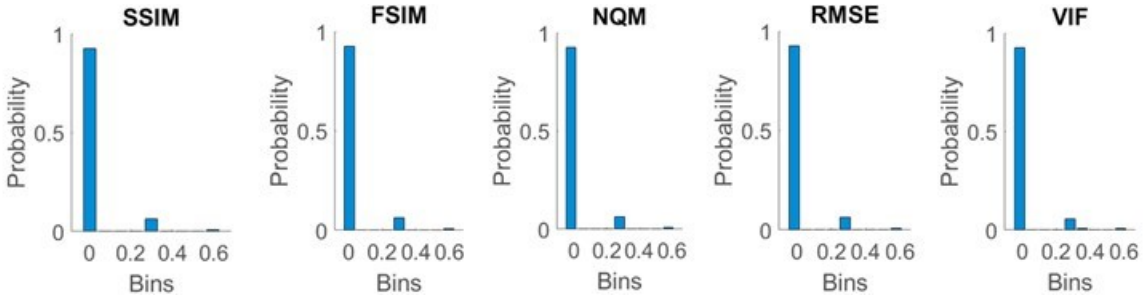


Fig. 32. Histogram plots of the raw signed residuals associated with the non-linear regression model fit to the objective IQM scores with respect to neuroradiologist raters' subjective diagnostic confidence scores in performing the acute stroke diagnostic task.

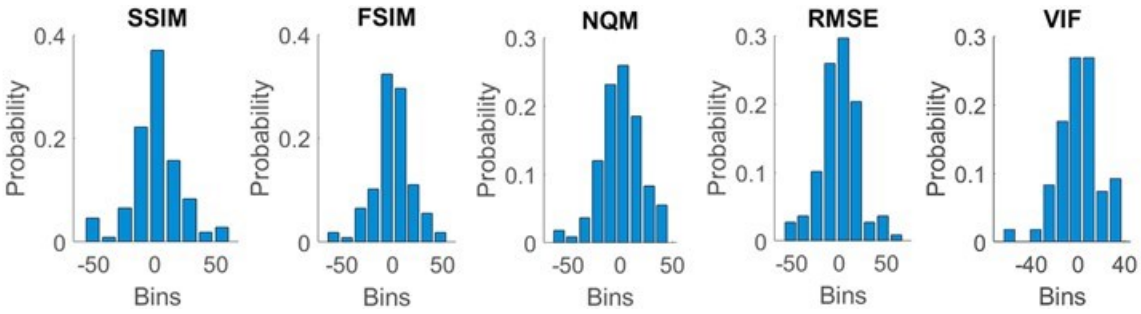


Fig. 33. Histogram plots of the raw signed residuals associated with the non-linear regression model fit to the objective IQM scores with respect to neuroradiologist raters' subjective diagnostic confidence scores in performing the chronic stroke diagnostic task.

Discussion

SSR represents the logistic model's goodness-of-fit to the plotted data. Based on the results in Table XII (a-b) (i), although there is no rank order in goodness-of-fit associated with the acute stroke diagnostic task, the rank order for the goodness-of-fit associated with the chronic stroke diagnostic task, from best to worst, is as follows: (1) VIF, (2) FSIM, (3) NQM, (4) RMSE, (5) SSIM.

SROCC, on the other hand, represents the correlation of the plotted data – in other words, the correlation between the neuroradiologist raters' diagnostic confidence scores in performing the diagnostic task and the IQM scores. Based on the results in Table XII (a-b) (iii), although there is no rank order in correlation associated with the acute stroke diagnostic task, the rank order in correlation associated with the chronic stroke diagnostic task, from best to worst, is as follows: (1) VIF, (2) FSIM, (3) SSIM, (4) RMSE, (5) NQM.

Although the logistic model attempts to fit the objective IQM scores for the acute stroke diagnostic task, the poor SROCC indicates that the objective and subjective scores are not correlated. A limitation to note with the non-linear regression model analysis of

the acute stroke diagnostic task is therefore the fact that the dataset is *approaching degeneracy* (*Sections 4.2.1 and 4.2.3*). Nonetheless, in the case of the acute stroke diagnostic task, the IQMs in question are all equally unrelated to diagnostic confidence due to highly uniform confidence scores across the given range of acceleration factors (R=2-7X). For the chronic stroke diagnostic task, however, not only do the IQMs provide predictability in diagnostic confidence via the logistic fit model (best to worst: VIF, FSIM, NQM, RMSE, SSIM), the plotted data is also correlated (best to worst: VIF, FSIM, SSIM, RMSE, NQM). Fig. 18 gives context to IQM performance, demonstrating that artifacts created by undersampling tend to mask diffuse chronic lesions, while often leaving focal acute lesions detectable. As such, neuroradiologist raters' diagnostic confidence scores were impacted for the chronic stroke diagnostic task, but were not impacted in the acute stroke task. Detectable acute lesions and resultant high diagnostic confidence, despite undersampling, demonstrate why IQMs previously shown to correlate with radiologists' scores of overall diagnostic image quality do not correlate with neuroradiologists' diagnostic confidence in the acute stroke diagnostic task. Less detectable chronic lesions and resultant lower diagnostic confidence scores due to undersampling demonstrate why IQMs previously shown to correlate with radiologists' scores of overall diagnostic image quality correlate with neuroradiologists' diagnostic confidence in the chronic stroke diagnostic task.

Finally, if the kurtosis value of the raw signed residuals was between 2.0 to 4.0 (inclusive) then the distribution of the residuals was deemed to be Gaussian. However, all kurtosis values for the acute stroke diagnostic task in Table XII (a) (ii), and 4 out of 5

kurtosis values for the chronic stroke diagnostic task in Table XII (b) (ii), were > 4.0, thereby indicating that 90 % of the distributions of the raw signed residuals were considered to be non-Gaussian distributions. Specifically, the raw signed residuals corresponding to the acute stroke diagnostic task plots (Fig. 32) are far from Gaussian, noting kurtosis values of >> 4.0, as shown in Table XII (a) (ii), while the raw signed residuals corresponding to the chronic stroke diagnostic task plots (Fig. 33) have almost-Gaussian (or Gaussian) distributions, noting the kurtosis values only slightly > 4.0, as shown in Table XII (b) (ii) for the 4 out of 5 non-Gaussian distributions.

The SSIM, FSIM, RMSE, and VIF histogram plots in Fig. 33 are close to a Gaussian distribution, as demonstrated by the associated kurtosis values in Table XII (b) (ii) being only slightly greater than 4.0, albeit non-Gaussian based on the aforementioned definition of Gaussianity. The NQM histogram plot in Fig. 33, however, is considered to be a Gaussian distribution since the associated kurtosis value is between 2.0 and 4.0 (inclusive) at 3.9. Alternatively, all of the histogram plots in Fig. 32 clearly do not represent a normal distribution, as justified by their kurtosis values in Table XII (a) (ii) being much greater than 4.0. Since 90 % of the plots were evidentiary of non-Gaussian distributions, a Wilcoxon signed-rank test was performed.

4.5.3. Wilcoxon Signed-Rank Tests

The results presented in this section (*Section 4.5.3.*) allow the completion of the testing of Hypothesis B via post-hoc Bonferroni adjusted Wilcoxon signed-rank (exact) p-values. To complete the testing of Hypothesis B, Tables XIII and XIV list the Wilcoxon signed-rank (exact) test results for pairwise comparisons of IQMs.

Results

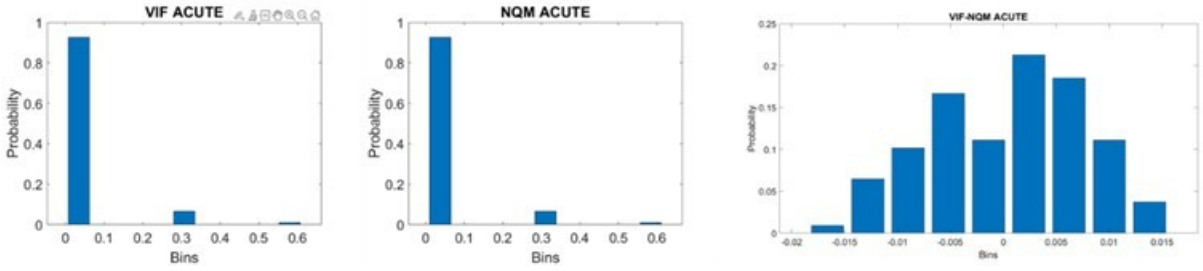


Fig. 34. Case: x – y symmetric about 0 (Wilcoxon signed-rank test). Plots correspond to the acute stroke diagnostic task. Histogram plots (left to right): absolute residuals for VIF plot (Fig. 29) = x, absolute residuals for NQM plot (Fig. 29) = y, and distribution of difference between absolute residuals for VIF plot – NQM plot = x – y, illustrating that the median is symmetric about zero (= 0). According to the Wilcoxon signed-rank test, VIF neither performs better nor worse than NQM for the acute stroke diagnostic task. VIF: visual information fidelity, NQM: noise quality measure.

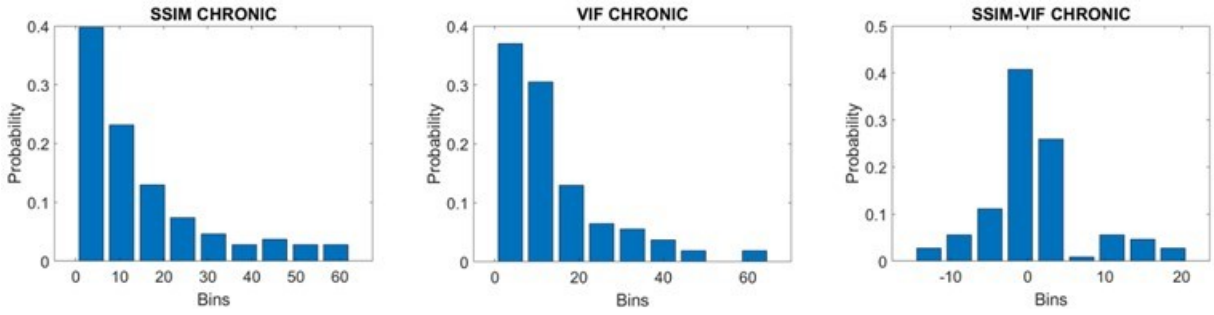


Fig. 35. Case: x – y > 0 (Wilcoxon signed-rank test). Plots correspond to the chronic stroke diagnostic task. Histogram plots (left to right): absolute residuals for SSIM plot (Fig. 30) = x, absolute residuals for VIF plot (Fig. 30) = y, and distribution of difference between absolute residuals for SSIM plot – VIF plot = x – y, illustrating that the median is greater than zero (> 0). According to the Wilcoxon signed-rank test, SSIM performs worse than VIF for the chronic stroke diagnostic task. SSIM: Structural SIMilarity; VIF: visual information fidelity.

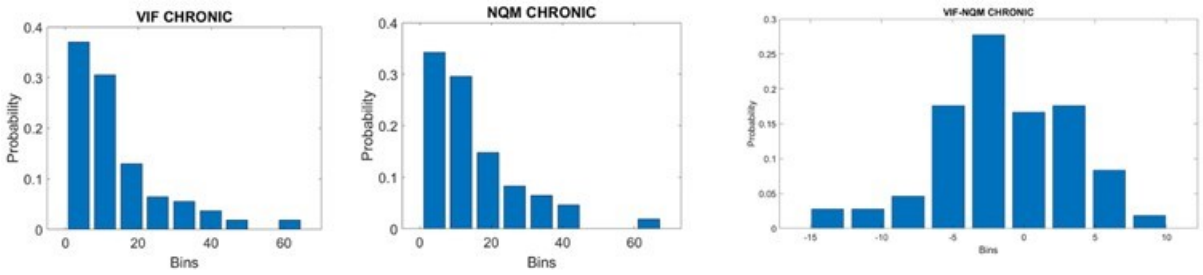


Fig. 36. Case: x – y < 0 (Wilcoxon signed-rank test). Plots correspond to the chronic stroke diagnostic task. Histogram plots (left to right): absolute residuals for VIF plot (Fig. 30) = x, absolute residuals for NQM plot (Fig. 30) = y, and distribution of difference between absolute residuals for VIF plot – NQM plot = x – y, illustrating that the median is less than zero (< 0). According to the Wilcoxon signed-rank test, VIF performs better than NQM for the chronic stroke diagnostic task. VIF: visual information fidelity, NQM: noise quality measure.

Figs. 34-36 show an illustrative example of the three cases corresponding to the results of the Wilcoxon signed-rank test, as follows: x – y symmetric about zero (Fig. 34), x – y > 0 (Fig. 35), and x – y < 0 (Fig. 36).

TABLE XIII: Wilcoxon signed-rank (exact) test results (alternative hypothesis: greater) for pairwise comparisons of absolute value residuals between subjective scores and the logistic fit for each IQM, corresponding to the (i) acute and (ii) chronic stroke diagnostic tasks, for all T2 FLAIR undersampled images. Clinically significant p-values are listed in bold text.

Pairwise Comparisons		Exact P-Values	
(x)	(y)	(i)	(ii)
SSIM	FSIM	0.1976	0.4160
SSIM	NQM	0.1172	0.5480
SSIM	VIF	0.0789	0.0601
SSIM	RMSE	0.3831	0.8077
NQM	SSIM	0.8828	0.4532
NQM	FSIM	0.9127	0.0504
NQM	RMSE	0.6650	0.5379
NQM	VIF	0.6337	**0.0010**
FSIM	SSIM	0.8024	0.5852
FSIM	RMSE	0.3554	0.9557
FSIM	VIF	0.1808	0.0771
FSIM	NQM	0.0873	0.9499
RMSE	FSIM	0.6446	**0.0446**
RMSE	NQM	0.3350	0.4621
RMSE	SSIM	0.6169	0.1932
RMSE	VIF	0.3122	**0.0043**
VIF	RMSE	0.6878	0.9957
VIF	SSIM	0.9211	0.9403
VIF	FSIM	0.8192	0.9234
VIF	NQM	0.3663	0.9990

IQM = image quality metric; T2 = transverse relaxation time; FLAIR = fluid attenuated inversion recovery; SSIM = Structural SIMilarity; NQM = noise quality measure; FSIM = Feature SIMilarity; RMSE = root mean square error; VIF = visual information fidelity

TABLE XIV: Post-hoc Bonferroni corrected Wilcoxon signed-rank (exact) test results (alternative hypothesis: greater) for pairwise comparisons of absolute value residuals between subjective scores and the logistic fit for each IQM, corresponding to the (i) acute and (ii) chronic stroke diagnostic tasks, for all T2 FLAIR undersampled images. Clinically significant p-values are listed in bold text.

Pairwise Comparisons		Bonferroni Adjusted Exact P-Values	
(x)	(y)	(i)	(ii)
SSIM	FSIM	0.7904	1.0000
SSIM	NQM	0.4688	1.0000
SSIM	VIF	0.3156	0.2402
SSIM	RMSE	1.0000	1.0000
NQM	SSIM	1.0000	1.0000
NQM	FSIM	1.0000	0.2017
NQM	RMSE	1.0000	1.0000
NQM	VIF	1.0000	**0.0042**
FSIM	SSIM	1.0000	1.0000
FSIM	RMSE	1.0000	1.0000
FSIM	VIF	0.7232	0.3084
FSIM	NQM	0.3490	1.0000
RMSE	FSIM	1.0000	0.1785
RMSE	NQM	1.0000	1.0000
RMSE	SSIM	1.0000	0.7728
RMSE	VIF	1.0000	**0.0172**
VIF	RMSE	1.0000	1.0000
VIF	SSIM	1.0000	1.0000
VIF	FSIM	1.0000	1.0000
VIF	NQM	1.0000	1.0000

IQM = image quality metric; T2 = transverse relaxation time; FLAIR = fluid attenuated inversion recovery; SSIM = Structural SIMilarity; NQM = noise quality measure; FSIM = Feature SIMilarity; RMSE = root mean square error; VIF = visual information fidelity

Wilcoxon, and Bonferroni adjusted Wilcoxon, signed-rank (exact) p-values are listed in Tables XIII and XIV, respectively, for the greater alternative hypothesis test corresponding to the acute and chronic stroke diagnostic tasks. Clinically significant p-values ($p < 0.05$) are listed in bold text.

TABLE XV: Results of Wilcoxon signed-rank (exact) test (alternative hypothesis: greater) based on Bonferroni adjusted p-values from Table XIV (i), corresponding to the acute stroke diagnostic task.

	SSIM	FSIM	NQM	RMSE	VIF
SSIM	---	= 0	= 0	= 0	= 0
FSIM	= 0	---	= 0	= 0	= 0
NQM	= 0	= 0	---	= 0	= 0
RMSE	= 0	= 0	= 0	---	= 0
VIF	= 0	= 0	= 0	= 0	---

SSIM = Structural SIMilarity; FSIM = Feature SIMilarity; NQM = noise quality measure; RMSE = root mean square error; VIF = visual information fidelity

TABLE XVI: Results of Wilcoxon signed-rank (exact) test (alternative hypothesis: greater) based on Bonferroni adjusted p-values from Table XIV (ii), corresponding to the chronic stroke diagnostic task.

	SSIM	FSIM	NQM	RMSE	VIF
SSIM	---	= 0	= 0	= 0	= 0
FSIM	= 0	---	= 0	= 0	= 0
NQM	= 0	= 0	---	= 0	< 0
RMSE	= 0	= 0	= 0	---	< 0
VIF	= 0	= 0	> 0	> 0	---

SSIM = Structural SIMilarity; FSIM = Feature SIMilarity; NQM = noise quality measure; RMSE = root mean square error; VIF = visual information fidelity

Based on the Bonferroni adjusted p-values in Table XIV (i-ii), the results of the Wilcoxon signed-rank (exact) test for the acute and chronic stroke diagnostic tasks are shown in Tables XV and XVI, respectively.

Discussion

Due to testing multiple comparisons, the Wilcoxon signed-rank (exact) p-values in Table XIII (i-ii) were adjusted, post-hoc, via the Bonferroni adjustment. To reiterate, the Wilcoxon signed-rank test tests whether the median of the difference between x and y is symmetric about zero (= 0), or statistically significantly greater than zero (> 0) or less

than zero (< 0). For Hypothesis B testing, x can be defined as the absolute value residuals corresponding to the IQMs listed in Table XIV (x) and column IQMs (IQM_{COLUMN}) in Tables XV and XVI, and y can be defined as absolute value residuals corresponding to the IQMs listed in Table XIV (y) and row IQMs (IQM_{ROW}) in Tables XV and XVI. Residuals are calculated by taking the absolute value of the difference between the IQM scores after regression and neuroradiologist raters' diagnostic confidence scores for the acute and chronic stroke diagnostic tasks. The null hypothesis is such that the distribution of the difference between x - y is symmetric about 0, as defined by '= 0' in Tables XV and XVI. The alternate hypotheses are such that either (a) the distribution of x - y is > 0 (i.e. true location shift is statistically significantly *greater* than zero), as defined by '> 0' in Tables XV and XVI, or (b) the distribution of x - y < 0 (i.e. true location shift is statistically significantly *less* than zero), as defined by '< 0' in Table XV and XVI.

Based on results from Tables XV and XVI, if the median of the difference between IQM_{COLUMN} and IQM_{ROW} is statistically significantly *greater* than zero then IQM_{COLUMN} performs worse than IQM_{ROW}, as this indicates that the residuals associated with IQM_{COLUMN} were larger than those associated with IQM_{ROW}. On the contrary, if the median of the difference between IQM_{COLUMN} and IQM_{ROW} is statistically significantly *less* than zero then IQM_{COLUMN} performs better than IQM_{ROW}, as this indicates that the residuals associated with IQM_{COLUMN} were smaller than those associated with IQM_{ROW}. If the median of the difference between IQM_{COLUMN} and IQM_{ROW} is symmetric about zero then IQM_{COLUMN} and IQM_{ROW} performs neither better nor worse in comparison to one another.

Table XV depicts that the medians of the difference between the absolute value of the residuals of any two IQMs were symmetric about zero; this supports the lack of rank order based on the SSR values in Table XII (a) (i) for the acute stroke diagnostic task. Alternatively, Table XVI provides statistical significance to the SSR values in Table XII (b) (i) for the chronic stroke diagnostic task, which yielded the rank order: (1) VIF, (2) FSIM, (3) NQM, (4) RMSE, (5) SSIM. Based on the results of the Wilcoxon signed-rank test, Hypothesis B (*Section 1.6.2.*) was proven only in part, in the sense that VIF performed better than RMSE. To explain, despite the SSR values of both RMSE and SSIM being greater than VIF, seemingly indicating that VIF performed better than both RMSE and SSIM due to smaller absolute residuals corresponding to the IQM plots in Fig. 30, VIF only performed better than RMSE from the statistical standpoint of the Wilcoxon signed-rank test. This means that, according to the Wilcoxon signed-rank test, VIF neither performs better nor worse than SSIM for the chronic stroke diagnostic task – but VIF performs better than RMSE for the chronic stroke diagnostic task. In all other cases, Hypothesis B (*Section 1.6.2.*) was disproven, such that FSIM did not perform better than RMSE and SSIM; in fact, according to the Wilcoxon signed-rank test, FSIM neither performs better nor worse than any of the other IQMs for the chronic stroke diagnostic task. Further, NQM did not perform better than RMSE and SSIM. Although NQM appeared to perform better than RMSE and SSIM as per the associated SSR values, from a statistical sense, the distribution of the absolute residuals of NQM minus RMSE (and NQM minus SSIM) was not statistically significantly different from zero, indicating that NQM did not perform better or worse than either RMSE or SSIM. Additionally, it is

interesting to note that, according to the Wilcoxon signed-rank test, NQM performed worse than VIF for the chronic stroke diagnostic task. To summarize, therefore, all IQMs performed neither better nor worse than one another, *except* for VIF, which performed better than RMSE and NQM.

Mason *et al.* [42] speculates, "NQM appears to perform particularly well for images degraded by noise" such as Gaussian noise, whereby it actually performed statistically better than VIF, which could lend a possible explanation for the poorer performance of NQM in this research work. If NQM possibly performs better for images degraded by noise (i.e. where noise is added), and performs statistically better than VIF in these cases, then it seems to logically follow suit that VIF may perform statistically better than NQM in cases where noise is "removed," such as with the CS reconstruction being inherently denoising. However, more research work would be required to make any of these speculations conclusively.

Although, [42] did mention that NQM performed statistically better than VIF in cases of undersampling, which would be counter to what was observed for the chronic stroke diagnostic task in this research work. However, the undersampling in [42] proceeded from image space, rather than raw fully-acquired k-space data as it were in this case.

To provide further detail, the results of this research work differ from Mason *et al.* [42] for a few reasons and may yield limitations for comparison. For example, [42] applied the logistic function to a single plot which combined results from different image alteration types onto the same axes, whereas this study applied the logistic function to a single plot with only a single type of image alteration which was

undersampling and CS reconstruction. The image alteration types studied by Mason *et al.* [42] were specifically those which would degrade the images, such as white noise, Gaussian blur, motion, Rician noise, undersampling, and wavelet compression. It is important to note, as well, that the version of undersampling in [42] was such that it was performed from the DICOM image, which means that pseudo k-space was undersampled (pseudo because a DICOM image only contains magnitude information, and therefore a FT from image space to k-space yields the same magnitude information, but not the same phase information from k-space acquisition – the acquired k-space phase information would have been lost and, therefore, upon reconstruction, the resultant image would contain an altered noise profile). In this study, raw k-space data were available and therefore undersampled, and thus the CS reconstruction had both magnitude and phase information available to perform the reconstruction. In addition, the results of [42] arise from various anatomies absent of pathology acquired using various MRI sequences. In this study, a single anatomy was imaged, pathology was present, and IQM results were obtained from T2 FLAIR images, only. Lastly, the expert rater participants in [42] comprised of two specialties of radiology (body and neuro) to perform confidence rankings, whereas only neuroradiologists were expert raters in this study. Overall, since confidence scores corresponding to all combinations of types of degradations, images and radiologists were combined into forming the single logistic regression per IQM for [42], it is therefore not a one-to-one comparison to the work of this thesis.

Further, although the progression of the IQM studies referenced in this thesis [42],

[68]-[69] that seek to understand the correlations between objective IQM scores and raters' subjective scores of image quality/diagnostic image quality have used the logistic function to model these correlations, it is understood that the logistic function may not be the preferred function in every situation going forward. Considering the relationship between objective and subjective scores depends on the diagnostic task, as was discovered through this research work in progression of the work by Mason *et al*. [42], it is likely that, if this work leads to other nuanced clinically translatable studies with specific anatomies, pathologies, and diagnostic tasks, the models themselves may need to become more nuanced, as well.

It is important to keep in mind that the results of this research work do not mean the results of Mason *et al.* [42] have been debunked, such than VIF, FSIM, and NQM do not actually perform better than RMSE and SSIM; this actually further solidifies the conclusions made by Mason *et al.* [42], one of which alluded to was that, despite RMSE and SSIM being common in MRI literature, there are other IQMs that may perform better depending on the purpose for their use. Therefore, to reiterate this exact point, for the specific CS reconstruction implementation at the given range of acceleration factors, all IQMs tested performed similarly for the purpose of performing the acute stroke diagnostic task, while VIF performed better than RMSE and NQM for the purpose of performing the chronic stroke diagnostic task.

Chapter 5: Conclusion

Chapter 5 discusses the conclusions of this thesis work, providing context to the clinical translation of these results, including opportunities for future work.

5.1. Hypothesis A

A reduced acquisition time up to R = 7X, in conjunction with neuroradiologists' remaining both accurate and confident in performing the acute stroke diagnostic task, yields promise for future work; exploring the prospective application of this research work to diagnosing acute stroke in emergency medicine is highly recommended. Despite a reduced acquisition time of only R = 2X where neuroradiologists' remained confident in performing the chronic stroke diagnostic task, it is important to remember that, at the POC, diagnosing chronic stroke is not the focus. As such, future work should investigate prospective acceleration up to R = 7X and not be limited by the fact that only R = 2X maintains diagnostic confidence scores for the chronic stroke diagnostic task.

The investigations of this thesis were mindful that the resultant accelerated images must remain diagnostically useful; however, the acceleration has yet to be maximized and subsequently optimized with diagnostic utility. Therefore, the determination of an optimized stroke imaging protocol that is *as fast as clinically useful* was out of the scope of this thesis, but should take precedence in future work in order to continue the trajectory of the important findings of this thesis. Since a future aim well beyond this thesis is direct clinical translation for using POC MRI as complementary to CT in the diagnosis of AIS in emergency medicine then imaging needs to be both as fast as possible and remain diagnostically useful.

The research work presented in this thesis was retrospective in nature because this is the first study performed with the 0.5 T system, so a retrospective study is necessary in order to provide a solid foundation of feasibility from which to pursue future, prospective accelerated image acquisition studies. Data acquisition is ongoing, thus continuously increasing *N* for future work. The population for future work will continue to include participants who are CT- AIS patients.

Future work includes investigation into determining the highest retrospective MR image acquisition acceleration factor (R_{MAX}) for which the diagnostic utility of images is maintained, including the determination of prospective clinical feasibility through the derivation of a diagnostic utility cost-function. The range for clinical significance should be fully explored in order to determine R_{MAX}. At R_{MAX}, diagnostic utility must remain at an acceptable level, as defined by expert radiologists.

Note that if R > 7X is to be examined, the sampling masks must be altered such that the central, fully-sampled portion of k-space is decreased – that is, if 2D linear Cartesian undersampling continue to be implemented. Undersampling and CS parameter tuning may therefore be part of future work.

Additional future work includes assessing neuroradiologists' diagnostic confidence scores corresponding to identification of diagnostically relevant features from images output from Objective 1. Diagnostically relevant features includes additional important aspects relevant to a neuroradiologist's diagnosis of stroke, and would therefore consist of not only the identification and corresponding confidence of presence/absence of AIS and chronic stroke, as was the case in the research work of this thesis, but also the

identification and corresponding confidence of: (1) the location of acute and chronic infarct, and (2) DWI-T2 FLAIR mismatch for AIS (*Appendix A.5.*).

If there is a DWI-T2 FLAIR mismatch then the patient may be in early AIS disease progression and therefore may not be in stable condition. Therefore, at this stage in the research work, DWI-T2 FLAIR mismatch was an irrelevant question given that the recruited and scanned patient participants were not in early AIS disease progression. Since NSH REB-approval was based on the pretense that scanned participants would be in stable condition, DWI-T2 FLAIR mismatch in the scans would not have (and should not have, for ethical reasons) existed. Nonetheless, even if DWI-T2 FLAIR mismatch was asked at this stage in the research work, the three board-certified neuroradiologist raters would have demonstrated bias in their answers, considering their a priori understanding of the participant recruitment criteria.

To conclude, Hypothesis A (*Section 1.6.2.*) was proven, and neuroradiologists' task-specific diagnostic confidence remained high at R = 7X ($p > 0.05$) for AIS, where chronic stroke was more sensitive to increasing R, decreasing at R = 3X ($p < 0.05$).

5.2. Hypothesis B

IQM calibration to radiologists' confidence scores is important for the following two reasons: (1) to narrow the search space for determining at what point radiologists' confidence in their identification of clinical stroke-related features drops past the point of being diagnostically useful for this particular use-case, and (2) as a result of narrowing this search space, improvements in study efficiency can be made by requiring less

iterations of questionnaire sessions with radiologists, in hopes of reducing the extent of the need for their valuable – yet limited – time.

Future work related to Hypothesis B (*Section 1.6.2.*) includes utilizing the logistic fit (non-linear regression model) results from this research work to assess the streamlining of images shown to radiologists. To this point, other models should be assessed, whereby the implementation of AI/machine learning models may be beneficial to address the case-by-case nature of correlating objective IQM scores with specific diagnostic tasks. Future studies should not be constrained to the logistic fit model, itself, but constrained, instead, to the diagnostic task at hand they wish to model. Ongoing work is investigating how IQMs behave with other types of image alterations, such as additive noise, Gaussian blur, etc. in order to provide more insight into the operations of these metrics.

To conclude, Hypothesis B (*Section 1.6.2.*) was only proven in part, such that VIF performed better than RMSE. However, important insights were gained that challenge habitual thinking. The fact that none of the IQMs tested – RMSE, SSIM, VIF, FSIM, or NQM – correlated with confidence for AIS but all correlated to various degrees with confidence for chronic stroke, suggests that IQM performance does not necessarily indicate an image's usefulness for a specific diagnostic task.

References

[1] C. Leiva-Salinas and M. Wintermark, "Imaging of ischemic stroke," *Neuroimaging Clin. N. Am.*, vol. 20, no. 4, pp. 455-468, 2010.

[2] T. D. Musuka, S. B. Wilton, M. Traboulsi, and M. Hill, "Diagnosis and management of acute ischemic stroke: speed is critical," *CMAJ,* vol. 187, no. 12, pp. 887-893, 2015.

[3] "Ischaemic stroke." Stroke Foundation. https://strokefoundation.org.au/about-stroke/learn/what-is-a-stroke/ischaemic-stroke-blocked-artery (accessed Mar. 20, 2022).

[4] "What is ischaemic stroke?" Health&. https://app.healthand.com/ca/topic/general-report/ischaemic-stroke (accessed Mar. 20, 2022).

[5] "The top 10 causes of death." World Health Organization. https://www.who.int/news-room/fact-sheets/detail/the-top-10-causes-of-death (accessed Dec. 9, 2020).

[6] M. Katan and A. Luft, "Global burden of stroke," *Semin. Neurol.*, vol. 38, no. 2, pp. 208-211, 2018.

[7] "Vascular dementia." Alzheimer Society of Canada. https://alzheimer.ca/en/about-dementia/other-types-dementia/vascular-dementia (accessed Aug. 22, 2021).

[8] "Stroke in Canada: Highlights from the Canadian Chronic Disease Surveillance System." Government of Canada. https://www.canada.ca/en/public-health/services/publications/ diseases-conditions/stroke-canada-fact-sheet.html (accessed Aug. 22, 2021).

[9] S. S. Virani *et al.*, "Heart disease and stroke statistics—2021 update: A report from the American Heart Association," *Circulation*, vol. 143, no. 8, pp. 254-743, 2021.

[10] C. R. Gomez, "Editorial: Time is brain!" *J. Stroke Cerebrovasc. Dis.*, vol. 3, no. 1, pp. 1-2, 1993.

[11] J. L. Saver, "Time is brain – quantified," *Stroke*, vol. 37, no. 1, pp. 263-266, 2006.

[12] W. G. Kunz *et al.*, "Public health and cost consequences of time delays to thrombectomy for acute ischemic stroke," *Neurology*, vol. 95, no. 18, pp. 2465-2475, 2020.

[13] J. M. Boulanger *et al*., "Canadian stroke best practice recommendations for acute stroke management: Prehospital, emergency department, and acute inpatient stroke care, 6th edition, update 2018," *Int. J. Stroke*, vol. 13, no. 9, pp. 949-984, 2018.

[14] W. J. Powers *et al.*, "Guidelines for the early management of patients with acute ischemic stroke: 2019 update to the 2018 guidelines for the early management of acute ischemic stroke: A guideline for healthcare professionals from the American Heart Association/American Stroke Association," *Stroke*, vol. 50, no. 12, pp. 344-418, 2019.

[15] W. J. Powers *et al.*, "2018 Guidelines for the early management of patients with acute ischemic stroke: A guideline for healthcare professionals from the American Heart Association/American Stroke Association," *Stroke*, vol. 49, no. 3, pp. 46-99, 2018.

[16] J. A. Chalela *et al.*, "Magnetic resonance imaging and computed tomography in emergency assessment of patients with suspected acute stroke: A prospective comparison," *Lancet*, vol. 369, no. 9558, pp. 293-298, 2007.

[17] D. Volders, private communication, Jun., 2022.

[18] "Stroke early management - guidelines - English version." Haute Autorité de Santé. https://www.has-sante.fr/jcms/c_830203/en/stroke-early-management (accessed Jun. 8, 2022).

[19] "Acute stroke guide." UCLA Health Resident Education System. https://res.mednet.ucla.edu/AcuteStrokeGuide (accessed Aug. 19, 2021).

[20] P. D. Schellinger, O. Jansen, J. B. Fiebach, W. Hacke, and Klaus Sartor, "A standardized MRI stroke protocol," *Stroke*, vol. 30, no. 4, pp. 765-768, 1999.

[21] R. Gasparotti, L. Pinelli, and R. Liserre, "New MR sequences in daily practice: susceptibility weighted imaging. A pictorial essay," *Insights Imaging*, vol. 2, no. 3, pp. 335-347, 2011.

[22] "1.5T SIGNA™ HDxt SIGNA™Works Edition." GE Healthcare. https://www.gehealthcare.com/products/magnetic-resonance-imaging/1-5t/1-5t-signa (accessed Mar. 20, 2022).

[23] "Revolution Apex." GE Healthcare. https://www.gehealthcare.ca/en-CA/products/computed-tomography/revolution-apex (accessed Aug. 16, 2021).

[24] C. Provost *et al.* "Magnetic resonance imaging or computed tomography before treatment in acute ischemic stroke: Effect on workflow and functional outcome," *Stroke*, vol. 50, no. 3, pp. 659-664, 2019.

[25] A. Panther *et al.*, “A dedicated head-only MRI scanner for point-of-care imaging,” in *Proc. Intl. Soc. Mag. Reson. Med.*, 2019, no. 3679.

[26] M. Lustig, D. Donoho, and J. M. Pauly, “Sparse MRI: The application of compressed sensing for rapid MR imaging,” *Magn. Reson. Med.*, vol. 58, no. 6, pp. 1182-1195, 2007.

[27] M. Buller and J. P. Karis, “Introduction of a dedicated emergency department MR imaging scanner at the barrow neurological institute,” *AJNR Am. J. Neuroradiol*., vol. 38, no. 8, pp. 1480-1485, 2017.

[28] “MRA parameter selection.” Questions and Answers in MRI. https://mriquestions.com/mra-parameters.html (accessed May 19, 2022).

[29] “DWI b-value.” Questions and Answers in MRI. https://mriquestions.com/what-is-the-b-value.html (accessed Jun. 5, 2022).

[30] D. J. Brenner and E. J. Hall, “Computed tomography - An increasing source of radiation exposure,” *N. Engl. J. Med.*, vol. 357, no. 22, pp. 2277-2284, 2007.

[31] W. A. Kalender, “Principles of Computed Tomography,” in Computed Tomography: Fundamentals, System Technology, Image Quality, Applications, 3rd ed., Erlangen, Germany: Publicis Publishing, 2011, ch. 1, sec. 1.1, pp. 18-23.

[32] S. Currie, N. Hoggard, I. J. Craven, M. Hadjivassiliou, and I. D. Wilkinson, “Understanding MRI: Basic MR physics for physicians,” *Postgrad. Med. J.*, vol. 89, no. 1050, pp. 209–223, 2013.

[33] S. D. Sharma, C. L. Fong, B. S. Tzung, M. Law, and K. S. Nayak, “Clinical image quality assessment of accelerated magnetic resonance neuroimaging using compressed sensing,” *Invest. Radiol.*, vol. 48, no. 9, pp. 638-645, 2013.

[34] S. Mönch, N. Sollmann, A. Hock, C. Zimmer, J. S. Kirschke, and D. M. Hedderich, “Magnetic resonance imaging of the brain using compressed sensing – quality assessment in daily clinical routine,” *Clin. Neuroradiol.*, vol. 30, no. 2, pp. 279-286, 2020.

[35] M. Kayvanrad, A. Lin, R. Joshi, J. Chiu, and T. Peters, “Diagnostic quality assessment of compressed sensing accelerated magnetic resonance neuroimaging,” *J. Magn. Reson. Imaging*, vol. 44, no. 2, pp. 433-444, 2016.

[36] M. Uecker *et al.*, “Berkeley Advanced Reconstruction Toolbox,” in *Proc. Intl. Soc. Mag. Reson. Med.*, 2015, no. 2486.

[37] M. Lustig, D. L. Donoho, J. M. Santos, and J. M. Pauly, “Compressed sensing: A look at how CS can improve on current imaging techniques,” *IEEE Signal Process. Mag.*, vol. 25, no. 2, pp. 72-82, 2008.

[38] C. Westbrook, C. K. Roth, and J. Talbot, “Parameters and trade-offs," in *MRI in Practice*, 4th ed., Chichester, UK: Blackwell Publishing Ltd., 2011, ch. 4, p. 132.

[39] L. Zhang, L. Zhang, X. Mou, and D. Zhang, “FSIM: A feature similarity index for image quality assessment,” *IEEE Trans. Image Process.*, vol. 20, no. 8, pp. 2378-2386, 2011.

[40] C. Westbrook, C. K. Roth, and J. Talbot, “Encoding and image formation," in *MRI in Practice*, 4th ed., Chichester, UK: Blackwell Publishing Ltd., 2011, ch. 3, p. 98-101.

[41] C. Westbrook, C. K. Roth, and J. Talbot, “Encoding and image formation," in *MRI in Practice*, 4th ed., Chichester, UK: Blackwell Publishing Ltd., 2011, ch. 3, p. 95.

[42] A. Mason, *et al.*, “Comparison of objective image quality metrics to expert radiologists’ scoring of diagnostic quality of MR images,” *IEEE Trans. Med. Imaging*, vol. 39, no. 4, pp. 1064-1072, 2020.

[43] C. Westbrook, C. K. Roth, and J. Talbot, “Pulse sequences," in *MRI in Practice*, 4th ed., Chichester, UK: Blackwell Publishing Ltd., 2011, ch. 5, p. 193

[44] D. L. Donoho, “Compressed sensing,” *IEEE Trans. Inf. Theory*, vol. 52, no. 4, pp. 1289-1306, 2006.

[45] D. J. Lin, P. M. Johnson, F. Knoll, and Y. W. Lui, “Artificial intelligence for MR image reconstruction: An overview for clinicians,” *J. Magn. Reson. Imaging*, vol. 53, no. 4, pp. 1015-1028, 2021.

[46] K. Hammernik *et al.*, “Learning a variational network for reconstruction of accelerated MRI data,” *Magn. Reson. Med.*, vol. 79, no. 6, pp. 3055-3071, 2018.

[47] M. Mardani *et al.*, “Deep generative adversarial neural networks for compressive sensing MRI,” *IEEE Trans. Med. Imaging*, vol. 38, no. 1, pp. 167-179, 2019.

[48] T. M. Quan, T. Nguyen-Duc, and W. Jeong, “Compressed sensing MRI reconstruction using a generative adversarial network with a cyclic loss,” *IEEE Trans. Med. Imaging*, vol. 37, no. 6, pp. 1488-1497, 2018.

[49] Y. Yang, J. Sun, H. Li, and Z. Xu, “Deep ADMM-Net for compressive sensing MRI,” in *Proc. 30th Int. Conf. NeurIPS*, Dec. 2016, pp. 10-18.

[50] J. Cao, S. Liu, H. Liu, and H. Lu, “CS-MRI reconstruction based on analysis dictionary learning and manifold structure regularization,” *Neural Netw.*, vol. 123, pp. 217-233, 2020.

[51] L. Gueddari, E. Chouzenoux, A. Vignaud, J. Pesquet, and P.Ciuciu, “Online MR image reconstruction for compressed sensing acquisition in T2* imaging,” in *SPIE Conf. Wavelets and Sparsity XVIII*, San Diego, CA, USA, Aug. 2019.

[52] R. Liu, Y. Zhang, S. Cheng, Z. Luo, and X. Fan, “A deep framework assembling principled modules for CS-MRI: Unrolling perspective, convergence behaviors, and practical modeling,” *IEEE Trans. Med. Imaging*, vol. 39, no. 12, pp. 4150-4163, 2020.

[53] A. C. Yang, M. Kretzler, S. Sudarski, V. Gulani, and N. Seiberlich, “Sparse reconstruction techniques in magnetic resonance imaging: Methods, applications, and challenges to clinical adoption,” *Invest. Radiol.*, vol. 51, no. 6, pp. 349-364, 2016.

[54] C. Westbrook, C. K. Roth, and J. Talbot, “Encoding and image formation," in *MRI in Practice*, 4th ed., Chichester, UK: Blackwell Publishing Ltd., 2011, ch. 3, p. 93-94.

[55] O. N. Jaspan, R. Fleysher, and M. L. Lipton, “Compressed sensing MRI: A review of the clinical literature,” *Br. J. Radiol.*, vol. 88, no. 1056, pp. 1-12, 2015.

[56] L. Feng, T. Benkert, K. T. Block, D. K. Sodickson, R. Otazo, and H. Chandarana, “Compressed sensing for body MRI,” *J. Magn. Reson. Imaging*, vol. 45, no. 4, pp. 966-987, 2017.

[57] V. M. Runge, J. K. Richter, and J. T. Heverhagen, “Speed in clinical magnetic resonance,” *Invest. Radiol.*, vol. 52, no. 1, pp. 1-17, 2017.

[58] C. Jaimes and M. S. Gee, “Strategies to minimize sedation in pediatric body magnetic resonance imaging,” *Pediatr. Radiol.*, vol. 46, no. 6, pp. 916-927, 2016.

[59] R. Ahmad, H. H. Hu, R. Krishnamurthy, and R. Krishnamurthy, “Reducing sedation for pediatric body MRI using accelerated and abbreviated imaging protocols,” *Pediatr. Radiol.*, vol. 48, no. 1, pp. 37-49, 2018.

[60] R. Krishnamurthy *et al.*, “Recent advances in pediatric brain, spine, and neuromuscular magnetic resonance imaging techniques,” *Pediatr. Neurol.*, vol. 96, pp. 7-23, 2019.

[61] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, “Image quality assessment: From error visibility to structural similarity,” *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600-612, 2004.

[62] N. Damera-Venkata, T. D. Kite, W. S. Geisler, B. L. Evans, and A. C. Bovik, “Image quality assessment based on a degradation model,” *IEEE Trans. Image Process.*, vol. 9, no. 4, pp. 636-650, 2000.

[63] H. R. Sheikh and A. C. Bovik, “Image information and visual quality,” *IEEE Trans. Image Process.*, vol. 15, no. 2, pp. 430-444, 2006.

[64] Z. Wang and A. C. Bovik, “Mean squared error: Love it or leave it? A new look at signal fidelity measures,” *IEEE Signal Process. Mag.*, vol. 26, no. 1, pp. 98-117, 2009.

[65] U. Hassan and M. S. Anwar, “Reducing noise by repetition: Introduction to signal averaging,” *Eur. J. Phys.*, vol. 31, no. 3, pp. 453-465, 2010.

[66] “Partial Fourier.” Radiology Key. https://radiologykey.com/partial-fourier/ (accessed
Jun. 20, 2022).

[67] “Magnetic resonance imaging (MRI) of the brain and spine: Basics neuroimaging in neurology.” Neuroimaging in Neurology: An Interactive Approach. https://case.edu/med/neurology/NR/MRI%20Basics.htm (accessed May. 19, 2022).

[68] H. R. Sheikh, M. F. Sabir, and A. C. Bovik, “A Statistical Evaluation of Recent Full Reference Image Quality Assessment Algorithms,” *IEEE Trans. Image Process*., vol. 15, no. 11, pp. 3441–3452, 2006.

[69] L. S. Chow, H. Rajagopal, and R. Paramesran, “Correlation between subjective and objective assessment of magnetic resonance (MR) images,” *Magn. Reson. Imaging*, vol. 34, no. 6, pp. 820–831, 2016.

[70] “Number of excitations (NEX/NSA) and image quality.” mrimaster.com. https://mrimaster.com/technique%20NEX.html (accessed Jun. 13, 2022).

[71] D. Puhr-Westerheide *et al.*, “Cost-effectiveness of short-protocol emergency brain MRI after negative non-contrast CT for minor stroke detection,” *Eur. Radiol.*, vol. 32, no. 2, pp. 1117-1126, 2022.

[72] A. de Havenon *et al.*, “Direct cost analysis of rapid MRI in the emergency department evaluation of patients suspected of having acute ischemic stroke,” *Neuroradiol. J.*, vol. 0, no. 0, pp. 1-6, 2022.

[73] X. Liu, J. Almast, and S. Ekholm, “Lesions masquerading as acute stroke,” *J. Magn. Reson. Imaging*, vol. 37, no. 1, pp. 15-34, 2013.

[74] C. S. Kidwell *et al.*, "Thrombolytic reversal of acute human cerebral ischemic injury shown by diffusion/perfusion magnetic resonance imaging," *Ann. Neurol.*, vol. 47, no. 4, pp. 462-469, 2000.

[75] H. Lee, Y. Yang, B. Liu, S. A. Castro, and T. Shi, "Patients with acute ischemic stroke who receive brain magnetic resonance imaging demonstrate favorable in-hospital outcomes," *J. Am. Heart Assoc.*, vol. 9, no. 20, pp. 1-6, 2020.

[76] W. A. Mehan Jr. *et al.*, "Optimal Brain MRI Protocol for New Neurological Complaint," *PloS one,* vol. 9, no. 10, pp. 1-10, 2014.

[77] D. Liang, B. Liu, J. Wang, and L. Ying, "Accelerating SENSE using compressed sensing," *Magn. Reson. Med.*, vol. 62, no. 6, pp. 1574-1584, 2009.

[78] U. Molnar, J. Nikolov, O. Nikolić, N. Boban, V. Subašić, and V. Till, "Diagnostic quality assessment of compressed SENSE accelerated magnetic resonance images in standard neuroimaging protocol: Choosing the right acceleration," *Phys. Med.*, vol. 88, pp. 158-166, 2021.

[79] "Data." Michael (Miki) Lustig. http://people.eecs.berkeley.edu/~mlustig/Software.html (accessed Dec. 6, 2020).

[80] R. G. González *et al.*, "The Massachusetts General Hospital acute stroke imaging algorithm: an experience and evidence based approach," *J. Neurointerv. Surg.*, vol. 5, pp. 7-12, 2013.

[81] S. E. Honig, L. S. Babiarz, E. L. Honig, S. Mirbagheri, V. Urrutia, and D. M. Yousema, "The impact of installing an MR scanner in the emergency department for patients presenting with acute stroke-like symptoms," *Clin. Imaging*, vol. 45, pp. 65-70, 2017.

[82] K. P. Pruessmann, M. Weiger, M. B. Scheidegger, and P. Boesiger, "SENSE: sensitivity encoding for fast MRI," *Magn. Reson. Med.*, vol. 42, pp. 952-962, 1999.

[83] M. A. Griswold *et al.*, "Generalized autocalibrating partially parallel acquisitions (GRAPPA)," *Magn. Reson. Med.*, vol. 47, no. 6, pp. 1202-1210, 2002.

[84] A. J. Viera and J. M. Garrett, "Understanding interobserver agreement: The kappa statistic," *Fam. Med.*, vol. 37, no. 5, pp. 360-363, 2005.

[85] S. Kojima, H. Shinohara, T. Hashimoto, and S. Suzuki, "Undersampling patterns in k-space for compressed sensing MRI using two-dimensional Cartesian sampling," *Radiol. Phys. Technol.*, vol. 11, no. 3, pp. 303-319, 2018.

[86] H. Zheng *et al.*, “Multi-contrast brain MRI image super-resolution with gradient-guided edge enhancement,” *IEEE Access*, vol. 6, pp. 57856-57867, 2018.

[87] A. R. Feinstein and D. V. Cicchetti, “High agreement but low kappa: I. The problems of the two paradoxes,” *J. Clin. Epidemiol.*, vol. 43, no. 6, pp. 543-549, 1990.

[88] K. L. Gwet, “Testing the difference of correlated agreement coefficients for statistical significance,” *Educ. Psychol. Meas.*, vol. 76, no. 4, pp. 609-637, 2016.

[89] C. E. Shannon, “Communication in the presence of noise,” *Proc. IRE*, vol. 37, no. 1, pp. 10-21, 1949.

[90] J. Cohen, “A coefficient of agreement for nominal scales,” *Educ. Psychol. Meas.*, vol. 20, no. 1, pp. 37-46, 1960.

[91] K. L. Gwet, “Computing inter-rater reliability and its variance in the presence of high agreement,” *Br. J. Math. Stat. Psychol.*, vol. 61, no. 1, pp. 29-48, 2008.

Appendix A: Ranking Study

A.1. Schedule

Dataset Series

100	013 max add	4000	015 max add	7900	007 r=1	11800	008 maxG	15700	012 minG	19600	010 r=6	**LEGEND**	**Dataset Series**
200	010 r=5	4100	002 r=1	8000	010 r=7	11900	015 r=6	15800	007 min add	19700	005ct- r=4	**Week 1**	100-3900
300	healthy1 max add	4200	009 r=7	8100	healthy2 r=5	12000	004 r=4	15900	013 r=2	19800	011 med add	**Week 2**	4000-7800
400	001ct- r=5	4300	010 max add	8200	015 r=5	12100	010 med add	16000	healthy2 r=6	19900	002ct- r=2	**Week 3**	7900-11700
500	healthy2 r=3	4400	004ct- r=3	8300	004ct- r=1	12200	healthy2 minG	16100	healthy1 r=2	20000	healthy1 min add	**Week 4**	11800-15600
600	005 r=1	4500	007 max add	8400	009 r=5	12300	healthy1 med add	16200	005 min add	20100	001ct- maxG	**Week 5**	15700-19500
700	008 max add	4600	healthy1 r=3	8500	001ct- r=1	12400	009 minG	16300	010 maxG	20200	008 r=6	**Week 6**	19600-23400
800	005ct- medG	4700	013 r=5	8600	healthy1 medG	12500	007 med add	16400	004 minG	20300	015 maxG		
900	002 r=7	4800	005ct- r=7	8700	005 r=7	12600	012 r=4	16500	001ct- r=2	20400	007 r=4		
1000	009 max add	4900	005 r=5	8800	014 medG	12700	004ct- minG	16600	008 med add	20500	009 min add		
1100	007 r=5	5000	healthy2 r=1	8900	002 max add	12800	011 r=4	16700	002ct- maxG	20600	005 r=4		
1200	011 r=1	5100	012 medG	9000	004 r=7	12900	005 med add	16800	014 r=4	20700	healthy2 min add		
1300	002ct- r=3	5200	014 r=1	9100	005ct- r=2	13000	005ct- r=6	16900	004ct- med add	20800	002 minG		
1400	014 r=7	5300	004 r=1	9200	008 r=7	13100	013 med add	17000	009 r=4	20900	013 maxG		
1500	004 r=3	5400	001ct- max add	9300	002ct- r=1	13200	002 r=2	17100	005ct- minG	21000	004ct- min add		
1600	004ct- r=5	5500	011 max add	9400	012 max add	13300	014 med add	17200	002 r=4	21100	012 maxG		
1700	012 r=3	5600	002ct- r=7	9500	011 r=7	13400	002ct- r=6	17300	015 r=2	21200	014 r=2		
1800	015 r=1	5700	008 r=3	9600	004ct- r=4	13500	001ct- minG	17400	011 minG	21300	004 min add		
1900	010 medG	5800	012 r=7	9700	013 r=7	13600	012 r=2	17500	012 min add	21400	011 r=2		
2000	011 r=5	5900	002ct- max add	9800	015 medG	13700	014 r=6	17600	013 r=4	21500	008 r=2		
2100	healthy1 r=7	6000	008 r=1	9900	005 r=3	13800	004ct- maxG	17700	008 minG	21600	012 r=6		
2200	008 r=5	6100	healthy2 max add	10000	007 r=7	13900	013 r=6	17800	014 minG	21700	005 r=6		
2300	014 r=3	6200	014 r=5	10100	012 r=5	14000	011 r=6	17900	002ct- r=4	21800	healthy1 r=6		
2400	005ct- r=1	6300	001ct- r=7	10200	healthy1 r=5	14100	009 med add	18000	007 minG	21900	001ct- r=4		
2500	015 r=3	6400	007 r=3	10300	009 medG	14200	healthy2 maxG	18100	010 min add	22000	002 med add		
2600	002ct- r=5	6500	healthy1 r=1	10400	012 med add	14300	healthy1 r=4	18200	015 minG	22100	010 r=4		
2700	healthy2 r=7	6600	005 max add	10500	004 max add	14400	002 r=6	18300	011 maxG	22200	004ct- r=2		
2800	007 medG	6700	011 r=3	10600	013 medG	14500	015 med add	18400	004ct- r=6	22300	015 r=4		
2900	009 r=1	6800	015 r=7	10700	002 maxG	14600	010 r=2	18500	009 maxG	22400	007 r=6		
3000	001ct- r=3	6900	004 r=5	10800	001ct- med add	14700	005 maxG	18600	001ct- min add	22500	002ct- med add		
3100	002 r=5	7000	002 r=3	10900	002ct- medG	14800	007 r=2	18700	002 min add	22600	004 maxG		
3200	005 medG	7100	004ct- max add	11000	010 minG	14900	008 r=4	18800	healthy2 r=4	22700	009 r=6		
3300	012 r=1	7200	010 r=3	11100	healthy2 medG	15000	005ct- min add	18900	005 minG	22800	014 min add		
3400	013 r=1	7300	009 r=3	11200	014 maxG	15100	001ct- r=6	19000	004 r=6	22900	healthy2 med add		
3500	004 medG	7400	013 r=3	11300	005ct- r=5	15200	004 r=2	19100	healthy1 minG	23000	013 min add		
3600	004ct- r=7	7500	001ct- medG	11400	008 medG	15300	002ct- min add	19200	013 minG	23100	007 maxG		
3700	010 r=1	7600	002 medG	11500	011 min add	15400	009 r=2	19300	002ct- minG	23200	015 min add		
3800	011 medG	7700	004ct- medG	11600	014 max add	15500	004 med add	19400	005ct- med add	23300	005ct- maxG		
3900	005ct- max add	7800	005ct- r=3	11700	005 r=2	15600	healthy2 r=2	19500	healthy1 maxG	23400	008 min add		

A.2. Questionnaire

Dataset Series:		**100**	**200**	**300**	**etc.**
Question 1:					
Using the DWI and T2 FLAIR images, would you report the presence of an acute stroke?	YES = **1** NO = **0**				
Question 2:					
On a scale from 1-5 (with 1 being no confidence and 5 being 100 % confident), how would you rate your clinical confidence in reporting the presence or absence of an acute stroke?	0 % confident = **1** 25 % confident = **2** 50 % confident = **3** 75 % confident = **4** 100 % confident = **5**				
Question 3:					
What Fazekas score (0-3) would you report for identification of chronic ischemic lesion burden?	absent = **0** punctate foci, or "caps" or pencil-thin lining = **1** beginning confluence, or smooth "halo" = **2** large confluent areas, or irregular periventricular signal extending into deep WM = **3**				
Question 4:					
On a scale from 1-5 (with 1 being no confidence and 5 being 100 % confident), how would you rate your clinical confidence in reporting the above Fazekas score?	0 % confident = **1** 25 % confident = **2** 50 % confident = **3** 75 % confident = **4** 100 % confident = **5**				
You have completed the Questionnaire for the particular **'Dataset Series'. Move on to answering the Questionnaire again for the next '**Dataset Series'**.*					

A.3. Weekly Calibration Email

The intent of the questions is to probe your confidence in your decision, whatever that might be, with respect to the clinical task you are performing. The rankings are neither a reflection of the quality of the images, nor of the clinical decision itself, but of your

confidence in identifying presence/absence of acute stroke and assigning a Fazekas score for chronic stroke.

Note:
-For all dataset series, T2 FLAIR images may/may not be degraded quality, but DWI/ADC will never be degraded quality.
-For the purposes of this study "Acute stroke" is considered to span both acute and subacute (no hyperacute).
-Fazekas scoring combines periventricular and deep white matter T2 hyperintense lesion(s).
-Likert scores (1-5) are not intended to reflect the decision itself, but rather **your confidence** in that decision with a 5 representing 100% confident and a 1 indicating no confidence in the relevant clinical decision.

- A **high** Likert rating indicates that you view findings to be **either a true positive or true negative**
- A **low** Likert rating indicates that you view findings to be **either a false negative or false positive.**

I.e. if these images arrived in your PACS feed for a patient with suspected stroke, how confident would you feel in using this data to make a diagnostic judgement, knowing that your findings may be used to triage this patient and potentially make choices with respect to treatment?

A.4. Raw Results

Using the DWI and T2 FLAIR images, would you report the presence of an acute stroke? YES = 1, NO = 0

	Patient ID																	
	TD 002			TD 004			TD 005			TD 007			TD 008			TD 009		
Acceleration Factor	Rater 1	Rater 2	Rater 3	Rater 1	Rater 2	Rater 3	Rater 1	Rater 2	Rater 3	Rater 1	Rater 2	Rater 3	Rater 1	Rater 2	Rater 3	Rater 1	Rater 2	Rater 3
R = 1X	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
R = 2X	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
R = 3X	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
R = 4X	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
R = 5X	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
R = 6X	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
R = 7X	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
	Patient ID																	
	TD 010			TD 011			TD 012			TD 013			TD 014			TD 015		
Acceleration Factor	Rater 1	Rater 2	Rater 3	Rater 1	Rater 2	Rater 3	Rater 1	Rater 2	Rater 3	Rater 1	Rater 2	Rater 3	Rater 1	Rater 2	Rater 3	Rater 1	Rater 2	Rater 3
R = 1X	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
R = 2X	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
R = 3X	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
R = 4X	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
R = 5X	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
R = 6X	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
R = 7X	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
	Patient ID																	
	CTneg 001			CTneg 002			CTneg 004			CTneg 005			STRHC 001			STRHC 002		
Acceleration Factor	Rater 1	Rater 2	Rater 3	Rater 1	Rater 2	Rater 3	Rater 1	Rater 2	Rater 3	Rater 1	Rater 2	Rater 3	Rater 1	Rater 2	Rater 3	Rater 1	Rater 2	Rater 3
R = 1X	1	1	1	0	0	0	0	0	0	1	1	1	0	0	0	0	0	0
R = 2X	1	1	1	0	0	0	0	0	0	1	1	1	0	0	0	0	0	0
R = 3X	1	1	1	0	0	0	0	0	0	1	1	1	0	0	0	0	0	0
R = 4X	1	1	1	0	0	0	0	0	0	1	1	1	0	0	0	0	0	0
R = 5X	1	1	1	0	0	0	0	0	0	1	1	1	0	0	0	0	0	0
R = 6X	1	1	1	0	0	0	0	0	0	1	1	1	0	0	0	0	0	0
R = 7X	1	1	1	0	0	0	0	0	0	1	1	1	0	0	0	0	0	0

LEGEND
Accurate w/ Calibration Score
Part of Verification (These are now Accurate)
Part of Verification (These remained Accurate)

On a scale from 1-5 (with 1 being no confidence and 5 being 100 % confident), how would you rate your clinical confidence in reporting the presence or absence of an acute stroke? 0 % confident = 1, 25 % confident = 2, 50 % confident = 3, 75 % confident = 4, 100 % confident = 5

	Patient ID																	
	TD 002			TD 004			TD 005			TD 007			TD 008			TD 009		
Acceleration Factor	Rater 1	Rater 2	Rater 3	Rater 1	Rater 2	Rater 3	Rater 1	Rater 2	Rater 3	Rater 1	Rater 2	Rater 3	Rater 1	Rater 2	Rater 3	Rater 1	Rater 2	Rater 3
R = 1X	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	4
R = 2X	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5
R = 3X	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5
R = 4X	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	4
R = 5X	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5
R = 6X	5	5	5	5	5	4	5	5	5	5	5	5	5	5	5	5	5	4
R = 7X	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	4
																Gave a Calibration score of 5		
																Gave a Verification score of 4		

	Patient ID																	
	TD 010			TD 011			TD 012			TD 013			TD 014			TD 015		
Acceleration Factor	Rater 1	Rater 2	Rater 3	Rater 1	Rater 2	Rater 3	Rater 1	Rater 2	Rater 3	Rater 1	Rater 2	Rater 3	Rater 1	Rater 2	Rater 3	Rater 1	Rater 2	Rater 3
R = 1X	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5
R = 2X	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5
R = 3X	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5
R = 4X	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5
R = 5X	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5
R = 6X	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5
R = 7X	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	3

	Patient ID																	
	CTneg 001			CTneg 002			CTneg 004			CTneg 005			STRHC 001			STRHC 002		
Degradation Level	Rater 1	Rater 2	Rater 3	Rater 1	Rater 2	Rater 3	Rater 1	Rater 2	Rater 3	Rater 1	Rater 2	Rater 3	Rater 1	Rater 2	Rater 3	Rater 1	Rater 2	Rater 3
R = 1X	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5
R = 2X	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5
R = 3X	5	5	5	5	5	4	5	5	5	5	5	5	5	5	5	5	5	5
R = 4X	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5
R = 5X	5	5	5	5	5	4	5	5	4	5	5	5	5	5	5	5	5	5
R = 6X	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5
R = 7X	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5
				Gave a Calibration score of 5														

LEGEND
Consistent w/ Calibration Score
Part of Verification (These remained Consistent)
Part of Verification (These remained Consistent)
Inconsistent w/ Calibration Score
Inconsistent w/ Verification Score

What Fazekas score (0-3) would you report for identification of chronic ischemic lesion burden? absent = 0; punctate foci, or "caps" or pencil-thin lining = 1; beginning confluence, or smooth "halo" = 2; large confluent areas, or irregular periventricular signal extending into deep WM = 3

	Patient ID																	
	TD 002			TD 004			TD 005			TD 007			TD 008			TD 009		
Acceleration Factor	Rater 1	Rater 2	Rater 3	Rater 1	Rater 2	Rater 3	Rater 1	Rater 2	Rater 3	Rater 1	Rater 2	Rater 3	Rater 1	Rater 2	Rater 3	Rater 1	Rater 2	Rater 3
R = 1X	0	1	1	3	3	3	3	3	3	0	1	1	2	2	2	0	1	1
R = 2X	0	1	2	3	3	3	3	3	3	0	1	1	3	2	2	0	1	1
R = 3X	0	1	1	3	3	3	3	3	3	0	0	1	2	2	2	0	0	1
R = 4X	0	1	1	3	3	3	3	3	3	0	1	1	2	2	2	0	1	1
R = 5X	0	1	1	3	3	3	3	3	3	1	1	1	2	1	2	0	0	1
R = 6X	0	1	1	3	3	3	3	3	3	0	1	1	2	2	2	0	0	1
R = 7X	0	1	2	3	3	3	3	3	3	0	1	2	1	2	2	0	0	0
	Gave a Calibration score of 2															Gave a Calibration score of 1		

	Patient ID																	
	TD 010			TD 011			TD 012			TD 013			TD 014			TD 015		
Acceleration Factor	Rater 1	Rater 2	Rater 3	Rater 1	Rater 2	Rater 3	Rater 1	Rater 2	Rater 3	Rater 1	Rater 2	Rater 3	Rater 1	Rater 2	Rater 3	Rater 1	Rater 2	Rater 3
R = 1X	0	0	1	0	0	1	0	0	1	2	3	3	1	1	1	1	1	1
R = 2X	0	0	1	0	0	1	0	0	1	2	2	2	0	0	1	2	1	1
R = 3X	0	0	1	0	0	1	0	0	1	2	1	2	1	1	1	2	1	2
R = 4X	0	1	1	0	0	1	1	0	1	2	2	2	0	1	1	2	2	2
R = 5X	0	0	0	0	0	1	1	0	1	2	3	2	0	1	1	1	2	2
R = 6X	0	1	1	0	0	1	0	0	1	2	2	2	0	0	1	2	1	2
R = 7X	0	0	1	0	0	1	1	0	0	2	2	2	0	1	1	1	1	1
	Gave a Calibration score of 1																	

	Patient ID																	
	CTneg 001			CTneg 002			CTneg 004			CTneg 005			STRHC 001			STRHC 002		
Acceleration Factor	Rater 1	Rater 2	Rater 3	Rater 1	Rater 2	Rater 3	Rater 1	Rater 2	Rater 3	Rater 1	Rater 2	Rater 3	Rater 1	Rater 2	Rater 3	Rater 1	Rater 2	Rater 3
R = 1X	2	2	3	2	2	3	1	0	1	0	0	1	0	0	0	0	0	1
R = 2X	2	3	2	2	3	3	0	0	1	0	0	1	0	0	0	0	0	1
R = 3X	2	3	3	2	2	3	0	0	1	0	0	1	0	0	1	0	0	1
R = 4X	2	3	2	2	2	2	1	1	1	1	0	1	0	0	1	0	0	1
R = 5X	2	2	3	2	2	3	0	0	1	0	0	1	0	0	1	0	1	1
R = 6X	2	2	2	2	2	3	0	1	1	0	1	1	0	0	1	0	0	1
R = 7X	2	2	2	2	2	3	0	1	1	0	1	1	0	1	0	0	0	1
	Gave a Calibration score of 2			Gave a Calibration score of 2												Rater 1: Gave a Calibration score of 1		
																Rater 2: Gave a Calibration score of 1		
																Rater 3: Gave a Calibration score of 0		

LEGEND
Consistent w/ Calibration Score
Inconsistent w/ Calibration Score

On a scale from 1-5 (with 1 being no confidence and 5 being 100 % confident), how would you rate your clinical confidence in reporting the above Fazekas score? 0 % confident = 1, 25 % confident = 2, 50 % confident = 3, 75 % confident = 4, 100 % confident = 5

	Patient ID																	
	TD 002			TD 004			TD 005			TD 007			TD 008			TD 009		
Acceleration Factor	Rater 1	Rater 2	Rater 3	Rater 1	Rater 2	Rater 3	Rater 1	Rater 2	Rater 3	Rater 1	Rater 2	Rater 3	Rater 1	Rater 2	Rater 3	Rater 1	Rater 2	Rater 3
R = 1X	5	5	4	5	5	5	5	5	5	5	5	5	5	5	4	5	5	4
R = 2X	4	5	4	5	5	5	5	5	5	4	5	4	5	5	4	5	5	4
R = 3X	4	5	4	5	5	5	5	5	5	4	4	3	5	5	4	5	5	4
R = 4X	4	5	4	5	5	5	5	5	5	5	4	3	5	5	4	4	5	3
R = 5X	4	5	3	5	5	5	5	5	5	4	4	3	5	5	4	4	5	3
R = 6X	4	4	4	5	5	5	5	5	5	4	4	3	5	5	4	4	5	2
R = 7X	3	3	4	5	5	5	5	5	5	3	3	3	3	4	3	2	3	2
	Gave a Calibration score of 5															Rater 2: Gave a Calibration score of 4 Rater 3: Gave a Calibration score of 5		

	Patient ID																	
	TD 010			TD 011			TD 012			TD 013			TD 014			TD 015		
Acceleration Factor	Rater 1	Rater 2	Rater 3	Rater 1	Rater 2	Rater 3	Rater 1	Rater 2	Rater 3	Rater 1	Rater 2	Rater 3	Rater 1	Rater 2	Rater 3	Rater 1	Rater 2	Rater 3
R = 1X	5	5	4	5	5	3	5	5	5	5	5	5	5	5	4	5	5	5
R = 2X	5	5	3	5	5	5	4	5	5	5	5	5	5	5	5	5	5	5
R = 3X	4	5	3	5	5	4	5	4	5	5	5	4	4	3	4	5	5	4
R = 4X	4	5	4	5	5	4	4	5	3	4	5	4	4	5	4	4	5	4
R = 5X	5	4	2	5	5	4	3	5	5	5	5	4	5	4	4	4	5	4
R = 6X	4	4	4	4	5	2	4	5	3	5	5	4	4	4	4	4	5	4
R = 7X	3	4	2	2	4	2	3	3	1	3	5	4	3	3	1	2	4	4
	Gave a Calibration score of 5																	

	Patient ID																	
	CTneg 001			CTneg 002			CTneg 004			CTneg 005			STRHC 001			STRHC 002		
Acceleration Factor	Rater 1	Rater 2	Rater 3	Rater 1	Rater 2	Rater 3	Rater 1	Rater 2	Rater 3	Rater 1	Rater 2	Rater 3	Rater 1	Rater 2	Rater 3	Rater 1	Rater 2	Rater 3
R = 1X	5	5	5	5	5	5	5	5	5	5	5	3	5	5	4	5	5	4
R = 2X	5	5	5	5	5	5	5	5	4	4	5	5	5	5	4	5	5	4
R = 3X	5	5	5	5	5	5	5	4	3	5	4	5	5	5	4	5	4	4
R = 4X	5	5	4	5	5	4	4	5	4	4	5	4	5	5	4	4	5	4
R = 5X	5	4	5	5	5	5	4	3	4	3	5	4	4	4	4	4	5	4
R = 6X	5	5	4	5	5	4	3	3	4	5	5	4	4	5	4	4	5	4
R = 7X	3	5	4	3	5	5	3	3	3	1	4	4	2	3	2	2	3	2
	Gave a Calibration score of 5															Rater 1: Gave a Calibration score of 1 Rater 2: Gave a Calibration score of 2		

LEGEND
Consistent w/ Calibration Score
Inconsistent w/ Calibration Score

A.5. Potential Questionnaire for Future Work

Case 1: Acute Ischemic Stroke		
	Question 1:	
a)	**Can you identify an acute infarct?**	YES = **1** NO = **0**
b	**What is your confidence in the accuracy of this identification?**	0% confident = **1** 25 % confident = **2** 50 % confident = **3** 75 % confident = **4** 100% confident = **5**
	If answer to 1. a) is 'YES = **1 ', continue. Otherwise, skip to ' **CASE 2: Chronic Ischemic Lesion Burden**' .*	
	Question 2:	
a)	**Where is the acute infarct located?**	*Written answer (be as **specific** as possible).*
b	**What is your confidence in the accuracy of this location?**	0% confident = **1** 25 % confident = **2** 50 % confident = **3** 75 % confident = **4** 100% confident = **5**
c)	**Which image slice allows for the best visualization of the acute infarct?**	*Specify slice based on **'Im'** (image index) in AW server.*
	Question 3:	
a)	**Can you identify a DWI-T2 FLAIR mismatch?**	YES = **1** NO = **0**
b	**What is your confidence in the accuracy of this mismatch?**	0% confident = **1** 25 % confident = **2** 50 % confident = **3** 75 % confident = **4** 100% confident = **5**
Case 2: Chronic Ischemic Lesion Burden		
	Question 1:	
a)	**Can you identify white matter T2 hyperintense lesion(s) in the periventricular and/or deep white matter?**	normal (no white matter disease) = **0** punctate white matter disease = **1** partial confluent white matter disease = **2** confluent white matter disease = **3**
b	**What is your confidence in the accuracy of this identification?**	0% confident = **1** 25 % confident = **2** 50 % confident = **3** 75 % confident = **4** 100% confident = **5**
c)	**Which slice(s) best represents the location of the lesion(s)?**	*Specify slice(s) based on **'Im'** (image index) in AW server.*
	You have completed the Questionnaire for the particular **'Dataset Series' . Move on to answering the Questionnaire again for the next ' **Dataset Series'** .*	

Appendix B: Kappa Paradox

The inter-rater reliability measurement Cohen's kappa (κ) can be unweighted (κ_{UW}) or weighted (e.g. quadratic weighted, κ_{QW}), and its computation uses the quantities observed agreement (Pa) and expected agreement by chance (Pe).

Despite known high agreement corresponding to the acute stroke diagnostic task between Rater 1 and Rater 2, and Rater 2 and Rater 3, "equivalent to chance" agreements (κ = 0) (Tables I-II (i)) were obtained (*Section 4.2.1.*). As calculated based on κ_{UW} pairwise-rater contingency tables [87] for both Rater 1 and Rater 3, and Rater 2 and Rater 3, Pa and Pe were identical values (e.g. Pa = Pe = 0.929). Pa and Pe were also identical values for κ_{QW}. Since $\kappa = (Pa - Pe)/(1 - Pe)$, $\kappa_{UW} = \kappa_{QW} = 0$ for these two pairwise raters. On the other hand, despite known high agreement corresponding to the acute stroke diagnostic task between Rater 1 and Rater 2, computational error (κ = not a number (NaN)) results (Tables I-II (i)) were obtained (*Section 4.2.1.*). As calculated based on the κ_{UW} pairwise-rater contingency table [87] for Rater 1 and Rater 2, Pe = 1. For κ_{QW}, Pe = 1, also. Therefore, for these pairwise raters, both κ_{UW} and κ_{QW} = NaN (division by zero, undefined).

Investigation into the *kappa paradox* led to the conclusion that κ – although a standard inter-rater reliability measurement reported in MRI literature – was not a useful metric for the given data. Part of the *kappa paradox* is such that, in cases of known high agreement, the resultant κ value is low [87]. A "paradox-resistant alternative to Cohen's kappa coefficient" [88] is Gwet's AC1 (unweighted) and its quadratic weighted counterpart, Gwet's AC2. In the case of the acute stroke diagnostic

task, this aspect of the *kappa paradox* was at play since agreement was known to be high between all three neuroradiologist raters. As such, Gwet's AC1 and AC2 were computed.

Although part of the *kappa paradox* is such that, in cases of known high agreement, the resultant κ value is low [87] [88], this is not the case for the chronic stroke diagnostic task since the raw inter-rater reliabilities were known to not be as high as they were in the acute case. An additional aspect of the *κ paradox* that is at play here, then, is such that κ is dependent on bias (i.e. the marginal distributions). If the marginal distributions are more homogeneous (i.e. similar ratings between pairwise raters), κ will be lower [87]. Considering the marginal distributions are similar between pairwise raters in the sense that the marginal distributions generally monotonically increase with increasing confidence score and the fact that κ penalizes similar ratings between pairwise raters, it is logical that the κ scores remained low for the chronic diagnostic task.

Nonetheless, dependency on bias is argued to be illogical, considering the expectation that any raters who are experts in their field would yield a higher level of observed agreeance, and thus their marginal distributions would be more homogeneous [87]. This aspect of the *kappa paradox* is therefore remediated by calculating Gwet's AC1 and AC2, instead, since it is not related to bias and therefore does not penalize similar ratings.

Lastly, "the determination of weights for a weighted kappa is a subjective issue on which even experts might disagree in a particular setting" [85], which is likely also the case for Gwet's AC2. Thus, it seems most appropriate that clinical experts, such as

radiologists, be consulted when making these types of decision for proper clinical translation of results.